7
Interlude: Collecting and managing data

Data is information. Data is the result of somebody's efforts to record and store information, often to provide an opportunity for insight. It can be used to discover patterns, test hypotheses, and support arguments, among other things. But the numbers themselves often cannot convey much meaning – it is through manipulation and interpretation of the data that those uses can be realized. We therefore need to be conscious about how data are structured and managed, so that they can be manipulated and interpreted to reveal insights. If data are badly structured and managed, the risks can be great. Optimistically, we risk wasting lots of time trying to restructure data to allow the kind of analysis we wish to perform; worse yet, we risk compromising the integrity of or, heaven-forbid, the complete loss of data through poor structuring and mismanagement.

First, lets be clear about what is meant by data structure and data management:

- **Data structure** refers to the organization and layout of data as it is stored. Whether it is handwritten in notebooks or stored in spreadsheets or text files, data usually has an architecture that reflects the intentions (or ignorance) of the data manager or sometimes the protocols of his/her organization or institution.
- **Data management** is the set of practices aimed at preserving the quality, integrity, and accessibility of the data. This can include all phases of data usage from collection and manipulation to storage, sharing, and archiving.

There are some great resources for data management out there in various forms, including some geared toward biologists. A few great ones are:

Data Carpentry www.datacarpentry.org.

Wickham, H., 2014, Tidy Data. *Journal of Statistical Software* 59(10).

Saltz, J.H. and J.M. Stanton, 2017, *Introduction to Data Science*, Sage publ.

7.1 Who is data for?

Unless you work with highly classified or proprietary information and are required to protect and encode your data, you likely need data to be readily understood and usable, not just to you but to others you work with or the broader public[1]. But we also need to realize

[1] Many government funding agencies such as the National Science Foundation, US Department of Agriculture, and National Institutes of Health now require that researchers develop a data management plan that includes strategies for structuring, archiving and indexing data in publicly-accessible repositories.

that the humans who need to make sense of the data will be using tools like computers to facilitate this approach. Therefore the data structure needs to also accommodate the demands of the computer hardware and software that it is used on as well as the humans. Thus, data should be organized in a logical, self-consistent way and it should be accompanied by documentation that helps to explain the content and and context of the data. Similarly, accessible data archiving, in principle, allows colleagues and competitors to test, verify, reproduce, and/or compare results with their own, ensuring that scientific advances that you make with the help of your data can also lead to advancement of science and management more broadly.

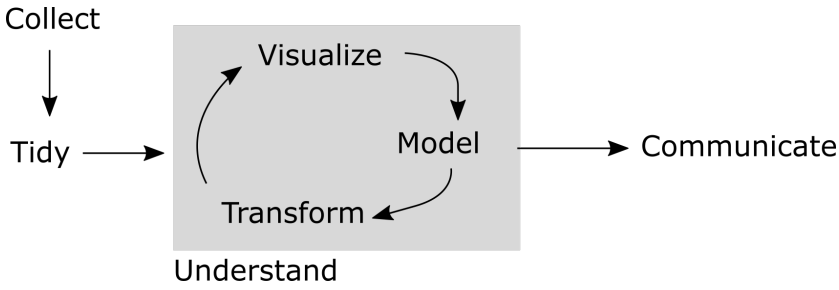

Figure 7.1: A typical workflow for data. After Grolemund and Wickham, 2017, *R for Data Science*, O'Reilly.

Consider the schematic workflow illustrated in Figure 7.1. Once collected, data must be organized and formatted in a way that facilitates their analysis on a computer. A popular term for data that are formatted to simplify computer manipulation is **tidy data** (more on this below). When this process is complete, the data may be analyzed as needed to address the problem or hypothesis at hand. This process of making sense of the data may then produce a result that needs to be communicated back to humans. When data is presented for consumption by the human eye and brain, the organization and structure should reflect the expectations and attention span of the humans. Unless the data set is small, the raw or transformed data may not be appropriate to display. Instead, summary data are more appropriate, either in narrative, table, or graphical format.

7.2 *Tidy data*

To understand the significance of tidyness, it is perhaps helpful to consider untidy or kludgy data. Below is a portion of a data table containing the weights and lengths of small fish captured during a population survey of Inch Lake, Wisconsin[2]. Let's unpack this data set. There are two different fish species listed, one observed both in 2007 and 2008, and one only in 2007. Lengths and widths are provided for all fish measured, but there are different numbers of

[2] This is a deliberately disorganized snippet based on a dataset from Derek Ogle's neat website http://www.derekogle.com/fishR/data

Table 7.1: A portion of an untidy dataset.

number	bluegill (2008)	bluegill (2007)	Iowa Darter (2007)
1	L:1.5 W:0.7	L:1.9 W:1.3	L:2.1 W:0.9
2	L:1.0 W:0.7	L:1.6 W:1.3	L:2.0 W:1.3
3	L:2.6 W:1.5	L:2.4 W:1.7	L:1.7 W:0.7
4		L:1.1 W:0.6	

fish in each column. There is also an index value in the first column that facilitates counting the number of fish of each species captured in each year. This is reasonably straight-forward for a human to interpret, particularly if we are told that L corresponds to a length in inches and W corresponds to weight in grams. However a larger dataset organized like this table would be miserable to analyze for a variety of reasons, including:

- columns don't have the same number of values
- the same species has data across multiple columns
- two variables (length and width) are listed together within each column, with numbers and letters mixed

One instructive question to ask is how many variables there are here. We notice that data span multiple years, so year could be a variable. There are also multiple species here, so species could be viewed as a variable. Then length and width should each be variables. Finally, if we wish to have an index or ID number for each fish, that might be a fifth variable[3]. In general, tidy data is organized in a rectangular array in which each column represents a variable and each row an observation. In most cases, the first row contains descriptive but simple column headers. This simple prescription seems unthreatening, but it is often surprising how pervasive untidy data is.

So with five variables, how many observations do we have? Each fish represents an observation, with a unique ID number, year of capture, species, weight and length. From Table 7.2, there appear to be ten fish listed among the three columns. so since each fish is an observation. According to the principles of tidy data, then, there should be ten rows of data with values in each of the five columns. Below is a tidy representation of this dataset. This table is now organized in a way that can easily be sorted, filtered, and summarized in common statistics and computational software packages.

[3] If there are multiple datasets derived from the same group of fish, assigning a fish ID number would be a simple way to connect these datasets using database methods.

Tidy data has:

- one column for each variable
- one row for each observation
- a header row

7.3 *Data management*

Because data is often the hard-won result of costly and time-consuming observations and measurements, its management should be deliber-

fishID	year	species	weight (g)	length (in)
1	2008	bluegill	1.5	0.7
2	2007	bluegill	1.9	1.3
3	2007	Iowa darter	2.1	0.9
4	2008	bluegill	1.0	0.7
5	2007	bluegill	1.6	1.3
6	2007	Iowa darter	2.0	1.3
7	2008	bluegill	2.6	1.5
8	2007	bluegill	2.4	1.7
9	2007	Iowa darter	1.7	0.7
10	2007	bluegill	1.1	0.6

Table 7.2: The data from Table 7.2 transformed to a tidy dataset.

ate and careful. Well-managed data can be stored, retrieved, analyzed, and used to develop insights or aid in management decisions without compromising the data itself, and without spending unnecessary time and energy in decoding and interpreting the raw data. Thus, proper management entails not only careful structure as described above, but also well-organized storage and full documentation.

First, important raw data should be stored redundantly. If it's only in hard-copy (e.g., in field notebooks), consider scanning or transcribing the hard-copy data to preserve a digital version that can be backed-up regularly. When the data results from original research that can be shared publically, it can be uploaded to data repositories[4] A complete set of data should be archived and never modified, while data reduction and analysis are done on copies of the formatted raw data.

When analysis or reduction occurs much later than the time of collection and storage, or is done by a different person or group than the researcher who collected the data, adequate documentation or metadata is essential. Metadata can include narrative descriptions of where the data was collected, when and how it was collected, and should include references or links to any published or publically-available research or information stemming from the data. The metadata should always include a data dictionary, fully describing the quantities represented by each of the variables collected (i.e., variable name, symbol, units, and procedural statement). These guidelines ensure that data remain safe, useful, and accessible.

[4] Data repositories like the LTER Data Portal require formatted data and metadata to ensure long-term accessibility and adequate documentation.