DATA MANAGEMENT: PLANS, RESOURCES, & SECURITY

TOM HOHENSTEIN – BU LIBRARIES
CHRISTINE CHAISSON – DATA COORDINATING CENTER
ERIC JACOBSEN – IS&T INFORMATION SECURITY
WHAT IS DATA MANAGEMENT?

A simple definition:

Being proactive about storing, sharing, and making your data accessible to others.
WHY MANAGE DATA?

I WANT YOU TO SHARE YOUR DATA

NEAREST FUNDING AGENCY
A FEW QUESTIONS TO ASK

What are the data sources?
Are there regulations you need to comply with?
How will you collect, manage, analyze your data?
How will these sets be stored and documented?
Where will it archived?
Will the data be shared?
HOW DO YOU MANAGE DATA?

Plan
Document
Update
Audit
THINGS CAN GO WRONG

• Data may not be saved or backed up
• Crucial data elements may be missing
• Data may be incorrect due to data collection or entry errors
• Data may be incorrectly identified
• May not have appropriate keys for merging data from different sources
• Data may be merged incorrectly
• Data files may be lost or corrupted
• Security issues
REAL WORLD EXAMPLES

• A few illustrations from popular news sources
OOPS: EXCEL ERROR CALLS INTO QUESTION…

• ‘This Time It’s Different’ a 2009 book by Harvard researchers (Reinhard & Rogoff) contained “Serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period” Inability to replicate work by investigators at UMass (among others) led to accusations of intentional fudging of the data or sloppiness

• The Authors admitted they forgot to include five rows in an Excel file resulting in exclusion of data from Australia, Austria, Belgium, Canada, and Denmark — a “coding error” which they said was “a significant lapse on our part.”
HOW BRIGHT PROMISE IN CANCER TESTING FELL APART

• Duke Cancer Center’s gene-based tests proved worthless, research behind them was discredited
• Statisticians from MD Anderson discovered errors such as columns moved over in a spread-sheet which Duke team “shrugged them off” as “clerical errors.”
• Four papers were retracted
• Duke shut down three trials
• Center leaders resigned or were removed.
• People died and relatives sued Duke
DATA MANAGEMENT PLAN

- Purpose: To help you manage and share your data and meet funder requirements.
- General elements include:
  - Project or study description
  - Documentation, organization, storage
  - Access and sharing
  - Archiving
Proposals submitted to NSF must include a supplementary document of no more than 2 pages labeled "data management plan" (DMP). This document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.

Proposals that do not include a DMP will not be able to be submitted.
The document should describe:

- Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced during course of project
- Standards to be used for data and metadata format and content
- Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property or other rights or requirements
- Policies and provisions for re-use, re-distribution, and the production of derivatives
- Plans for archiving data, samples, and other research products and for preservation of access to them
Data Management and Access Plan

This project will produce a variety of data types on a rolling basis. To efficiently receive, clean, process, manage, archive and share this data both among the research team and the wider community, we will hire a 30% FTE data manager/IT support specialist. The data will be made available to the research team for the life of the collaborative group, and particular data types will be made publicly available as described below. Data types will include:

1. Quantitative and qualitative data from quarterly, nationally representative surveys of American adults.
2. Quantitative and qualitative data from content analyses of source material from diverse domains (news media articles, press releases, reports, Congressional debates, etc.).
3. Quantitative and qualitative data from in-depth interviews with key actors and organizations in the social amplification system, including interview transcripts and social network surveys.

All survey data will be initially stored and analyzed as SPSS or SAS files. All content analysis data will be stored as Excel spreadsheet files and transcripts will be stored as Microsoft Word files. Since these formats could become unreadable over time as software systems change, final versions of all datasets and documents will also be exported to and made available as ASCII and/or CSV data files, with accompanying command/syntax files, so future users will still be able to access the data, even if this proprietary software is no longer supported. Each data file will be cataloged in a single database, with accompanying metadata (e.g., filename, author, abstract, producer, geographic coverage, temporal period of collection, response rate, etc.) using Data Documentation Initiative standards.

National survey datasets will be publicly released and archived 18 months after final versions are completed, to allow the research team time to publish initial results. All survey responses will be voluntary, anonymous, confidential, and unidentifiable. All results will be released only as aggregate statistics. Our research collaborators at Knowledge Networks will remove all identifiers from the dataset before release to the rest of the research team. Survey datasets deposited in public archives will thus have no individually identifiable information attached. Content analysis data will be acquired from publicly available sources (e.g., news media stories, press releases, etc.), thus will not require special confidentiality protection procedures. Any publicly released interview data will also be voluntary, anonymous, confidential, and unidentifiable, unless the source has provided written permission to be quoted by name. Interview transcripts will not be publicly released without the express written permission of the interviewee. All archived survey and content analysis data and accompanying metadata will be deposited and made publicly available through ICPSR at the University of Michigan.
Data Management Plan

Effects of temperature and salinity on estuarine copepod, Eurytemora affinis.

1. Products of Research
   Every two days, we will subsample E. affinis poe and use a microscope to identify the life-stage and document the information first in a laboratory spreadsheet. For quality control, values will be double-checked for accuracy. The Excel spreadsheet will be saved up to a server. After all data are collected, the Excel file will be imported into the program R for statistical analysis.

2. Data Storage and Preservation
   Our short-term data storage plan, which will be 1) the .txt metadata file and 2) the Excel spreadsheet.
   We will use the University’s local network to store all data and metadata files daily on the University of Alberta Mathematics Department server. We will also have a laboratory notebook as a hard copy backup that will be stored in a fire-proof cabinet.
   The data set will be submitted to the Knowledge Network for Biocomplexity (KNB) data repository for long-term preservation and storage. The authors will submit metadata in EML format along with the data to facilitate its reuse. The data manager will be responsible for updating metadata and data author contact information in the KNB.

3. Data Formats and Metadata
   We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a .txt document that will be stored with the data file. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. EML is one of the accepted formats used in ecology, and works well for the types of data we will be producing.
   We will create these metadata using Morpho software, available through KNB (http://knb.ecoinformatics.org/morphoportal.jsp). The metadata will fully describe the data files and the context of the measurements.

4. Data Dissemination and Policies for Data Sharing and Public Access
   We are required to share our data with the CAISN network after all data have been collected and metadata have been generated. This should be no more than 6 months after the experiments are completed. In order to gain access to CAISN data, interested parties must contact the CAISN data manager (data@caisn.ca) or the authors and explain their intended use. Data requests will be approved by the authors after review of the proposed use.
   The authors will retain rights to the data until the resulting publication is produced, within two years of data production. After publication (or after two years, whichever is first), the authors will open data to public use. After publication, we will submit our data to the KNB enabling discovery and use by the wider scientific community. Interested parties will be able to download the data directly from KNB without contacting the authors, but will be encouraged to give credit to the authors for the data used by citing a KNB accession number either in the publication’s text or in the references list.

5. Roles and Responsibilities
   The PI will be responsible for all data management during and after data collection.
WHERE TO START?
DATA MANAGEMENT PLAN: BASIC ELEMENTS

- Study design
- Data types and sources
- Storage format and location
- Software used for manipulation
- File naming conventions, data manipulation documentation
- Project Staff – who has permission to what
- Identifiers (if applicable)
- Back ups -- frequency and location
- Testing systems, cleaning programs if applicable
- Security
TIMELINE

Design Project
Draft DMP (2 pages)
GET FUNDED!
Detailed DMP
Analyze Data
Collect Data
Publish
Share Data
Re-budget
Collect Data
Detailed DMP
GET FUNDED!
Draft DMP (2 pages)
Identify Data sources
BOSTON UNIVERSITY

GRANT PROPOSAL

PLAN AHEAD
EXAMPLE: DETAILED DMPs

Detailed plan for clinical trial to be submitted to FDA from the NIH website.
### DCC’s DMP

#### Summary/Overview
- **Personnel/Responsibilities**
- **Data sources & handling**
- **Tasks & Identifiers**
- **Analytics & Data changes**

#### Personnel/Responsibilities
- **Study Responsibilities**
  - Initials of person responsible
- **Monitoring DMP**
- **Tracking progress of external datasets**
- **Tracking progress of TQG database**
- **Testing database**
- **Testing website**
- **Basic data cleaning procedures**
  - Check for duplicates
- **Communication with start on data problems**
- **Randomization plan**

#### Data sources & handling
- **Data sources**
  - DCC-managed data
  - DCC data
  - Study available data
- **Data entry methods**
  - DCC-managed data
  - Other data
- **Database/Web Development**
  - Data cleaning
  - Data change indicators
  - Data change calculations

#### Tasks & Identifiers
- **Data Management Tasks**
  - Data transformations
  - Data cleaning
  - Data storage
  - Data version control

#### Analytics & Data changes
- **Data change calculations**
  - Data change correlations related to patient data, etc.
- **Data change indicators**
  - Data change history
- **Data change table**
  - Table of data changes

#### Documentation
- **Database Specifications**
  - Database specifications
  - Documentation of database changes
- **Web site Specifications**
  - Web site specifications
  - Documentation of web site changes

#### Data Type & Location
- **Data sources**
  - Type: DCC
  - Location: DCC SQL
- **Database/Web Development**
  - Web site address
  - Web site URL

#### Data/Doc
- **Study documents**
  - Type: DCC
  - Location: DCC SQL
- **Electronic data sets**
  - Type: Excel
  - Location: DCC SQL

#### FUNDING:
- **IRB Protocol Number**
  - Initial approval date
- **IRB Protocol Number**
  - Initial approval letter
- **DSMB**
  - Related documents

#### Study Type:
- **Type:** Case control
  - Cohort
  - Case crossover
  - X-sectional
  - Survey
- **RCT:**
  - Specific Academic
  - Industry
  - FDA
- **Other:**
  - Specific

#### Study Information:
- **Project title:**
  - Description
  - Duration
- **Timeline:**
  - Start date
  - End date
The Data Management Plan should be an evolving, written document. It will help you:

- Manage in event of staff absences/turnover
- Answer questions about infrequently occurring events
- Prepare for/survive audits by IRB or funding agency

Plan to revise it often – things change

Everyone on the team should be familiar with it and know where it is
DMP SUMMARY

• Think of DMPs in 2 stages: short for proposal; detailed for management of the project
• There is no one-size fits all DMP
• Write the plan that is appropriate for your project
• Write the plan that is appropriate for your funding agency or sponsor
• Draft a plan before you budget so you make sure you budget appropriately
• Consider compliance issues when writing the DMP and budgeting
• Plan ahead whenever possible
• Plan to revise your DMP often – things change
INFORMATION SECURITY

**Confidentiality** - Ensuring that we meet regulatory and contractual obligations to protect the secrecy/privacy of data

**Integrity** - Ensuring your data sets are stored in a manner that is hardened against loss or modification

**Availability** - Ensuring that data is available to yourself and the parties you intend to share it with

My job, and the job of my team, is to help our community address those concerns for administration, academics, and research.
INFORMATION SECURITY SERVICES FOR RESEARCH

Security Consulting

Compliance assistance:

- HIPAA and FISMA have long lists of required controls.
- We cannot make you compliant, but we are happy to assist you in becoming compliant.
- Compliance can be a lengthy process – start early!

Network and Security Architecture: Firewalls, VPNs, private networks

One on one: Meet with you to discuss your data and its specific requirements and formulate a security plan.

Policy

- Working with the Information Security and Business Continuity Governance Committee, we set institutional policy for security.
- Security Policy is often helpful in guiding activities towards more secure behaviors.
BU DATA CLASSIFICATION

Data Classification

*Restricted Use* Data: Has reporting obligations if the data is disclosed to unauthorized parties: PCI, HIPAA, MGL 201 CMR 17 (SSN, Driver’s License, financial account numbers)

*Confidential* Data: Data must be held in secret but does not trigger reporting requirements if breached (student data)

*Internal* Data: Data you might not consider a secret but you wouldn’t post (contact lists, meeting agendas)

*Public* Data: Data that can be published freely without constraint (web pages)

DATA PROTECTION STANDARDS AND MINIMUM SECURITY GUIDELINES

Data Protection Requirements lay out non-technical specifications for accessing, sharing, sending (transmitting), storing, auditing, destroying, and reporting on a breach of information.  

http://www.bu.edu/policies/information-security/data-protection-requirements/

The Minimum Security Standards provide the technical description of the controls that should be put in place for data:  

http://www.bu.edu/policies/information-security/minimum-security-standards/

If you have HIPAA data:  

http://www.bu.edu/policies/information-security/hipaa-policy/
DATA SECURITY - GENERAL

• Keep paper records should be kept in locked cabinets and/or offices
• Store identifiers like names and addresses separate from clinical data
• Keep particularly sensitive data apart from other identifiers (e.g., SSN) – in a separate file, by ID
• Do not collect sensitive data unless you really need it
DATA SECURITY - HARDWARE

- Password protect all computers
- Set to automatically timeout if inactive
- Encrypt laptops, flash-drives and other storage devices when possible
- Do not put identifiable data on portable media unless password protected, preferably encrypted
INFORMATION SERVICES FOR RESEARCH

Tools
- Anti-virus, Anti-malware
- Backup solutions (personal or university, laptops to servers)
- Disk/File/Folder Encryption
- Varonis Data Governance (file access control auditing)

Services
- Shared Computing Cluster
  - Can provide dbGaP compliant solution (genomics), may be good enough for others.
- Storage and Archive
  - General Purpose Network Attached Storage (GP-NAS) solution for data up to Confidential Level
  - Restricted Use Network Attached Storage (RU-NAS) solution for Restricted Use data
- Microsoft/BU OneDrive can be used for HIPAA data
- Note that there is more to Restricted Use than how you store it!
- Media Destruction
- VPN
BREACH REPORTING

The Information Security Incident Response Team (IRT)
• Responsible for coordinating the response to cyber security incidents at Boston University.
• When an incident occurs, the IRT helps system owners and administrators respond properly to an incident, collects and analyzes evidence, assists in the recovery of systems to a secure state, and develops remediation plans.
• The IRT is also responsible for conducting computer forensic investigations.

Responding to a breach:
• Do not power off, log in to, continue to use, or alter any system suspected of a data breach.
• Contact our Incident Response Team at irt@bu.edu or 617-358-1100
QUESTIONS?

CHAISSON@BU.EDU

JACOBSEN@BU.EDU

TAH@BU.EDU
INFORMATION RESOURCES

http://www.bu.edu/infosec/howtos/
http://www.bu.edu/infosec/howtos/securing-your-devices/
http://www.bu.edu/infosec/howtos/mobile-device-security/
http://www.bu.edu/tech/about/security-resources/bestpractice/

External Resources
http://www.library.illinois.edu/sc/services/data_management/
Find out what may be available to you

HTTP://WWW.BU.EDU/CTSI/

Computing Resources
- Overview
- Shared Computing Cluster (SCC)
- Storage
- Network
- Technical Summary

Resources
- Funding
  --Pilot Grant Programs
  --Other Funding Opportunities

BU CTSI
CLINICAL AND TRANSLATIONAL SCIENCE INSTITUTE
Advancing Translational Research

Tools

Software available to BU investigators:

No Cost
BU-2b2
OpenClinica
REDCap

Fee Charged
FreezerPro
StudyTRAX

The CTSI provides limited consulting services for software available to BU investigators.

BU-2b2 is an easy-to-use aggregate data query tool which is accessible via the internet via the 2b2 "web client" and is currently being evaluated by pilot researchers at BUMC. Researchers at BUMC and BMC have spent the past 3 years developing and implementing 2b2 to better understand the health and healthcare of Boston's residents. The BU-2b2 database is currently available for research and quality improvement activities.

View a video providing an overview of the BU-2b2 web client here.