Dataset Harmonization

2019-03-12

Environmental Data Initiative (EDI)

Harmonization Goals

1) Flexible intermediate format so common scripts can streamline later analysis
2) Does not interfere with local needs and processes
3) Mechanism for dataset preparers to know
    a) Data elements that are the most important
    b) Data arrangements that are the easiest to use

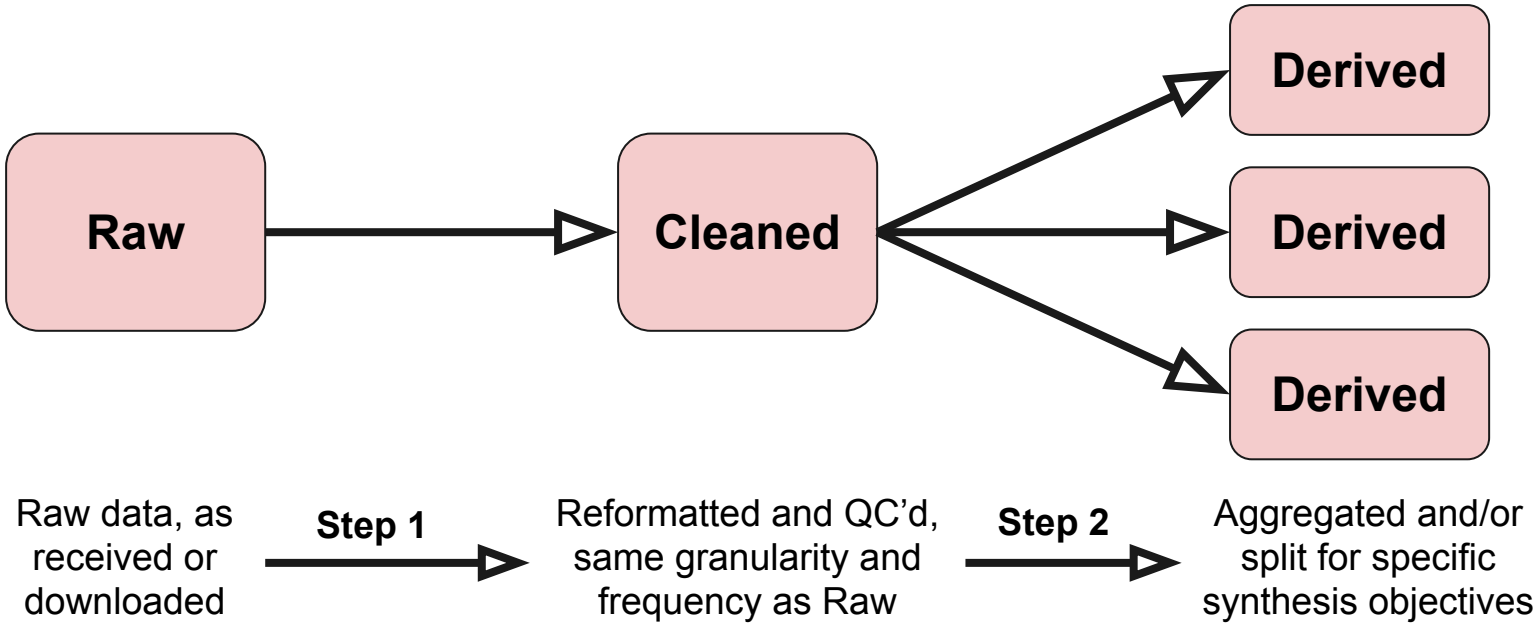
Thematic approach

Work with scientists currently engaged in a scientific domain

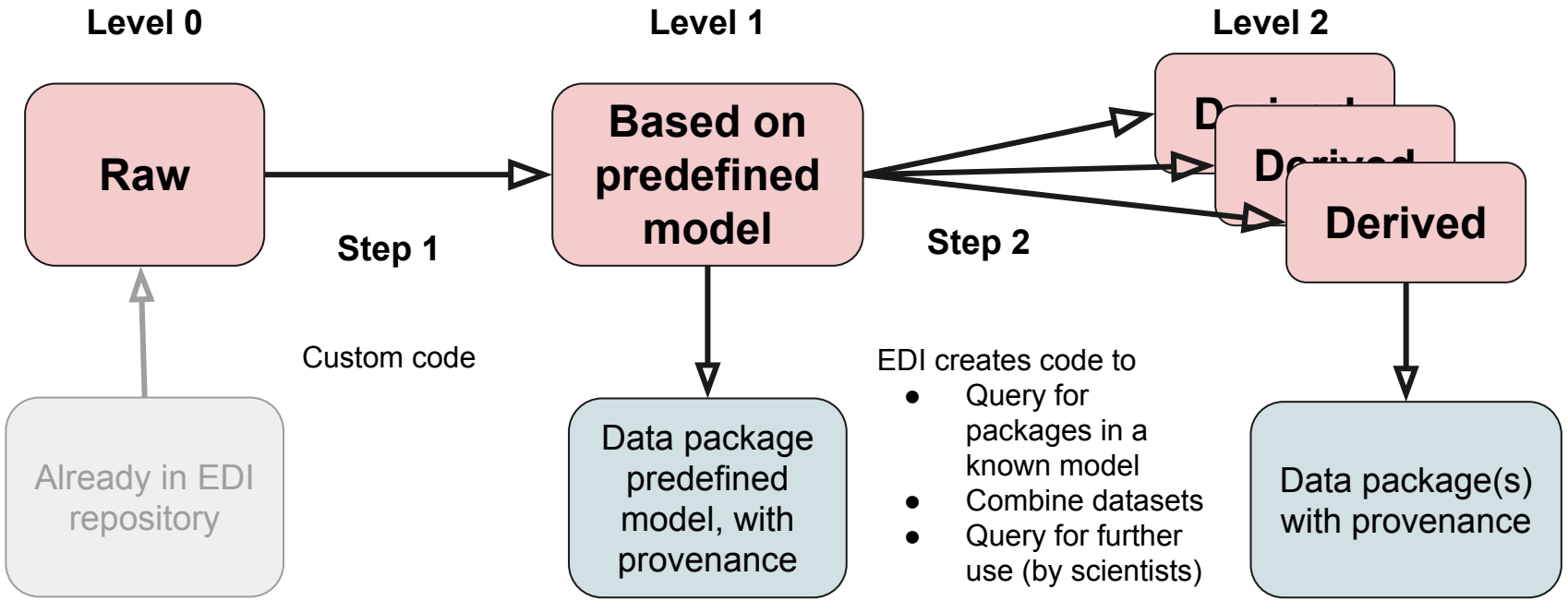
Template process

For reuse in other scientific domains

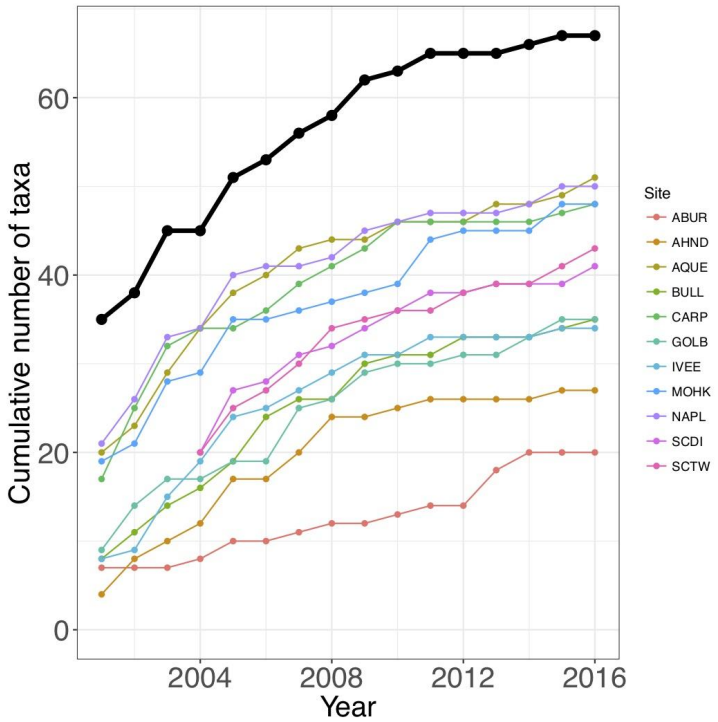
Typical Reuse Workflow

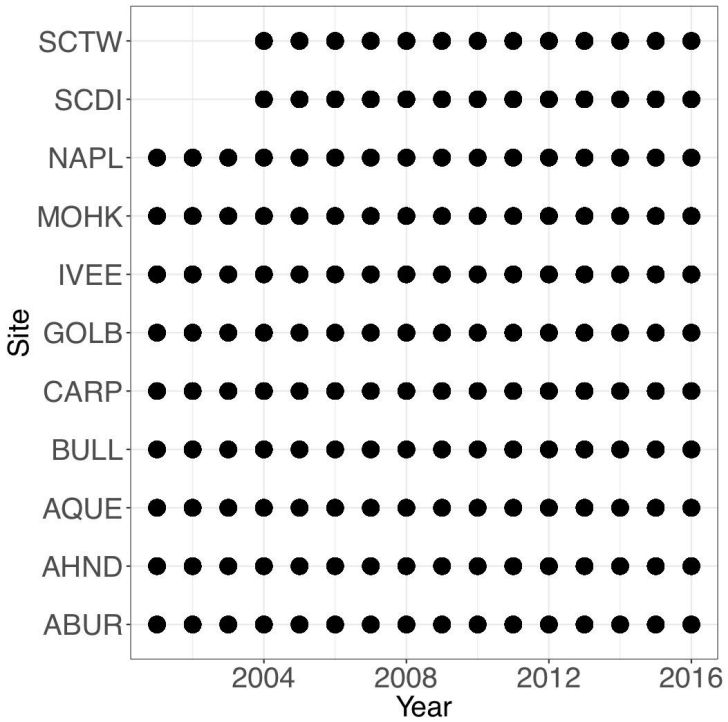


Ideal Reuse Workflow



Harmonized Format -> Harmonized Plots





From: Lany et al, 2018. A reproducible workflow for synthesizing disparate LTER data (ASM 2018, Poster session)

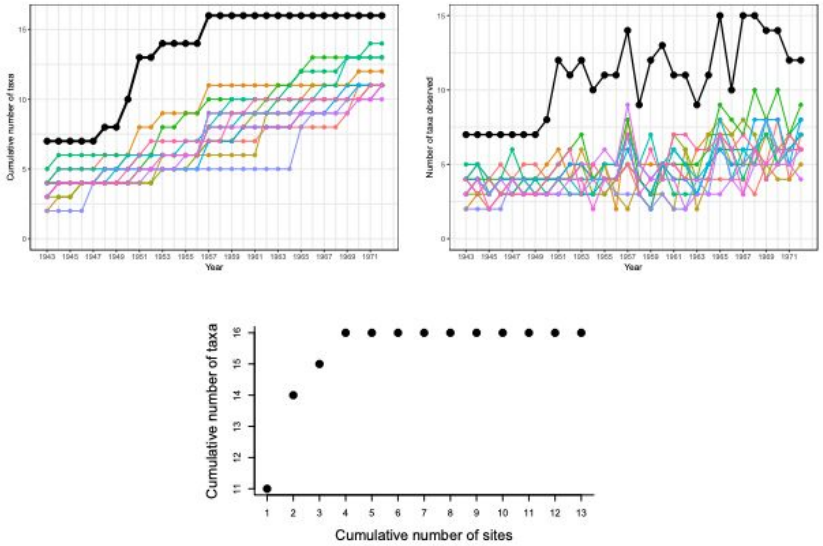


Figure 1: Temporal species accumulation curves (upper left), annual richness (upper right), and spatial species accumulation curve (lower) for 13 plots at Hay, Kansas (1943-1972). The black lines represent total site-level values across all plots.

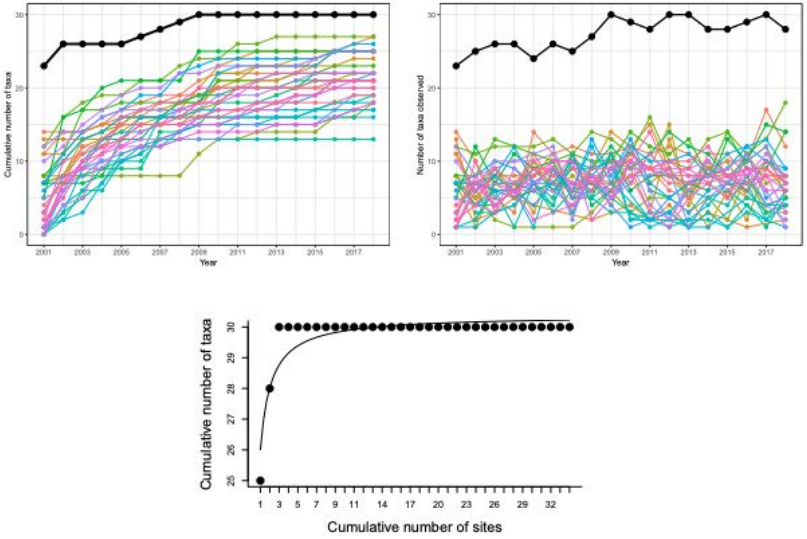


Figure 14: Temporal species accumulation curves (left), annual richness (right), and spatial species accumulation curve (lower) for sessile invertebrate and algal taxa at 34 plots at Santa Barbara Coastal LTER. The black lines represent site-level values.

Objective - Design Pattern for Level 1 Dataset

Flexible format, for multiple types of measurements and synthesis projects

Metadata in EML

Reformat only, no calculations or aggregations

“Derived product”; original data referenced

Database-style linking between tables

Complete; all original material is present

Basic Process

Examine available models currently in use

Find and describe patterns

Define (or adopt) design pattern (e.g., tables, typing)

Test pattern against data of interest

Convert datasets (L0 > L1)

Create utility scripts for QC, metadata generation

Create recommendations for L0 data submitters

Basic Process

Examine available models currently in use -- ***this workshop***

Find and describe patterns -- ***this workshop***

Define (or adopt) design pattern (e.g., tables, typing) -- ***this workshop***

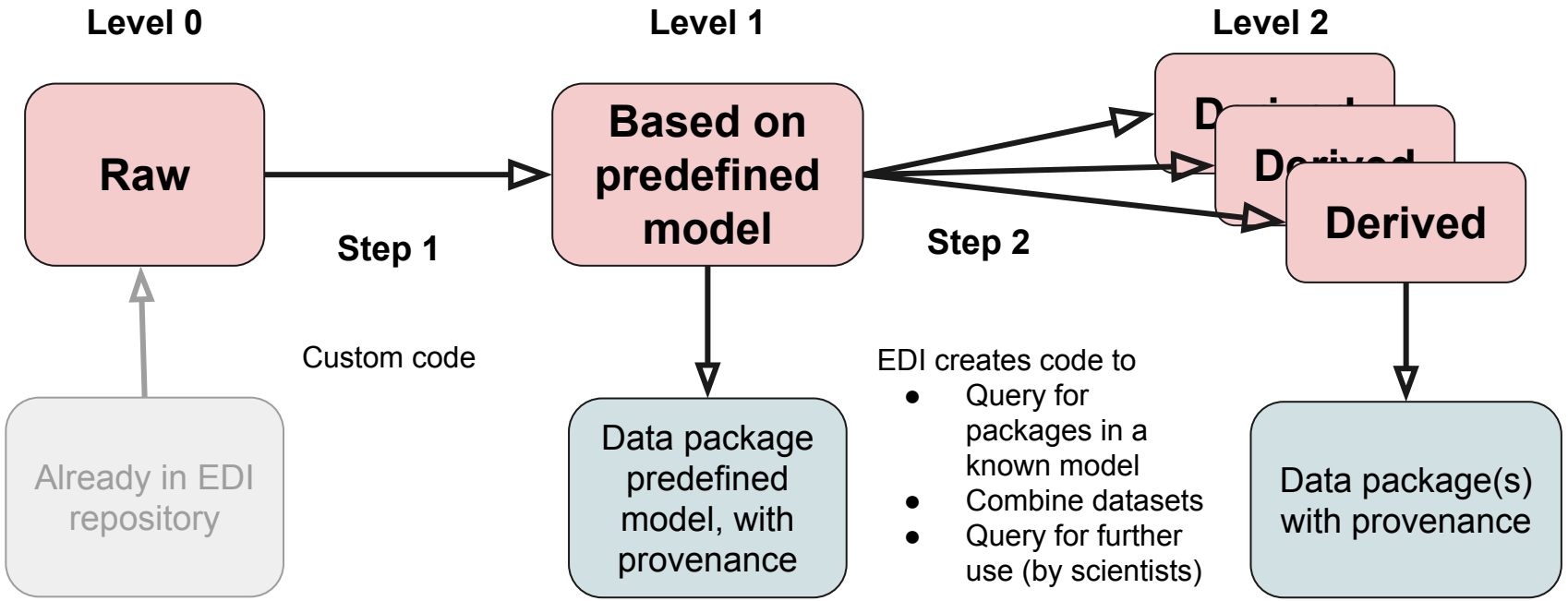
Test pattern against data of interest -- ***this workshop***

Convert datasets (L0 > L1) -- ***EDI***

Create utility scripts for QC, metadata generation -- ***EDI***

Create recommendations for L0 data submitters -- ***EDI***

Ideal Reuse Workflow



Scripts

Step 1	Step 2

● Validate tables
  ○ Referential integrity
  ○ Unique constraints
● Create EML metadata
  ○ Using EML R library
  ○ Metadata templates
	● Query for packages in a known model
	● Combine datasets
	● Query for further use (by scientists)

Example: Ecological Community Surveys

https://github.com/EDIorg/ecocomDP/

… documentation/examples/user_workflow.R

Progress

- Model
- Utility Scripts
- Dataset conversions - 36
- Collaborations - NEON, Popler (GBIF)

Community Survey Data Workshop 2017, Albuquerque

Next Steps - ecocomDP

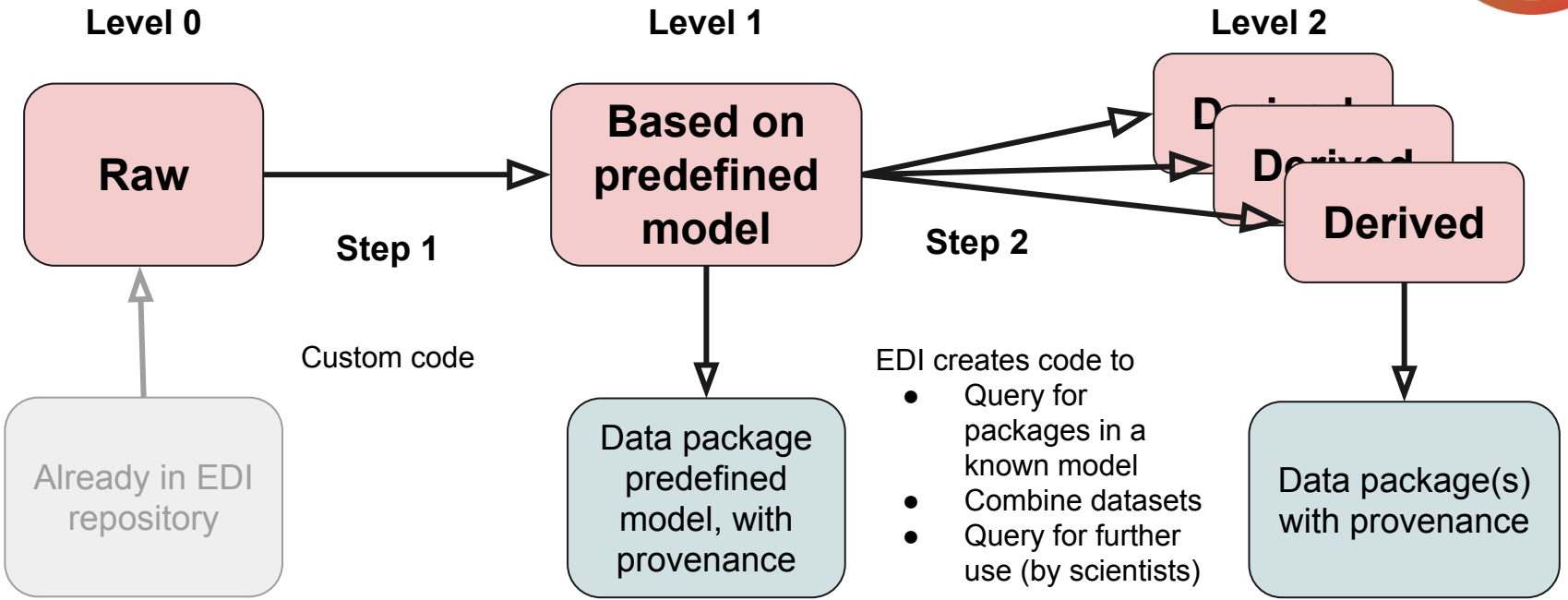
Dataset conversions
- Prioritized by LTER sites

Conversion/creation resources
- Mapping/planning template, “Best Practices”
- Additional QC and validation
- Manipulations with `gather()`, `spread()` from the **tidyr** package

Use
- Aggregation scripts
- Visualizations

Model enhancement
- Linkages to measurement vocabularies (following example in Taxon)
- Renaming (suggested: “Taxon” > “Organism”)

Questions?


Level 0
Level 1
Level 2
Raw
Step 1
Based on predefined model
Step 2
Derived
Custom code
EDI creates code to
Query for packages in a known model
Combine datasets
Query for further use (by scientists)
Already in EDI repository
Data package predefined model, with provenance
Data package(s) with provenance