Best Practices for Data Management

RMACC HPC Symposium, 8/13/2014

Presenters

Andrew Johnson
Research Data Librarian
CU-Boulder Libraries

Shelley Knuth
Research Data Specialist
CU-Boulder Research Computing


What do we mean when we talk about “data”?

Some research data definitions

White House Office of Management and Budget:
“the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

CU-Boulder Data Management Task Force:
“the digital representation of information generated at any stage of the research process in a formalized manner suitable for communication, interpretation, or processing.”

CU-Boulder Research Data Advisory Committee:
“digital outputs, which include content in structured forms including but not limited to: text files, word processing documents, spreadsheets, websites, calibration information, or simulation outputs; digital information artifacts, such as images, vector-based map products, audio and video products; digital outputs may also include the code used to decode those products, the metadata describing such information artifacts, and required source code that runs computer instructions.”

Why do we care about research data management?

Good for science

- Reproducibility
- Efficiency
- Innovation

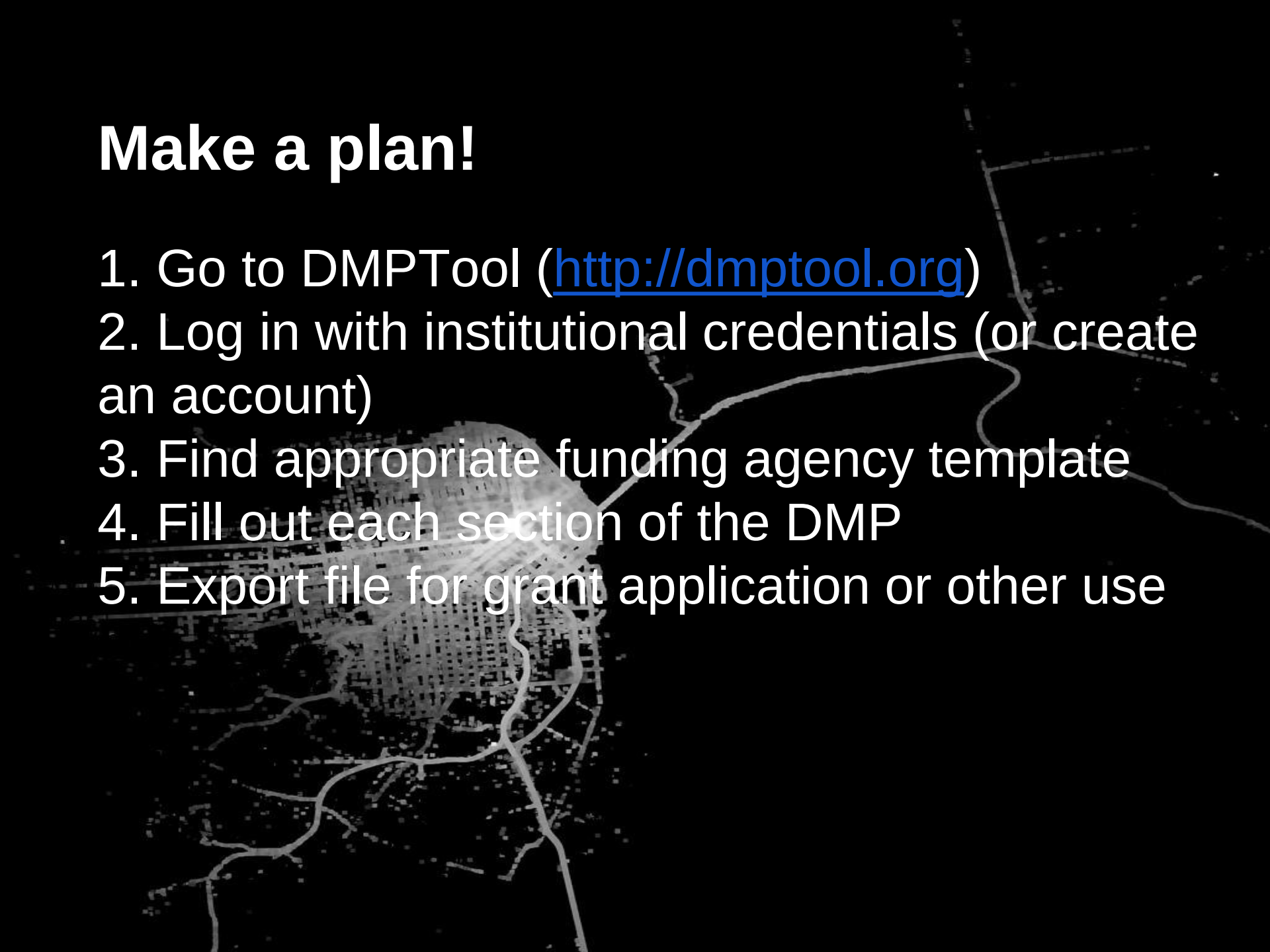
Good for you

- Increased visibility (including citations)
- Reduced time/effort
- More competitive grant applications

How do we get started with research data management?

Make a plan!

1. Go to DMPTool (http://dmptool.org)
2. Log in with institutional credentials (or create an account)
3. Find appropriate funding agency template
4. Fill out each section of the DMP
5. Export file for grant application or other use



NSF general DMP requirements

1. What types of data will you produce?
2. What (if any) standards will you use?
http://www.dcc.ac.uk/resources/metadata-standards
3. When and how will you share data?
http://figshare.com/
4. What can people do with your data?
5. How will you archive and preserve data?
http://databib.org/

DOE suggested DMP elements

1. Data types and sources
2. Content and format
http://www.dcc.ac.uk/resources/metadata-standards
3. Sharing and preservation
http://figshare.com/
http://databib.org/
4. Protection
5. Rationale

DOE requirements (some highlights)

"At a minimum, DMPs must describe how data sharing and preservation will ***enable validation of results***, or how results could be validated if data are not shared or preserved."

"DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research ***open, machine-readable, and digitally accessible to the public*** at the time of publication."

What is metadata?

It depends...

...who you're talking to:

- Researchers
- Data managers
- Librarians

...what you're trying to do:

- Describe your data for purposes of replication
- Make your data discoverable via a repository
- Enable your data to interoperate with various systems
- Preserve your data in an archive
- Cite data

Basic metadata should include:

Title
Creators
Date(s)
Version
Identifiers (for published data)
Location (online and/or physical)
Archive/Distributor (for archived data)

Metadata can include:

Discipline-specific information
Instrument settings
Image/video/audio specifications
Provenance
Spatial information
Environmental conditions
Etc., etc., etc.

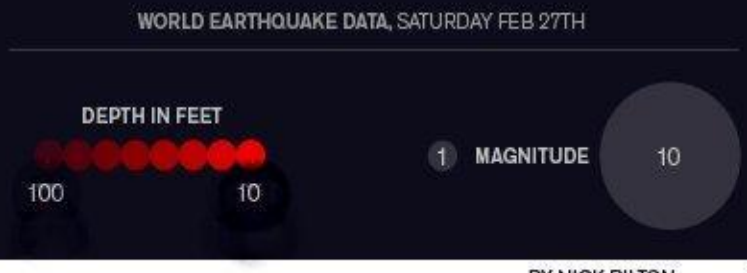
Good metadata should:

Follow community- or discipline-based standards
http://www.dcc.ac.uk/resources/metadata-standards

Use consistent and documented conventions in the absence of standards

Allow intended users to:

- Discover data
- Understand data
- Reuse data



Some examples

Good metadata:

- Microscopy image with embedded OME-XML metadata
- Survey data in spreadsheets with accompanying codebook using DDI metadata

Bad metadata:

- Ambiguous/absent column labels on a spreadsheet
- Image data with no information about what/where/when it represents
- Observation data with no record of where/when it was collected

Data Archiving and Preservation

Proper archiving for long-term preservation of data critical

What should be archived?

- Data
- Metadata
- Research products
- Scripts
- Anything required to reproduce the results of research

What Components are Important to Archiving Data?

- Where will the data be held?
- What is the infrastructure where the data will be housed?
- Which data will be archived?
- How long will the data be archived?
- Will there be multiple copies of the data?
- How do you ensure long term preservation?

Good Practices for Long Term Data Archiving

Trusted repository is best!

- Only storing data on thumb drives – bad

Store multiple copies!

Active management

Backups!

Security?

Review schedule for preservation

- Ten years?

CU Data Storage: PetaLibrary

- NSF Major Research Instrumentation grant
- Data collections from faculty and students
- Deposition and archival of data
- Researchers pay for the medium (disk or tape)
- No HIPAA, FERPA, ITAR data
- Infrastructure guaranteed for 5 years

PetaLibrary Storage Options

- ***Active***: for data that is written or read frequently
  - Always stored on spinning disk
  - Mounted on CU-RC compute resources (NFS, GPFS)
  - Accessible via certain protocols from outside CU-RC
  - Option for second copy on tape or on disk in a different building
- ***Archive***: for data that is accessed infrequently
  - Stored on a combination of disk and tape
  - Not mounted on compute resources
  - Accessible via certain protocols from outside CU-RC
  - Option for additional copy on tape

Data Availability – Why?

Allowing data to be discoverable by

- General public
- Other researchers

Beneficial to researchers by:

- Promoting transparency
- Increasing awareness of research
- Meeting funding agency requirements
- Enhancing discovery

Data Availability – Create a plan

How will you make the data available?

To whom will you make the data available?

Is the data secure? What special considerations need to be made?

Will there be an embargo period? How long?

Data Availability – How?

- Personal website
- Domain repository (datadryad.org)

  “DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable”

- Domain specific repository
- Institutional Repositories

Intellectual Property

Who owns the data you are sharing?

- Do you have the right to share it?

Is any of the material copyrighted? Under patent protection?

Who maintains long-term preservation?

Consult legal experts if necessary

Data Publication

- I can publish my data? YES!
- Difference between:
  - Publishing data
  - Publishing research that uses data

- Second is the more traditional method
- No direct link to the data

Data Publication

To publish data, can write an article about the data collection process and link to a repository

Earth System Science Data

Geoscience Data Journal

Can also publish data in FigShare

figshare.com

And soon as part of Globus

globus.org

Data Publication and Citation

- Publishing data is a great way to get cited
- Add a DOI to dataset
- Depending on type of publication and field standards might improve changes of tenure/promotion

- Remember to ALWAYS CITE DATA!!!

  Yours, or others!!

Best Data Management Plans and Practices Competition

Internal grant this summer offered by the Vice Chancellor's Office

August 15 close

Winners announced at Research Fair on 9/17

Grad students, postdocs, and faculty

Five people $2000 each

Arts and Humanities, Engineering, Life Sciences, Physical Sciences, and Social Sciences

General funds

Best Data Management Plans and Practices Competition

Awards for data plans that describe best practices for their current data or a plan for the future

http://data.colorado.edu

Please don’t hesitate to ask questions

Contact Shelley.Knuth@colorado.edu

Thank you!

Image credits

Slide 3: “Data”, https://flic.kr/p/bJDnpB
Slide 4: “Earth's City Lights 1994”, https://flic.kr/p/dywxTR
Slide 5: “Wired UK - NDNAD Infographic”, https://flic.kr/p/6tV5SZ
Slide 6: “Milky Way”, https://flic.kr/p/9q6g4U
Slide 7: “Social graph”, https://flic.kr/p/fnzLPk
Slide 8: “20 minutes from 4th and Market by car”, https://flic.kr/p/8FDzna
Slide 9: “Landsat data of Beijing in 2010”, https://flic.kr/p/cE8fns
Slides 10-11: “The Cat's Paw Remastered”, https://flic.kr/p/ctRpEw
Slide 12: “Metadata”, https://flic.kr/p/9BZ7rG
Slide 13: “Card catalogs at Sterling Memorial Library, kept only for appearances”, https://flic.kr/p/cKHzQ
Slide 14: “Pure Data File Killer - Bliss (sgi)”, https://flic.kr/p/h4RV3y
Slide 15: “Collecting Data from the Air”, https://flic.kr/p/jRk9Fp
Slide 16: “Earthquake Data, Feb 27th”, https://flic.kr/p/7GaoYw
Slide 17: “Geography of Twitter @replies”, https://flic.kr/p/avgZqq