

Ten Simple Rules for Data Storage

Edmund Hart ^{1,*}, Pauline Barmby ², Jeffrey Hollister ³, David LeBauer ⁴, Naupaka Zimmerman ⁵,

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, National Center for Supercomputing Applications and Institute for Genomic Biology

5 University of Arizona, School of Plant Sciences

* 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.

2 Rule 3: Rule {Keep raw data raw}

Since analytical and data processing procedures may improve or otherwise change over time, having access to the 'raw' or unprocessed data helps to facilitate future re-analysis and analytical reproducibility. Data should always be kept in a raw format whenever possible, within the constraints of technical limitations. In addition to being the most appropriate way to ensure transparency in analysis, having the data stored and archived in its original state given a common point of reference for derivative analyses. There's clearly some discussion about what constitutes "raw" (ohms off a temperature sensor?). However the focus here is really on the spirit of the rule. Data should be as "pure" as possible when it's stored. If derivations occur, they should include storage of relevant code and subsequent data sets. NEON handles this with a schema for various "levels" of data products that pertain to the amount of processing that happens, here's a brief overview. We defined raw data as things like voltage, or unprocessed lidar returns. The issue of course is that this is a tremendous amount of

data that NEON still hasn't quite figured out how to share and document (e.g. if you ask for L0 they'll ship you an HDD). The way around it is to write detailed "Algorithm Theoretical Basis Documents" (ATBD's) detailing the different processing "levels". This is mostly borrowed from the NASA EOSDIS program.

3 Rule 4: Rule {Store data in open formats}

To maximize accessibility and long-term value, data should be stored in file formats whose specifications are freely-available. The exact file type will depend on the type of data being stored (e.g. numeric measurements, text, images, video) but the key idea is that data should not require proprietary software or hardware to access. Proprietary formats can change, maintaining organizations can go out of business, and changes in license fees could make access to data in proprietary formats simply unaffordable. Examples of open data formats include * CSV for tabular data * HDF5 for (??) * ?? for images * (help me out here folks, it's been a long week) and examples of closed formats include XLSX, DICOM, (again need more examples). At a minimum, data being stored for archival purposes should be stored in open formats, even if day-to-day processing uses closed formats.

Rule 5: Rule

Rule 6: Rule

Rule 7: Rule

Rule 8: Rule

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

- #39 and related GH issues #16, #19, #25
- Size classes:
 - larger than RAM
 - larger than HD space
 - larger than data storage server
- Storage method depends on the size of data; storage costs, transfer time, and computing costs can become substantial.
 - data generated by simulation and derived data should consider cost of storage vs. the cost of re-generating output.
 - For analyses of large data sets, the speed of reading and writing data can limit the speed of computation.
- Larger data sets that are actively used in analysis should be stored on a disk that is attached to a computer rather than being moved around between analysis and storage.
 - inactive data can be put in longer-term storage; this is less expensive, but can be slow to retrieve. Archiving of 'stale' files can be automated (and is at HPC centers).

- Data that is larger than memory can handle,
 - can be handled by ‘big memory’ nodes.
 - Computing can also be done ‘in the database’
- Don’t move (large data) around more than you have to - it can become inefficient, and make storage slower than necessary.
 - New tools make it easier to find and download data combined with reproducible scripts can lead to excessive and careless abuse of resources.
 - subset and compute on the server, in the database where possible. The dplyr R package does lazy eval; SQL can perform a wide range of data summaries, by groups, etc. On the other hand, it may be quicker to transfer normalized (e.g. ‘flattening’ a relational database can increase the size of data by orders of magnitude)
 - Use tools to store local ‘cached’ copies, instead of writing scripts that always download archived data. Only update data if there are changes. * knitr has a cache argument that saves time in re-computing and in re-downloading.
- For data larger than a single hard drive disk, up to multiple servers
 - requires a meta-data server to allow fast access to distributed across many disks
- For very large data
 - it is not practical to store data
 - there are trade offs among cost, information content, and accessibility.

Rule 10: Rule

5 Acknowledgements

National Center for Supercomputing Applications. Software Carpentry Foundation.

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.