Data management 101

General Guidelines for Effective Data Management

Rolando Garcia-Milian and Hannah Norton
UF Health Sciences Center Library
Rolando.milian@ufl.edu / nortonh@ufl.edu
UF Research Computing Day - April 25, 2012
General Guidelines / Best Practices

- Planning (DMP – Norton’s presentation)
- Metadata
- Formatting
- Storing
- Security
- Copyright
- Sharing
Benefits of proper data management

- Data is evidence supporting/refuting models in science
- Efficient use of resources
- Effective protection
- Preservation and re-use through data sharing and collaboration
- High quality results
- Research excellence
- Advancing science
Challenges of data management

- Planning
- Organization
- Documenting
- Formatting
- Submitting
- Answer questions?
- Data errors/mistakes?
- Being scooped?
- Public resistance?
Tools for data management

Data Asset Framework

The Data Asset Framework (formerly the Data Audit Framework) provides organisations with the means to identify, locate, describe and assess how they are managing their research data assets.

Welcome to the Data Curation Profiles community!
Results of poor data management

Table 2. Data reporting problems in the scientific literature, according to Marco and Larkin (2000).

- Failing to include the number of eligible participants
- Reporting of missing data points inaccurately
- Failing to report all pertinent data
- Failing to report negative results
- Allowing research sponsors to influence reporting of results
- Labeling graphs inappropriately
- Reporting percentages rather than actual numbers
- Reporting results of inappropriately applied statistical tests
- Reporting differences when statistical significance is not reached
- Reporting no difference, when power is inadequate
- Performing multiple comparisons without correction
- Splitting data into multiple publications
- Using terminology without precise definitions
- Reporting conclusions not supported by the data
- Ignoring citations of prior work that challenge stated conclusions
- Inflating research results for the media

Metadata
Annotation
Documenting
Metadata (Annotation/Documenting)

Metadata
Information about data: the information required to understand data, context, quality, structure, and accessibility (Michener et al., 1997)

-Who, what, when, where, and how about every aspect of the data.
Metadata (Annotation/Documenting)

Benefits of proper metadata
- Reuse and data sharing are facilitated
- Data discovery
- Expand the scale of study
- Addresses unanticipated questions
- Integrate data

http://www.flickr.com/photos/boojee/3743753784/in/photostream/
Use standardized taxonomies and controlled vocabularies including domain, national, and international standards in the capture, management and archiving of data.
Metadata (Annotation/ Documenting)

Automatic addition of metadata
-Some is automatically added during the data collection or analysis process- i.e. date, time

-Some software (e.g. R statistical package, MATLAB, SAS, Galaxy) provide analysis scripts - records of the various steps involved in processing and analyzing data, and provide a form of “analytical metadata.” always leave record of what you did with your data,
Metadata (Annotation/ Documenting)

User interface-driven analysis - changes to data are made by selecting steps from drop-down menus, followed by a “run” or “execute” or “ok” button rarely leave a clear accounting of exactly what you have done
Metadata (Annotation/Documenting)

Manually added metadata

About the project
-Title, people, key dates, funders and grants

About the data
-Title, key dates, creator(s), subjects, rights, included files, format(s), versions
-Interpretive aids: codebooks, data dictionaries, algorithms, code
Metadata (Annotation/Documenting)

Keep a README file for each data file

- Plain text files
- Short description of what data it includes
- Who collected the data and whom to contact with questions
- Column headings for any tabular data
- Units of measurement used
- Symbols used
- Specialized formats or abbreviations used

http://datadryad.org/handle/10255/dryad.8525
Formatting Your Data

http://www.ehow.co.uk/how_8510149_make-excel-spreadsheets-look-good.htm
Formatting Your Data

File formats in which data is created depend on:
- Software in which research data are created and digitized
- How researchers plan to analyze data
- Hardware used
- Availability of software
- Discipline-specific
Formatting Your Data

Organizing Files and Folders:
- Essential for accessibility
- Makes it easier to find and keep track of data files.
- Develop a system that works for your project
- Be consistent

http://jdorganizer.blogspot.com/2008/03/file-folders-declare-that-you-are.html
Formatting Your Data

File names:
- Use file names to classify broad types of files
- Create meaningful but brief names
  “Year01” or “Fall03” vs “Corvallis_VegBiodiv_2007”
- Capitalize each word to differentiate it.
- Avoid using special characters in a file name.
  \/:*?“<>|[]\&$
Formatting Your Data

File names:
- Use underscore or hyphen symbols instead of spaces "_" or "-"
- Capture place, time, and theme – extremely useful, even if done in a highly abbreviated manner
- Reverse dates so they sort usefully YYYYMMDD e.g. filenaming_20080507
- Capture document version control
  v01, v02, v03 instead of filenaming_lastestversion
Formatting Your Data for Storage

Store data in nonproprietary software formats (e.g., comma delimited text file, .csv); proprietary software (e.g., Excel, Access) may become unavailable, whereas text files can always be read.

NOTE: When data are converted from one format to another, certain changes may occur to the data. After conversions, data should be checked for errors or changes that may be caused by this process.
## Formatting Your Data for Storage

### Recommended File Formats for Preservation

<table>
<thead>
<tr>
<th>Textual Formats</th>
<th>File Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrobat PDF/A</td>
<td>.pdf</td>
</tr>
<tr>
<td>Comma-Separated Values</td>
<td>.csv</td>
</tr>
<tr>
<td>Open Office Formats</td>
<td>.odt, .ods, .odp</td>
</tr>
<tr>
<td>Plain Text (US-ASCII, UTF-8)</td>
<td>.txt</td>
</tr>
<tr>
<td>XML</td>
<td>.xml</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image/Graphic Formats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>.jpg</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>.jp2</td>
</tr>
<tr>
<td>PNG</td>
<td>.png</td>
</tr>
<tr>
<td>SVG 1.1 (no Java binding)</td>
<td>.svg</td>
</tr>
<tr>
<td>TIFF</td>
<td>.tif, .tiff</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Audio Formats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIFF</td>
<td>.aif, .aiff</td>
</tr>
<tr>
<td>WAVE</td>
<td>.wav</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video Formats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AVI (uncompressed)</td>
<td>.avi</td>
</tr>
<tr>
<td>Motion JPEG2000</td>
<td>.mj2, .mjp2</td>
</tr>
</tbody>
</table>

Recommended File Formats for Preservation. University of Texas [http://repositories.lib.utexas.edu/recommended_file_formats](http://repositories.lib.utexas.edu/recommended_file_formats)
Storing Your Data

Storing Your Data

-Store data in nonproprietary hardware formats

Formats can rapidly become obsolete valuable data that are essentially lost because they are trapped on old formats, 5.25” floppy disks

CD/DVD experiential life expectancy is 2 to 5 years even though published life expectancies are often cited as 10 years, 25 years, or longer

Manufacturers claim that CD-R and DVD-R discs have a shelf life of 5 to 10 years before recording on them (U.S. National Archives)
Storing Your Data

Always store an uncorrected (the original data set) data file version or **master version**:

- Do not make any corrections to this file
- Make corrections using a scripted language.
- Consider making your original data file read-only
- Limit access to this file
Storing Your Data

- Whenever possible, use online storage (i.e. Dropbox) or institutional resources

http://www.hpc.ufl.edu/about/newStorage.php
Storing Your Data

Regular back-ups protect against accidental data loss:

- hardware failure
- software or media faults
- virus infection or malicious hacking
- power failure
- human errors

Ensure that areas and rooms for data storage are structurally sound, and free from the risk of flood and fire.

Data Security

Security

Unrestricted Data
If available to the public, will not harm an individual, group, or institution

Sensitive Data
If available to unauthorized users, may harm an individual, a group or institution

Restricted Data
Highest level of protection: i.e. Patient data, student data, security-related data such as passwords and risk assessments, and intellectual property

UF IT Data Security Standard
http://www.it.ufl.edu/policies/security/uf-it-sec-data.html
Security

DATA SECURITY AND ACCESS

- Physical security

- Network security

- Security of computer systems and files

Security

When working with Restricted Data
AVOID:

-Storing data on workstations, portable devices or removable media.
-Sending data in email or instant messages.
-Using data on unapproved web sites.
-Removing data from UF premises.

Security

Information Technology
Security

Kathy Bergsma, UF Information Security Manager

392-2061
ufirt@ufl.edu
http://infosec.ufl.edu/

About the HSC

Contact Information

Colleen Ebel
HSC Chief, Information Security

Phone: 352-273-5014
Mailing Address: P.O. Box 100152
Gainesville, FL 32610-0152
E-mail: cebel@ufl.edu
Web Site: security.health.ufl.edu

› back to HSC Organization
Security

Information Privacy

The University of Florida values individuals' privacy and actively seeks to preserve the privacy rights of those who share information with us. Your trust is important to us and we believe you have the right to know how information submitted to the University of Florida is generally handled.

We are dedicated to preventing unauthorized access to information, maintaining the accuracy of information, and ensuring the appropriate use of information. We strive to put in place appropriate physical, electronic, and managerial safeguards to secure the information we collect in all formats: on paper, electronically, and verbally. These security practices are consistent with the policies of the university and with the laws and regulatory practices of the State of Florida and multiple federal agencies.

Privacy Incidents

University of Florida officials notify affected individuals in the case of a privacy breach.

- State of Florida's Unclaimed Property Website Breach
- College of Engineering Hard Drives Stolen
- Physics Department Server Breach
- Cardiothoracic Patient Research Breach
- Epidemiology and Health Policy Research Breach
Security

DATA DISPOSAL

For hard drives, simply deleting does not erase a file on most systems. Files need to be overwritten to ensure they are effectively scrambled.

Shredders certified to an appropriate security level should be used for destroying paper and CD/DVD discs.

External hard drives at the end of their life can be removed from their casings and disposed of securely through physical destruction.

Contact your IT person.

Security

UF Restricted Data Required Training

Kathy Bergsma
UF Information Security Manager
ufirt@ufl.edu
http://infosec.ufl.edu

Copyright

Give credit to the data source used, the data distributor and the copyright holder

In the case of collaborative research, copyright may be held jointly by various researchers or institutions.

Secondary users of data must obtain copyright clearance from the rights holder before data can be reproduced.

Data can be copied for non-commercial teaching or research purposes without infringing copyright, under the fair dealing concept, providing that the owner of the data is acknowledged.
Copyright

UF Office of Technology Licensing
http://www.research.ufl.edu/otl/index.html

UF Intellectual Property Policy

Christine Ross – Copyright on Campus
http://guides.uflib.ufl.edu/copyright
Sharing Your Data

http://www.amazon.com/Sharing-Toddler-Tools-Elizabeth-Verdick/dp/1575423146/ref=sr_1_1?ie=UTF8&qid=1335134736&sr=1-1
Sharing Your Data

WHY SHARE RESEARCH DATA

- Encourage scientific debate
- Promotes potential new uses of data
- New collaborations
- Improvement and validation of research methods
- Increases impact and visibility of research
- Promotes the research study and its outcomes
- Required by journals/funding agencies
- Provide direct credit to the researcher
Sharing Your Data

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Frisimna
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

To add a note, highlight some text. Hide notes
Make a general comment

http://www.amazon.com/Sharing-Toddler-Tools-Elizabeth-Verdick/dp/1575423146/ref=sr_1_1?ie=UTF8&qid=1335134736&sr=1-1
Sharing Your Data

HOW TO SHARE YOUR RESEARCH DATA

- Depositing with a specialist or discipline-specific data repository
- Submitting to a journal to support a publication
- Depositing in an institutional repository
- Available online via a project or institutional website
- Available informally between researchers on a peer-to-peer basis
Sharing Your Data

A comprehensive list of data repositories by disciplines
http://oad.simmons.edu/oadwiki/Data_repositories
Sharing Your Data
Sharing Your Data

Advantages of depositing data with a data repository

- Assurance that data meet set quality standards
- Safe-keeping of data in a secure environment with the ability to control access where required
- Standardized citation mechanism to acknowledge data
- Promotion of data to many users
- Online resource discovery of data through data catalogues
- Monitoring of the secondary usage of data
References


Data Repositories. http://oad.simmons.edu/oadwiki/Data_repositories


References

Jones, S., Ross, S., and Ruusalepp, R., Data Audit Framework Methodology, draft for discussion, version 1.8, (Glasgow, HATII, May 2009)

Kruse R.L., and Mehr D.R. 2008. Data management for prospective research studies using SAS® Software. BMC Medical Research Methodology 8: 61-


References

Recommended file formats for long-term preservation. University of Texas http://repositories.lib.utexas.edu/recommended_file_formats


UK - Joint. Info. Sys. Comm.- Choosing a file name
www.jiscdigitalmedia.ac.uk/crossmedia/advice/choosing-a-file-name

University of Edinburgh Records Management Section, Standard Naming Conventions For Electronic Records: The Rules,
www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RMprojects/PP/FileNameRules/Rules.htm