ODTÜ TEKPOL

RESEARCH CENTER *for* SCIENCE *and* TECHNOLOGY POLICIES

Research Data Management Getting Your Organization Started

Arsev Umur Aydınoğlu, Ph.D.

MIDDLE EAST TECHNICAL UNIVERSITY
ORTA DOĞU TEKNİK ÜNİVERSİTESİ

Agenda

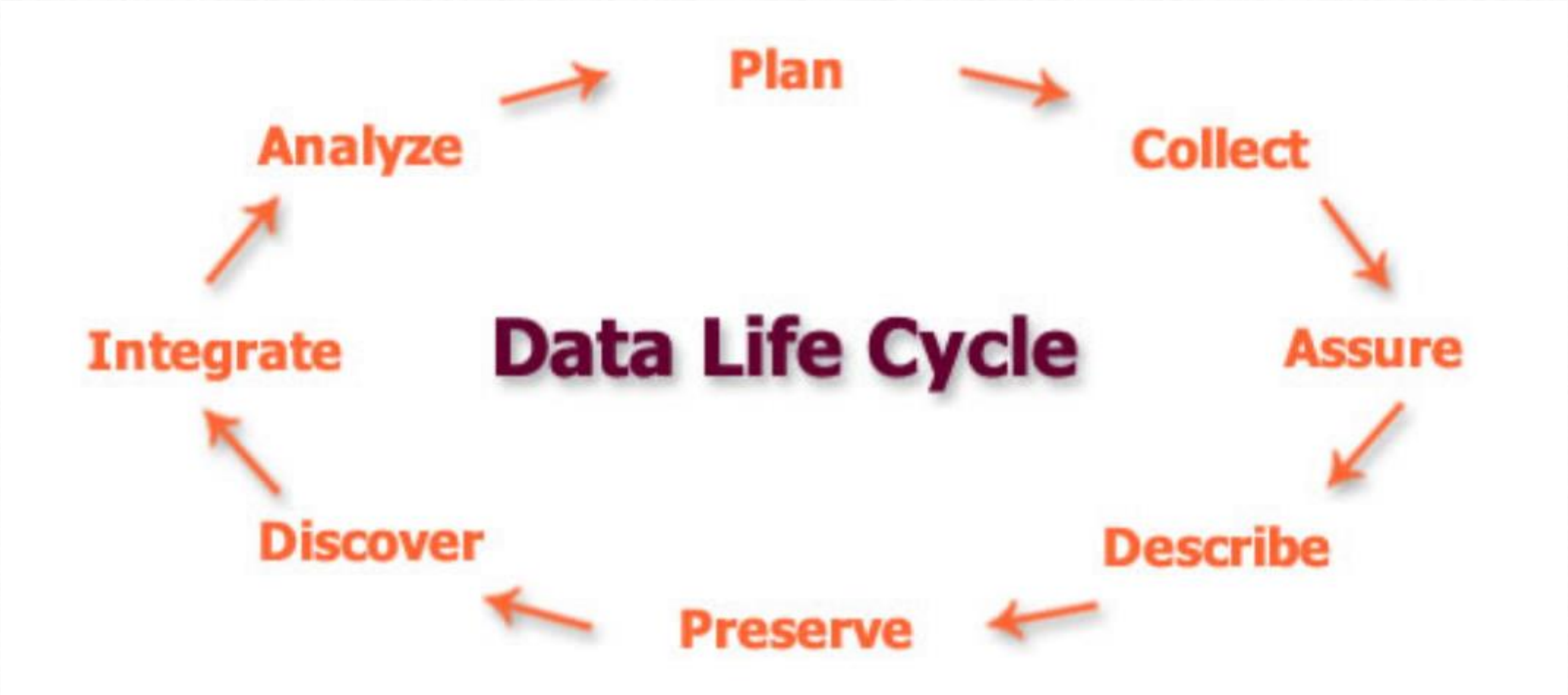
- Data Management Plan
- Storage
- Sharing

- - Data Management Plan
  - Why manage data?
  - How to evaluate data management needs and resources
  - Planning to encourage buy-in from the management and users
  - How to create a data management plan
- - Storage
  - What practitioners need to know
  - The basics of data organization
  - Understanding metadata & documentation
  - Planning for data security
- - Sharing
  - How to find data repositories
  - How to meet journal requirements
  - Legal Basics(intellectual property, confidentiality, sensitive data, etc.)

What is DMP

- a formal document that outlines what you will do with your data during and after a research project.

Data life cycle



Why manage data?

- It will benefit your collaborators
- It will benefit the scientific community
- Journals & sponsors want you to share your data.

Who requires a plan?

- Everybody

Data types

- By source
- By format
- By stability
- By volume

Types of data by source

- Observational
- Experimental
- Simulation
- Derived/compiled

Types of data by form

- **Text:** field or laboratory notes, survey responses
- **Numeric:** tables, counts, measurements
- **Audiovisual:** images, sound recordings, video
- **Models, computer code**
- **Discipline-specific:** FITS in astronomy, CIF in chemistry
- **Instrument-specific:** equipment outputs

Types of data by stability

- **Fixed datasets:** never change after being collected or generated
- **Growing datasets:** new data may be added, but the old data is never changed or deleted
- **Revisable datasets:** new data may be added, and old data may be changed or deleted

By volume

Questions to ponder:

- Are you manually collecting and recording data?
- Are you using observational instruments and computers to collect data?
- Is your data collection highly iterative?
- How much data will you accumulate every month or every 90 days?
- How much data do you anticipate collecting and generating by the end of your project?

File formats

Formats likely to be accessible in the future are

- Non-proprietary
- Open, with documented standards
- In common usage by the research community
- Using standard character encodings (i.e., ASCII, UTF-8)
- Uncompressed (space permitting)

Examples of discouraged format choices and better alternatives:

Discouraged Format	Alternative Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv)
Word (.doc, .docx)	plain text (.txt), or if formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)

Tabular data

Your spreadsheets will be easier to understand and to export if you follow best practices when you set them up, such as:

- Don't put more than one table on a worksheet
- Include a header row with understandable title for each column
- Create charts on new sheets- don't embed them in the worksheet with the data

Organizing Files

Top-level directory/folder

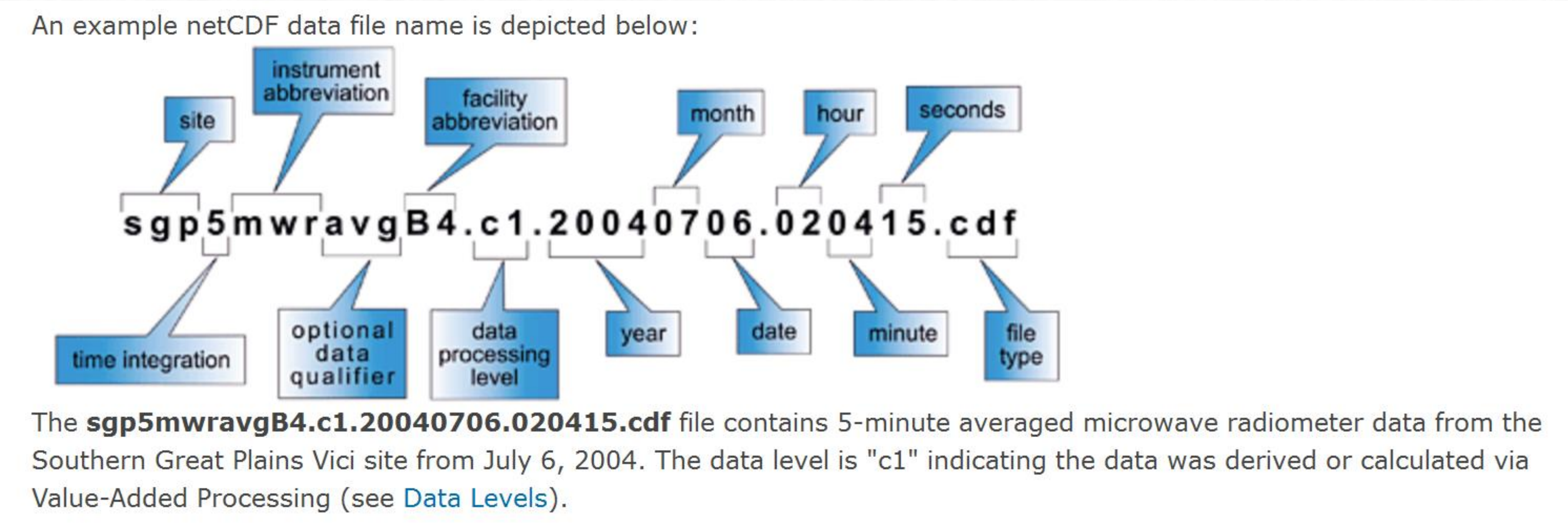
- Project title
- Unique identifier
- Date

The sub-directory structure

- Reserve three letter file extension for the file format
- Identify the activity or project in the file name
- Identify separate version of files and datasets
- Record all changes to a file no matter how small

File naming conventions

DoE’s Atmospheric Radiation Measurement (ARM) Program



An example netCDF data file name is depicted below:

The **sgp5mwravgB4.c1.20040706.020415.cdf** file contains 5-minute averaged microwave radiometer data from the Southern Great Plains Vici site from July 6, 2004. The data level is "c1" indicating the data was derived or calculated via Value-Added Processing (see Data Levels).

Metadata: Data Documentation

- Research project documentation
    - Rationale and context for data collection
    - Data collection methods
    - Structure and organization of data files
    - Data sources used (see citing data)
    - Data validation and quality assurance
    - Transformations of data from the raw data through analysis
    - Information on confidentiality, access & use conditions
- Dataset documentation
    - Variable names and descriptions
    - Explanation of codes and classification schemes used
    - Algorithms used to transform data (may include computer code)
    - File format and software (including version) used

How will you document your data?

General Overview (Title, creator, identifier, date, method, processing, source, funder)

Content Description (Subject, place, language, variable list, code list)

Technical Description (File inventory, file formats, file structure, version, checksum, necessary software)

Access (rights, access information)

Persistent Identifiers

Actionable / Globally unique / Persistent

Identifier Schemes

- ARK
- DOI
- HTTP
- InChI
- LSID
- NCBI
- PURL
- URL
- URN

Security & Storage

Data Security

- Network security
- Physical security
- Computer systems and files

Encryption & compression

Unencrypted data will be more easily read by you and others in the future, but you may need to encrypt sensitive data.

- Use mainstream encryption tools (e.g., PGP)
- Don't rely on 3rd party encryption alone
- Keep passwords and keys on paper (2 copies)

Uncompressed data will be also be easier to read in the future, but you may need to compress files to conserve disk space.

- Use a mainstream compression tool (e.g., ZIP, GZIP, TAR)
- Limit compression to the 3rd backup copy

Backups & storage

Data backup options

- Hard drive using software
- Tape backup system
- Cloud storage

Other data preservation considerations

- Who is responsible for managing and controlling data?
- For what or whom are the data intended?
- How long should the data be retained?

Sharing & Archiving

Why share your data?

- Required by publishers (e.g., Cell, Nature, Science).
- Required by government funding agencies (e.g., NIH, NSF)
- Allows data to be used to answer new questions
- Makes research more open
- Makes your papers more useful and citable by other researchers

Considerations when preparing to share data

- File formats for long term access
- Don’t forget the documentation
- Ownership and privacy

Ways to share data

- Email to individual requesters
- Post online via a project or personal web site
- Submit as supplemental material to be hosted on a journal publisher's website
- Deposit in an open repository or archive
- Deposit in an open repository and publish a "data paper" describing the data

Finding a data repository

- Discipline specific
- Institutional

Citing data

Citing data is important in order to:

- Give the data producer appropriate credit
- Allow easier access to the data for re-purposing or re-use
- Enable readers to verify your results

Citation elements

Core elements (Creator, title, publication year, publisher, identifier)

Common additional elements (Version, access date, subset, verifier, location)

Copyright & Privacy

Sharing data that you produced/collected yourself

- Data is not copyrightable.
- Data can be licensed.

Copyright & Privacy

Sharing data that you collected from other sources

- You may or may not have the rights to do so, depending upon whether that data were accessed under a license with terms of use.
- Most databases to which the UC Libraries subscribe are licensed and prohibit redistribution of data outside of UC. For more information on terms of use for databases licensed by the Libraries, contact UC3.

Copyright & Privacy

Confidendiality and Ethical Concerns

- Evaluate the anonymity of your data.
- Obtain a confidentiality review.
- Comply regulations.

Copyright & Privacy

To ethically share confidential data, you may be able to

- Gain informed consent for data sharing (e.g. deposit in a repository or archive)
- Anonymize the data by removing identifying information. Be aware, however, that any dataset that contains enough information to be useful will always present some risk.
- Restrict the use of your data. The ICPSR DSDR provides a tool for Designing a Restricted Data Use Contract.

Any questions?

arsevu@gmail.com & aaydinoglu@metu.edu.tr

References

- Data Management Plan dmptool.org
- DataONE Primer www.dataone.org
- **Aydinoglu, A.U.**, Suomela, T., Malone, J. (2014). Data management practices among astrobiology researchers. *Astrobiology, 14(6)*, 451-461.
- Allard, S. & **Aydinoglu, A.U.** (2012). Environmental researchers' data practices: An exploratory study in Turkey. In *E-Science and Information Management* (pp. 13-24). Springer Berlin Heidelberg
- Tenopir, C., Allard S., Douglass K., **Aydinoglu A.U.**, Wu L., Read E., Manoff, M., Wilson, B. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE, 6 (6)*.