

Living in an Ivory Basement

Stochastic thoughts on science, testing, and programming.

[misc](#)[personal](#)[python](#)[science](#)[teaching](#)[testing](#)

My Data Management Plan - a satire

Dear NSF,

I am happy to respond to [your request](#) for a 2-page Data Management Plan.

First of all, let me say how enthusiastic I am that you have embraced this new field of "large scale data analysis". Ever since I started working with large Avida data sets in 1993, then with large meteorological data sets in 1995, and then again with large sequence data sets in 1999, I have seen the need for a systematic plan to manage the data. It is nice to see NSF stepping up to the plate in such a timely manner, and I am happy to comply.

Now, as to my actual data management plan, here is how I plan to deal with research data in the future.

I will store all data on at least one, and possibly up to 50, hard drives in my lab. The directory structure will be custom, not self-explanatory, and in no way documented or described. Students working with the data will be encouraged to make their own copies and modify them as they please, in order to ensure that no one can ever figure out what the actual real raw data is.

Backups will rarely, if ever, be done.

When required to make the data available by my program manager, my collaborators, and ultimately by law, I will grudgingly do so by placing the raw data on an FTP site, named with UUIDs like 4e283d36-61c4-11df-9a26-edddf420622d. I will under no circumstances make any attempt to provide analysis source code, documentation for formats, or any metadata with the raw data. When requested (and ONLY when requested), I will provide an Excel spreadsheet linking the names to data sets with published results. This spreadsheet will likely be wrong -- but since no one will be able to analyze the data, that won't matter.

Did I mention the click-through license? "You are provided this data for the sole purpose of reproducing our published results. Any attempt to publish your own analyses of this data will be rejected, if necessary during the anonymous review process, by pointing out all of the data cleanup steps you forgot to do correctly in your analysis. (We don't remember all of them ourselves, but there sure were a lot!) Give up now."

We will provide a short note -- in a Word document -- detailing the licensing restrictions, as above.

We will make sure that any CSV files we do eventually produce will have format errors, such as missing or extra commas. They will also be encoded in ISO 8859-16, "by accident".

On the off chance that we do choose to provide the source code, it will be in a file named 'source.tar.gz' that unpacks in to the current directory. There will be no explanation of contents, instructions on how to run it, or any enabling information -- it was hard to write, and it should be hard to run! Old, patched, or otherwise impossible-to-obtain versions of Redhat Linux, Perl 5, and associated CPAN libraries will be required before the code runs, even if it doesn't actually use any of them. No source code documentation will be present, of course -- we don't need it ourselves, after all! Automated tests will also not be present (we don't have any of those, either). New versions of the code will be published under the identical file name, with no indication of what changes were made. (We'll be sure to use mixed

Mon 17 May 2010

By [C. Titus Brown](#)

In [science](#).

tags: [science](#)

DOS and Unix EOL editors for our files, so 'diff' won't work to figure out what has changed.)

Note, we didn't use a version control system, either. Or if we did, we made sure to use svn branching and merging profligately, with extremely obscure commit messages (our main programmer only speaks Chinese, so that's how she enters her commit notes. Wouldn't have it any other way). And our repository is not publicly available - you have to beg for permission. Note, I only answer e-mail on every other Tuesday.

Any design notes on the data analysis are in our private e-mail, and we will fight to the death -- up to and including ignoring FOIA requests -- to prevent you from obtaining them.

Meanwhile we will continue publishing exciting sounding (but irerproducible) analyses, and submitting grants based on them, because that's the only thing that the reviewers care about.

sincerely yours,

--titus

(representing every computational scientist in the world.)

Legacy Comments

Posted by Hector on 2010-05-18 at 07:56.

One word - Brilliant ... I came from a crystallography lab and am now in a environmental microbiology lab doing metagenomics. I have these discussions with the other post-doc in the lab on an almost daily basis. Particularly when trying to compare our results to what is published in the literature. Yes, It makes you almost want to throw in the towel.

Posted by Madelaine on 2010-05-18 at 10:11.

So funny, and so unfortunately true.

Posted by C. Titus Brown on 2010-05-18 at 11:04.

Greg Wilson pointed out: <http://www.sc2000.org/bell/twelve-ways.txt>

Posted by Greg Tyrelle on 2010-05-18 at 11:33.

I'm not sure what to make of this. Based on the description of your data management plan, it appears you have very precise and detailed information regarding **our** data management plan!

Posted by nolley on 2010-05-18 at 20:33.

what? that's not how you're supposed to manage software products?

Posted by John Comeau on 2010-05-18 at 23:26.

That's no satire... It's SOP.

Comments!

[\(Please check out the comments policy before commenting.\)](#)

0 Comments

1

2

[Share](#)

Sort by Best



ALSO ON LIVING IN AN IVORY BASEMENT - BLOG

WHAT'S THIS?

On licensing bioinformatics software: use the BSD, Luke.

32 comments

— This is a complex issue, there are good arguments on both sides, and there are good reasons to use each of the three main options ...

Docker workshop at BIDS - post-workshop report

18 comments

— Well... user experience design principles, documentation, and being easy to install and run on whatever system they want ...

DIB jclub: Fast and sensitive mapping of error-prone nanopore sequencing reads with ...

3 comments

— See also Heng Li's 'review' on his blog: lh3.github.io/2015/07/30/a-few...

What do you think about the term & practice of "hardening software"?

24 comments

a — I agree -- "productize" is the verb I already use to describe this process.

Proudly powered by [pelican](#), which uses [python](#).

The theme is subtly modified from one by [Smashing Magazine](#), thanks!

For more about this blog's author, see [the main site](#) or [the lab site](#)

While the author is employed by the University of California, Davis, his opinions are his own and almost certainly bear no resemblance to what UC Davis's official opinion would be, had they any.