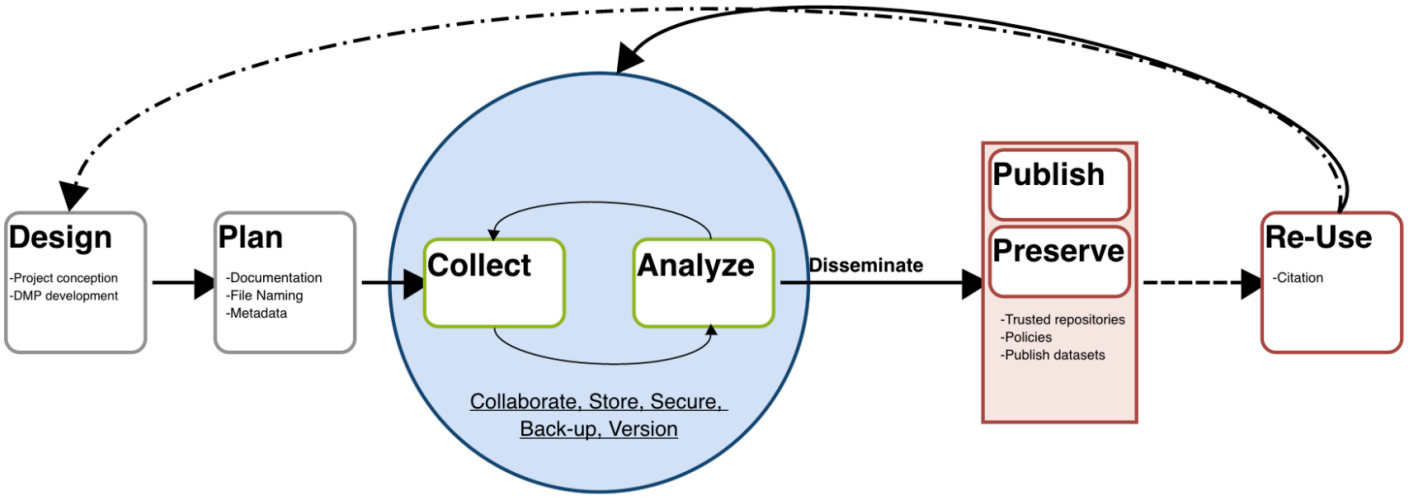

The DMPTool
and other useful tools to
support data management
Daina Bouquin
Harvard-Smithsonian Center for Astrophysics
@dainabouquin



http://www.library.cmu.edu/RDM

DMP

- Should be **thorough**
  - The two-page plan you submit to your funder will likely be a high-level overview of what you will actually do
- Different funders will ask different questions, but no matter what the same issues should be considered
  - Before, during, after mission considerations are important
- Make sure your whole team understands the plan
  - **execution** of the plan is most important
  - flexibility should be assumed
    - changes should be documented
- The DMP is a living document

Living in an Ivory Basement

Stochastic thoughts on science, testing, and programming.

misc | personal | python | science | teaching | testing

My Data Management Plan - a satire

Mon 17 May 2010
By C. Titus Brown
In science.
tags: science

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.

http://ivory.idyll.org/blog/data-management.html

Defining "Data"

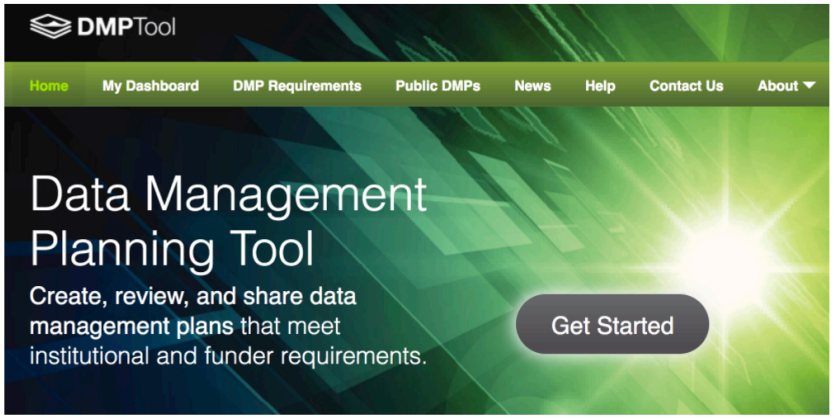
ex: Chandra X-Ray Observatory

(best case scenario)

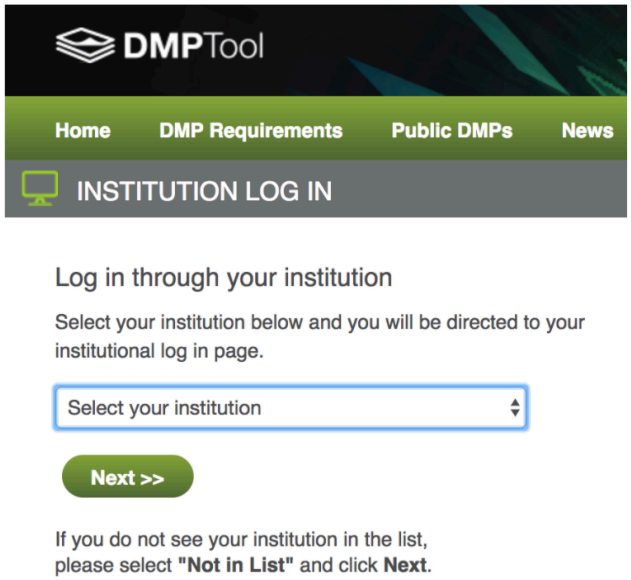
X-ray: NASA/CXC/PSU/L.Townsley et al; Optical: UKIRT; Infrared: NASA/JPL-Caltech

This composite contains X-ray data from Chandra (purple) plus infrared (orange) and optical data (blue)

- Many instruments give readings - different geographically and organizationally distributed teams involved
- Many data sources with different protocols
- Publications: http://lanl.arxiv.org/abs/1403.2576
- Non-image data associated with the image
- Format standards, preservation, and sharing are essential



https://dmptool.org/



My selection of tools are focused on

Open Reproducible Science

Consider:

Culture of practice

Institutional guidelines

Funder policies

Collaboration and Sharing

- Systematic version control is essential
- Tools should be useful at many/all parts of research workflow
  - Record provenance, reproduce your workflows
- Make your work citable and archived

- Git repo hosting
- Wikis, issues, social media features
- Distributed Version Control
- Terminal based with available GUIs
- **Alternatives - BitBucket, GitLab, etc.**

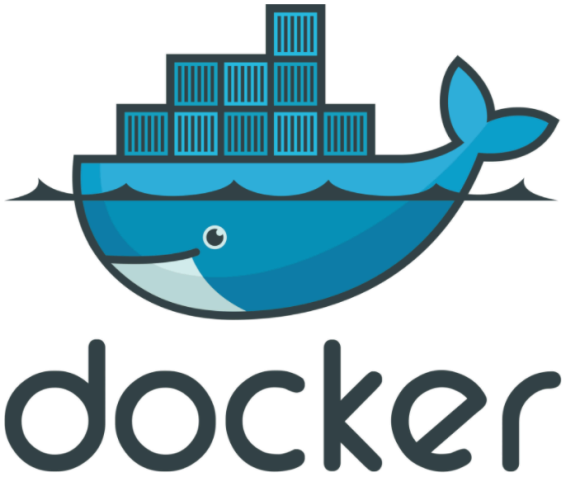
http://swcarpentry.github.io/git-novice/

- Cern project - discipline agnostic
- Create curated collections of software, data, papers
- Links with GitHub to allow you to mint DOIs for entire git repos of code releases
- **Alternatives - FigShare, Dataverse, etc.**

https://guides.github.com/activities/citable-code/

- Originally "IPython Notebooks" but now useful for many languages
  - R, Julia, Python, Ruby and expanding
- Capture full workflows and run your code in a browser interface
- NBviewer - easy rendering of notebooks on GitHub
- Similar features for R in RMarkdown via R Studio
  - http://jupyter.readthedocs.io/en/latest/index.html
- Example: LIGO Gravitational waves-

  https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html



- Containers for your code
  - bundle all of your dependencies to make your work more easily reproducible
  - more scalable than virtual machine
    - https://docs.docker.com/engine/getstarted/
- Acknowledge limitations: http://ivory.idyll.org/blog/2017-pof-software-archivability.html

Remember and Encourage Licenses!

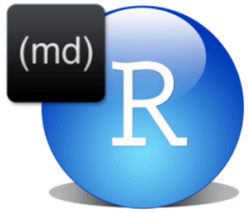
http://choosealicense.com/

Publishing/Proposing/Planning

Collaborative, browser-based LaTeX and
Markdown editors with Version control

https://www.authorea.com/

https://www.overleaf.com/



http://rmarkdown.rstudio.com/

Analysis/Data manipulation/Visualization

- Open source programming languages
- "libraries" and "packages" of useful functions

Bring it all together

https://osf.io/

https://www.globus.org/