Lesson 9: Analysis and Workflows

View all Education Modules at https://www.dataone.org/education-modules

Typical data analyses

Data processing: may include selecting a subset of data for analysis, merging multiple data sets, manipulating data for usability, or data transformation

Graphical analysis: makes it easier to see patterns and can aid in the identification of outliers

Statistical analysis: conventional statistics are used to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data

Science is iterative: the process that results in the final product can be complex.

Reproducibility..

...is at the core of the scientific process. If results are not reproducible, they lose credibility.

Good documentation of the data and the analysis are essential!

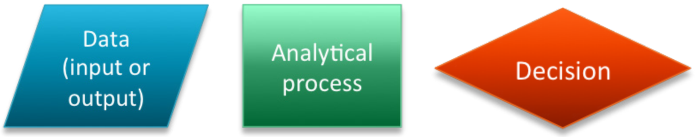
Workflows

Definition: Precise description of the procedures used in a project. Can be formal or informal.

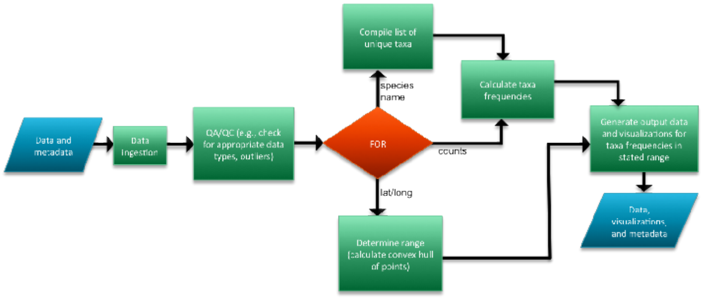
Informal workflow

No special software is needed to create workflow diagrams. Workflow diagrams include:

- Inputs and outputs
- Transformation rules or analytical processes
- Decision points
- Arrows indicating direction of process flow



Informal Workflow Example

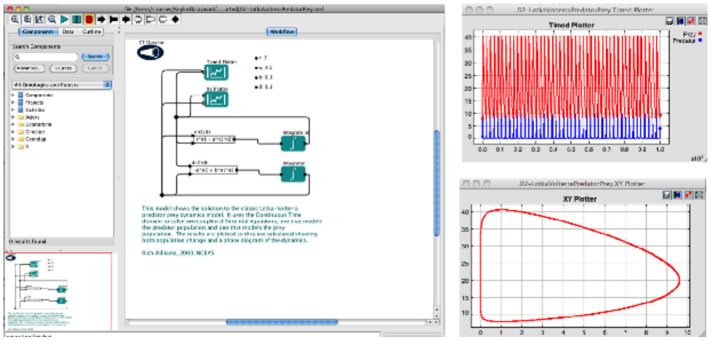

Data and metadata
Data Ingestion
QA/QC (e.g., check for appropriate data types, outliers)
FOR
species name
Compile list of unique taxa
counts
Calculate taxa frequencies
lat/long
Determine range (calculate convex hull of points)
Generate output data and visualizations for taxa frequencies in stated range
Data, visualizations, and metadata

Formal Workflow

Analytical pipeline where each step can be implemented in different software systems. Parameters and requirements for each step are formally recorded.

- Single access point for multiple analyses across software packages
- Keeps track of analysis and provenance to better enable reproducibility
- Workflow can be stored
- Allows sharing and reuse of individual steps or overall workflow

Formal workflow example: Kepler software



Best practices for data analysis

Formally or informally document the workflows used to create results. Include:

- Data provenance
- Analyses and parameters used
- Connections betweeen analyses via inputs and outputs

Document the code you write for analyses.

- Well-documented code is easier to review and share and enables repeated analyses
- Include project level information; script dependencies, inputs, and outputs; parameters; and what happens in individual sections

Construct end-to-end scripts that run the entire process from start to finish without intervention.

Local contact information