Data Management, Ethics, and Data Sharing

ENHANCED RESEARCH SKILLS CERTIFICATE PROGRAM - WORKSHOP 4

FEBRUARY 3, 2016

VICTORIA MITCHELL – VMITCH@UOREGON.EDU
BRIAN WESTRA – BWESTRA@UOREGON.EDU

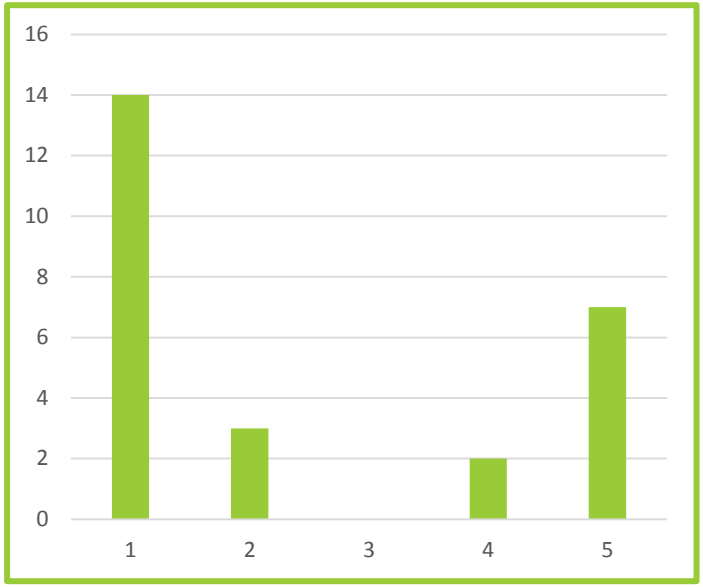
Review

Assessment questions:

I feel confident that I can…

- describe one reason why data management is important…
- begin using a file naming system…
- read and interpret basic metadata...
- Etc.

~60% agree but **~30% disagree**



Data Ownership

Who owns the data you are working with?

- You
- Your advisor
- The research team
- UO
- The funding agency

Research Integrity…

…depends on you and the integrity of the data

Data Integrity…

Data having a complete or whole structure

- Accuracy
- Consistency

Data that follows a predefined set of rules (constraints).

(more often considered in the context of social science and science research)

Data Integrity rules/constraints

Entries for US zipcode have 5 (+4) digits: _ _ _ _ _ - _ _ _ _

All dates are in the following format: YYYY-MM-DD

Numeric columns/cells do not contain alphabetic data

No two participants in the research study may have the same ID

Each row in the table contains only one observation

Data Integrity – after collection

Data should not be transformed or modified in such a way as to invalidate accuracy or completeness of the data.

Unpack Data Integrity

Planning/Selection:
Determination of appropriate data type, source and instruments that allow investigators to adequately answer research questions

Collection:
Process of gathering and measuring information in an established systematic fashion that makes it possible to answer research qs, test hypotheses…

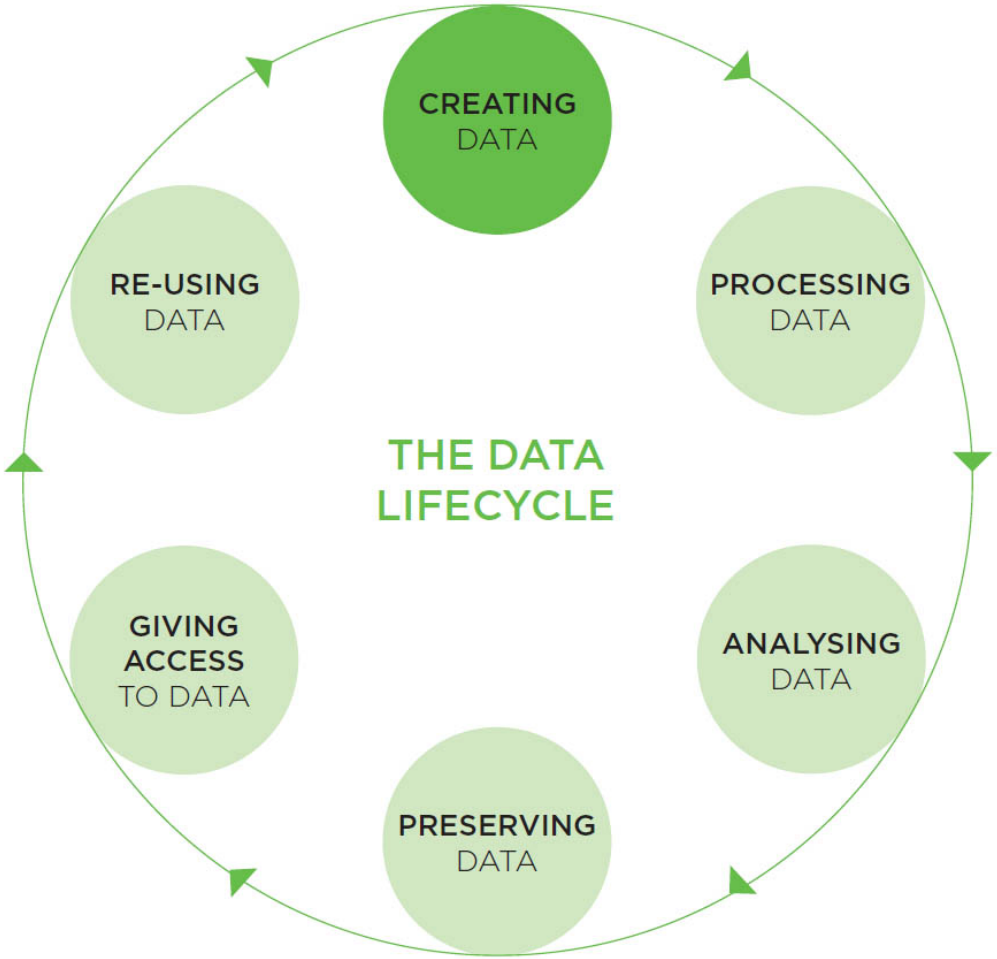
Handling:
Processes to ensure research data is stored, archived, disposed in a safe and secure manner, during and after research project

Analysis:
Process of systematically applying techniques to describe and illustrate, recap and condense, and evaluate data

Reporting and publication: preparing and disseminated research findings

Ownership: possession of, and responsibility for, information

From http://www.slideshare.net/JISC/paul-jeffreys-research-integrity-institutional-responsibility/22


CREATING DATA
PROCESSING DATA
ANALYSING DATA
PRESERVING DATA
GIVING ACCESS TO DATA
RE-USING DATA
THE DATA LIFECYCLE

Quality Control in Practice

Quality Control in Practice

- Set of processes, procedures, and activities associated with monitoring, detection, and action during and after data collection.
- Examples:
  - Errors in individual data fields
  - Systematic errors
  - Violation of protocol
  - Staff performance issues
  - Fraud or scientific misconduct

Source: Dept of Biostatistics –Data Management, IUSM

Data Integrity and Quality Control

Activity: http://bit.ly/1SYRLkK

Lessons learned from Doorn article, Retraction Watch page, and Konczal blog post:

- Name and describe three specific data practices that caused problems for the researchers in one or more of these articles.
- What was the impact to the authors, fellow researchers, and the public?
- Describe alternatives that would have helped them avoid the problems.

Data Restrictions and Protections

- Protection of privacy, confidentiality
- Intellectual Property
- Embargo (thesis data) before article publication

Why might data use or sharing be restricted?

- Threatened and endangered species
- National security and classified research
- Export controls
  - Can apply to technologies and data
- Use of Human Subjects
  - Personally identifiable information of any kind
    - *E.g.*, HIPAA as governing law for personal health information

Slide adapted from DataONE

Privacy and Security

- What we can collect and how
- How we share data, results and outcomes
- Reuse of human subject data
- Data storage and destruction
- IRB interpretations and review across institutions are not always consistent

Slide adapted from DataONE

Privacy vs. Confidentiality vs. Anonymity

Privacy

- Protects access to individuals (or entities)

Confidentiality

- Protects access to information *about* individuals
- Can be thought of as information privacy

- Anonymity
  - Means that either the project does **not collect** identifying information of individual subjects or the project **cannot link** individual responses with participants' identities.

Slide adapted from DataONE

Disclosure Risk & Mitigation

- What are the risks?
- Informed consent
- Vulnerable populations
- Direct identifiers
- Indirect identifiers

Direct and Indirect Identifiers

Direct	Indirect
Name	Detailed geographic information (e.g., state, county, province, or census tract of residence)
Address	Organizations to which the respondent belongs
Relatives' names or addresses	Educational institutions (from which the respondent graduated and year of graduation)
Phone/fax numbers	Detailed occupational titles
E-mail address	Place where respondent grew up
Social security numbers or member/account numbers	Exact dates of events (birth, death, marriage, divorce)
Full face photos	Offices or posts held by respondent
Fingerprints	Detailed income

Identifiers Activity

Variable	Direct Identifier	Indirect Identifier	??
Name			
Gender			
Age			
Birthdate			
Income			
Occupation			
State			
Zip Code			
Census Tract			
Address			

Data Sharing Options

Name one pro and one con for sharing your data by:

- Posting on a web site
- Publishing in a journal
- Making it available on request

Sharing Data – Considerations

1. Discoverability – put it where others (people and search engines) will find it
2. Description – describe it so others can find it
3. Utility/applicability – include annotation (metadata)
4. Format – software considerations and standards
5. Citability (DOI)
6. Permission (Creative Commons) and ownership.

Data Repositories

- Open data repositories
  - A relatively new phenomenon
  - Mostly in the sciences and social sciences
- Data enclaves
  - Sensitive data can be deposited in these – researchers have to apply for access
  - Can be physical or virtual
- Other types of repositories
  - Can search and find data, but not necessarily use it – e.g. must be a member (ICSPR) or submit a specific request

Terms and restrictions on use

- Data use agreements
  - Set terms for using (sensitive) data, including security measures, who can access the data, what to do with the data when you're finished with your research
- Creative Commons licenses
  - Very handy for stating what you want other people to be able to do with your data (or other creative or research products)
  - An array of levels available from no restrictions at all to several

You can search or browse for repositories with data you can use at:

Re3data http://service.re3data.org/search

UO Library Data & Statistics Guide http://researchguides.uoregon.edu/data-stats

Activity

Explore one or more of these example repositories and look for requirements regarding access, reuse and deposit:

Arts and Humanities

Jewish Databank http://www.jewishdatabank.org/

National Archive of Data on Arts & Culture http://www.icpsr.umich.edu/icpsrweb/NADAC/index.jsp

ART-Dok http://archiv.ub.uni-heidelberg.de/artdok/?source_opus=&la=en

Sciences

Dryad http://datadryad.org/

DOE Data Explorer http://www.osti.gov/dataexplorer/

Social Sciences

ICPSR http://www.icpsr.umich.edu/icpsrweb/

Harvard Dataverse https://dataverse.harvard.edu/dataverse/harvard

Citing Data

Activity

Resources

1. Department of Biostatistics –Data Management Team, Indiana University School of Medicine (2013), from http://www.slideshare.net/goldenphizzwizards/ensuring-data-quality
2. DataONE exercises: https://www.dataone.org/education-modules
3. Data lifecycle diagram: http://www.data-archive.ac.uk/create-manage/life-cycle