




The Good, The Bad, and the Ugly of Data Management

David T. Redden, PhD
Biostatistics, School of Public Health

The Center for Clinical & Translational Science

 205.934.7442

 ccts@uab.edu

 www.uab.edu/ccts

 [@cctsnetwork](https://twitter.com/cctsnetwork)



Let us set the expectations

- This presentation is designed to discuss DM across all types of research (lab studies to multi-site clinical trials). It is about common principles.
- PLEASE, PLEASE, be ready to make suggestions for future discussions.
- This presentation is a joint effort between many groups (CCTS, Lister Hill Library , SPH)





The Team

- Dorothy Ogdon, Assistant Professor, UAB Libraries
- Catherine Smith, Professor, UAB Libraries
- Madeline Gibson, Program Manager, CCTS
- David Chaplin, Professor, Microbiology and CCTS Training
- David Schneider, Associate Dean, Graduate School-Biomedical Science





Outline

- Motivating Example (Apollo 11)
- What is Data?
- What is a Data Management?
- What is a Data Management Life Cycle?
- Tools and Resources at UAB to Guide You.



Motivating Example

- Do you realize that the original telemetry data recordings, which included high resolution images, for the Apollo 11 moon landing was lost due to poor data management?
- Do you know, that NASA spend almost \$500,000 just to discover ‘We are almost certain the data is lost forever?’
- **MAINTAINING GOOD DATA REQUIRES BOTH PLANNING AND MANAGEMENT.**



For the curious

- www.nasa.gov/mission_pages/apollo/apollo_tapes.html
- www.hs.nasa.gov/all/Apollo_11_TV_Tapes_Report.pdf



Outline

- The United States government defines data, specifically Research Data, as ‘the recorded factual material commonly accepted in the scientific community as necessary to **validate** research finding.’ White House OMB 2013.



Another Data Definition

- The Organisation for Economic and Cooperation and Development (OCED) defined data as ‘**factual records** (numerical scores, text records, images, sounds) used as primary sources for research findings.’



Data Management

- Data Management is a series of practices designed to maximize and protect the quality, utility, and completeness of your research data. These practices are documented in a data management plan.



A Complete Data LifeCycle (The Good!)

- Project Planning.
- Writing a Data Management Plan.
- Data Acquisition and Entry
- Data Quality Control/ Data Version Control/ MetaData.
- Data Analysis.
- Publication and Data Sharing.
- Data Preservation and Archiving.
- Data Reuse



Project Planning (A Good Place to Start)

- What are you trying to show?
- What are you trying to measure?
- How will you measure it?



Project Planning (A Good Place to Start)

- Observational Data (Medicare Claims, Chart Review)
- Experimental Data (Laboratory Study, Clinical Trials, Behavioral Experiments)
- Simulation Data (Climate Forecasting, Financial Modeling)
- Compiled Data (Meta-analysis, Systematic Reviews)



Project Planning (A Good Place to Start)

- Nominal Data (Race, Sex, Education)
- Ordinal Data (Ratings)
- Quantitative Data (Temperature, Height, Weight)

- **START THINKING AND BUILDING A DATA DICTIONARY. A DATA DICTIONARY CONTAINS ALL YOUR VARIABLES, HOW THEY WERE MEASURED, HOW TO RECORD MISSING DATA, AND ACCEPTABLE VALUES**



Writing a Data Management Plan

- Key components include:
 - a) Data Types
 - b) Contextual Details (MetaData)
 - c) Storage, Backup, and Security
 - d) Provisions for Protection/Privacy
 - e) Policies for Re-use
 - f) Policy for Accessing and Sharing
 - g) Policy for Archiving and Preservation



Writing a Data Management Plan

- There are numerous websites that post examples of Research Data Plans.
- Because they are so important and useful, most researchers are happy to share them.
- A great webpage to visit is www.dmptool.org.



Data Acquisition and Entry

- Key Components:
 - a) Naming conventions for both files and variable names (Give meaningful names).
 - b) Naming conventions for both files and variable names (Avoid spaces, special characters, and long names).
 - c) What software system will you use? Is it proprietary? Is it common? Can it export you data to other formats easily?
 - d) Can logic and range checks included as data is entered?



Version Control/MetaData

- Over the past several decades, version control and metadata files have evolved.
- It is imperative that you save the ‘iterative history’ of your data. There is no other way than to do this with version control.
- Old strategies were monthly, or weekly, or daily ‘backup archives.’ Newer strategies include version control software such as Git (<https://git-scm.com>).



Version Control/MetaData

- Metadata : data about data. Imagine a page/file/map that describes all the files (raw data, processed data, programs, images) from your research.
- Metadata allows someone to rediscover your 'path' or 'system' within your data.
- Types of metadata: 1) Descriptive, 2) Structural, and 3) Administrative



Data Analyses

- Be aware of GUI (Graphical User Interfaces) when performing statistical analyses.
- Unless you are using asking for syntax to be written, your results may be hard to reproduce.
- Use a syntax recorder, work with a methodologist who program it, or create a metadata file of the analytic steps.



Publication and Data Sharing

- If all the previous steps are successfully implemented, data sharing and publishing with confidence becomes much easier.
- However, when publishing anything (abstracts, figures, papers), build an archive that contains all components (DMP, Raw Data, Processed/Analysis Ready Data, MetaData file, Figures)
- Building an archive increases REPRODUCIBILITY.



Data Preservation and Archiving

- Short Term Storage: Where is the data being stored as the study is ongoing?
- Long Term Storage: Where is the data going to be stored in 10 years.



Data Reuse

- Data Management is labor intensive. It is perpetual task that requires constant attention to version control of electronic case report forms, ensure completeness of the data, accuracy of data (logic and range checks), scheduled back-ups of data, and documentation and archiving.



A very incomplete Data Life Cycle (The BAD)

- Plan the Project.
- Data Acquisition.
- Data Analysis.
- Publish Results.
- All data in one office or repository but little documentation.



The Ugly of Data Management

- Lack of DMP
- Poor Record Management
- Lack of Data Dictionary/MetaData
- Lack of Version Control and Back-up
- Lack of Long-Term Plans
- Not knowing who has the last version of the data.



Expertise on Campus

- Preventive Medicine's Biostatistics and Bioinformatics Shared Facility (bsejong@uab.edu, 4-6887).
- Department of Medicine's Information Technology (nislam@uab.edu, 4-3616)
- School of Public Health's Biostatistics Department (bstchair@uab.edu, 4-4905).
- Research Commons, Center for Clinical and Translational Research (ccts@uab.edu, 4-7442)
- Lister Hill Library (dogdon@uab.edu, 4-2231)



THE MOST IMPORTANT SLIDE

- You can find a large amount of information for Data Management here:
- <https://www.uab.edu/faculty/rdm>