Data Curation Workshop: Tips and Tools for Today

Matthew M. Benzing
Miami University
matt.benzing@miamioh.edu

Abstract

The current state of research data is like a disorganized photo collection: a mix of formats scattered across different media without a lot of authority control. That is changing as the need to make data available to researchers across the world is becoming recognized. Researchers know that their data needs to be maintained and made accessible, but often they do not have the time or the inclination to get involved in all of the details. This provides an excellent opportunity for librarians. Data curation is the process of preparing data to be made available in a repository with the goal of making it FAIR: Findable, Accessible, Interoperable, and Reusable. This workshop walks attendees through the steps in the process and gives them hands on experience in data curation activities.

I used to own a Pentax K-1000 35mm. A great camera. I took a lot of pictures, but was limited by the need to buy film in rolls of 24 or 36; I had to be a little judicious in what I shot. Then I would get the exposures developed, and be given a set of prints and strips of negatives. The best prints went into albums, the less than great and the negatives tended to go into boxes.

Then one day I purchased my first digital camera, a little Samsung point and shoot. The pictures were not close to the quality of the 35mm print camera, but they were good enough. More importantly, the camera was so convenient. No buying film, no loading film, no developing film. Smart cards stored the pictures, and they held a lot. I also had a little flip phone at this point. It took pictures too, although they looked really bad. I always had it with me though, so I took pictures with it when something caught my eye and I didn't have a real camera handy.

I could choose which photos were worth printing. I started taking multiple pictures of the same scene so that I could choose the best shot later. I told myself I would delete the inferior shots but never did, so I began to accumulate a lot of files. I began to digitize some of my older print photos for sharing and for preservation; even more material for my growing collection.

I changed to different makes and models of point and shoot cameras over the years, and also picked up a DSLR. The DSLR took really good pictures, comparable to my old Pentax, but it was big and bulky next to the point and shoots, so I used it mostly on special occasions and trips. Meanwhile I was upgrading phones along the way, and eventually graduated to smartphones. They took pictures almost as good as the point and shoots and yet had the portability and convenience of a phone, so eventually an iPhone became my everyday camera

Now when I look over the photos I've taken over my life I'm confronted with a mess. I have prints, and some of the negatives they were printed from. I have image files on smart cards. I have image files on my pc. I have image files on two different external hard drives. I have image files on CDs as part of a halfhearted attempt several years ago to back, and unfortunately some of these CDs have become corrupt and unreadable. I have image files on a number of different servers on the Internet: printing services like Walgreens and Shutterfly that allow you to store your images and print from them whenever you want, social media like Facebook and Twitter, and dedicated photo storage sites like Google Photos and Amazon Photos. Most of my images are in JPG format, although some are TIFF. The various cameras I've used over the years sometimes have different file naming conventions, sometimes the same. This can result in having several different photo files with the same name, which can in turn lead to photos being overwritten or excluded from a group upload. There are often multiple copies of the same file, and sometimes multiple photographs of the same thing with only the slightest of variations. iPhone Live Photos can be backed up with or without the motion intact. Different devices provide different levels of metadata. All give the date and time taken at least but if the camera's clock wasn't programmed the dates will be inaccurate. Even some of the old prints have descriptions or dates written on the back, but some of the writing on the old family photographs is barely legible.

I've taken us on this long digression in order to showcase the challenges of data curation. I am now trying to come up with a comprehensive plan to preserve all my photographs and make them findable and accessible for myself and anyone that I might want to share them with. I'm sure most of you are dealing with the same situation with your photos. When we talk about the data curation of research, all of the complications that I have described in regards to my photo collection apply. The complications that come

with are even more extensive. After all, most consumer cameras save in only a handful of formats; research data can take a myriad of forms.

Data curation guru Lisa R. Johnson says that the practice is about:

"…applying the archival principles of library and information sciences to a wide variety of data objects from all disciplines and prepare them for ingest, access, and long term preservation within an environment that facilitates discovery and access while not diminishing their content, authenticity and value" (Johnston L. R., Curating Research Data. Volume One: Practical Strategies for Your Digital Repository, 2017)

There is quite a bit of jargon in there but all of these concepts were in our initial discussion about photographs. Data objects are our photos; ingesting is uploading them to a server (called a repository), discovery is the ability for us or others we allow to find specific photographs that we are interested in, and access is the ability to view and/or download the photographs. "Content, authenticity, and value" = the pictures are intact, labelled correctly, with duplicates and erroneous and redundant shots removed.

Research data can come in many forms including measurements from scientific instruments, observations, surveys, computer programs (the output and/or the application), media files, handwritten notes, and so on. Any material that the researcher has generated in some way to support the thesis of their study.

Research data is often classified into three groups, according to the feasibility of reproduction: observational, computational, and experimental (National Science Board, 2005). Observational data is data that has been collected at a certain point in time and can never be reproduced. Climate readings for a particular date cannot be retaken. When a researcher observes wildlife and makes field notes it is impossible for that researcher or anyone else to go back in time to the occasion of the research and make new notes. Computational data consist of computer programs. The output may or not be important, but the program itself needs to be kept in a state so that another researcher can enter their own data or the same data and get results; this often requires information about hardware and operating systems, as well as a plan for migrating to a new platform if the program is meant to be viable indefinitely. Experimental data is that gathered through an experiment. Some experiments, like subjecting a material to a stress test, are reproducible. Others, like many social science experiments, rely on an irreproducible set of conditions and participants.

The type of data being dealt with determines how it is to be handled, what exactly needs to be loaded into a repository and what ancillary information is necessary to make future use of it. With observational data it is especially important to follow proper archival practice, because once the data is lost there is no regaining it. "Data rescue" is the term used to describe efforts to try to save data that is disappearing due to media failure (which can be anything from paper stock or VHS tape disintegrating to CDs "rotting" or simply the scarcity of hardware to read obsolete media). Historical handwritten climate records are disappearing at an alarming rate due to the fragility of paper (Eveleth, 2014); even more alarming are the threats by ideological groups to send any data that challenges their worldview down the memory hole (Rosen, 2017). Making sure that data is properly backed up and redundantly stored is crucial. Efforts to replicate government data on other "rescue servers" is an attempt to guard against deliberate erasure of inconvenient data (Lief, 2017), even as guerilla data curation in the field works to save data from crumbling media (Eveleth, 2014).

As was the case with the photographs, advances in information technology have caused the amount of data that is stored, the number of places that it can be stored, and the ability to share with colleagues to explode exponentially. The fact that data is now so easily accessible has resulted in publishers and grant endowing institutions creating policies for data sharing for authors and grant recipients. Since data can now be shared with researchers down the hall or halfway around the world, the publishers and grant funders are demanding that researchers make the data necessary to reproduce their results or base new studies on easily accessible. The data curator's job is to aid the researcher is making their data discoverable and accessible, which involves more than just dropping it in an open folder online.

The curators job is to act as a go between facilitating the transfer of data between the researcher and the repository or repositories it ends up in. Often a curator will be the first person other than the researcher to examine the data, therefore in the first stages of the process they act as a proofreader. Often we are unable to catch mistakes in our own writing because we know what we were trying to say, and our brains will try to be helpful by filling in the missing blanks as we read, rendering errors invisible to us. A second set of eyes can pick out mistakes and omissions. With data, the researcher may think that documentation is appropriate and clear, while the curator may note that the readme file does not give enough information to be able to run a piece of software that is being archived, or that a spreadsheet has cryptic labels that make little sense to anyone who isn't working in the same lab. It is the curators job to work with the researcher to clear any matters like this up. While many people may think of the data curator as a "techie" position, there is a definite need for people skills and the ability to communicate effectively yet tactfully. Data management is typically not the biggest priority for a faculty member, and getting all the necessary answers from them involves respecting their time and asking succinct questions.

It helps to explain at the outset the value that good data management provides not only to the scholarly community but to the researchers themselves. Having their data accessible means that others will be able to verify their research, increasing the trustworthiness of any articles based on the data. Proper identifiers, metadata and repository indexing not only improve discoverability but and more use of the data but also more citations of the data, thus boosting the researcher's impact factors.

Many researchers recognize the value of data curation, but feel that it is too daunting a task. A study by the Data Curation Network found that many researchers felt that they were not being given the support necessary in order to satisfactorily carry out the data curation activities that they found most important. The DCN concluded that "These may be areas of opportunity for libraries to invest in new services and/or heavily promote new services that may already exist but are not reaching the researchers who value them." (Johnston L. R., et al., 2018) Assisting researchers with their data is yet another way for libraries and librarians to maintain relevance in an era when advances in technology can quickly make many human activities redundant. While many aspects of curation can be reduced to algorithms, the crucial element of being able to communicate with a client and grasp their needs will probably be best done by humans for years to come. There is a reason that some music services advertise that their playlists as "curated", i.e. compiled by people; sometimes only a human can understand another human.

The goal of the curation process is to create data sets that are FAIR: Findable, Accessible, Interoperable, and Reusable (Wilkinson, et al., 2016). These principles were defined by a group representing academia, industry, grant funding bodies, and publishers. The group met at an International Loretz Center workshop in 2014 in order to hammer out a set of best practices that would make sure that data was being utilized in a way that would increase its usefulness (Wilkinson, et al., 2016). It has quickly become an unofficial standard and has become second nature to many working in the field.

The F in FAIR stands for findability. The essential factor in findability is a persistent identifier. DOIs are by far the most common identifiers, although there are other standards. DOIs are typically assigned by publishers when an article and its related data set are published, but there are other providers such as DataCite and the Open Science Framework. A DOI ensures that each dataset has a unique thumbprint so that can always be used to access it, no matter how many links pointing to it are broken. The other essential factor in findability is indexing. To be findable a dataset must reside in a repository that is indexed and searchable, either by web search engines or more specialized data search engines.

A is for accessible. Accessible data is data that resides on a server that is reachable through what the FAIR guidelines call a "standardized communications protocol"; in most cases this will be the Internet. In order for data to be useful, researchers need to be able to not just find it, but also to download it over an open network. The information should be free and not behind a paywall.

The I stands for Interoperability. The guidelines state that "data use a formal, accessible, shared, and broadly applicable language for knowledge representation." This refers to the use of recognized metadata standards. For data to be useful in much be described in such a way that anyone who has any familiarity with the discipline will be able to understand what the data is, and at least the basic details about its scope, how it was created, who created it, and how to best use it. This in turn requires the use of a standard descriptive language, either a general purpose metadata schema like Dublin Core, or one of the many specialized metadata standards. A list of metadata standards for different fields can be found at http://www.dcc.ac.uk/resources/subject-areas/general-research-data .

The last guideline is R, which stands for Reusable. The data must have a clearly understood license which grants permission for the data to be used by others. In most cases this will be a Creative Commons license. All of the four Creative Commons licenses grant permission for reuse of data with attribution, however they differ in whether the data can be modified, if it is available for commercial use, and if any resulting research must also be under the same terms (Network, 2018).

The Data Curation Network was founded through an ILMS grant in 2016. The purpose of the group is to promote awareness of good data curation practice as well as to set up a network of data curation support between libraries, so that curators in different institutions will be assist each other according to their respective expertise (Johnston L. R., et al., 2016). To further those goals DCN has been working on standardizing the data curation process.

DCN has developed a workflow for the data curation process called CURATE (Data Curation Network, 2018). This workflow describes the process through which a curator works with a researcher to prepare data for ingest into a repository. The letters stand for

- C – Check files and read documentation
- U – Understand the data
- R – Request missing information
- A – Augment documentation and metadata
- T – Transform file formats for reuse
- E – Evaluate for FAIRness

Checking involves examining the files contained in the data set and asking questions. What types of files are there? What applications are they affiliated with? If the application is not readily available are there

viewers available to check the content of the files? Are the files intact? Depending on the type of research data there may be extra steps that need to be taken. If the files are code, can they be run? If the files are the result of medical or social science research has any sensitive data been removed and has the data been sufficiently anonymized? It is not always as easy as it might seem. One frequently quoted statistic claims that 87% of Americans can be identified by zip code, gender, and date of birth (Sweeney, 2000). This step also involves locating and reading any documentation.

Understand is the next step. Is the curator able to understand the dataset? In the case of a spreadsheet, are idiosyncratic or ambiguous labels used? Is there adequate documentation? Is the documentation where it should be? For example, if you are dealing with code, is all the information necessary to run it in a readme file? Or is some of it hidden in comments within the code itself?

In an ideal world every data set would be matched to a curator with expertise in the data's subject area. In the real world, however, curators are often forced to deal with data that they are not familiar with. There are resources, called Data Curation Profiles, which can assist in those cases (Witt, Carlson, Brandt, & Cragin, 2009). A Data Curation Profile is a document that describes the sorts of data that are used in a particular subject area and tells a curator how to work with it. These are assembled by curators or subject liaison librarians who interview researchers about their data and then document the findings. There is a Data Profile Toolkit at http://datacurationprofiles.org/ that gives detailed instructions on how to compose these reports. The toolkit includes templates and guides for each stage of the process. The website also has a directory of existing profiles.

After the curator has checked the nature of the files and documentation and has attempted to understand the dataset the next step is to Request further information from the researcher based on questions that were generated during the Check and Understand steps. This step can be difficult and time consuming, so it is best to simplify as much as possible. It is usually a good idea to contact the researchers via email so that a record of the conversation exists that can be added to your documentation at the end of the process. If there are tasks that you can do yourself, ask permission and carry them out, rather than asking the researcher to make the changes and then having to wait for them to be completed. For example, if one of the problems with the data is that the column headings in a spreadsheet are ambiguous, suggest a change to standardized ontology and offer to do it yourself.

Augmenting the dataset consists of carrying out any changes that have been signed off on by the researcher, creating metadata to accurately describe the data or checking and correcting existing metadata, and standardizing the data and metadata; for example, making sure that all dates conform to ISO 8601 (YYYY-MM-DD). Lastly a curation log should be compiled that lists everything that has been done with the data and the supporting correspondence with the researcher.

All files in the dataset should be Transformed to archival formats if necessary. The formats should be open, not proprietary (for example, CSV rather than XSLX), widely adopted to ensure longevity, and independent (not tied to any particular software, hardware, or operating system that could cease production leaving the files orphaned). Media files should be in lossless formats like FLAC or TIFF rather than compressed formats like MP3 or JPG to maintain fidelity even in the case of future migration. The Library of Congress has recommendations for formats that meet these criteria here: https://www.loc.gov/preservation/resources/rfs/ .

The last step is to Evaluate for FAIRness; ensuring that the data set is Findable, Accessible,

Interoperable, and Reusable as described above. If the data set does not have a persistent identifier that will need to be corrected, the researcher will have to make a decision on which Creative Commons license to accept, and a suitable repository or repositories will need to be found that indexes content and makes it available online. A database of general and subject specific data repositories can be found here: https://www.re3data.org/ .

To sum up, data curation is an activity that requires a combination of technical and people skills; to be a successful curator an individual needs to be able to evaluate digital files yet also be able to work with and enlist the aid of researchers in correcting those files. Curation is a means toward providing usable data to the world in order to facilitate more efficient and insightful research; by following the CURATE model and creating FAIR data sets the curator is engaging in the best practices to ensure the perpetuation and continued usage of the data in their care.

References

Data Curation Network. (2018). Specialized Data Curation Workshop 1. Las Vegas. Retrieved from https://sites.google.com/site/datacurationnetwork/workshops/workshop-1---october-17-18

Eveleth, R. (2014, August 25). The Quest to Scan Millions of Weather Records. The Atlantic. Retrieved 8 16, 2018, from https://www.theatlantic.com/technology/archive/2014/08/the-quest-to-scanmillions-of-weather-records/378962/

Johnston, L. R. (2017). Curating Research Data. Volume One: Practical Strategies for Your Digital Repository. Chicago: ALA.

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication., 6(1), eP2198. doi:https://doi.org/10.7710/21623309.2198

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, Heidi, Kozlowski, W., Olendorf, R., & Stewart, C. (2016). Grant Narrative for the Data Curation Network project.

Lief, L. (2017, March 7). Universities Race to Safeguard Government Data Under Trump. Columbia Journalism Review. Retrieved from https://www.cjr.org/covering_trump/government-datapreservation-trump.php

National Science Board. (2005). The Elements of the Digital Data Collections Universe. In N. S. Board, Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century. Alexandria, Virginia: National Science Foundation. Retrieved 10 29, 2018, from https://www.nsf.gov/pubs/2005/nsb0540/nsb0540_4.pdf

Network, T. C. (2018, 10 26). Share Your Work. Retrieved from Creative Commons: http://creativecommons.org/share-your-work

Rosen, J. (2017, April 27). Turbulence ahead: US Science Faces a Political Storm, and Early-Career Researchers Should Prepare Themselves. Nature, 544(7651). doi:doi:10.1038/nj7651-509a

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Pittsburgh. Retrieved 10 10, 2018, from http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3. doi:doi:10.1038/sdata.2016.18

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. (2009, December 7). Constructing Data Curation Profiles. International Journal of Digital Curation. doi:https://doi.org/10.2218/ijdc.v4i3.117