**Module 2: Types, Formats, and Stages of Data**

Jen Ferguson (Northeastern University)

Learner Objectives:

1. Explain what research data is and the range of data types

2. Identify stages of research data

3. Identify common potential storage formats for data

4. Identify relevant quality control techniques/technical standards

5. Identify methods of recording data that are specific to researchers’ disciplines and research interests

**What is research data?**

Data is often thought of in quantitative terms. Much research data is indeed quantitative, but ‘number’ and ‘data’ are not synonymous. Data is typically considered in two broad categories: quantitative and qualitative.

Quantitative data can include experimental measurements, e.g. lab instrument data, sensor readings, survey results, and test/simulation models. Qualitative data can include text, audio, images, and video. Some definitions of data are quite broad, and include objects such as laboratory specimens.

**What types of research data are associated with major disciplines?**

In social sciences disciplines, data can include survey responses, video, economic indicators, demographics, opinion polling, climate change, literacy data, and/or data related to topics such as crime and transportation.

Data in the ‘hard’ sciences can include measurements generated by a sensor or laboratory instrument, computer modeling, simulations, images and video, observations, and/or field studies.

Of course, research data is not limited just to these two broad disciplinary areas. For example, digital humanities scholars may generate research data by text mining historical documents, analyzing the prevalence of particular words.

While many types of data are generated in the course of research, existing data can also be used in research projects. Examples include social scientists using census data for their studies, or scientists comparing published gene sequences from various organisms.

Some types of research data and data files are fairly ubiquitous and can be found distributed across disciplines. Images and video are data types that are widely generated and used by researchers regardless of discipline. Mapping/GIS data is also used across a wide range of disciplines. Similarly, researchers in many disciplines make use of spreadsheets to store, manipulate, and visualize data.

**What are the stages of research data, and how do these stages relate to the** [**research data life cycle**](http://www.data-archive.ac.uk/create-manage/life-cycle)**?**

Read over the research data life cycle tutorial on the UK Data Archive site: <http://www.data-archive.ac.uk/create-manage/life-cycle> and Ball, Alex. 2012. “Review of Data Management Lifecycle Models.” <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>.

For context, we’ll illustrate the various stages of research data by using this purely hypothetical research question: Water temperatures in Lake Superior are now significantly warmer than in previous years. Is there evidence to link this to global warming?

Raw data

What is being measured or observed? This is the data that is being generated during the research project. An example of raw data in our hypothetical research question might be daily measurements of temperature in Lake Superior. Raw data coincides with the [creating data](http://www.data-archive.ac.uk/create-manage/life-cycle) section of the research data life cycle.

Processed data

How can the raw data be made useful/manipulable?

To continue with the example above, lake temperature data may become processed once researchers remove clearly erroneous temperature measurements from the data set, and enter the remaining temperatures into a spreadsheet for manipulation and analysis. Processed data coincides with the [processing data](http://www.data-archive.ac.uk/create-manage/life-cycle) section of the research data life cycle.

Analyzed data

What does the data tell us? Is it significant? How so?

For example, we could analyze daily lake temperature data by finding average temperatures, looking at seasonal fluctuations, and generating graphs that demonstrate these changes. Analyzed data coincides with the [analyzing data](http://www.data-archive.ac.uk/create-manage/life-cycle) section of the research data lifecycle.

Finalized/published data

How does the data support your research question?

For example, a plot of our average lake temperatures for 2013 may show statistically significant differences when compared to the same data from 1913 and 1963. Finalized/published data coincides with a few sections of the research data lifecycle including [preserving, giving access to, and re-using data](http://www.data-archive.ac.uk/create-manage/life-cycle).

Note that our lake temperature scenario is a very simplified example. In many cases, raw data from different sources will be collected, combined, and analyzed in support of a research question. Data does not always flow simply from raw to published stages in a single, linear direction. For example, raw data from one experiment may be repurposed for analysis in another. Processed or raw data may also be subjected to multiple rounds of different methods of analysis, whether by original researchers, collaborators, and/or by those they share data with.

**What format characteristics and file types are best choices for long-term access to data?**

These file format characteristics ensure the best chances for long-term access:

* Non-proprietary
	+ Non-proprietary or open formats are readable by more than just the equipment and/or program that generated it. Sometimes proprietary file formats are unavoidable. However, proprietary formats can often be converted to open formats. Please see the following section for more on this topic.
* Unencrypted / uncompressed
	+ Unencrypted and uncompressed files offer the best prospects for long-term access. If files are encrypted and/or compressed, the method of encryption/compression used will need to be both discoverable and usable for file access in the future.

Here are some examples of preferred formats for various data types (from <https://lib.stanford.edu/data-management-services/file-formats>):

Containers: TAR, GZIP, ZIP

Databases: XML, CSV

Geospatial: SHP, DBF, GeoTIFF, NetCDF

Moving Images: MOV, MPEG, AVI, MXF

Audio: WAVE, AIFF, MP3, MXF

Numbers/statistics: ASCII, DTA, POR, SAS, SAV

Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP

Text: PDF/A, HTML, ASCII, XML, UTF-8

Web Archive: WARC

For a list of common file formats and evaluations of format quality and long-term sustainability see <http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml>

Some archives specify the optimal data formats they use for long-term preservation of data. One example is the UK Data Archive file formats (<http://data-archive.ac.uk/create-manage/format/formats-table>).

This table contains guidance on file formats accepted by the UK data archive for deposited data.

**How can I help ensure long-term access to my research data files? What if I have to use proprietary formats?**

Proprietary systems and file formats can resist attempts at data integration, reuse, and sharing. These barriers can often be addressed by converting proprietary formats to open formats. The protocols and solutions for doing so can be discipline-specific (e.g. <http://loci.wisc.edu/bio-formats/formats>), but some general guidelines apply.

In file format conversion, information can be lost. In one fairly common example, converting large image files from TIFF format to JPEG (or JPG) format results in a much smaller file size, but at the expense of information lost from the file. This lost information may be visible as a decrease in the converted file’s quality (as is the case when TIFF images are converted to JPEG). Format conversions may also result in lost contextual information such as creation date, creator, system, or instrument. You may have heard the term ‘metadata’ before; this contextual information is a type of metadata.

Data may need to be converted from the original format to a preferred data preservation format in preparation for long-term storage. When data is converted from one format to another - through export or by using data translation software - certain changes may occur to the data:

• for data held in statistical packages, spreadsheets or databases, some data or internal metadata such as missing value definitions, decimal numbers, formulae or variable labels may be lost during conversion to another format, or data may be truncated; and

• for textual data, editing such as highlighting, bold text or headers/footers may be lost.

After conversion data should be checked for errors or changes (see <http://data-archive.ac.uk/create-manage/format/formats>).

Since information can be lost in format conversion, when it is necessary to convert or migrate one file format to another, it’s important to carefully note the steps taken during the conversion/migration. Whenever possible, it is also best practice to keep the original file as well as the converted one, especially for data that would be problematic or prohibitively expensive to reproduce.

As in our previous example with an original image in TIFF format, you might choose to convert and/or compress this image, perhaps just to save hard drive space, or to make a version of the image that will load more quickly on a web page. When compressing this file, it’s best to a) choose open compression software programs and b) make note of which program – and even which specific version of the program - was used to compress the image. Whenever possible, it’s best to keep a copy of the image in the original TIFF format as well. This original can be kept in a separate archive if the goal is to save hard drive space.

**How can I help maintain quality and standardization of my research data, and make it easy to understand and interpret?**

Data documentation explains the who/what/where/when/why of data:

* Who collected this data? Who/what were the subjects under study?
* What was collected, and for what purpose? What is the content/structure of the data?
* Where was this data collected? What were the experimental conditions?
* When was this data collected? Is it part of a series, or ongoing experiment?
* Why was this experiment performed?

Good data documentation helps you, the researcher. Clear documentation makes it easier to interpret your findings later, helps facilitate collaboration, sharing, and reuse, and can also help ensure successful long-term preservation of your research findings.

Documentation takes place at the study level – that is, the level of your research question itself – as well as the more granular data level.

Study level documentation includes contextual information about the project – what are its goals and hypotheses? It also includes information on investigation design such as sampling, methods of data collection, and time series and geographic range used. Any information on confidentiality and access restrictions to the project data would also be part of this study-level documentation.

For documentation at the data level, a best practice is to embed this documentation somewhere within the data file itself. This approach ensures that the data description is always collocated with the data. Documentation could be placed in a readme file, or included in a spreadsheet containing the data.

For more on data-level documentation, please see these examples:

[Documentation for structured, tabular data (e.g. Excel, SPSS, GIS)](http://www.data-archive.ac.uk/create-manage/document/data-level?index=1)

[Documentation for qualitative data](http://www.data-archive.ac.uk/create-manage/document/data-level?index=2)

**What practices are used by other disciplines to document data?**

Data documentation practices vary by discipline. These methods include lab notebooks [data dictionaries and codebooks](http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html) in the social sciences, and well-documented/commented code for computer science (or for really any project that uses code and/or scripting).

While formats and methods for documentation differ, the general idea is always to describe:

* what the data is
* when the data was collected
* where it was collected
* how it was collected
* notes about the data characteristics (including file formats/potential format conversions), and
* any pertinent notes about experimental conditions.

Note that for collaborative research projects, it’s important to come to some agreement among members of the project team that will help ensure consistent data documentation practices by all. This can be particularly challenging when team members are geographically dispersed. While specific data documentation practices vary greatly from one discipline to the next, one aspect is fairly universal: file-naming and organizational conventions. Well-organized and clearly named files can help ensure better organization and more efficient retrieval. It’s difficult to recommend hard and fast rules in this area, as different methods work for different people. A few disciplines also have prescribed file-naming conventions.

Generally speaking, a good protocol is to give files meaningful, descriptive names, while avoiding certain problematic characters (such as ampersands) and spaces. Overly long file names (>25 characters) should also be avoided, as these may not cooperate well with different operating systems. Consider, too, strategies you might use for version control; how will you keep track of which version of a file or document is the most current? Version control can be particularly challenging for collaborative projects in which several people may have access to, or even the ability to, change, a file. Often multiple people are collaborating and it can be difficult to establish provenance of versions. If each older version is not saved or if it is merged into new version, valuable data could be lost.

Please see the resources below for more detail on:

Data organization and naming conventions:

<http://libraries.mit.edu/guides/subjects/data-management/organizing.html>

<http://www.records.ncdcr.gov/erecords/filenaming_20080508_final.pdf>

Methods of version control:

<http://chronicle.com/blogs/profhacker/consider-revision-control-methods-for-documents/25683>

***Module 2 Activities***

***Activity 1: Please read this case study\* and answer the following questions.***

Researchers were interested in impacts of potential future climate change and coastal flooding on two Massachusetts communities. These communities would both be affected by coastal flooding, but in one community flooding would mostly affect residential areas, while the other community’s commercial areas would be most vulnerable. Researchers held a series of workshops with residents of both communities to explore possible responses to increased coastal flooding resulting from sea level rise. Researchers designed the earlier workshops in the series to gain understanding of the residents’ existing knowledge of climate change, with the goal of linking this knowledge to the communities’ potential to adapt to climate change in later workshops. The workshops were organized by local community and non-governmental organizations, and were held over the course of several evenings. Audio recordings were made of the workshops with the consent of the participants.

The goal of the first two community workshops was to better understand the residents’ existing understanding of climate change. In the first workshop, residents were asked to list all words they could think of when they considered the term ‘climate change’. The words were called out by residents and recorded on flip charts by organizers. These words included terms related to environmental impacts (e.g. droughts, deforestation, global warming) as well as public health impacts (e.g. pollution, famine, epidemics). Residents were then asked to vote on the words they thought represented the potential climate change impacts that were most important. This voting process reduced the word list to 47.

In the second workshop, participants were given a pile sorting exercise. They were provided with the 47 words generated from the previous workshop, each on a separate piece of paper, and asked to sort the words into piles of related terms. After the workshop, the pile sorting results were entered into a spreadsheet. Multidimensional scaling analysis was performed on this data set, and the residents’ pile sorting results were used to generate a model of perceived proximities between terms. The ‘relatedness’ of these terms was then visualized.

In the third and final workshops, the researchers presented potential 100-year coastal flood maps of the communities that had been generated based on various emissions scenarios for the years 2030 and 2070. Researchers then presented various options that would address sea level rise and help residents adapt to coastal flooding in their communities. Option choices were informed by the residents’ knowledge of climate change as elucidated in the first two workshops. These options included a modular sea wall, construction of natural buffers such as a beach/dune system, evacuation plans and routes to move residents away from the flood-prone areas of their communities, and various types of structural flood-proofing. After the presentations, residents were divided into four groups. Each group was assigned a moderator/facilitator who transcribed responses to these discussion questions:

1. Which of the options seems most feasible/attractive?

2. Which options would you object to and why?

3. What obstacles are in the way to getting the options in place?

4. What needs to happen in the community to adapt to sea level rise?

Following the workshop, researchers consolidated and summarized the discussion responses, and looked for similarities and differences in responses between the communities. They also evaluated residents’ levels of preparedness and engagement, based in part on outcomes of the word generation and pile sorting exercises.

***Questions:***

*1. Briefly list the data generated and used in the course of this study.*

*2. What file types/formats might be involved in the study?*

*3. For each of the items on your data list, match it to the data stage/life cycle continuum described in the module.*

\* Case study adapted from Douglas, E.M., P.H. Kirshen, M.J. Paolisso, C. Watson, J.Wiggin and M. Ruth. 2012. Coastal Flooding, Climate Change, and Environmental Justice: Identifying Obstacles and Incentives for Adaptation in Two Metropolitan Boston Massachusetts Communities, Mitigation and Adaptation Strategies for Global Change, Vol. 17, No. 5, pp. 537-562.

***Activity 2: This screenshot depicts various files generated & assembled by a researcher working on a journal article.***



***Questions:***

*1. Would you organize these files differently? If so, please briefly describe your method.*

*2. What suggestions would you make to this researcher about his or her file naming and organization practices?*

Readings

1. University of Edinburgh. Research Data Management Guidance.

<http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-mgmt/research-data-definition>

2. University of Virginia Library. Steps in the Research Life Cycle, Scientific Data Consulting.

 <http://www2.lib.virginia.edu/brown/data/lifecycle.html>

3. Ball, Alex (2012). Review of Data Management Lifecycle Models.

 <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>

4. UK Data Archive. Research Data Life Cycle.

 [http://www.data-archive.ac.uk/create- manage/life-cycle](http://www.data-archive.ac.uk/create-%20manage/life-cycle)

5. MIT Libraries. File Formats for Long-Term Access.

<http://libraries.mit.edu/guides/subjects/data-management/formats.html>

6. UK Data Archive. Formatting your Data.

<http://www.data-archive.ac.uk/create-manage/format>

7. Stanford University Libraries. File Formats for Long-Term Storage.

 <https://lib.stanford.edu/data-management-services/file-formats>

8. Library of Congress. Sustainability of Digital Formats.

 <http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml>

9. UK Data Archive. Data Formats Table.

 [http://data-archive.ac.uk/create- manage/format/formats-table](http://data-archive.ac.uk/create-%20%20manage/format/formats-table)

10. OME. Bio-formats.

 <http://www.openmicroscopy.org/site/support/bioformats4/supportedformats.html>

11. UK Data Archive: File Formats & Software.

 <http://data-archive.ac.uk/create-manage/format/formats>

12. UK Data Archive. Qualitative and Tabular Data.

 <http://www.data-archive.ac.uk/create-manage/document/data-level?index=1>

13. ICPSR. Best Practices in Creating Metadata.

 <http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html>