9-15-2016

What Could Possibly Go Wrong? The Impact of Poor Data Management

Chris Eaker
University of Tennessee - Knoxville, ceaker@utk.edu

Chapter 4

What Could Possibly Go Wrong? The Impact of Poor Data Management

Introduction

A Tibetan monk lost his life's work after posing for a photograph with London Mayor Boris Johnson.[1] A long-time Flickr user lost thousands of original digital photographs when the photo sharing service erroneously deleted them.[2] The programmers of the movie *Toy Story 2* nearly lost the entire movie file when someone accidentally typed a wrong command.[3] A data set containing mistakes from careless data entry forced a scientist to request retraction of seven articles.[4] What do these unfortunate situations all have in common? These are all situations in which poor data management practices caused problems that could have been avoided.

Why are data management skills so important? Should they matter more now than in the past? The answers to those questions may not always be apparent. Conceivably, some of the problems possible now were not possible when research data was primarily in paper form. On the one hand, the changes in the makeup of research and data from analog to digital have made collecting, processing, and analyzing data easier than ever.[5] On the other hand, the improvements offered by digital research bring ways to collect, process, and analyze data poorly, thereby creating more opportunities for problems.[6] If rates of article retractions are any indication, one study found that problems have increased ten-fold since 1975.[7] Although about two-thirds (67.4%) of retractions in that study were caused by scientific misconduct, including fraud, duplicate publications, and plagiarism, the authors found retractions caused by error have also increased, including errors related to analysis and reproducibility.[8] Article retractions represent lost time, effort, and money in research projects, many of which were funded with public money

through federal grants. Freedman, Cockburn, and Simcoe[9] estimate that the lack of reproducibility in scientific research costs $28 billion per year. The authors did not posit why these errors occurred or how they could have been avoided; it is possible rigorous data management practices may have mitigated some of the errors.

This chapter will highlight the importance of good data management practices by providing examples of problems a researcher may encounter when research data is poorly managed. It will provide examples of actual situations when bad data management led to serious problems with data loss, research integrity, and worse. It will also provide tips on how data management could have been done differently to encourage a more positive outcome.

Literature Review

As research produces higher volumes of digital research data, effective management of these data is very important. Federal grant funding agencies require researchers to submit data management plans with grant proposals and share the results of their research, including the data, in part to increase the return on their investment. Stewardship of the data is the responsibility of the researcher.[10] However, data stewardship is typically not the first thing on researchers' minds. Busy researchers are often more focused on getting the research project finished, the data analyzed, and the articles published than they are on making sure the data are described and preserved for later reuse.[11]

Librarians have recognized the need for sound data management practices to support preservation and sharing of data. Within the last decade, in order to understand researchers' current data management practices and find opportunities to help, library and information science researchers have studied different groups of researchers' data management practices[12]

and found that they often employ inconsistent practices.[13] In response to the need to improve data management practices among these researchers, librarians have developed training programs at different universities[14] to help improve these practices among faculty and students. These training programs are often framed around the data life cycle, such as the DataONE Data Life Cycle shown in Figure 4.1, which puts the skills into the context of the research process.

Librarians are not the only ones who see the need for data management training; scientific researchers have published primers on data management for their fellow scientists in fields such as ecology[15] and earth sciences,[16] and even for the general public involved in citizen science initiatives.[17] Researchers are trained in the art of conducting research in their fields, but specific skills related to managing the data they collect are not always covered in their curricula.[18] The sheer proliferation of studies of researchers' data management skills and the number of training programs, books, and articles about data management best practices suggest a tacit admission that these skills need improvement.

Potential Problems from Poor Data Management

There are many opportunities for error to be introduced during a research project, even when data are managed well. However, when data are mismanaged, even inadvertently, the risks of errors are multiplied. The DataONE Data Life Cycle,[19] shown in Figure 4.1, will be used as a model in this discussion of some of the things that can go wrong during a research project.

<Insert Figure 4.1 here>

Of course, not every research project flows through every step of the life cycle, nor is each step of the life cycle passed through in the order shown in the model. For example, the analysis phase is often completed concurrently with or subsequent to the data collection and assurance phases, not after the preservation phase as the life cycle model shows. Nevertheless, the point of the life cycle model is to encourage a researcher to think about the issues involved in managing research data throughout a project. Each of these steps involves specific tasks associated with managing data, and when these tasks are not completed effectively, the potential for error arises. In this chapter, the following six phases will be discussed: planning, data collection, quality assurance, documentation, preservation, and analysis.

The Planning Stage

As the first step in the research data life cycle and the step upon which all the remaining steps are built, the planning stage is arguably the most important. With proper planning, research is more organized and leads to better quality data. Researchers should allow ample time before they begin a project to write a data management plan that spells out who will be involved in various aspects of data collection, documentation, and analysis, and how the data will be preserved.

What Can Go Wrong?

Without a clear plan for how the research project will proceed, one might describe the situation as “flying by the seat of your pants.” Researchers may never have a clear idea of how data should be collected which can lead to data being collected inconsistently among project members. Different people may describe the variables and data files differently. People may save

data in different places and in different formats, which makes it more difficult to locate data when needed. Roles among researchers may not be clearly defined; without clear roles, important tasks may be overlooked as no one claims responsibility. This threat is higher in larger labs, as the possibility for human error is higher.[20] In a survey of graduate students, who are often tasked with data collection within laboratories and research groups,[21] Doucette and Fyfe found that almost 15% of their respondents had to collect data again that they knew had already been collected because the file was lost or corrupted. Worse, just over 17% indicated the data were permanently lost because they could not collect them again.[22] In each of these cases, lost data led to lost time and money. A thorough data management plan might have eliminated these loses. Lastly, as many labs constantly deal with students graduating and new ones becoming involved in a project, without a clear transfer protocol, that changeover can be disorganized and may lead to missed tasks and lost data.

What Can Be Done Differently?

Data management planning is an important step of the research process. It is the first step any researcher must complete in a research project. Most grant funding agencies, both public and private, now require that researchers complete a data management plan to accompany their grant proposals. However, many agencies require only brief or limited data management plans. Therefore, all researchers, even those not applying for grant funding, should consider completing a longer, more in-depth data management plan that covers detailed processes, steps, and roles. This planning step forces the researcher to think through the issues surrounding data, such as who will be involved in the project, what their roles will be, how often and where data will be

backed up, how data will be cleaned and processed, how the data and processes will be described, and where that data will be shared upon completion of the project.

The Data Collection Stage

After the planning stage, during which processes and systems are established, data collection begins. The collected or generated data are the foundation upon which all future analysis and conclusions are built; therefore, it is important to collect the data consistently. As is the case with each consecutive step of the research data life cycle (Figure 4.1), success during the data collection step depends on proper planning in the previous step.

What Can Go Wrong?

Without proper planning for data collection, a number of problems can occur. If the data collection steps and processes are not properly planned, the research project can ultimately end up with a data set that does not serve the purpose for which it was intended. For example, if more than one person is involved in the data collection, but data collectors do not follow consistent data collection practices, they can end up with data with different units, collection processes, and variable names. One person may collect temperature using one device while another collects it using a different one. The difference in data collection device may not cause problems in later data analysis, especially if these differences are known and planned for. However, researchers should attempt to minimize these differences and collect data consistently among all members of the research team. If differences in data collection are not planned for, researchers may discover they have incompatible data sources. Problems of incompatibility are especially common when

dealing with geospatial data of different coordinate projections.[23] If this incompatibility goes undetected, errors in analysis may occur.

In addition to consistency, data collection problems can be exacerbated by poor data entry techniques. Some popular data entry tools, such as Microsoft Excel and other spreadsheet software, make data entry easy. However, this ease of data entry can bring consequences, as these spreadsheet programs do not enforce any rules on data entry unless specifically told to do so. Without enforcement, people can input data into wrong fields, use incorrect formats, or leave data fields empty where there should be a value. It is important for researchers to be aware of the limitations of data entry in spreadsheet software so they can take precautions to eliminate opportunities for error.

What Can Be Done Differently?

Data collection processes, procedures, and standards should be put in place early in the research process, preferably during the planning stage, so that all people involved collect data consistently. Examples of processes that should be established early on include consistent data collection procedures, an agreed-upon naming convention for all variables to be collected during the project, and a preferred unit convention and geodetic frame of reference. Researchers should document these standards in the data management plan, and periodically check that the research team is adhering to established procedures.

When using spreadsheets for data entry, three features in Excel improve the quality of data entry validity: dropdown lists, data validation, and data input forms. Dropdown lists of preset values make data entry easier by reducing the need to manually type repeated values and eliminate variation in the ways data collectors may record the same value. For example, if one of

the pieces of information to be collected is the name of a particular species of plant and the set of species is already known and constant, the researcher can create a dropdown list of species' names to be selected from the list rather than typed repeatedly for each observation. Another helpful tool in data entry is data validation. Excel will allow the researcher to specify what type of information can go in a specific cell. For example, for a column of weights where the researcher wants two decimal places and knows the weights will always be within a certain range, the cells can be set to accept only numerical values with two decimal places within a certain numerical range. If a number that is input is out of that numerical range, Excel will display a warning. The last tool for more accurate data input is forms, which provide an easy way for data to be input into the spreadsheet. An example of an Excel input form is shown in Figure 4.2.[24]

<Insert Figure 4.2 here>

The Quality Assurance Phase

Once data are collected and steps have been taken to reduce the opportunity for error during data input, the researcher still must undergo steps to assure the quality of the data. During data collection, two types of errors can occur: errors of omission and errors of commission.[25] Errors of omission occur when data or metadata are omitted from the data set. These errors often occur inadvertently during data entry, such as when someone simply forgets to enter data for a specific observation. Errors of commission occur when incorrect data or metadata are entered. These errors also often occur inadvertently, such as when someone enters an incorrect value into a cell in the spreadsheet. In both cases, the researchers must take steps to eliminate these kinds of

errors. These steps are discussed later in "What Can Be Done Differently?" It is important to note the goal of data quality assurance processes is not to eliminate legitimate outliers in the data, but to eliminate incorrect data. Legitimate outliers should be maintained and explained thoroughly in the documentation.

What Can Go Wrong?

Poor quality data can have serious effects on later analysis. Data containing errors of commission or omission have the potential of throwing off analytical calculations, which may then lead to incorrect conclusions. In addition to errors of commission or omission, careless handling of spreadsheet data can cause one column to be sorted out of order with the others, which is not always apparent at first glance.

Ultimately, poor quality data sets can have far-reaching implications and can lead to multiple article retractions. In one case, careless data entry caused retraction of seven articles.[26] In another case, a researcher who used a homemade computer model that erroneously reversed two columns of data had to request retraction of five articles.[27] Other researchers who had used the erroneous results from the original researcher then had to request retraction of several more articles. Additionally, other researchers who attempted to publish *correct* results in contradiction to the original researcher's *incorrect* results had difficulty getting their articles published.[28]

What Can Be Done Differently?

There are several techniques to check the quality of data once they have been entered, two of which are discussed here. One way to reduce error during data input is for two people to

input the same data into separate files. Once the data are entered twice, the researcher can compare the two files and identify and resolve any discrepancies.

Another powerful way to check data quickly is to use visualization techniques. For example, for geographic data, a simple visualization of all data points on a map will quickly identify any data that are geographically out of place. Then the researcher can flag those data and go back to check them for accuracy. Visualization can also be useful for identifying errors in data that can be plotted on a graph. If one data point shows up far away from the rest of the data points, it can be flagged for later verification.

The Documentation Phase

Although documentation is shown as one of many "steps" in the research life cycle in Figure 4.1, in reality, it should be an ongoing process throughout the project. Data, people, instruments, processes, and more, should be described thoroughly using a standard metadata schema. Metadata is "structured information that describes the attributes of information resources[29] for the purposes of identification, discovery, selection, use, access, and management."[30] The data management plan must explain how this process will be completed and who is responsible for it.

What Can Go Wrong?

Many problems can occur when data are not documented and described properly. Reproducibility is an important cornerstone of scientific research, and without explicitly described methods and data, research projects are difficult to replicate. Without metadata, other researchers cannot know how the data were collected, processed, and analyzed, and therefore

cannot replicate the study. This lack of reproducibility in scientific research has prompted the editors of the journal *Nature* to gather a list of articles about how to fix the problem[31] and strengthen their requirements for the methods sections for authors publishing in their journal.[32]

Data reuse also suffers when data and methods are not sufficiently described. Other researchers who were not involved in the data collection will lack important information necessary to reuse the data, such as the meaning of variable names, identification of instruments used to collect the data and their calibration, the spatial and temporal coverage of the data, and the accuracy of the data set. Additionally, researchers wishing to reuse the data will not know the conditions under which the data were collected. These pieces of information are important when integrating data from several sources into one data set for reuse.

Additionally, without documentation, it is even difficult for the researchers who conducted the research to reproduce their own efforts, should that become necessary, such as if data are lost. If analysis and processing steps were not adequately documented, re-creation of the lost data set is much more difficult and time consuming.

Lastly, a researcher's recollection of the details of a research project are lost quickly after the end of the project. Michener, et al., demonstrate in Figure 4.3 a phenomenon they call "Information Entropy." Soon after the article is published, researchers forget specific details about the conditions under which the data were collected and processed. As time goes on, they forget more general details about the data. Catastrophic losses of data can occur at any time when the media on which they are stored are lost. Later, as the researchers change positions or retire, their ability to remember details about the project drop substantially. Finally, if the researcher dies and there is no metadata for the project, the information dies along with the researcher.[33]

<Insert Figure 4.3 here>

What Can Be Done Differently?

Metadata standards are established to provide a standardized way for researchers within a field to describe their projects and data sets. Metadata standards also make data sets machine-readable so they can be indexed and searched. Many disciplines have standardized metadata schemas. Examples of discipline-specific metadata schemas are Ecological Metadata Language for the ecological sciences,[34] Darwin Core for the biological sciences,[35] and ISO 19115 for the geographical sciences.[36] In cases where no standard format for metadata exists, researchers may describe their projects clearly and accurately using a simple Dublin Core metadata record. To provide more detailed information not contained in a metadata record, the researcher should include an accompanying "ReadMe" file.

Metadata records and other documentation, such as ReadMe files, contain different levels of granularity. Generally, the metadata should include information such as the overall purpose of the research, the people involved in the research, conditions on the use of the data, and structure of the data files, including how they are related to one another.[37] More detailed information may also be appropriate, including information on the research design, the data collection processes and methods, data processing processes and methods, and the spatial and temporal coverages of the data set.[38]

Lastly, in the most granular, or detailed, view of the research project, information should be included about the variables within the data set and how they relate to one another, the types of instruments used to collect each variable and the instruments' calibration, description of any

codes used for missing values, explanation of any derived values, explanations of errors within the data files, and documentation of any outliers.[39]

The Preservation Phase

Preservation ensures that all the previous hard work is not lost and is thus perhaps even more important than data collection itself. Preservation includes both short-term back-ups of data files and long-term preservation of those files beyond the end of the project. During the planning stage, the researcher should devise a procedure for a regular backup schedule and location, as well as determine the most suitable format and location for long-term preservation.

What Can Go Wrong?

During the active research stage of a project, the researcher's primary concern is maintaining access to the data being collected. A study of 724 National Science Foundation grant awardees found that half of them had suffered a loss of data of some form or another ranging from human error to equipment error.[40] Therefore, redundancy of copies is crucial to maintaining access to important research data and supporting documents. Lack of a backup plan can result in the loss of data when hard drives fail or laptops are stolen; placing all of a project's data on one computer is risky. Lelung Rinpoche, the Tibetan monk mentioned in this chapter's introduction, exited the London Tube at his stop after snapping a photograph with London Mayor Boris Johnson. He accidentally left his laptop, and it was stolen. Rinpoche's computer contained "900 pages of rare Tibetan Buddhist scriptures he had travelled the world to find."[41] As they were his only copies of the material, his life's work was gone.

Many researchers are turning to cloud storage to maintain working copies of their current and past research data; however, cloud storage is not without faults. In 2014, Dedoose, a cloud storage system for academic research, suffered a major failure resulting in the loss of researchers' work over a three-week period prior to the crash.[42] These data were never recovered. Some researchers estimated the lost time to be about 100 hours.[43] Unfortunately, this type of problem is not unique to this particular cloud storage service. Other cloud storage services also have had failures that caused users to lose valuable information. One Box.com user lost his files when the service gave access to his account to someone else and that new user deleted his files.[44] Likewise, Flickr erroneously deleted all (about 4,000) of one user's original digital photographs when the service mistook his account with one containing stolen photographs.[45]

While short-term storage of research data is of immediate importance to most researchers, long-term storage solutions are not always on their minds. Vines, et al., attempted to obtain data sets from 516 scholarly articles from 1991 to 2011. They found the older the publications were, the more likely that the data were not available. In fact, they found the data availability dropped 17% per year. They report one main reason the data were not available was because they were on inaccessible media.[46]

Digital files on electronic media are notorious for becoming inaccessible, both because of bit rot[47] and because the file formats and media themselves are highly susceptible to obsolescence.[48] Bit rot happens when physical storage media degrade, causing loss of access to the files stored on them. This degradation is a breakdown of the electrical, optical, or magnetic properties of the storage media, which causes them to lose their ability to hold the digital information. File format and media obsolescence is caused when software and hardware advancements cause older versions to no longer be accessible. Lotus 1-2-3, which was an

extremely popular spreadsheet software throughout the 1980s and 1990s, is a perfect example of how file formats become obsolete. Researchers who have data in this file format from decades ago are no longer able to open them in modern spreadsheet packages. Moreover, those files may have been stored on floppy disks, which most modern computers lack the hardware to read.

What Can Be Done Differently?

During the planning stage, the researcher should devise a plan for short-term and long-term preservation of the digital files from a research project. The first concern is to develop a regular backup schedule and a suitable location for the backups in order to maintain access to files throughout the research project and to ensure data are not lost. See Table 4.1 for important questions to answer in developing a backup plan.[49] Ideally, three backup copies should be maintained to safeguard against the possibility of losing one copy. One copy can be local and internal, such as a hard drive on a laboratory or office computer, allowing easy access to the files most often used. A second copy can be local, but should be on an external device, such as an external hard drive. It is not recommended to save backup copies on devices such as CDs, DVDs, and USB drives, as they may suffer from bit rot over time. A third copy should be at an external, geographically separate location from the place the research is taking place. This location can be cloud storage or an off-site, physical server. If using cloud storage as an off-site backup copy, turn off automatic synchronization, as deletion or corruption of one copy will automatically duplicate that change in the other. Off-site backup is important to prevent loss due to fire or natural disasters.

<Insert Table 4.1 here>

Long term preservation is the second important consideration. The main goal of preserving data sets for the long-term is to facilitate reuse by other researchers. Researchers may be able to use data from another researcher to answer new research questions. When accompanied by adequate metadata and verified for accuracy, data sets have a higher potential to serve future research,[50] and as data sets are reused for additional research, their value increases.[51]

To facilitate long-term preservation and reuse, whenever possible, files should be saved in non-proprietary file formats. Proprietary file formats have the potential of becoming obsolete over time, as the software needed to read them may no longer be available, while non-proprietary, open file formats are readable by many software packages. If the file format used to create the data must be preserved as is, the software required to open and use the data should be preserved along with the data set. Additionally, a digital object identifier (DOI) should be assigned to the data set for ease of discovery and citation.

The Analysis Phase

Once the data are collected and have been cleaned by eliminating errors of commission and correcting errors of omission, the data set is ready for analysis. If the researcher has planned the project thoroughly, made efforts to reduce the possibility for error during data collection, and checked the data for accuracy during the quality assurance phase, the chance for error during the analysis phase is greatly reduced. However, there are still techniques the researcher can employ to reduce that chance even further.

What Can Go Wrong?

During the analysis phase, the researcher is processing and manipulating the data set to find the information of interest to the research project. During this processing, the data set may be transformed into a new form, such as converting a raw data file to a more usable spreadsheet format. This transformation is important for the analysis but can cause problems if not managed properly. A problem that can occur during data set processing and analysis is that the data set can be transformed to the wrong form, thereby requiring the researcher to revert to an earlier version. Geospatial data is especially susceptible to incorrect transformations when projecting a data set of one coordinate projection to another. If earlier versions of data files were not backed up, reverting to an earlier version may be difficult or impossible.

What Can Be Done Differently?

Borer, et al., recommend two best practices in maintaining proper versioning of data sets.[52] First, using a scripted software program, such as R, for processing will make a record of the steps necessary to recreate what has been done or make changes and reprocess the files. Second, the original, uncorrected file should always be saved, so that it will always be possible to go back to the beginning of the process and start over. Additionally, as files are processed and certain milestones are reached, those versions should be backed up in case it is necessary to revert to an earlier version. Milestones are points the researcher wants to preserve for easy retrieval. The first milestone that should be preserved is the original raw data generated by the research equipment. A subsequent important milestone may be set when the raw data is initially converted to a usable format, such as a spreadsheet. A final milestone may be set when the data is in its final format that supports a published journal article and that the researcher wants to share with other researchers.

Conclusion

The purpose of this chapter has been to highlight the importance of good data management practices from the viewpoint of what can go wrong if data are poorly managed. The examples in this chapter show a range of problems from minor to severe. Potential issues usually arise from neglectful or careless treatment of the data sets. While it is impossible to reduce the potential for error to zero, it is clear from these examples that managing data before, during, and after a research project will substantially reduce the chance for error. Estimates of the costs of irreproducible research range from $20 billion per year in one study of medical research[53] to $28 billion per year in one study of biological research.[54] Much of this research is irreproducible because of poor data management and lack of adequate metadata.

In addition to the financial costs, both researchers' and their institutions' reputations are on the line. Academic institutions both in the United States[55] and abroad[56] recognize how good data management practices ultimately help improve researchers' and institutions' reputations.

Pearls

1. Plan as many details of your research as possible—from collection to processing to preservation—prior to beginning the project.
2. Use data input tools such as data validation and input forms to reduce the chance for error during data collection.
3. Stay current on documentation of processes and description of project details throughout the project; this work is more difficult to do at the end of a project.

4. Always maintain three current backup copies of important work, such as the original unprocessed data set and milestone versions of processed files.
5. Give adequate attention to cleaning errors from the data set prior to analysis, but maintain legitimate outliers.
6. Understand the limitations of common statistical tests and provide as much supporting information as possible to support your claims.

Recommended Reading and Resources

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and Sharing Research Data: A Guide to Good Practice*. London: Sage Publications Ltd, 2014.

Pryor, Graham, ed. *Managing Research Data*. London: Facet Publishing, 2012.

Pryor, Graham, Sarah Jones, and Angus Whyte, eds. *Delivering Research Data Management Services: Fundamentals of Good Practice*. London: Facet Publishing, 2014.

Ray, Joyce, ed. *Research Data Management: Practical Strategies for Information Professionals*. West Lafayette, IN: Purdue University Press, 2014.

Bibliography

Adamick, Jessica, Rebecca Reznik-Zellen, and Matt Sheridan. "Data Management Training for Graduate Students at a Large Research University." *Journal of eScience Librarianship* 1, no. 1 (2012). doi:10.7191/jeslib.2012.1022.

Akers, Katherine G., and Jennifer Doty. "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives." *International Journal of Digital Curation* 8, no. 2 (2013): 5-26. doi:10.2218/ijdc.v8i2.263.

Baker, Monya. "Irreproducible Biology Research Costs Put at $28 Billion Per Year." *Nature* (2015). Published electronically June 9, 2015. doi:10.1038/nature.2015.17711.

"Bit Rot." *The Economist*. Published electronically April 28, 2012. http://www.economist.com/node/21553445.

Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* 90, no. 2 (2009/04/01 2009): 205-14. doi:10.1890/0012-9623-90.2.205.

Carlson, Jacob, Michael Fosmire, CC Miller, and Megan Sapp Nelson. "Determining Data Information Literacy Needs: A Study of Students and Research Faculty." *portal: Libraries and the Academy* 11, no. 2 (2011): 629-57.

Carlson, Jake, Lisa Johnston, Brian Westra, and Mason Nichols. "Developing an Approach for Data Management Education: A Report from the Data Information Literacy Project." *International Journal of Digital Curation* 8, no. 1 (2013): 204-17. doi:10.2218/ijdc.v8i1.254.

Casadevall, Arturo, R Grant Steen, and Ferric C Fang. "Sources of Error in the Retracted Scientific Literature." *The FASEB Journal* 28, no. 9 (2014): 3847-55.

Chandler, Adam. "A Warehouse Fire of Digital Memories." *The Atlantic* (2015). Published electronically February 13, 2015. http://www.theatlantic.com/technology/archive/2015/02/google-forgotten-century-digital-files-bit-rot/385500/.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and Sharing Research Data: A Guide to Good Practice*. London: Sage Publications Ltd, 2014.

D'Ignazio, John, and Jian Qin. "Faculty Data Management Practices: A Campus-Wide Census of Stem Departments." 2008. doi:citeulike-article-id:8241850.

DataONE. "Dataone Data Management Education Modules: Data Quality Control and Assurance." (2012). https://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx.

Doucette, L., and B. Fyfe. "Drowning in Research Data: Addressing Data Management Literacy of Graduate Students." *ACRL 2013 Proceedings* (2013).

Eaker, Christopher. "Educating Researchers for Effective Data Management." *Bulletin of the American Society for Information Science and Technology* 40, no. 3 (2014): 45-46. doi:10.1002/bult.2014.1720400314.

———. "Planning Data Management Education Initiatives: Process, Feedback, and Future Directions." *Journal of eScience Librarianship* 3, no. 1 (2014). doi:10.7191/jeslib.2014.1054.

Eaker, Christopher, Peter Fernandez, Shea Swauger, and Miriam Davis. "Data Sharing Practices of Agricultural Researchers: Implications for the Land-Grant University Mission." Paper

presented at the Special Libraries Association Food and Agriculture Division Virtual Contributed Papers Session, May 13, 2015.

Fang, Ferric C., R. Grant Steen, and Arturo Casadevall. "Misconduct Accounts for the Majority of Retracted Scientific Publications." *Proceedings of the National Academy of Sciences* 109, no. 42 (2012): 17028-33. doi:10.1073/pnas.1212247109.

Fischer, Manfred, Henk Scholten, and David Unwin. *Spatial Analytical Perspectives on Gis*. London: Taylor & Francis, 1996.

Freedman, Leonard P., Iain M. Cockburn, and Timothy S. Simcoe. "The Economics of Reproducibility in Preclinical Research." *PLoS Biol* 13, no. 6 (2015): e1002165. doi:10.1371/journal.pbio.1002165.

Group, Darwin Core Task. "Darwin Core." http://rs.tdwg.org/dwc/.

Henty, Margaret, Belinda Weaver, Stephanie Bradbury, and Simon Porter. "Investigating Data Management Practices in Australian Universities." Australian Partnership for Sustainable Repositories, 2008.

"How Toy Story 2 Nearly Vanished." *Tor.com*. Published electronically June 25, 2012. http://www.tor.com/2012/06/25/how-toy-story-2-nearly-vanished/.

ISO/TC 211 Geographic Information/Geomatics Committee. "Iso 19115: Geographic Information -- Metadata." International Standards Organization, http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798.

Janke, Lori, Andrew Asher, and Spencer Keralis. "The Problem of Data." *Council on Library and Information Resources* CLIR Publication No. 154. Published electronically August 2012.

Johnston, Lisa, Meghan Lafferty, and Beth Petsan. "Training Researchers on Data Management: A Scalable, Cross-Disciplinary Approach." *Journal of eScience Librarianship* 1, no. 2 (2012). doi:10.7191/jeslib.2012.1012.

Kervin, Karina, William Michener, and Robert Cook. "Common Errors in Ecological Data Sharing." *Journal of eScience Librarianship* (2013). doi:10.7191/jeslib.2013.1024.

Knowledge Network for Biocomplexity, The. "Ecological Metadata Language." https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

Koebler, Jason. "Our Digital Memories Are Languishing on Obsolete Cd-Rs in Our Closets." *Motherboard* (2015). Published electronically January 2, 2015. http://motherboard.vice.com/read/our-digital-memories-are-languishing-on-obsolete-cd-rs-in-our-closets.

Kolowich, Steve. "Hazards of the Cloud: Data-Storage Service's Crash Sets Back Researchers." *The Chronicle of Higher Education* (2014). Published electronically May 12, 2014. http://chronicle.com/blogs/wiredcampus/hazards-of-the-cloud-data-storage-services-crash-sets-back-researchers/52571.

Kowalczyk, Stacy. "Before the Repository: Defining the Preservation Threats to Research Data in the Lab." Paper presented at the Joint Conference on Digital Libraries, Knoxville, TN, June 24, 2015 2015. doi:10.1145/2756406.2756909.

LaTrobe University. "The Benefits of Data Management." http://www.latrobe.edu.au/research-infrastructure/eresearch/services/data-management/benefits.

McLure, Merinda, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson. "Data Curation: A Study of Researcher Practices and Needs." *portal: Libraries and the Academy* 14, no. 2 (2014): 139-64. doi:10.1353/pla.2014.0009.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Applications* 7, no. 1 (1997): 330-42. doi:10.2307/2269427.

Miller, Greg. "A Scientist's Nightmare: Software Problem Leads to Five Retractions." *Science* 314, no. 1856 (2006). doi:10.1126/science.314.5807.1856.

Morgan, Ben, and Rashid Razaq. "Tibetan Monk Loses His Life's Work in Tube Laptop Theft after Taking Selfie with Mayor Boris Johnson." *The London Evening Standard*. Published electronically April 25, 2014. http://www.standard.co.uk/news/crime/tibetan-monk-loses-his-lifes-work-in-tube-laptop-theft-after-taking-selfie-with-mayor-boris-johnson-9285082.html.

Nature. "Availability of Data, Material and Methods." Macmillan Publishers Limited, http://www.nature.com/authors/policies/availability.html.

———. "Challenges in Irreproducible Research." Macmillan Publishers Limited, http://www.nature.com/nature/focus/reproducibility/index.html.

O'Brien, Chris. "Crash at Academic Cloud Service Dedoose May Wipe out Weeks of Research." *Los Angeles Times* (2014). Published electronically May 12, 2014. http://www.latimes.com/business/technology/la-fi-tn-dedoose-crash-academic-cloud-20140512-story.html.

Palmer, Carole L, Nicholas M Weber, and Melissa H Cragin. "The Analytic Potential of Scientific Data: Understanding Re-Use Value." *Proceedings of the American Society for Information Science and Technology* 48, no. 1 (2011): 1-10. doi:10.1002/meet.2011.14504801174.

Piorun, Mary, Donna Kafel, Tracey Leger-Hornby, Siamak Najafi, Elaine Martin, Paul Colombo, and Nancy LaPelle. "Teaching Research Data Management: An Undergraduate/Graduate Curriculum." *Journal of eScience Librarianship* 1, no. 3 (2012). doi:10.7191/jeslib.2012.1003.

Ray, Joyce, ed. *Research Data Management: Practical Strategies for Information Professionals*. West Lafayette, IN: Purdue University Press, 2014.

Rekers, Hans, and Biran Affandi. "Letter to the Editor." *Contraception* 70, no. 5 (October 26, 2004): 433. doi:10.1016/j.contraception.2004.07.004.

Royal Holloway University of London. "Research Data Management Policy." 2014.

Scott, Mark, Richard Boardman, Philippa Reed, and Simon Cox. "Research Data Management Education for Future Curators." *International Journal of Digital Curation* 8, no. 1 (2013): 288-94. doi:10.2218/ijdc.v8i1.261.

Strasser, C., R.B. Cook, W.K. Michener, and A. Budden. *Primer on Data Management: What You Always Wanted to Know, but Where Afraid to Ask*. Albuquerque, NM: DataONE, 2012.

Taylor, Arlene G., and Daniel N. Joudrey. *The Organization of Information*. Third ed. Westport, CT: Libraries Unlimited, 2009.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. "Data Sharing by Scientists: Practices and Perceptions." *PLoS ONE* 6, no. 6 (2011): e21101. doi:10.1371/journal.pone.0021101.

Tynan, Dan. "How Box.Com Allowed a Complete Stranger to Delete All My Files." *IT World*. Published electronically October 23, 2013. http://www.itworld.com/article/2833267/it-management/how-box-com-allowed-a-complete-stranger-to-delete-all-my-files.html.

Uhlir, Paul F. "Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data." *Data Science Journal* 9 (2010): ES1-ES5. doi:10.2481/dsj.Essay-001-Uhlir.

University of California Santa Barbara. "Data Curation and Management." http://www.library.ucsb.edu/scholarly-communication/data-curation-management.

University of Leeds. "University of Leeds Research Data Management Policy." http://library.leeds.ac.uk/research-data-policies.

Vines, Timothy H, Arianne Y K. Albert, Rose L Andrew, Florence Débarre, Dan G Bock, Michelle T Franklin, Kimberly J Gilbert, *et al.* "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (1/6/ 2014): 94-97. doi:10.1016/j.cub.2013.11.014.

Ward, C., L. Freiman, S. Jones, L. Molloy, and K. Snow. "Making Sense: Talking Data Management with Researchers." *International Journal of Digital Curation* 6, no. 2 (2010).

Wauters, Robin. "Flickr Accidentally Wipes out Account: Five Years and 4,000 Photos Down the Drain." *Techcrunch*. Published electronically February 2, 2011. http://techcrunch.com/2011/02/02/flickr-accidentally-wipes-out-account-five-years-and-4000-photos-down-the-drain/.

Wiggins, Andrea, Rick Bonney, Eric Graham, Sandra Henderson, Steve Kelling, Gretchen LeBuhn, Richard Littauer, *et al. Data Management Guide for Public Participation in Scientific Research*. Albuquerque, NM: DataONE, 2013.

Woo, Kara. "Abandon All Hope, Ye Who Enter Dates in Excel." *Data Pub*. Published electronically April 10, 2014. http://datapub.cdlib.org/2014/04/10/abandon-all-hope-ye-who-enter-dates-in-excel/.

XLCalibre. "The Seven Deadly Sins of Data Entry (or How Not to Use Excel)." *DataScopic*. Published electronically n.d., http://datascopic.net/xlcaliber-7deadlysins/.

---

[1] Ben Morgan and Rashid Razaq, "Tibetan Monk Loses His Life's Work in Tube Laptop Theft after Taking Selfie with Mayor Boris Johnson," *The London Evening Standard*, http://www.standard.co.uk/news/crime/tibetan-monk-loses-his-lifes-work-in-tube-laptop-theft-after-taking-selfie-with-mayor-boris-johnson-9285082.html.

[2] Robin Wauters, "Flickr Accidentally Wipes out Account: Five Years and 4,000 Photos Down the Drain," *Techcrunch*, http://techcrunch.com/2011/02/02/flickr-accidentally-wipes-out-account-five-years-and-4000-photos-down-the-drain/.

[3] "How Toy Story 2 Nearly Vanished," *Tor.com*, http://www.tor.com/2012/06/25/how-toy-story-2-nearly-vanished/.

[4] Hans Rekers and Biran Affandi, "Letter to the Editor," *Contraception* 70, no. 5.

[5] Joyce Ray, ed. *Research Data Management: Practical Strategies for Information Professionals* (West Lafayette, IN: Purdue University Press, 2014), 1.

[6] Kara Woo, "Abandon All Hope, Ye Who Enter Dates in Excel," *Data Pub*, http://datapub.cdlib.org/2014/04/10/abandon-all-hope-ye-who-enter-dates-in-excel/; XLCalibre, "The Seven Deadly Sins of Data Entry (or How Not to Use Excel)," *DataScopic*, http://datascopic.net/xlcaliber-7deadlysins/.

[7] Ferric C. Fang, R. Grant Steen, and Arturo Casadevall, "Misconduct Accounts for the Majority of Retracted Scientific Publications," *Proceedings of the National Academy of Sciences* 109, no. 42 (2012).
[8] Arturo Casadevall, R Grant Steen, and Ferric C Fang, "Sources of Error in the Retracted Scientific Literature," *The FASEB Journal* 28, no. 9 (2014).
[9] Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe, "The Economics of Reproducibility in Preclinical Research," *PLoS Biol* 13, no. 6 (2015).
[10] Louise Corti et al., *Managing and Sharing Research Data: A Guide to Good Practice* (London: Sage Publications Ltd, 2014).
[11] Christopher Eaker et al., "Data Sharing Practices of Agricultural Researchers: Implications for the Land-Grant University Mission" (paper presented at the Special Libraries Association Food and Agriculture Division Virtual Contributed Papers Session, May 13, 2015).
[12] Katherine G. Akers and Jennifer Doty, "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives," *International Journal of Digital Curation* 8, no. 2 (2013); John D'Ignazio and Jian Qin, "Faculty Data Management Practices: A Campus-Wide Census of Stem Departments" (2008); L. Doucette and B. Fyfe, "Drowning in Research Data: Addressing Data Management Literacy of Graduate Students," *ACRL 2013 Proceedings* (2013); Margaret Henty et al., "Investigating Data Management Practices in Australian Universities," (Australian Partnership for Sustainable Repositories, 2008); Merinda McLure et al., "Data Curation: A Study of Researcher Practices and Needs," *portal: Libraries and the Academy* 14, no. 2 (2014); Carol Tenopir et al., "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE* 6, no. 6 (2011).
[13] C. Ward et al., "Making Sense: Talking Data Management with Researchers," *International Journal of Digital Curation* 6, no. 2 (2010).
[14] Jessica Adamick, Rebecca Reznik-Zellen, and Matt Sheridan, "Data Management Training for Graduate Students at a Large Research University," *Journal of eScience Librarianship* 1, no. 1 (2012); Jake Carlson et al., "Developing an Approach for Data Management Education: A Report from the Data Information Literacy Project," *International Journal of Digital Curation* 8, no. 1 (2013); Christopher Eaker, "Educating Researchers for Effective Data Management," *Bulletin of the American Society for Information Science and Technology* 40, no. 3 (2014); "Planning Data Management Education Initiatives: Process, Feedback, and Future Directions," *Journal of eScience Librarianship* 3, no. 1 (2014); Lisa Johnston, Meghan Lafferty, and Beth Petsan, "Training Researchers on Data Management: A Scalable, Cross-Disciplinary Approach," ibid.1, no. 2 (2012); Mary Piorun et al., "Teaching Research Data Management: An Undergraduate/Graduate Curriculum," ibid., no. 3; Mark Scott et al., "Research Data Management Education for Future Curators," *International Journal of Digital Curation* 8, no. 1 (2013).
[15] Elizabeth T. Borer et al., "Some Simple Guidelines for Effective Data Management," *Bulletin of the Ecological Society of America* 90, no. 2 (2009); Karina Kervin, William Michener, and Robert Cook, "Common Errors in Ecological Data Sharing," *Journal of eScience Librarianship* (2013).
[16] C. Strasser et al., *Primer on Data Management: What You Always Wanted to Know, but Where Afraid to Ask* (Albuquerque, NM: DataONE, 2012).
[17] Andrea Wiggins et al., *Data Management Guide for Public Participation in Scientific Research* (Albuquerque, NM: DataONE, 2013).

[18] Lori Janke, Andrew Asher, and Spencer Keralis, "The Problem of Data," *Council on Library and Information Resources* CLIR Publication No. 154.
[19] DataONE Data Life Cycle (online at https://www.dataone.org/best-practices).
[20] Stacy Kowalczyk, "Before the Repository: Defining the Preservation Threats to Research Data in the Lab" (paper presented at the Joint Conference on Digital Libraries, Knoxville, TN, June 24, 2015 2015).
[21] Jacob Carlson et al., "Determining Data Information Literacy Needs: A Study of Students and Research Faculty," *portal: Libraries and the Academy* 11, no. 2 (2011).
[22] Doucette and Fyfe, "Drowning in Research Data: Addressing Data Management Literacy of Graduate Students."
[23] Manfred Fischer, Henk Scholten, and David Unwin, *Spatial Analytical Perspectives on Gis* (London: Taylor & Francis, 1996).
[24] Created by Christopher Eaker from a sample Microsoft Excel data set
[25] DataONE, "Dataone Data Management Education Modules: Data Quality Control and Assurance," (2012), https://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx.
[26] Rekers and Affandi, "Letter to the Editor."
[27] Greg Miller, "A Scientist's Nightmare: Software Problem Leads to Five Retractions," *Science* 314, no. 1856 (2006).
[28] Ibid.
[29] Or in this case, a research project and its data.
[30] Arlene G. Taylor and Daniel N. Joudrey, *The Organization of Information*, Third ed. (Westport, CT: Libraries Unlimited, 2009), 89.
[31] Nature, "Challenges in Irreproducible Research," Macmillan Publishers Limited, http://www.nature.com/nature/focus/reproducibility/index.html.
[32] "Availability of Data, Material and Methods," Macmillan Publishers Limited, http://www.nature.com/authors/policies/availability.html.
[33] William K. Michener et al., "Nongeospatial Metadata for the Ecological Sciences," *Ecological Applications* 7, no. 1 (1997).
[34] The Knowledge Network for Biocomplexity, "Ecological Metadata Language," https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
[35] Darwin Core Task Group, "Darwin Core," http://rs.tdwg.org/dwc/.
[36] ISO/TC 211 Geographic Information/Geomatics Committee, "Iso 19115: Geographic Information -- Metadata," International Standards Organization, http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798.
[37] Corti et al., *Managing and Sharing Research Data: A Guide to Good Practice*, 39.
[38] Ibid.
[39] Ibid., 41.
[40] Stacy Kowalczyk, "Defore the Repository: Defining the Preservation Threats to Research Data in the Lab" (paper presented at the Joint Conference on Digital Libraries, Knoxville, TN, June 24, 2015 2015).
[41] Morgan and Razaq, "Tibetan Monk Loses His Life's Work in Tube Laptop Theft after Taking Selfie with Mayor Boris Johnson".
[42] Chris O'Brien, "Crash at Academic Cloud Service Dedoose May Wipe out Weeks of Research," *Los Angeles Times* (2014), http://www.latimes.com/business/technology/la-fi-tn-dedoose-crash-academic-cloud-20140512-story.html.

[43] Steve Kolowich, "Hazards of the Cloud: Data-Storage Service's Crash Sets Back Researchers," *The Chronicle of Higher Education* (2014), http://chronicle.com/blogs/wiredcampus/hazards-of-the-cloud-data-storage-services-crash-sets-back-researchers/52571.
[44] Dan Tynan, "How Box.Com Allowed a Complete Stranger to Delete All My Files," *IT World*, http://www.itworld.com/article/2833267/it-management/how-box-com-allowed-a-complete-stranger-to-delete-all-my-files.html.
[45] Wauters, "Flickr Accidentally Wipes out Account: Five Years and 4,000 Photos Down the Drain".
[46] Timothy H Vines et al., "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology* 24, no. 1 (2014).
[47] "Bit Rot," *The Economist*, http://www.economist.com/node/21553445.
[48] Adam Chandler, "A Warehouse Fire of Digital Memories," *The Atlantic* (2015), http://www.theatlantic.com/technology/archive/2015/02/google-forgotten-century-digital-files-bit-rot/385500/; Jason Koebler, "Our Digital Memories Are Languishing on Obsolete Cd-Rs in Our Closets," *Motherboard* (2015), http://motherboard.vice.com/read/our-digital-memories-are-languishing-on-obsolete-cd-rs-in-our-closets.
[49] Modified from Lamar Soutter Library, University of Massachusetts Medical School, licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License (online at https://creativecommons.org/licenses/by-nc/3.0/deed.en_US).
[50] Carole L Palmer, Nicholas M Weber, and Melissa H Cragin, "The Analytic Potential of Scientific Data: Understanding Re-Use Value," *Proceedings of the American Society for Information Science and Technology* 48, no. 1 (2011).
[51] Paul F. Uhlir, "Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data," *Data Science Journal* 9 (2010).
[52] Borer et al., "Some Simple Guidelines for Effective Data Management."
[53] Freedman, Cockburn, and Simcoe, "The Economics of Reproducibility in Preclinical Research."
[54] Monya Baker, "Irreproducible Biology Research Costs Put at $28 Billion Per Year," *Nature* (2015).
[55] University of California Santa Barbara, "Data Curation and Management," http://www.library.ucsb.edu/scholarly-communication/data-curation-management.
[56] LaTrobe University, "The Benefits of Data Management," http://www.latrobe.edu.au/research-infrastructure/eresearch/services/data-management/benefits; Royal Holloway University of London, "Research Data Management Policy," (2014); University of Leeds, "University of Leeds Research Data Management Policy," http://library.leeds.ac.uk/research-data-policies.