Data Management Overview for Instruction Librarians

Sarah Sheehan
George Mason University
ssheehan@gmu.edu
@sarah4libraries
Libraries Collect Data

- Door Count
- Circulation Stats
- Reference Desk Stats
- Instruction Stats
- Collection Stats

- Institutional Research & Reporting
- Accreditation Agency
- ARL/ACRL
This isn't about me.

It's about the data.

Data Management & Sharing Mandates

- Journals – PLOS, Nature, JDAP partners
- Funders – NSF, NIH
- Office of Science & Technology Policy mandate, February 2013

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
Data Life Cycle

Creating data
- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Source: Boston University Libraries, Research Data Management
http://www.bu.edu/datamanagement/background/data-life-cycle/
Benefits of Managing Data

• Ensuring research integrity and reproducibility
• Ensuring research data and records are accurate, complete, authentic and reliable
• Saving time and resources
• Enhancing data security and minimizing the risk of data loss
• Preventing duplication of effort by enabling others to use your data
• Facilitating the analysis of change, by providing data with which data at other points in time can be compared.

Source: http://datalib.edina.ac.uk/xerte/play.php?template_id=2
What is a Data Management Plan?

A document that describes what you will do with your data during your research and after you complete your project.

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
DMP is a Living Document

• Start with what you know NOW
• Use as a guide for daily activities
• Share your plan with co-workers/supervisors
• Will not be perfect the first time through
• Keep your plan current & incorporate changes

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
Components of a Basic DMP

1. Types of Data
2. Data & Metadata Standards
3. Policies
4. Plans for Preservation
5. Budget

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
1. Types of Data

- Types of data produced
- Format and volume of data produced
- How/when/where will the data be captured or created?
- How will the data be processed?
- Quality assurance, quality control measures, version control
- Who will be responsible for data management during/after project?
1. Types of Data: Organization

- **File name structure**
  - include file name structure in DMP
  - Share file name structure with team

- **Date is important in file names**
  - ISO 8601 Standard, yyyymmdd

- **File name is dependent on users**
  - Too long or nested names can be great for some/not for others
1. Types of Data: Documentation

- Create a Read-Me file
  - Lives outside the data
  - Why & How was the data created
  - What is the data about
  - How was the data analyzed/what did you do
  - Includes file names, variable names and abbreviations
  - Version of software used
  - Can be basic to complex
1. Types of Data: Documentation

- Code Books – common in social sciences
- Data Dictionary – very structured, can be a table
- Lab Notebooks
- Programming Code
2. Data & Meta Data Standards

• What metadata are needed to make the data meaningful?
  – Describe the resource
  – Enable discovery

• How will you create or capture these metadata?

• Why have you chosen particular standards and approaches for metadata, i.e. Dublin Core

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
2. Data & Meta Data Standards

• Talk to your Cataloger!
• Good example: Itunes
• Colectica – add on for Excel
• More Info
  http://www.dcc.ac.uk/resources/metadata-standards/list
3. Policies for access, sharing, re-use

• Does your institution have a data management policy?
• Are you under any obligation to share data?
• How, when, & where will you make the data available?
• What is the process for gaining access to the data?
• Who owns the copyright and/or intellectual property?
• Privacy issues, embargo periods for political/commercial/patent reasons
• How should your data be cited?

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
3. Policies for access, sharing, re-use

Never work with the original data
3. Policies for access, sharing, re-use

• Keep a pristine, untouched document in a safe place; always work with a copy

• Work with the most recent version

• Use version control
  – Subversion; Beanstalk, Git, Mecurial
  – Allows roll back to previous version

• Document what you did to the data and why
3. Policies for access, sharing, re-use

• Excel Do
  – Keep it simple
  – Create separate variables for each element
  – Include missing data as a blank cell
  – Keep variable names to 8 characters
  – Save in CSV

• Excel Don’t
  – Don’t use Tabs
  – No strange ch@4@cters
  – No merged cells
  – No formatting in cells
4. Plans for archiving & preservation

- What data will be preserved for the long term? For how long?
- Where will data be preserved?
- What data transformations need to occur before preservation?
- What metadata will be submitted alongside the datasets?
- Who will be responsible for preparing data for preservation?
- Who will be the main contact person for the archived data?

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
3-2-1 Archiving and Data Protection Best Practice

3 Copies of all data
- Optimized IT Resources
- Reduced Capital Acquisition
- Backup Elimination

2 Different types of storage media
- Fast Data Access
- Secure Data Retention
- Archive Resilience

1 Copy offsite on removable media
- Lower Power Consumption
- Foundation for Compliance
- Robust Disaster Recovery

Source: http://www.dicom.com/de/products/storage.php#
4. Plans for Archiving: The Cloud

- Not always the best choice
  - How much are storing
  - What are you storing, ie: sensitive data, HIPAA or SS#
  - Obligations from grant or institution
  - Encryption for download, upload & stored
  - Access to data, various levels of permissions

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
4. Plans for Archiving: Formats

• Optimal format, ISO Standard, PDF/A-1
  – Tiff or PNG uncompressed okay
  – Florida State University Library Services Table of FDA-supported File Formats Table [http://fclaweb.fcla.edu/node/795](http://fclaweb.fcla.edu/node/795)

• Data in SPSS, SAS, other commercial products are dependent on software environments

• R Open source alternative

Source: Getting Started with Data Management & DMPTool, [https://dmptool.org/promote](https://dmptool.org/promote)
4. Plans for Archiving: Repositories

- Institutional Repository
- Discipline – specific repositories
  - ICPSR – Inter-University Consortium for Political and Social Research
  - DataONE – earth observation data
  - Dryad – bioscience
  - FigShare - science

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
5. Budget

- Costs of data preparation & documentation?
  - Hardware, software
  - Personnel?
  - Archive fees?

- How costs will be paid?
  - Request funding!

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
6. Process

- Name a person responsible for implementing the plan
- Assign a person to each DMP activity
- Schedule training for software or other expertise required
- Update as changes are made
KEEP CALM AND ASK A LIBRARIAN
Additional Resources

New England Data Collaborative Data Management Curriculum
http://library.umassmed.edu/necdmc/index

Mantra: Research Data Management Training
http://datalib.edina.ac.uk/mantra/

Virginia Data Management Boot Camp
http://guides.lib.odu.edu/VADMBC
References & Additional Resources

• E-Science Portal for New England Librarians: A Librarians Link to e-Science Resources
  http://esciencelibrary.umassmed.edu/

• Research Data Curation Bibliography, version 4 June 23rd, 2014
  http://digital-scholarship.org/rdcb/rdcb.htm

• Strasser, Carly. (29 May 2014). New DMPTool Released Today blog post.
  http://blog.dmptool.org/2014/05/29/new-dmptool-released-today/

• University of Virginia, Data Management Consulting Group, Data Management Competents
  http://dmconsult.library.virginia.edu/plan/
Build your Data Management Plan

Email: uc3@ucop.edu
Twitter: @TheDMPTool
Facebook: dmptool
Blog: blog.dmptool.org

Source: Getting Started with Data Management & DMPTool, https://dmptool.org/promote
DMPTool version 1 released in October 2011
Self-Funded by Partners
Built for researchers; limited functionality for editors or administrators
• DMPTool version 2 released May 2014
• What was in DMPTool1 converted to DMPTool2
• External funding
  – Alfred P. Sloan Foundation
  – Institute for Museum and Library Services
• More functionality for researchers and institutional administrators
DMPTool2 New & Improved

- New user interface with embedded tips and help throughout
- Library of publicly available data management plans
- Visibility of Plans – 3 levels
  - Public = everyone
  - Institution
  - Private
- Can copy an existing DMP

DMPTool2 New & Improved

• Assigning plan co-owners for better collaboration
• New help on data management in general
• Frequently Asked Questions, where users can submit questions and get answers
• Quick-start guide for creating a DMP
• Up-to-date data management funder requirements

DMPTool2 Customization for Administrators

- Template includes questions that the funder wants answered
- Templates can be created by institutions
- Resource questions
  - Added by institution
  - DMPTool – wide resources
  - Requirements specified by funders
DMPTool2 Customization for Administrators

• Administrator Wiki including glossary of terms

• Review Functionality

• Visibility of Plans – 3 levels
  – Public = everyone
  – Institution
  – Private
States of the Plan

• New
• Complete
• Submitted
• Reviewed
• Approved
• Rejected
• Revised

All Plans

Used during the review phase