16: Data Management

Environmental Data Analytics | Kateri Salk

Spring 2020

Objectives

1. Discuss data challenges in the environmental field
2. Evaluate how data management fits into the pipeline of data analysis
3. Access data via online and R-based databases

Contemporary Data Challenges

Environmental fields are experiencing massive changes related to data. Many new challenges exist today, including:

- Big data: volume, variety, frequency
- Open data
- Long-term records
- Interconnected networks
- Verifying accuracy and integrity

Data Management Workflow

Adapted from the DataOne data management primer and the University of Alabama Library guide on data management

1. Choose and assemble data management toolbox
2. Create (and revisit) your data management plan

- Volume and type of data
- File and folder structures/formats
- Roles and responsibilities of personnel
- Version control
- Access
- Preservation

3. Collect data
4. Quality assurance/quality control (QA/QC)
5. Describe and document data
6. Store and preserve data in a repository

How does data management fit into the pipeline of data analytics?

Accessing Data via Databases

Today's activity: Choose an online or R-based database and search for a dataset that is interesting and/or relevant to you. If you have not yet chosen a dataset for your course project, this would be a good opportunity to do some exploring.

When you have found a dataset you like, familiarize yourself with it (basic exploration, visualizations if time permits) and then post information about your dataset to the **Slack forum** under the channel #forum-databases.

Online Databases

Various disciplines

re3data

DataOne

Google Dataset Search

Environmental Data Initiative Portal

National Ecological Observatory Network

Long Term Ecological Research Network

Water

CUAHSI HydroClient

CUAHSI HydroShare

Spatial Data

ArcGIS

R Packages

- NHANES: National Health and Nutrition Examination Survey
- TidyCensus: U.S. Census data
- FedData: Geospatial data from federal sources
- dataRetrieval: USGS and EPA water quality, streamflow, and metadata
- LAGOSNE: Multiscaled geospatial and temporal data for U.S. lakes