Data management
for postdocs and research scientists

Ece (pronounced AJ) Turnator
Phoebe Ayers
MIT Libraries
data-management@mit.edu
August 3, 2017
Workshops

Web guide: http://libraries.mit.edu/data-management

Individual assistance/consultations
  - includes assistance with creating data management plans

Contact: data-management@mit.edu

Help is here when needed!
Workshop outline

• Why data management?
• Describing your data
• Collaborating
• Managing your data for reuse
• Managing your data for long-term access
• Starting and ending projects
• Grants and data management plans
Why data management?
Shared data has impact: big data
Shared data: smaller datasets

data and tools

We have made the following datasets and tools available to the public.

Data on coal-fired electricity

Data here.
Paper here.

Greenhouse gas equivalency metrics tool

Tool here.
Fact Sheet here.
Paper here.
## Funder Requirements

### Research Funder Open Access Requirements

The table below summarizes major US Research funder open access requirements for publications and data, and links to information on related journal requirements.

For questions, please contact:
- For publications: Ellen Finnie, Head, Scholarly Communications & Collections Strategy
- For data: data-management@mit.edu
- See also more information on the Research Funder Policies and Related Legislation page.

List of major funder requirements: [http://libraries.mit.edu/scholarly/publishing/research-funders/research-funder-open-access-requirements/](http://libraries.mit.edu/scholarly/publishing/research-funders/research-funder-open-access-requirements/)

<table>
<thead>
<tr>
<th>Agency / Entity</th>
<th>Publications</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>White House Office of Science and Technology Policy Directive</td>
<td>All federal agencies over $100M annually in R&amp;D required to make peer-reviewed manuscripts available within 12 months. 2013, with implementation beginning October 2014 and timetable varying by agency.</td>
<td>All federal agencies over $100M annually in R&amp;D required to make data stored and publicly accessible search, retrieve, and analyze. Scope: data necessary to validate research findings, including data used to support scholarly publications. (Lab notebooks not included)</td>
</tr>
<tr>
<td>Office of Science &amp; Technology Policy</td>
<td>View the Directive</td>
<td></td>
</tr>
<tr>
<td>DoD US Department of Defense</td>
<td>Public Access Plan Draft in 2015, takes effect FY16. Authors will be required to upload their final accepted peer-reviewed journal manuscripts (or the final published article, if the author has sufficient rights) to the Defense Technical Information Center (DTIC) at the time of acceptance. Articles will be available no later than 12 months following publication.</td>
<td>Public Access Plan Draft in 2015, compliance begins FY17. Data Management Plans (DMPs) will be required, focusing on data “necessary to validate research findings.” Metadata for each data set, including subject, characteristics, and location, will be shared via DTIC’s DoD data set catalog.</td>
</tr>
</tbody>
</table>
Describing your data
## Research data: Common types by discipline

<table>
<thead>
<tr>
<th>General</th>
<th>Social Sciences</th>
<th>Hard Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>images</td>
<td>survey responses</td>
<td>measurements generated by sensors/laboratory instruments</td>
</tr>
<tr>
<td>video</td>
<td>focus group and individual interviews</td>
<td>computer modeling</td>
</tr>
<tr>
<td>mapping/GIS data</td>
<td>economic indicators</td>
<td>simulations</td>
</tr>
<tr>
<td>numerical measurements</td>
<td>demographics</td>
<td>observations and/or field studies</td>
</tr>
<tr>
<td>software &amp; code</td>
<td>opinion polling</td>
<td>specimen</td>
</tr>
</tbody>
</table>
Research Data: Stages

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data</td>
<td>raw txt file produced by an instrument</td>
</tr>
<tr>
<td>Processed Data</td>
<td>data with Z-scores calculated</td>
</tr>
<tr>
<td>Analyzed Data</td>
<td>rendered computational analysis</td>
</tr>
<tr>
<td>Finalized/Published Data</td>
<td>polished figures appear in <em>Cell</em></td>
</tr>
</tbody>
</table>
Digital data + Complex workflows

Software logos for VT, Kepler, FileMaker, zotero, Asymptote, 7, SPSS, R, Mendeley, ArcGIS, and SAS.
Questions funders ask

• **Data description**
  - What is it?
  - How will it be collected?
  - What format is it in?
  - How much data will be generated?
  - Is any PII or confidential data?

• **Secondary data**
  - Are you using data that someone else produced? If so, where is it from?
Exercise

• What data products has your latest project produced? What do you anticipate generating?

• How many different formats do you use?

• What programs or code is needed to read or understand these files?
Collaborating on data
Who are the custodians of your data?

Who has access
• To read & download
• To write
• **Who has responsibility**
  • For organization and documentation
  • For storage & backup
• **Who has intellectual property rights**
  • Created by you or someone else?
  • Funder restrictions?
Questions funders ask

• Roles and Responsibilities
  - Who is responsible for managing the data?
  - Time allocation, personnel, cost staff responsibilities?
  - Who takes it over at the end of the project?

• Intellectual property
  - Who holds intellectual property rights for the data and other information created by the project?
  - Are there any patent- or technology-licensing-related restrictions on data sharing associated with this grant?
  - Will you permit re-use, redistribution, or the creation of new tools, services, data sets, or products?
Some tools for collaboration

- Dropbox for MIT
- Open Science Framework (OSF) - osf.io
- Lab Archives and other electronic lab notebooks
- Collaborative writing tools (Overleaf/ShareLaTeX/Google Docs)
- README files and documentation (for instruments/procedures/file structures/etc)
- Versioning
Exercise

• Is there anyone in your lab responsible for data management?

• Is there anyone on your latest project responsible for data management?

• Does your group use collaborative tools, and if so are they useful?
Managing your data for reuse
Considerations for reusability

• File formats for long-term access
• Versioning
• Metadata
• Documentation
• Copyright
In the best case, your data files are both:

• **Non-proprietary (also known as open)**, and
• Unencrypted and uncompressed
## Formats: Preferred Examples

<table>
<thead>
<tr>
<th>Proprietary Format</th>
<th>Alternative/Preferred Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excel (.xls, .xlsx)</td>
<td>Comma Separated Values (.csv)</td>
</tr>
<tr>
<td></td>
<td>ASCII</td>
</tr>
<tr>
<td>Word (.doc, .docx)</td>
<td>plain text (.txt), or if formatting is needed, PDF/A (.pdf)</td>
</tr>
<tr>
<td>PowerPoint (.ppt, .pptx)</td>
<td>PDF/A (.pdf)</td>
</tr>
<tr>
<td>Photoshop (.psd)</td>
<td>TIFF (.tif, .tiff)</td>
</tr>
<tr>
<td>Quicktime (.mov)</td>
<td>MPEG-4 (.mp4)</td>
</tr>
</tbody>
</table>
Versioning: Basic Practices

In some cases, it may make sense to log the changes so that you can quickly assess and access the versions.

It’s good to document:
- What was changed?
- Who is responsible?
- When did it happen?
- Why?

CHANGELOG.md

CHANGELOG.txt

git

Google Drive
Versioning: File Naming Conventions

Naming conventions make life easier!

Naming conventions should be:

- **Descriptive**
- **Consistent**

Consider including:

- Unique identifier (ie. Project Name or Grant # in folder name)
- Project or research data name
- Conditions (Lab instrument, Solvent, Temperature, etc.)
- Run of experiment (sequential)
- Date (in file properties too)
- Version #
Versioning: File Naming Conventions

Naming conventions make life easier!

Naming conventions should be:

- Descriptive
- **Consistent**

<table>
<thead>
<tr>
<th>YYYYMMDD</th>
<th>MMDDYYYY</th>
</tr>
</thead>
<tbody>
<tr>
<td>YYMMDD</td>
<td>MMDDYY</td>
</tr>
<tr>
<td>MMDD</td>
<td>DDMM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TimeDate</th>
<th>DateProjectID</th>
</tr>
</thead>
<tbody>
<tr>
<td>TimeProjectID</td>
<td></td>
</tr>
</tbody>
</table>

| Sample001234 |
| Sample01234 |
| Sample1234  |

Include the same information

Maintain order
Versioning: File Naming Conventions

Resources:

- Check for Established File Naming Conventions in your discipline
  - DOE's Atmospheric Radiation Measurement (ARM) program
  - GIS datasets from Massachusetts
  - The Open Biological and Biomedical Ontologies

- File Renaming Tools
  - Bulk Rename Utility
  - Renamer
  - PSRenamer
  - WildRename

- Contact data-management@mit.edu!
Exercise

- How do you currently organize your files? Do you find what you want, when you need it?

- Define the types of data and file formats you use in your project/work

- How do you want your files to sort?

- What’s the folder hierarchy for where your files will live?
Metadata should tell you...

- **What** do the data consist of?
- **Why** were the data created?
- What **limitations**, if any, do the data have?
- What does the data **mean**?
- How should the data be **cited**?
Metadata: Things to Document

- Title: datasetName
- Creator: Malinowski, Christine
- Identifier: dataID
- Funders: NIH
- Dates: 20140123-20150114
- Rights: We own this data.
- Processing: Normalized
- Location: This file is located in this directory
  MyProject_NSF_2014
Document your workflow

• Workflow: how you get from raw data to the final product of research
• Documentation could be a flowchart or document
• Comment your code and scripts
• Well-commented code is easier
  • to review
  • share
  • and use for repeat analysis
Data Sharing: Copyright / Licensing

Except where otherwise noted, this work is licensed under
http://creativecommons.org/licenses/by-sa/3.0/
Questions funders ask

• **Documentation (aka metadata), Organization**
  - What types of information are important for describing, discovering & using the data?
  - Which mechanisms for metadata capture make sense given your research workflow?
  - Are you using documentation standard to your field?
Exercise

• What metadata elements are important for understanding your data? (eg, author, date created, type of instrument, etc).

• Are there any file naming conventions in your project?
Managing your data for long-term access
Considerations

- Long-term storage
- Managing confidential data
- Discovering data
- Citation
Preservation of data = active management

- **Backup** - 3 distributed copies, different media (cloud, tape, disk)
- **Fixity checks** - checksums, hashes
- **Format migration** - due to obsolete software / media
- **Security/permissioning** - physical & virtual access to storage
Long-term Storage

Institutional resources

Backup services - http://ist.mit.edu/backup
Storage - http://ist.mit.edu/managed-servers

Grant/project funding

Repositories: a great solution for many challenges!
Long-term Storage

MIT’s institutional repository

DSpace@MIT http://dspace.mit.edu/

Other repositories:

- Harvard Dataverse https://dataverse.harvard.edu/
- Zenodo, https://zenodo.org
- Inter-university Consortium for Political and Social Research (ICPSR) http://www.icpsr.umich.edu

Resources for finding a repository

- Registry for Research Data Repositories: re3data.org
- Our website: http://libraries.mit.edu/data-management/share/find-repository/
Data Sharing: Managing Private / Confidential Data

Things to consider:

- de-identification / anonymization
- segregation of sensitive information
- adherence to relevant laws & policies

http://informatics.mit.edu/classes/managing-confidential-data
Discovering data

Data as article supplementary material:

- **Pros:**
  - Associates data with published articles
  - Provides a citable source
  - Journal requirements

- **Cons:**
  - Limits to number and sizes of files
  - Possible format limitations
  - Reduced metadata
  - Fragments your dataset
Data journals:

- Publish “data papers”
- Help make data sets discoverable and citable
- Peer-reviewed
- Data usually stored in a repository

Data journal examples:

*Scientific Data* [http://www.nature.com/sdata/about](http://www.nature.com/sdata/about)
*Journal of Chemical and Engineering Data* [http://pubs.acs.org/journal/jceaax](http://pubs.acs.org/journal/jceaax)
*Open Health Data* [http://openhealthdata.metajnl.com/](http://openhealthdata.metajnl.com/)
Data Sharing: Citation

- Facilitates discovery of data
- Gives credit to the researcher
- Recognizes data as substantial output of the research process
- Allows for citation/impact analysis, as with article publications
Data Sharing: Citation

Important components:
- Creator/author
- Title
- Publisher
- Publication date
- Version
- Persistent ID
Data Sharing: Persistent IDs

- DOI - Digital Object Identifier
- ARK - Archival Resource Key
- Researcher identifier
  - ORCID - Open Researcher and Contributor ID
Questions funders ask

- Will you store the data in an archive or repository for long-term access? If not, how will you preserve access to the data?
  - Is a discipline-specific repository available?
  - Or is there a funder-mandated repository?

- **Storage**
  - What are your local storage and backup procedures?
  - Will this data require secure storage?
  - What tools or software are required to read or view the data?
Exercise

• How long does your data need to be stored for?

• Where is your data stored? Who is responsible for that storage? Is it backed up, and if so where?

• How is data usually cited in your field? How do you find datasets?
Starting and ending projects
Start-up checklist

• Sources of data
  • File naming
  • Metadata collected
  • Storage
  • IP restrictions
• Funder and grant requirements and restrictions
• Personnel involved
Closing checklist

• README files
  • Data
  • Code
  • Instruments/protocols
• Who has access? Will you have access to these accounts after you leave?
• Storage and storage commitments
Exercise

• If someone else had to take over your project for you, what would they need to know?

• What are the storage plans when the grant ends or if the responsible person leaves the project?
Grants & data management plans

DMPs now required by all major federal funders & many private funders
Append to your grant
Part of your grant approval & reporting
General content of a DMP

1. Project, experiment, and data description
2. Documentation, organization, and storage
3. Access, sharing, and re-use
4. Archiving

Always check for and follow a funder’s specific requirements

Details for each of these areas can be found on our website at:
http://libraries.mit.edu/data-management/plan/write/
The DMPTool

https://dmptool.org
Exercise: creating a DMP

• Log into the DMPtool: dmptool.org
• Choose a likely funding agency for your work
• Create a ‘test’ DMP
Resources for Data Management

Scholarly Publishing @ MIT Libraries

- [http://libraries.mit.edu/scholarly/](http://libraries.mit.edu/scholarly/)
- [Research Funder Open Access Requirements](http://libraries.mit.edu/scholarly/research-funder-open-access-requirements) (includes data information)

Data Management Services @ MIT Libraries

- [data-management@mit.edu](mailto:data-management@mit.edu)