February 22, 2014

Best Practices for Research Data Management

Max Eckard, Grand Valley State University
BEST PRACTICES FOR RESEARCH DATA MANAGEMENT

Max Eckard
Metadata & Digital Curation Librarian
University Libraries
Outline

• Preliminaries
  – What is it research data management? Why is it important?
• Best Practices
• Resources at GVSU
PRELIMINARIES

What is it? Why is it important?
What is Research Data Management?
CREATING DATA

RE-USING DATA

PROCESSING DATA

ANALYZING DATA

PRESERVING DATA

GIVING ACCESS TO DATA

Creating data

• Design research
• Plan data management (formats, storage, etc.)
• Plan for consistent sharing
• Locate existing data
• Collect data (experiment, observe, measure, simulate)
• Capture and create metadata
Processing data

- Enter data, digitize, transcribe, translate
- Check, validate, clean data
- Anonymize data where necessary
- Describe data
- Manage and store data
Analyzing data

• Interpret data
• Derive data
• Produce research outputs
• Author publications
• Prepare for data preservation
Preserving data

- Migrate to best format
- Migrate data to suitable medium
- Back-up and store data
- Create metadata and documentation
- Archive data
Giving access to data

- Distribute data
- Share data
- Control access
- Establish copyright
- Promote data
Re-using data

• Follow-up research
• New research
• Undertake research reviews
• Scrutinize findings
• Teach and learn
CREATING DATA

RE-USING DATA

PROCESSING DATA

ANALYZING DATA

PRESERVING DATA

GIVING ACCESS TO DATA

Why is Research Data Management Important?
Why is Research Data Management Important?

- Saves time in the long run
- Ensures that others can understand your data
- Makes sharing/preserving data easier
- Helps to avoid data loss in the event of a disaster or user error
- Lets your work live into the future
Why else?

• Funders now require data management and sharing plans
“Proposals must include a supplementary document of no more than two pages labeled ‘Data Management Plan’. This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results...”
“Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, primary data, samples, physical collections and other supporting materials...”
And...

• There are horror stories
Violent crime victim and witness data stolen from laptop, hard drive

University researchers collected information on the 119 people for a study on interrogation techniques.

By Rebecca Harrington  January 21, 2014

The theft of a University of Minnesota professor’s laptop, portable scanner and external hard drive last February resulted in a data breach of the personal information of 119 victims and witnesses of violent crimes.

Law professor Barry Feld and several law students were researching interrogation techniques in violent crimes prosecuted in Hennepin and Ramsey County District Courts during January and February 2005.

The researchers had collected the crime data but hadn’t analyzed it yet, so it took the county attorneys’ offices and the University almost a year to determine whose information was on the stolen devices and where to contact them.

While much of the data was public in police reports and court records, Feld had received permission from the county attorney and the victims before he began the research.
California Company Settles FTC Charges Of Exposing Consumers’ Medical Transcripts Online

By Ashlee Kieler  January 31, 2014

Personal information doesn't get much more personal than your medical history. Ensuring your medical records are secure remains a top priority for the Federal Trade Commission as they settled their 50th data security case on Friday.

The FTC announced that GMR Transcription Services, Inc. has agreed to settle charges that its inadequate data security measures resulted in thousands of consumers’ medical transcripts being indexed by a major search engine and available publicly online.

Affected transcripts were prepared by hired contractors between March 2011 and October 2011. Some of the files contained data for a medical organization of children and adolescents.
Retraction: Spontaneous pre-stimulus fluctuations in the activity of right fronto-parietal areas influence inhibitory control performance

A retraction of the Original Research Article:
Spontaneous pre-stimulus fluctuations in the activity of right fronto-parietal areas influence inhibitory control performance

The authors and the journal wish to retract the 12 March 2013 article cited above.

While applying the same analyses to another dataset, the authors discovered that a systematic human error in coding the name of the files had been made during the extraction of the EEG template topographic maps best differentiating the two experimental conditions at the single subject level. Because the subsequent processing steps were based on these EEG maps, this error has ultimately modified the final result of the paper, which is therefore not correct. For this reason, the authors request to
BEST PRACTICES FOR RESEARCH
DATA MANAGEMENT
• Data Types
• File Formats
• Organizing Data
• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data
• Re-use, Re-distribution
• Budget
Data Types

- File Formats
- Organizing Data
- Storage & Security
- Documentation & Metadata
- Copyright & Privacy/Confidentiality
- Archiving & Sharing Data
- Re-use, Re-distribution
- Budget
Sciences

- **Observational data**, like sensor readings and weather research
- **Experimental data**, like clinical trials
- **Model or simulation data**, like climate or economic models
- **Derived, compiled or computational data**, like text or data mining
Social Sciences

• **Quantitative data**, like data collected from surveys
  – Recorded in spreadsheets
  – Manipulated with R, SPSS, SAS or Excel

• **Qualitative data**, like audio or video interviews and coded transcripts
• *Anything* could be data to a humanist!
• Digital humanities are an emerging field
File Formats

- Organizing Data
- Storage & Security
- Documentation & Metadata
- Copyright & Privacy/Confidentiality
- Archiving & Sharing Data
- Re-use, Re-distribution
- Budget
File formats should be...

- Relatively common
- Uncompressed (or at least lossless compressed)
- Non-proprietary
Other considerations

• Transformations, conversions
  – May affect “look and feel” of a digital file, but should not affect content
  – Exercise extreme caution when transforming “complex” digital objects
• Data Types
• File Formats

Organizing Data

• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data
• Re-use, Re-distribution
• Budget
File naming conventions

Be consistent

• Conventions for:
  – Directory structure
  – Folder names
  – File names

• Tip: Retain a nice sort order
  – YYYY-MM-DD, not MM-DD-YYYY
  – 01-10 or 001-100, not 1-10 or 1-100

Be descriptive

• Don’t use system-generated names (i.e., 001.jpg, 002.jpg, Document1.doc, etc.)
• Unique identifier
• Project or data name
• Conditions
• Run of experiment
• Date
Other considerations

• No special characters
  – & , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? <> \/

• Use 3-letter file extension and lowercase

• Periods
  – name_paper.doc, not name.paper.doc or name_paper..doc

• Batch renaming
Version control

- Manual
  - Sequential numbering system like v01, v02
  - Don’t use labels like revision, final, final2, etc.

- Version control software

- Raw or semi-processed/cleaned data vs. intermediate working files vs. final product
• Data Types
• File Formats
• Organizing Data

**Storage & Security**

• Documentation & Metadata
• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data
• Re-use, Re-distribution
• Budget
Storage

• Consider your options:
  – Network data storage
  – Personal computers or laptops
  – External storage devices
  – Cloud computing

• **Tip:** 3-2-1 rule
  – Redundancy, redundancy, redundancy
  – Different storage media
  – Keep one copy in a separate geographic location
Security

- Network security
- Physical security
- Computer systems & files
Network security

• Keep confidential information off the Internet
• Put sensitive materials on computers not connected to the Internet
Physical Security

- Restrict access to buildings and rooms not connected to the Internet
- Only let trusted individuals troubleshoot computer problems
- Do not share passwords
- Lock your office
- Store external hard drives with highly sensitive data in a safe
Computer systems & files

• Keep anti-virus protection up-to-date
• Don’t send confidential data via e-mail or FTP
• Data encryption
Other considerations

• Confidential disposal of research data

• **Disclaimer**: This may not be all you need to know!
• Data Types
• File Formats
• Organizing Data
• Storage & Security

Documentation & Metadata

• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data
• Re-use, Re-distribution
• Budget
Documentation

• Type of unstructured metadata
• Supplementary file or document intended to accompany data
• Two common forms:
  – README.txt
  – Codebook
• Definitions of column headings and row labels, data codes, and measurement units
• For each file name, a short description of what data it includes
• Any processing steps, especially if not described in the publications, that may affect interpretation of results
• A description of what associated datasets are stored elsewhere, if applicable
• Whom to contact with questions
Codebook

• Column locations and widths for each variable
• Definitions of different record types
• Response codes for each variable
• Codes used to indicate nonresponse and missing data
• Other indications of the content and characteristics of each variable
metadata

/nəʊn/
noun: metadata; noun: meta-data

1. a set of data that describes and gives information about other data.

Translate metadata to

Choose language

Use over time for: metadata
Metadata

- Metadata is “data about data”
- It’s a little more complex than documentation...
Republican Congressman Harold (Hal) Sawyer represented Michigan's Fifth District in the U.S. House of Representatives from 1977 to 1984. He was a member of the House Appropriations Committee and served on the House Rules Committee. Sawyer was known for his work to promote social justice and civil rights, particularly during the civil rights movement of the 1960s and 1970s. He was also a strong advocate for environmental issues and supported the creation of the National Environmental Policy Act (NEPA). Sawyer was a member of the Democratic Party and worked closely with other liberal lawmakers to achieve their policy goals. He was succeeded in Congress by his former colleague, Minority Whip James Oberstar.
Metadata

• Comes in “schemes”
• Range from the very simple...
  – Dublin Core, DataCite
• To the very complex...
  – FGDC Content Standard for Digital Geospatial Metadata, DDI
Metadata

Biology

Earth Science

General Research Data

Physical Science

Social Science & Humanities
• Data Types
• File Formats
• Organizing Data
• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality

• Archiving & Sharing Data
• Re-use, Re-distribution
• Budget
Copyright & Privacy/Confidentiality

- Do you have the right to make data available?
- Should the data be embargoed for a certain period of time?
- How will you protect privacy, security, confidentiality and intellectual property?
- Can you think of any privacy, ethical or confidentiality concerns, particularly for human subjects data?
- Do you need to anonymize data before sharing it?
- Do any regulations, such as HIPAA, apply to your data?
Other Considerations

• Creative Commons license
• Several types:
  – Attribution
  – Share-alike
  – Non-commercial
  – No derivative works
• CC0 for data
• Data Types
• File Formats
• Organizing Data
• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality

Archiving & Sharing Data

• Re-use, Re-distribution
• Budget
Archiving

Backups
• Short-term data recovery solution
• Assume that data loses value over time
• Older data eventually overwritten
• On their own are insufficient for long-term preservation

Long-term digital preservation
• Much more comprehensive
  – Authenticity
  – Metadata
  – Security
  – Documentation
  – Auditing
  – Weeding
  – Sharing and discovery
  – Format obsolescence
  – Media corruption and failure
  – Organizational risks
Sharing Data

• Allows others to...
  – Do follow-up research
  – Do new research
  – Scrutinize your findings

• Provision of the NSF Award and Administration Guide, and others
Data Repositories

• Allow you to focus on your research rather than user requests
• Make your data more findable through SEO
• Track data impact metrics
• Help you receive credit for re-use of published data via data citation
• More effective than personal website
• Definitely more effective than email
ScholarWorks@GVSU

Enter search terms:

Search

in this repository

Advanced Search

Browse

Collections
Open Education Materials
Conferences and Events
Journals
Selected Works Galleries
Disciplines
Authors

About ScholarWorks

ScholarWorks@GVSU is a service of the Grand Valley State University Libraries. The 8,348 papers to date represent research and scholarly output selected and deposited by individual university departments and centers on campus. These items have had 395,968 full-text downloads in the past year and 850,727 downloads to date, helping our researchers increase the visibility and impact of their work. Contact ScholarWorks staff to find out how your work can be included!

Faculty Research

ScholarWorks makes GVSU faculty publications available to researchers around the world. Contact us to find out how to include your publications.
Other considerations

• Sponsor likely has terms and conditions for sharing
• Look for discipline-specific data repositories for archiving, sharing and re-use
• Trustworthy (with a capital “T”)

• Data Types
• File Formats
• Organizing Data
• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data

Re-use, Re-distribution

• Budget
Re-use, Re-distribution

• Will any permission restrictions need to be placed on the data?

• Who is your audience? Which bodies/groups are likely to be interested in the data?

• Who may be interested in your data in the future and what might it be used for?
Other considerations

• PI is responsible for sharing data without violating federal law or regulation
  – Export controls
• Or compromising individual rights
  – FERPA
  – FOIA
  – HIPAA
  – Intellectual Property
• Data Types
• File Formats
• Organizing Data
• Storage & Security
• Documentation & Metadata
• Copyright & Privacy/Confidentiality
• Archiving & Sharing Data
• Re-use, Re-distribution
Budget

• Funders recognize that it takes time and money to prepare data for sharing
• Put data sharing costs in your budget proposal
RESOURCES AT GVSU
You don’t have to do this alone.
Other Stakeholders

Outside GVSU
- Funders
- Industrial Partners
- Collaborators

GVSU
- Office of Sponsored Programs
- University Libraries
- IT
- University Counsel
Data Management Planning
QUESTIONS?
Thank you!