Data Legacy Rescue Guide

Luis Francisco Henao Diaz, Freek de Haas, Diane S. Srivastava

Introduction

The goal of data rescue is to retrieve data from a (retiring) senior researcher who is interested in sharing her or his lifework with the scientific community. In order to retrieve and analyze the data, you will be working in close contact with a senior researcher and an intermediate faculty coordinator at UBC. You are responsible for collating and arranging the data in a "long-format" data structure, to be uploaded to an online repository. We suggest you use R to wrangle and reshape the data, however other tools can be useful too.

The data, once you are done, should be "clean" and "tidy", meaning that all variables are clearly defined in a *metadata file*, that you have checked extensively for errors, and that the structure of the data is long-format. Long format refers to a data structure where the rows represent the measurements/ examples and the columns refer to the measures/features/variables of the data. Keep in mind that the final product must be understandable by anyone, meaning the database should become self-explanatory.

The scope of the project can vary greatly depending on the amount, quality and complexity of the data you will receive. This is why you should plan a mid-project evaluation meeting with your faculty coordinator. Here we provide you with (1) a general workflow that will guide you through the basic steps. However, keep in mind that projects vary greatly and that deviations from this plan are acceptable whenever appropriate. At the end of this document, you will find (2) tips based on the results of a pilot study. Also, this tutorial will give you (3) the basic information about the metadata requirements to publish and achieve.

Goals

The goal of data rescue is to ensure that valuable datasets in ecology and evolution are preserved permanently - or at least as far into the future as we could ever imagine. These datasets can be very valuable as historical baselines (i.e. surveys that can be repeated to establish rates of change), as part of a species range projection, as a part of a long term population or evolutionary time series, or as a classic dataset that can be combined with others or re-analysed in different ways to draw out generalities in process and pattern. There are a number of data repositories that provide the infrastructure needed for data to be permanently archived but also discoverable.

This may seem like a straightforward goal, but to do it well, you need to be a bit of a time traveller. First, you have to imagine a researcher in the future that you will never meet, and you need to convey to them all the complexities and subtleties in the data that will allow them to effectively use it. This is rather like being on a desert island, and putting a message in a bottle about your location so that someone could rescue you...all the information has to be on the message (data) in that bottle (repository). You also need to be able to travel backwards in time, to when the data was collected, There may be a set of field notebooks where the data was recorded, full of all sorts of valuable bits of information that cannot be effectively captured in the digital version of the data - so you need to also think about how to make sure to maintain all those links so that other people (including the senior researcher) can cross-reference the original and archived datasets.

Repositories

The end product is a dataset archived on a digital repositories. We recommend using one of the following: Knowledge Network for Biocomplexity (free), Zenodo (free), Figshare (free), Dryad ($100, requires an accompanying publication), Dataverse (free, this is operated by the UBC Library, and soon by all universities across Canada). Note that there are also Canadian government repositories that may be mandatory for government scientists: the Federated Research Data Repository brings together a number of department / ministry specific repositories.

Data and Metadata

You will upload two different types of information on the digital repository: metadata and the dataset. Metadata is a collection of supportive data that provides additional information to a dataset. It accompanies the dataset giving context information about the data that can not be included or inferred from the database. It's a helping document to anyone who wants to use the dataset. The dataset is composed usually of one or more files saved in a bare bones file format such as a tab-delimited text or csv file. Spatial data often requires higher dimensionality, and SQL databases are much more complex, but you will probably not be working with such data. We avoid any proprietary software (e.g. Excel), as there is no provision for this software to be available in the future - and consequently such data is usually not accepted by repositories.

General workflow

1. Familiarize yourself with the data.

   The goal of this step is to prepare you to become an expert in this dataset, This may be challenging since you were not involved in the collection process. Take a look at the complete dataset and all the additional information you have. Be aware of dataset properties (e.g. colours) that can be lost when saving a spreadsheet into a plain text format (.csv). Once you read the data into R, a useful first command is “summary()”.

   Go through the following questions as a preparation for your first meeting with the senior researcher:
   - Can you already find a long-format structure in the data?
   - Are there any missing values?
   - Are there acronyms? Is it clear what they mean?

   Take some time to understand the nature of the data by asking yourself:
   - What was the original intention for collecting this data?
   - How was the data collection performed?
   - How many people were involved in collecting the data? Was it collected and then transcribed or was it directly entered into a database/machine?

   Research the possible repositories, so you can provide the senior researcher with some pros and cons of each.

2. First meeting with the senior researcher

Have a first meeting with the senior researcher on Skype or face-to-face (we do not recommend e-mail). Be aware that the senior researcher might consider his or her data an important part of his lifework: be respectful, and make sure that the researcher always feels in control of the destiny of their data.

Things to discuss on this meeting:

- What is the goal of the senior researcher in archiving their data?
- What are the senior researcher's reservations about archiving this data? Often there are solutions. Data that is still being analyzed can sometimes be embargoed to a future data, either by the repository or by us. Data that might be misinterpreted can be clarified in the metadata or by adding extra columns giving information on data quality.
- Review the questions from the previous step. Are all these points clear to you?
- PI's sometimes have a lot of anecdotal data that they have not documented. Make sure things that are not immediately obvious when looking at the data file (e.g. it was a rainy year causing a batch of missing data). Note this! If necessary incorporate it in the data- or in the metadata file.
- Make a rough draft of the metadata. This should include:
    (1) Dataset description in terms of where, when, who collected the data
    (2) Purpose of data collection
    (3) Publications that used some of all of the data
    (4) For each column: what is the variable measured? What units? How?
- Make sure you have **all** the data belonging to the same project. Do not use any subsets unless the projects are completely uncorrelated. The reason for this: (1) fragmenting data is easy but putting it back together is hard, (2) the same issues that you are fixing in one file can easily be extended with R to be fixed in **all** files.
- If you have relational tables try to identify with the senior researcher the best "key" in order to merge properly. What columns would increase the number of unambiguous matches?
- Which columns should be included? You should include columns that:
    (1) Represent all x and y variables
    (2) Describe how, where, when and by whom the x and y variables were collected
    (3) Link the digital dataset to other versions of the data, especially field notebooks. For example, if a code AMRO was used to represent American Robin in the field notebooks or early spreadsheets, there is value in including both American Robin (fully explanatory name) and AMRO (link to original fieldnotes).
    (4) In general, preserve as much of the data as possible, throw away only information that the senior researcher judges as unreliable, biased, misleading or confusing. Variables that seem unimportant now may be very critical in the future: 40 years ago, recording the timing of budbreak seemed unimportant, now it is critical to understanding how climate change affects phenology.
    (5) Incomplete data is fine, we will simply indicate this with "NA" - see below.

- What repository does the senior researcher prefer - if any preference.
- You will encounter problems as you deal with the data. Therefore establish a way of communication with the senior researcher.. Does the senior researcher prefer meetings once a week or can communications be "on the go"?

3. <u>Merge all data into one single dataframe</u>

You might receive multiple datasets. In this case, it is probably a good idea to read in all the datafiles (ideally they are already in csv format) and as soon as possible merge the files into one single large dataframe. This requires the use of data wrangling functions such as dplyr. Before merging by a "key" column, check for

duplicate values within the columns used to merge; the dataset should increase just in one dimension, not in two.

Useful functions are:

- dplyr::bind_rows(data1,data2).
- tidyverse::left_join(core_data, distance_code, by=c("Distance"="distance.code")].

4. Column cleanup
    - Make sure the columns are recognized by R in its appropriate data format (as.numeric, as.date, as.character, as.factor). Often this requires that you first work on the data in a column.
    - Use table(data$colname) to get a sense of the values that are present in each column. Especially useful for categorical data to find typos. This includes: upper/lower case, spaces, etc.
    - Check for redundant information and preserve just one: different types or formats of coordinates, dates, descriptions (e.g. 4/08/05 and/or 4-Aug-2005). It is preferable to have one value per cell.
    - Replace acronyms or category abbreviations by interpretable data (e.g. "Points" instead of "P"). To that task may be useful to familiarize with the regular expressions in R:
    - Clean up the column (potentially write small functions if the data is more complicated). Replace missing values with NA, e.g. dataframe$colname[dataframe$colname==""] <- NA

5. Outlier detection

Since most data has been entered by hand, typos and outliers won't be unusual. There are several ways to check the data for outliers. Here we provide a couple, depending on the data type:

- *Model-based outlier detection.* If the data is approximately normally distributed then you can define outliers by their Z scores less than or greater than 3. In this case 3 is arbitrarily chosen, but you can use your own cut-off value. Outlier detection is an inherently non-trivial problem so there are no clear rules. Make sure you talk to your senior researcher before removing data!
- *Graphical outlier detection.* If you are working with coordinates try to get a simple summary or plot them and check if they are consistent with localities. Another option is to make a bar-graph or scatter plot of your data. The downside here is that you can only look at 2 variables at a time. Take a look at the scatterplot array options in R.
- *Cluster-based outlier detection.* ML models like DBSCAN cluster your data based on proximity. Some data-points might not get assigned to a cluster, indicating they could be outliers. These are more advanced techniques: K-means, Hierarchical clustering.

6. Output dataframe and write metadata file.

Once the dataset is complete, review once more with the senior researcher that the columns included are relevant (informative) and sufficient for archiving. Finalize the metadata file which should, most importantly, be self-explanatory and complement the data. Once this is done, present your final work to the senior researcher. The last step is to upload the metadata and dataset to the data repository.

Tips

- In case of doubt: always ask the senior researcher first, do not assume anything!
- Keep records of your changes, fixes (e.g. make notes on a notebook or at your cleaning/working code)
- Plan your **mid-project meeting** with the faculty coordinator to ensure that the project is not getting too big.
- Data management: https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
- Regular expressions: https://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf