Data Management Best Practices

Ryan Womack and Aletia Morgan
Rutgers University Libraries
Research Data Services

February 26, 2015
The Data Lifecycle
Why Data Management?

• Best Case Scenarios, Data is...
  – Robust
  – Recoverable
  – Reliable
  – Reusable
  – Reproducible
  – Reputable
  – Renowned
So What Can Go Wrong?

• Worst case scenarios
  – **Lost data** can happen anywhere, even **in the cloud**
  – **Stolen data**
  – **Unreadable data**
  – **Unverifiable data**
  – And many other **horror stories**
Data Preservation Considerations

• Access Control

• Versioning

• Backup
  – Local
  – Shared
    • Campus
    • Offsite

See [http://datalib.edina.ac.uk/mantra/storageandsecurity/](http://datalib.edina.ac.uk/mantra/storageandsecurity/) for more information
Understand your data

• Quantity (MB, GB, TB)
  – will affect your decisions about how to store, package, transport, and backup your data
  – Large numbers of discrete files, regardless of size, may be harder to handle
  – Multiple versions? Frequent updates?

• Format
  – Seek out open formats or at least commonly used formats (.xlsx is ok, consider .csv)
  – Both for preservation and reuse, dependency on a single option can be a problem, even if it is the most convenient in the short-term.

• Rights and permissions
  – Reuse of other data sources (licensed or otherwise) may require investigation and restrict data sharing options
  – Confidentiality of human subjects or guarding of information for patent protection may be reasons to restrict data
  – Give appropriate credit to others’ contributions
Classifying your Data

What types of data need to be managed and stored over the course of your project?

• Raw Data

• Working Data

• Processed or Final Data

• Preserved Data for Possible Re-Use
Security

Security includes both

- Physical Security
- Access Security (Logical/Network/Password)

Sensitive data (e.g., involving personally identifiable information) requires a more careful approach to security
Organizing your Data

Key Considerations

• Who controls the data for which phases of the project?

• A logical structure and plan will help during and after data creation, for directories and filenames (e.g. project directory has standard locations for raw data, analysis, code, graphs, etc.)

• Project folders with well-documented naming conventions (e.g., date of data creation as part of file name in a structured format (ISO is yyyy-mm-dd). See ARM example.

• Versioning scheme should be agreed on and documented.

• Goal is to have unique and understandable identifiers so that different parts of a project can merge and be handled easily over time, even if participants and conditions change.
Documenting your Data

• Documentation helps
  — with your own reuse of the data at a later date,
  — others in your research group working with the data
  — with the long-term reuse potential of your data

• A Codebook is a structured way of explaining the contents of your data file

• Readme files are a well-understood way of communicating information about the contents and setup of your data

• Documentation can also take the form of a simple text or Word document

• Include any explanations of experimental methods, software code run on the data, and any other tools needed to work with the data

• Your previous work on creating and describing a consistent organization for your data will help here.

• See http://datalib.edina.ac.uk/mantra/documentation_metadata_citation/ for more information
Reproducible Research

• Ideally, someone else can grab your data project as a complete bundle of data, documentation, and software code, and recreate the analysis to get exactly the same results.

• Reports and data can be integrated so that live analysis run on actual data can be placed in reports (some R packages do this).

• Many initiatives are advancing the concept of reproducible research.

• This high standard of evidence and validation is an assurance that data and conclusions are not flawed (or faked).

• Good data management practices lay the groundwork for success in reproducible research.

• Upcoming workshop on March 12 delves into reproducible research in more detail.
Data Management Plans

• Many sponsored research agencies now require a "Data Management Plan" (DMP) as a component of any proposal.

• The DMP is a formal document that outlines what the PI will do with data during and after completion of a funded research project.

• The specific requirements for the DMP vary by funder, and by research subject, with most emphasizing preservation and data sharing.

• Data Management Best Practices are the foundation of a good DMP proposal and effective Data Sharing.

• **Workshop on March 5** goes in depth into DMP and Data Sharing.
The Data Lifecycle

Reuse and Citation
Further References

- Australian National Data Service: Data Management for Researchers
- Australian National University: Data Management
- CIESIN: Geospatial Electronic Records
- ICPSR Guide to Social Science Data Preparation and Archiving (pdf):
- Oak Ridge National Laboratory: Best Practices for Preparing Environmental Data Sets to Share and Archive