Yale Harvey Cushing/ John Hay Whitney MEDICAL LIBRARY

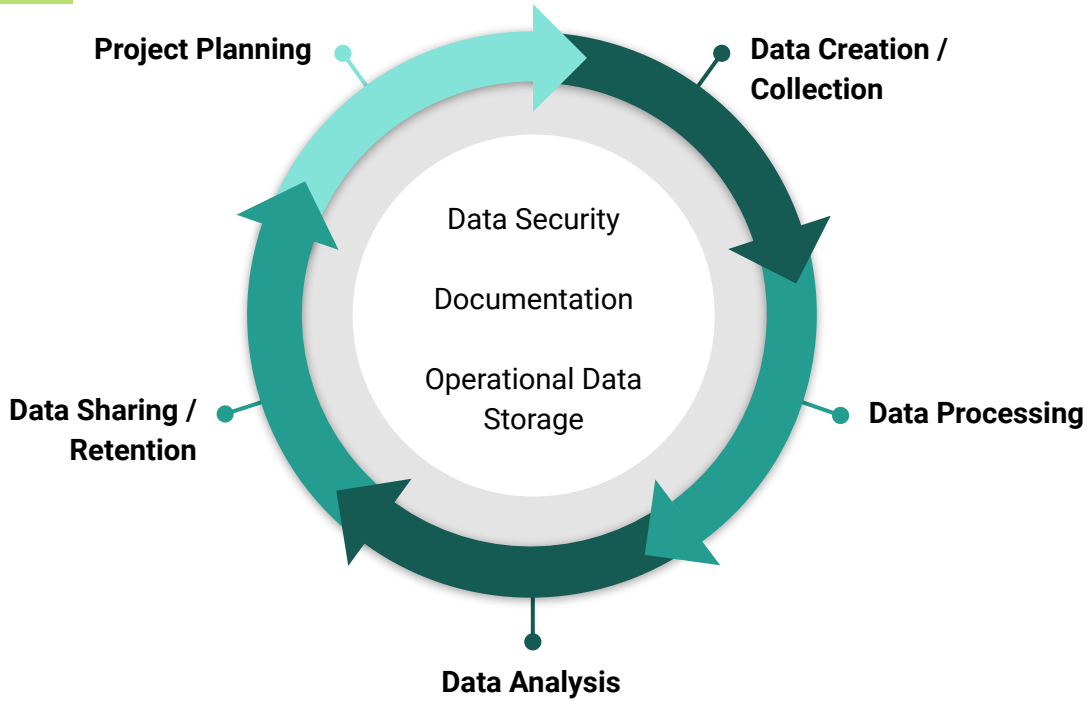
Research Data Management & Relational Database Concepts

Defining research data management.

Core RDM Topics Include:

- Organization
- Storage
- Preservation
- Sharing

RDM is Often Represented by the Research Data Lifecycle



Organization

Levels of Organization

Storage Locations

How many places is your data stored? Does having multiple locations improve or hamper your ability to work?

Folder Structures

Within a storage location, how do you organize your files and folders?

Data Structures

Data structures can include spreadsheets or databases

Folder Organization Considerations

- Follow a consistent organization pattern
- Create documentation descripting the organization pattern
- Think about what you are organizing by (per grant, per analysis, per dataset, etc.)
- Make your folder structure user friendly

Folder and File Naming Considerations

- Maintain a consistent naming patterns
- Be descriptive
- Think about how you will want to filter, sort or search through your folders

Data Structure Organization

- Is your data multiple discrete spreadsheets? One spreadsheet with multiple tabs? Custom pulls from a database? A folder of focus group transcripts?
- What structure does your data need to be in for you to conduct your analysis?

Spreadsheet Organization

- Each column should be a single variable
- Each row should contain a single observation
- Data formats should be consistent
- Document spreadsheet structures using data dictionaries
- Always keep master copies of your data

Storage

Cloud storage

Box Secure (and Box)

- Box Secure is the only HIPAA compliant cloud solution
- Version control / file history

Google Drive

- Integrates text documents, spreadsheets, slides, etc.
- Version control / file history

Microsoft Teams

- Includes robust project tracking and communication tools

Storage @ Yale

Standard

Daily/typical use storage and compute power

Enhanced

Higher performance compute with storage

Archive

Preservation storage

Preservation and Sharing

Preservation and Retention

- Data must be retained by a researcher for 3 years after publication (Yale Policy 6001 Research Data & Materials Policy)
-

Data Sharing

- NIH Data Sharing Policy requires data to be made accessible
- Research data includes raw data + any code used during your research
- You may deposit data into a data repository

Relational Databases.

What is a Database?

- Databases are data structures that allow you to organize data with more control and ability to query against it than using a spreadsheet
- Databases are useful in understanding RDM concepts because they follow strict and predefined parameters

What are Relational Databases (RDBs)?

- Collective set of data organized in tables, where relationships are defined between these tables, typically through identifiers

Quick Visualization of Database Relationships

subject_id	DOB	Insurance_Type
00111	1985-03-17	Medicare
00112	2000-09-21	Medicare
00113	1991-08-20	Private
00115	1980-01-04	Medicare

admit_id	admit_date	subject_id
4737	2020-03-01	00111
4332	2020-01-17	00111
4555	2020-02-12	00112
9822	2020-02-01	00113

Documentation and RDBs

- As RDBS are designed and constructed, with data formats, sizes, and other considerations in mind
- Data dictionaries are used to interpret data fields within databases

Data Dictionaries

subject_id	data_type	format	Nulls
subject_id	numerical	XXXXX	Not Null
DOB	date	YYYY-MM-DD	Not Null
Insurance_Type	character	Medicare Private None	Not Null

Interacting with Databases

- Front end access through a user interface
  - CDC Wonder Cancer Incidence (https://wonder.cdc.gov/cancer-v2016.HTML)
- Front end SQL queries
- Backend SQL queries

Technology to Build Databases

- MySQL
- Microsoft Access
- Python and R
- RedCap

Structured Query Language (SQL)

A language used to build, query, and modify relational databases

Demo and Exercises

- The jupyter notebooks containing the demo and exercises leverage a python libraries for handling data and running SQL queries (pandas and sqlite3)

SQL Structure

```
SELECT column_name
FROM table1_name
    JOIN table2_name
        ON table1_name.linking_column_name =
        table2_name.linking_column_name
WHERE contraint_expression
GROUP BY column_name
ORDER_BY column_name ASC/DESC
```

Yale services that can help.

Cushing/Whitney Medical Library

- Data consultations and workshops
  - Core concepts in research data management
  - Computational strategies for working with data
  - Finding datasets for reuse
  - Data visualization
- Bioinformatics Hub
  - Consultations and workshops on “-omics” data interpretation and analysis
- Systematic review services
- Research publication copyright

Yale IT

- Database configuration
- Security assurance
- Data storage solutions (3 Tiered Storage @ Yale, Google Drive, Box, Microsoft Teams)

Yale Center for Research Computing

- Transferring large datasets
- Optimizing code
- Utilizing high performance computing clusers

Joint Data Analytics Team (JDAT)

- Access to EHR data stored within Epic

Next Steps

- Jupyter Notebook Demo
- Office hours
- Afternoon Exercises

Contact medicaldata@yale.edu