Data Management Plans and You: It’s not just the NSF, NIH, Science, Nature, American Economic Review...

E-Teaching Day
May 16, 2012

Barbara Mento, Data/GIS Librarian
Sally Wyman, Chemistry/Physics/Environmental Studies/General Science Librarian
Why now?

• Required by funding agencies (NSF and NIH, etc.)
  – Proposals *have* been rejected for inadequate DMPs!

• Scholarly journal (*Nature, American Economic Review...*) policies that data must be
  – clearly documented
  – available for sharing
  – detailed enough to permit replication of analysis
Why now?

Additional reasons:

• Risk (to you and the University) of data loss
• Retention planning fits into “responsible conduct of research” regardless of whether data is shared
• Shared data (“open access”) → higher citation rate!
Summary Reasons for Managing Research Data

• May be required by your funding agency
• Essential to the responsible conduct of research
• Ensures future accessibility/preservation
• Will generate more article citations
• Will make your life easier!
What do you consider to be “data”?

... and what do the funding agencies consider “data”?

- Varies by NSF Directorate
- See OMB Circular A-110
• Amended 1999
  – Federal awarding agencies require data produced will be available to the public under the Freedom of Information Act (FOIA).

• *Research data* is defined as

  The recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues.
The Digital Research Data Lifecycle

- Funding
- Data Collection
- Data Sharing
- Research Output

Arrows indicate the flow from one stage to another.
Stage 1: Funding

Many federal funding agencies now require *Data Management Plans* in the grant application.
The Data Management Plan

• 1-2 pages describing how data will be:
  • Collected
  • Documented
  • Stored
  • Secured
  • Analyzed
  • Preserved
  • Shared
  • And other relevant details

• Even if not required, being aware of data management concepts will help you to be a better data manager and ensure the long term access and preservation of your data.
Visit the **Boston College Libraries’ Data Management LibGuide** when you begin writing a DMP for:

- Guidance on content
- Templates/examples
- Additional resources
- To arrange a consultation with a subject specialist

[Visit the Boston College Libraries’ Data Management LibGuide](http://libguides.bc.edu/dataplan)
Elements of a DMP: a few more details

- **Varies by NSF directorate** and other funders
- Common elements include:
  - Description of the project
  - Description of the data to be collected, including formats, size
  - Access/sharing
    - Potential audience(s) for the data
    - How access will be provided and how others will find it
      - “Access” freely available, no specific request needed
      - “Sharing” of data per specific request
    - Stipulations for privacy, confidentiality, IP or other rights
    - Allowed re-use of the data, derivative products
  - Metadata standards to be used
  - How long data will be retained
  - Provisions for archiving and preserving for the long-term
  - Plan for format migration
How Can the Libraries Help?

- Consultations with faculty and/or RA for
  - Creating a data management plan
  - Raising awareness of data management best practices

- Creating Metadata
  - What is it?
  - Why is it important in providing access/discoverability

- Help with Sharing Your Data:
  - E-Scholarship@BC
  - Identifying relevant repositories
BC Librarians Can Help with Data Sharing

• **E-Scholarship@bc**
  – A repository for data sets
  – A portal for pointing to your data wherever it is stored (at BC or beyond)

• Assistance in identifying relevant repositories (subject, institutional)
Stage 2: Data Management in Action

Follow best practices while collecting or generating your data (you and/or your graduate students).

- Storage
- Documentation
- Loss Prevention
- Security

To start the discussion: What kind of data do you collect? What is the format?
Best Practices: Handling/Storing/Back up Data

Data Storage Elements to Consider:

• File Formats
• File Naming
• Directory Structure
• Version Control
• Assign Responsibility

Librarians are happy to meet with you or your Research Assistants to talk about best practices!
File Formats

Consider formats for sharing, re-use and preservation.

Open standards support sharing and archiving for use in the future.

Chart from *Managing and Sharing Data*, The UK Data Archive 2011.

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative tabular data with extensive metadata</td>
<td>SPSS portable format (.sav)</td>
</tr>
<tr>
<td>a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data</td>
<td>delimited text and command (&quot;setup&quot;) file (SPSS, Stata, SAS, etc.) containing metadata information</td>
</tr>
<tr>
<td></td>
<td>some structured text or mark-up file containing metadata information, e.g. DDI XML file</td>
</tr>
<tr>
<td>Quantitative tabular data with minimal metadata</td>
<td>comma-separated values (CSV) file (.csv)</td>
</tr>
<tr>
<td>a matrix of data with or without column headings or variable names, but no other metadata or labelling</td>
<td>tab-delimited file (.tab)</td>
</tr>
<tr>
<td></td>
<td>including delimited text of given character set with SQL data definition statements where appropriate</td>
</tr>
<tr>
<td>Geospatial data</td>
<td>ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn)</td>
</tr>
<tr>
<td>vector and raster data</td>
<td>geo-referenced TIFF (.tif, .tiff)</td>
</tr>
<tr>
<td></td>
<td>CAD data (.dwg)</td>
</tr>
<tr>
<td></td>
<td>tabular GIS attribute data</td>
</tr>
<tr>
<td>Qualitative data</td>
<td>eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xsl)</td>
</tr>
<tr>
<td>textual</td>
<td>Rich Text Format (.rtf)</td>
</tr>
<tr>
<td></td>
<td>plain text data, ASCII (.txt)</td>
</tr>
<tr>
<td>Digital image data</td>
<td>TIFF version 6 uncompressed (.tif)</td>
</tr>
<tr>
<td>Digital audio data</td>
<td>Free Lossless Audio Codec (FLAC) (.flac)</td>
</tr>
<tr>
<td>Digital video data</td>
<td>MPEG-4 (.mp4)</td>
</tr>
<tr>
<td></td>
<td>motion JPEG 2000 (.jpg)</td>
</tr>
<tr>
<td>Documentation</td>
<td>Rich Text Format (.rtf)</td>
</tr>
<tr>
<td></td>
<td>PDF/A or PDF (.pdf)</td>
</tr>
<tr>
<td></td>
<td>OpenDocument Text (.odt)</td>
</tr>
</tbody>
</table>
Examples of preferred file formats

• TXT, PDF/PDF Archival, not Word (doc, docx)
• ASCII, not Excel (xls, xlsx)
• MPEG-4, not Quicktime (qtff)
• TIFF or JPEG2000, not GIF or JPG
• XML or RDF, not RDBMS

Ideally, save files in both original format AND one of the preferred ones listed above.
Why bother with non-proprietary file formats?

• No restrictions on their use
• Open source code → future migration easier
• Propriety formats are offered by companies that may go out of business, carrying the code knowledge with them
Organization

File Naming Conventions Best Practices
- Consistent and descriptive
- Avoid spaces and special characters
- Use brief names
- Can contain:
  - Project acronyms
  - Researchers’ initials
  - File type information
  - Version number
  - Date
  - File Status

IUS_v02_092011_final.csv
Internet Usage Study version 2, Sept 2011, final draft, in csv format
Directory Structure

• Use folders!

• Three ways to organize:
  – By types of data
    • databases, text, images, models, sound
  – By research activities
    • interviews, surveys, focus groups
  – By materials
    • data, documentation, publications
Version Control

- What is a version and why should it be controlled?
- How do you ensure authenticity?

**VERSION CONTROL TABLE FOR A DATA FILE**

<table>
<thead>
<tr>
<th>Title:</th>
<th>Vision screening tests in Essex nurseries</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Name:</td>
<td>VisionScreenResults_00_05</td>
</tr>
<tr>
<td>Description:</td>
<td>Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007</td>
</tr>
<tr>
<td>Created By:</td>
<td>Chris Wilkinson</td>
</tr>
<tr>
<td>Maintained By:</td>
<td>Sally Watsley</td>
</tr>
<tr>
<td>Created:</td>
<td>04/07/2007</td>
</tr>
<tr>
<td>Last Modified:</td>
<td>25/11/2007</td>
</tr>
<tr>
<td>Based on:</td>
<td>VisionScreenDatabaseDesign_02_00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VERSION</th>
<th>RESPONSIBLE</th>
<th>NOTES</th>
<th>LAST AMENDED</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_05</td>
<td>Sally Watsley</td>
<td>Version 00_03 and 00_04 compared and merged by SW</td>
<td>25/11/2007</td>
</tr>
<tr>
<td>00_04</td>
<td>Vani Yussu</td>
<td>Entries checked by VY, independent from SK</td>
<td>17/10/2007</td>
</tr>
<tr>
<td>00_03</td>
<td>Steve Knight</td>
<td>Entries checked by SK</td>
<td>29/07/2007</td>
</tr>
<tr>
<td>00_02</td>
<td>Karin Mills</td>
<td>Test results 81-120 entered</td>
<td>05/07/2007</td>
</tr>
<tr>
<td>00_01</td>
<td>Karin Mills</td>
<td>Test results 1-80 entered</td>
<td>04/07/2007</td>
</tr>
</tbody>
</table>

Chart from *Managing and Sharing Data*, The UK Data Archive 2011.
Data Documentation

• What is metadata?

• Benefits of good documentation

• What elements should be documented?

ISO suggested Minimum Data Elements

- Title
- Creator (Principal Investigators)
- Date Created (also versions)
- Format (and software required)
- Subject
- Unique Identifier
- Description of the specific data resource
- Coverage of the data (spatial or temporal)
- Publishing Organization
- Type of Resource
- Rights
- Funding or Grant

To contact your subject specialist:
www.bc.edu/libraries/help/askalib.html
What is a schema?
What standards do organizations use?

• Metadata schema
  • Set of specific elements used to record information about an item, object, data set, etc.
  • A **standard** is a schema that has been *standardized*
    • An organization in a discipline determined which elements were appropriate characterizing data in their field

• A few examples:
  • Dublin Core – a basic, all-purpose schema
  • MARC – widely used in library cataloguing
  • The Federal Geographic Data Committee’s Geospatial Metadata
Finding Your Schema ... Or, Identifying Relevant Metadata Standards

• If your discipline has no standard
  – ISO’s suggested minimal metadata elements
    OR
  – use the schema in use by your disciplinary repository

• Consult with your library subject specialist
Data Documentation – What do you do with it once you have it?

• Record it in a readme.txt file
• Keep definitions of your metadata terms; these can be compiled into a “codebook”
  • Codebooks are also used to record methodology and other data management notes (e.g. IRB compliance statements, etc.)
• Inserted with deposited data it facilitates “discovery” of your data on the Web
Data Loss Prevention

• Regular back-ups protect against data loss
• Back up strategy will depend on your needs:
  – Back up all versions of the files or certain ones?
  – How often will you back up files?
• Have at least two back up locations
  – internal (your computer)
  – external (i.e. the BC Research Data Archive)
## Physical Storage Options

<table>
<thead>
<tr>
<th>Local</th>
<th>Centralized</th>
<th>Remote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convenient but less secure (especially external media)</td>
<td>More secure, with automatic back-up ... and more space</td>
<td>Permanent, someone else takes responsibility for future migration</td>
</tr>
<tr>
<td>• On your own computer’s hard drive</td>
<td>• ITS</td>
<td>• Disciplinary Repositories, e.g. GenBank, Cambridge Structure Database</td>
</tr>
<tr>
<td>• External media (hard drive, CD/DVD, flash drive)</td>
<td></td>
<td>• Secure cloud options are in use at other institutions</td>
</tr>
<tr>
<td>• Departmental server, local network access</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[1] On your own computer’s hard drive
[3] Departmental server, local network access
**Data Storage**

*ITS* offers a remote, automated backup of faculty and staff computers using a product called Connected Backup by Autonomy (formerly owned by Iron Mountain). Users of the service are offered automated backup of their computer and have the ability to recover files from any location using a web browser.

http://www.bc.edu/offices/help/essentials/backup/ironmtn.html

*Research Services* provides secure archive space for research data that is backed up nightly.

http://www.bc.edu/offices/researchservices/dataresources/archive.html
There are Cost Implications in Long-term Data Storage

• Who will pay for this?

• NSF DMP guidelines encourage inclusion of information on costs (and grants may pay for them):

  “Cost of documenting, preparing, publishing, disseminating and sharing research findings and supporting material are allowable charges against the grant. (See AAG Chapter V.B.7.)”

  From the NSF Award and Administration Guide (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)
# Data Security

<table>
<thead>
<tr>
<th>Physical Data Security</th>
<th>Network Security</th>
<th>Computer Systems Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Access to rooms and buildings where data is held is controlled</td>
<td>• Do not store confidential data on computers or servers connected to an external network</td>
<td>• Access to computers is controlled with passwords.</td>
</tr>
<tr>
<td>• Access to data is logged</td>
<td>• Computers where data is stored have firewalls and virus protection</td>
<td>• Implementing password protection of, and controlled access to, data files, e.g. no access, read only, read and write or administrator-only permission</td>
</tr>
<tr>
<td>• Data is moved only when necessary</td>
<td></td>
<td>• Access to restricted materials is controlled with encryption</td>
</tr>
</tbody>
</table>
Data Security

For additional assistance with security planning, consult the Computer Policy & Security Office of the IT Assurance Department.

Director: David Escalante

http://www.bc.edu/offices/its/depts/assurance/policysecurity.html
Data Retention

Generally data must be retained for three years from the date of project submission or final financial report. For additional assistance, contact Dr. Stephen Erickson at the Boston College Office for Research Integrity and Compliance:

http://www.bc.edu/content/bc/research/oric/compliance.html
Stage 3: Data Access and Sharing

Options include:

- Personal website
- Institutional repository, e.g. eScholarship@bc
- Journal “supplementary materials”
- Disciplinary (or multidisciplinary) repository
- Or, a combination of above: journal-designated repository – *Nature* example)
Data Sharing Options Beyond BC

• Find subject-based archives – ask your subject librarian

• Repository Directories:
  – DataCite
    • http://datacite.org/repolist
  – DataBib (Beta)
    • http://databib.org/index.php#
  – Simmons Data Repositories Listing
    • http://oad.simmons.edu/oadwiki/Data_repositories
Some Examples of Disciplinary Repositories

• The Social Sciences
  – ICPSR (Interuniversity Consortium for Political and Social Research)
• Biomedicine
  – GenBank
  – RSCB Protein DataBank
• Chemistry:
  – Cambridge Structural Database
  – PubChem
• Environmental Sciences
  – Dryad
• Humanities
  – Cultural Policy and the Arts National Database (CPANDA) (Princeton University)
**Sample GenBank Record**

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>SCU99855</th>
<th>5028 bp DNA</th>
<th>PLN</th>
<th>21-JUN-1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Saccharomyces cerevisiae TCP1-beta gene, partial cts, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U49845</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>U49845.1</td>
<td>GI:1293613</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Saccharomyces cerevisiae (baker's yeast)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ORGANISM    | Saccharomyces cerevisiae  
Eukarya; Fungi; Ascomycota; Saccharomycotina; Saccharomyces;  
Saccharomyceseae; Saccharomyceseae; Saccharomyces. |
| REFERENCE   | 1 (bases 1 to 5028) |
| AUTHORS     | Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W. |
| TITLE       | Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae |
| JOURNAL     | Yeast 10 (11), 1503-1509 (1994) |
| PUBLMED     | 78716390 |
| REFERENCE   | 2 (bases 1 to 5028) |
| AUTHORS     | Roemer,T., Madden,K., Chang,J. and Snyder,H. |
| TITLE       | Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein |
| JOURNAL     | Genes Dev. 10 (7), 777-793 (1996) |
| PUBLMED     | 8546515 |
| REFERENCE   | 3 (bases 1 to 5028) |
| AUTHORS     | Roemer,T. |
| TITLE       | Direct Submission |
Stage 3: (Not just) Data Sharing ... but also Archiving

What does the Data Sharing Policy Mean?

Example NSF: “plans for archiving data, samples, and other research products, and for preservation of access to them.”

Archiving Data means not just preserving the data in the original format but also in a format that is non-platform reliant, using a standard that ensures that the data can be re-used in the future.

Metadata is vital to insure data is findable using terminology accepted in the field.
Ethics and Privacy

• Sensitive data should be redacted before depositing in a public archive or repository.

• Access to data may need to be embargoed (limited for a certain amount of time) for confidentiality or other reasons.

• Dark archives ensure permanent protection of confidentiality.
Institutional Review Board

The Boston College Institutional Review Board's mission is to protect the rights and welfare of people who take part in research at Boston College.

Any research involving human subjects must be reviewed and approved by the IRB.

http://www.bc.edu/research/oric/human.html
Data Ownership

You may have copyright or ownership concerns when planning to share your data. For assistance and more information, please contact the Boston College Office for Research Integrity and Compliance:

http://www.bc.edu/content/bc/research/oric/compliance.html
The NSF expects that you will share your data within a reasonable amount of time...

However, it also recognizes the need to protect intellectual property rights and potential commercial value.

The DMP should describe your plans to protect those rights.

If you have concerns/questions, plan to meet with the Boston College Office for Technology Transfer as part of your DMP writing process.
Intellectual Property/Technology Transfer Concerns

From a sample NSF DMP (Social and Behavioral Sciences) posted on the Rice University Website (http://osr.rice.edu/forms/dataManagementPlans.pdf)

“We do not anticipate that significant intellectual property issues involved with these data will arise. However, in the event that discoveries or inventions are made in direct connection with these data, access to the data will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made.”
Why should I cite data?

• Ensures that original producers of the data (you!) are credited in citation indexes.*
• Allows researchers to locate research data used in an article.
• May be required by the archive that stored the data you have repurposed.

How to Cite Data Sets

A data set citation should include at least the following elements, which will be arranged depending on the style you use:

- author or creator
- title or description
- year of publication
- publisher and/or the database/archive from which it was retrieved
- the URL or DOI if the data set is online


Always review a journal's guidelines before formatting your data citations, for they may want you to use their house style.

For example, if you are submitting an article to Nature, data sets are not cited in the references list. Rather, authors should cite an accession number and URL in the main text where they are discussed.
The Digital Research Data Lifecycle

- Funding
- Data Collection
- Data Sharing
- Research Output

Arrows indicate the flow or cycle of the processes.
Additional Support

The Data Management LibGuide
http://libguides.bc.edu/dataplan

Subject Specialists
www.bc.edu/libraries/help/askalib.html

Institutional Review Board
http://www.bc.edu/research/oric/human.html

ITS/Research Services
http://www.bc.edu/offices/researchservices/

Office for Research Integrity and Compliance:
http://www.bc.edu/content/bc/research/oric/compliance.html
Questions?