Data structures for Open Science

Published February 9, 2017

Citation

Brembs, B. (2017, February 9). Data structures for Open Science. *Bjoern.brembs.blog*. https://doi.org/10.59350/f4pjb-9ky40

Keywords

Own Data, Data Structure, Metadata, Open Data, Open Science

For the last few years, we have been working on the development of new *Drosophila* flight simulators. Now, finally, we are reaching a stage where we are starting to think about how to store the data we'll be capturing both with Open Science in mind, but particularly keeping in mind that this will likely be the final major overhaul of this kind of data until I retire in 20 years. The plan is to have about 3-5 such machines here and potentially others in other labs, if these machines remain as 'popular' as they have been over the last almost 60 years. So I really want to get it right this time (if there is such a thing as 'right' in this question).

Such an experiment essentially captures time series with around 70-120k data points per session, where about 3-6 variables are stored, i.e., a total of at most ~500-800k table cells per session, each with 8-12bit resolution. There will be at most about 8-16 such sessions per day and machine, so we're really talking small/tiny data here.

Historically (i.e., from the early 1990s on), these data were saved in a custom, compressed format (they needed to fit on floppy disks) with separate meta-data and data files. We kept this concept of separated meta-data from data also in other, more modern set-ups such as our Buridan experiments. For these experiments, we use XML for the meta-data files (example data). One of our experiments also uses data files where the meta-data are contained as a header at the beginning of the file with the actual time-series data below (example data). That of course makes for easy understanding of the data and makes sure the meta-data are never separated from the raw data, i.e., less potential for mistakes. In another, newer, experiment we are following some of the standards from the Data Documentation Initiative (no example data, yet).

With all of these different approaches over the last two decades, I thought I ought to get myself updated on by now surely generally agreed on conventions for data structure, meta-data vocabularies, naming conventions, etc. I started looking around and got the impression that the different approaches we have used over time are still being used and then some new ones, of course. I then asked on Twitter and the varying responses confirmed my impression that there isn't really a "best-practice" kind of rule.

Given that there was quite a lively discussion on Twitter, I'm hoping to continue this discussion here, with maybe an outcome that can serve as an example use case someday.

What do we want to use these data for?

Each recording session will be one animal experiment with different phases ("periods") during the experiment, for instance some "training" sessions and some "test" sessions with experimental conditions differing between training and test. The data will be saved as time series data continuously throughout the experiment, so the minimal data would be a timestamp, the behavior of the animal and a variable (stimulus) that the animal is controlling with its behavior. Thus, in the simplest case, three columns of integers.

The meta-data for each experiment has to contain a description of the columns, of course, as well as date and time at the start of the experiment, genotype of the animal, text description of

the experiment, DOI of the code used to generate the data, sequence and duration of periods, temperature, and other variables to be recorded or set on a per session or per period level.

A dataset or small project can consist of maybe three to four groups of experiments, let's say one experimental genotype and two control groups. Traditionally the way we handled this grouping in most of our experiments, is to keep a text file in which the experimenter lists which file belongs to which group. That way, anybody can read the text file and get an understanding of the experimental design. The file also contains comments and notes about user observations during the experiment and a text description of the project. In a way, this text file is like a meta-data file for a data-set, rather than an individual experiment and thus should probably also contain some minimal mark-up. This text file is then read by either custom software or an R script to compile summary data for each group, e.g. means and standard errors of some variables we extract on a per period basis, plotted and compared between groups. As there are numerous ways to evaluate an animal's behavior if we have the full time series, there is any number of different parameters one can want to extract from the data and plot/compare.

This is where the open science part would come in. Whenever the user runs the script that evaluates, plots and compares the data, the entire dataset is automatically made publicly accessible. Along with the dataset (raw data, meta-data and grouping text file), all the evaluations should also be deposited. Currently, we do this as a PDF file, but that is all but useless – only for human use. Ideally, I'd like this evaluation file to contain all the content of the grouping text file, as well as the DOI of the script that generated it and (semantic?) markup that structures the evaluation document. Such an evaluation document would be both machine and human (with a reader, which is why we started by using the PDF format) readable and provide an overview of exactly what was done to what data.

One eventual goal is to also use these evaluation documents during manuscript authoring. Instead of copying the figures, pasting them into a manuscript and then trying to describe the statistics, I'd like to just link the different evaluations from inside the manuscript. Each figure in a manuscript would then just be a link to one of the evaluations in the evaluation document, the one I want readers to see so they can follow my line of arguments. Any reader who wants to see other aspects of the data has single-click access to the entire evaluation document, with all our evaluations for this data-set, as well as access to all the code used to generate and evaluate the data, if they so wish. For this, all the data and meta-data in each dataset has to be linked to both each other, and the code and the text. Of course, all the data in a manuscript should also be linked together, even though they likely come from different datasets/projects.

With the data and code solutions we're currently developing, this should allow us to just write code, collect data and link both into our manuscripts. Everything else (data management, DOI assignment, data deposition, etc.) would be completely automatic. Starting at the undergraduate student level, users would simply have to follow one protocol for their experiments and have all their lab-notebooks essentially written and published for them – they'd have a collection of these evaluation documents, ready to either be used by their supervisor, or to be linked in a thesis or manuscript.

So, what would be the best data structure and meta-data format with these goals in mind?