DATA MANAGEMENT

Jezryl Jaeger L. Garcia
Project Assistant
BIFA6_026

OBJECTIVE

Learn key concepts or principles of data maintenance/cleaning

DATA CLEANING

According to Arthur Chapman (2005)

"A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions."

THE DATA CLEANING FRAMEWORK (MALETIC & MARCUS, 2000)

1. Define and determine error types
2. Search and identify error instances
3. Correct the errors
4. Document the error instances and error types
5. Modify data entry procedures to reduce future errors

WHY CLEAN DATA?

Errors are common and expected

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

12 KEY PRINCIPLES OF DATA CLEANING

Plan	Organize	Prevention	Responsibility
Partnership	Prioritization	Performance measures	Optimization
Feedback	Training	Transparency	Documentation

SPOTTING ERRORS

- TECHNICA L ERRORS
  - Completeness
  - Bounds
  - Data Type
  - Data Format
- CONSISTENCY ERRORS
  - Taxonomic
  - Currency
  - Outliers
  - Geographic
  - Collecting Patterns
  - Accuracy and Precision
  - Collecting Methods

DATA CLEANING TOOLS

- Considerations in choosing:
  - Price, availability, and licensing
  - Ease of use
  - Documentation and support
  - Flexibility
  - Performance
- Technical considerations:
  - Use an exchange format
  - CSV and TSV
  - Encoding: UTF-8
  - Documenting formats and options used when creating files

DATA CLEANING TOOLS

BDI software tools database:

https://bit.ly/BDISoftwareDB

TOOLS - TAXONOMIC

- GBIF Names Parser
- Global Names Resolver
- CoL Checklist Bank Name Match
- iPlant Taxonomic Name Resolution Service
- World Register of Marine Organisms (WoRMS)

TOOLS - DMS TO DECIMAL FORMAT

- Degree Minutes Second Hemisphere to Decimal Degrees

$$DD = (D + M/60 + S/3600) * (H)$$

Hemisphere longitudinal: West = 1; East = 1

- Canadensys Coordinate Conversion

TOPIC - GEOREFERENCING

- CRIA's Species Link
- infoXY - locality information form coordinates
- Georeferencing Calculator - georeferencing of descriptive localities (Museum of Natural History Collections)
- Google Earth, Google Maps, QGIS

TOOLS - AUTOMATING

- Regular expressions (powerful, all platforms)
- Scripts (e.g. Perl, Python)
- Bash, DOS (shell scripts)
- R (command line), RStudio (GUI)

END OF PRESENTATION