Ten Simple Rules for Data Storage

Edmund Hart [1,*], Pauline Barmby [2], Jeffrey Hollister [3], David LeBauer [4],

1 Univeristy of Vermont, Department of Biology, Burlington
2 University of Western Ontario, Department of Physics and Astronomy
3 US Environmental Protection Agency, Atlantic Ecology Division
4 University of Illinois at Urbana-Champaign, Institute for Genomic Biology
5 National Center for Supercomputing Applications,

∗ **E-mail: emh@emhart.info**

Abstract

Introduction

Some example text with a citation [1]

Rule 1: Rule

1 Rule 2: Rule {Know your use case}

Researchers should know their use case and store data appropriately. Is this data collected and just being archived? Will it change regularly? How will those changes be logged (e.g. provenance if any)? Will this be shared via an API? Linked to a paper? What are the institutional restrictions? Can you use a commercial service like Dropbox or use a personally maintained system? Knowing the reason why you're sharing your data will constrain your choices here.

Rule 3: Rule

Rule 4: Rule

Rule 5: Rule

Rule 6: Rule

Rule 7: Rule

Rule 8: Rule

2 Rule 9: Rule {Data size matters / requires special considerations}

- #39 and related GH issues #16, #19, #25

Data that is larger than memory can handle, or ...

raw data on server, cloud, or long term storage. Data for analysis (if small enough) can go on a fast access disk, like SSD.

Can be handled by 'big memory' nodes.

In many cases fast access between disk and compute server is needed, so that I/O doesn't bog down the entire process. If you have a big dataset and now I need to do some big computation with it" in which case you often want the disk with the data to be close to the machine doing the computation.

Data size matters. As **???** points out there's a vast difference between data created by NEON and a data set you might archive on Figshare from your PhD (if you're an ecologist anyway). How you store your data has to take these size considerations. It will determine where you archive your data, how it's backed-up and how it's shared.

lemma 1: storage costs and transfer / compute time are trivial if data is small, but can add up if data is large. At some size, it is not practical to store data - there are trade offs among cost, information content, and accessibility.

- put data where and how you can compute on it, rather than moving raw datasets around

Covers proximity / mounting and architecture. Don't hog active space if slow storage is sufficient - scan for untouched files to put in longer term storage.

Don't move it around if you can help it. If you must, use appropriate tools, Store local 'cached' copies (eg use knitr argument) instead of writing scripts that always download archived data. Only do so if there are changes.

Do sub setting server-side, computing in database (dplyr lazy eval) etc.

Basically, the format for long-term archival is not the same format that is needed for day-to-day analysis. I'd say that in the opposite direction, actually: don't archive your data in the format that makes sense for you in your day-to-day use.

Rule 10: Rule

Figure Legends

Figures here: Will need to figure out numbering...

Tables

Tables here: Will need to figure out numbering...

References

1. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. PLoS computational biology. Public Library of Science; 2014;10: e1003542.