ecocomDP

Dataset Design Pattern for Ecological Community Surveys

2017-09-26
Environmental Data Initiative (EDI)

Introduction

Goals

1) Flexible intermediate format so common scripts can streamline their analysis
2) Mechanism for preparers to know
    a) Data elements that are the most important
    b) Presentations are the easiest to use

Thematic approach

Work with scientists currently engaged in synthesis of primary "Metacommunities", "Synchrony" - LTER working groups

3) Template for a process that can be reused in other scientific domains

Summer 2017, EDI workshop, Albuquerque

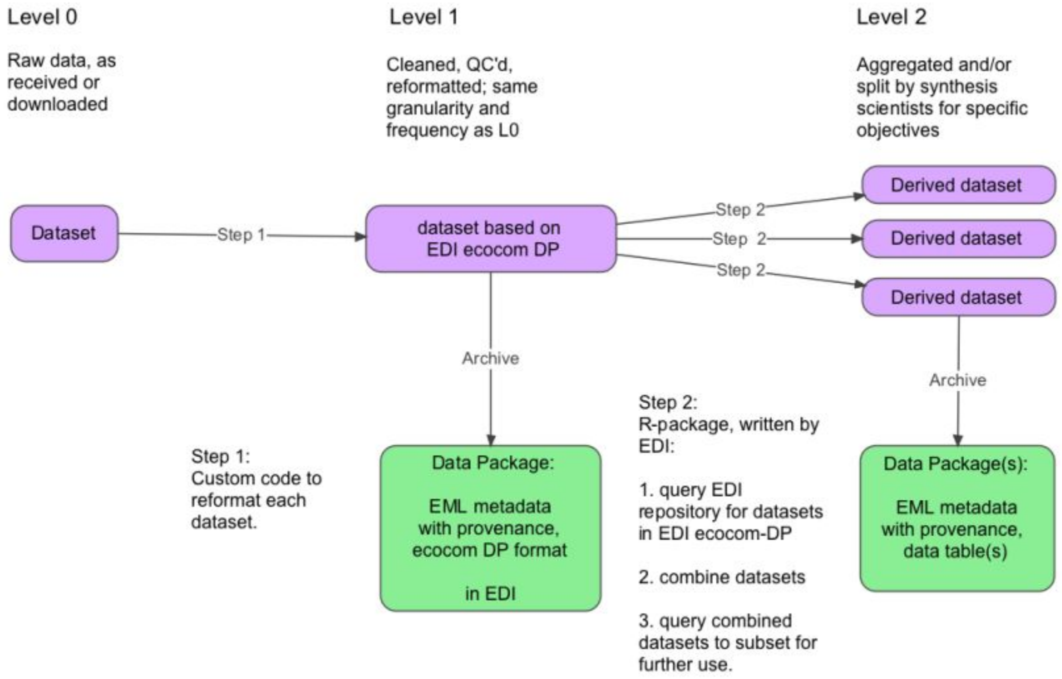
Background

	Popler	Darwin Core (Archive)
Developed by	Miller, Compagnoni, Bibian, et al	Biodiversity community
Supported by	NSF	GBIF/TDWG
Since	2015 (funded)	1998 (coined), 2009 (ratified)
Description	Relational DB and associated R code	Vocabulary of terms and dataset format
In a nutshell	Optimized for LTER time series Describes community-level abundance Effect of environmental fluctuations on populations	Optimized for organism occurrences No inherent concept of a time series; time-series data are contributed to GBIF, and a query can infer a time series from a group of records

Workflow

Steps

1. Custom code for reformatting, because datasets are designed with a project-specific sampling plan
2. If data are repackaged into a common format, Step 2 can be streamlined



Level 0

Raw data, as received or downloaded

Level 1

Cleaned, QC'd, reformatted; same granularity and frequency as L0

Level 2

Aggregated and/or split by synthesis scientists for specific objectives

Step 1:
Custom code to reformat each dataset.

Step 2:
R-package, written by EDI:

1. query EDI repository for datasets in EDI ecocom-DP

2. combine datasets

3. query combined datasets to subset for further use.

Objective - Design Pattern for Level 1 Dataset

Flexible format, for multiple types of measurements and synthesis projects

Metadata in EML

Reformat only, no calculations

Original data referenced

Complete; original records can be recreated

Database-style linking between tables

Model Overview

Observation table for data related to

Count, biomass, abundance, density

Primary organization

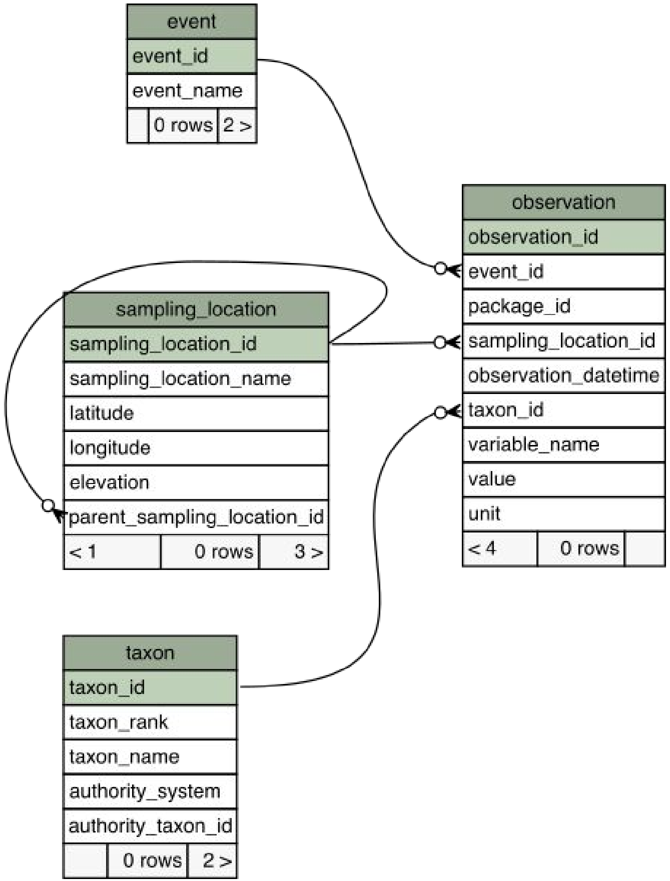
Entity, name, value, unit (EAV, U)

Linked to tables for

Sampling location

Organism

Event



event
event_id
event_name
0 rows | 2 >

observation
observation_id
event_id
package_id
sampling_location_id
observation_datetime
taxon_id
variable_name
value
unit
< 4 | 0 rows

sampling_location
sampling_location_id
sampling_location_name
latitude
longitude
elevation
parent_sampling_location_id
< 1 | 0 rows | 3 >

taxon
taxon_id
taxon_rank
taxon_name
authority_system
authority_taxon_id
0 rows | 2 >

Model Overview

Ancillary tables

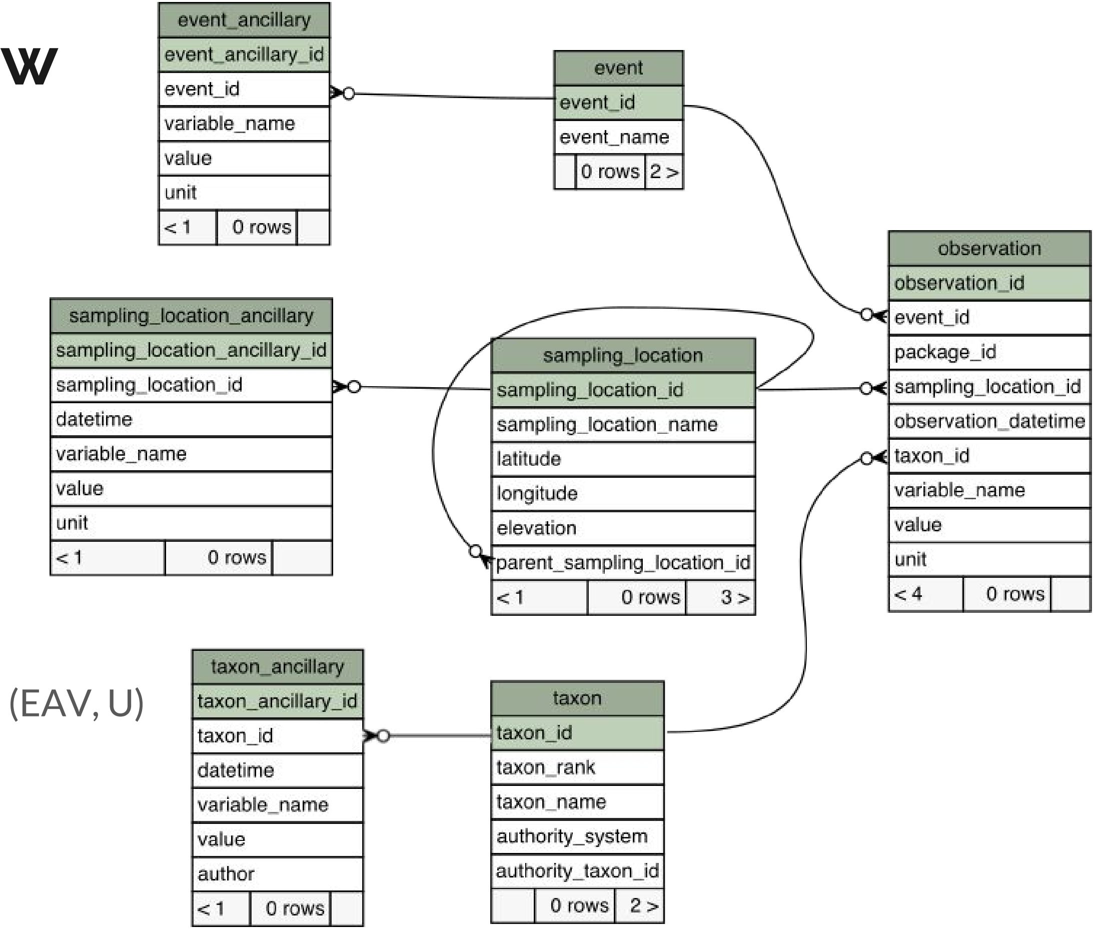
Event

Location

Organism

Primary organization

Entity, name, value, unit (EAV, U)



event_ancillary
event_ancillary_id
event_id
variable_name
value
unit
< 1 | 0 rows

event
event_id
event_name
0 rows | 2 >

observation
observation_id
event_id
package_id
sampling_location_id
observation_datetime
taxon_id
variable_name
value
unit
< 4 | 0 rows

sampling_location_ancillary
sampling_location_ancillary_id
sampling_location_id
datetime
variable_name
value
unit
< 1 | 0 rows

sampling_location
sampling_location_id
sampling_location_name
latitude
longitude
elevation
parent_sampling_location_id
< 1 | 0 rows | 3 >

taxon_ancillary
taxon_ancillary_id
taxon_id
datetime
variable_name
value
author
< 1 | 0 rows

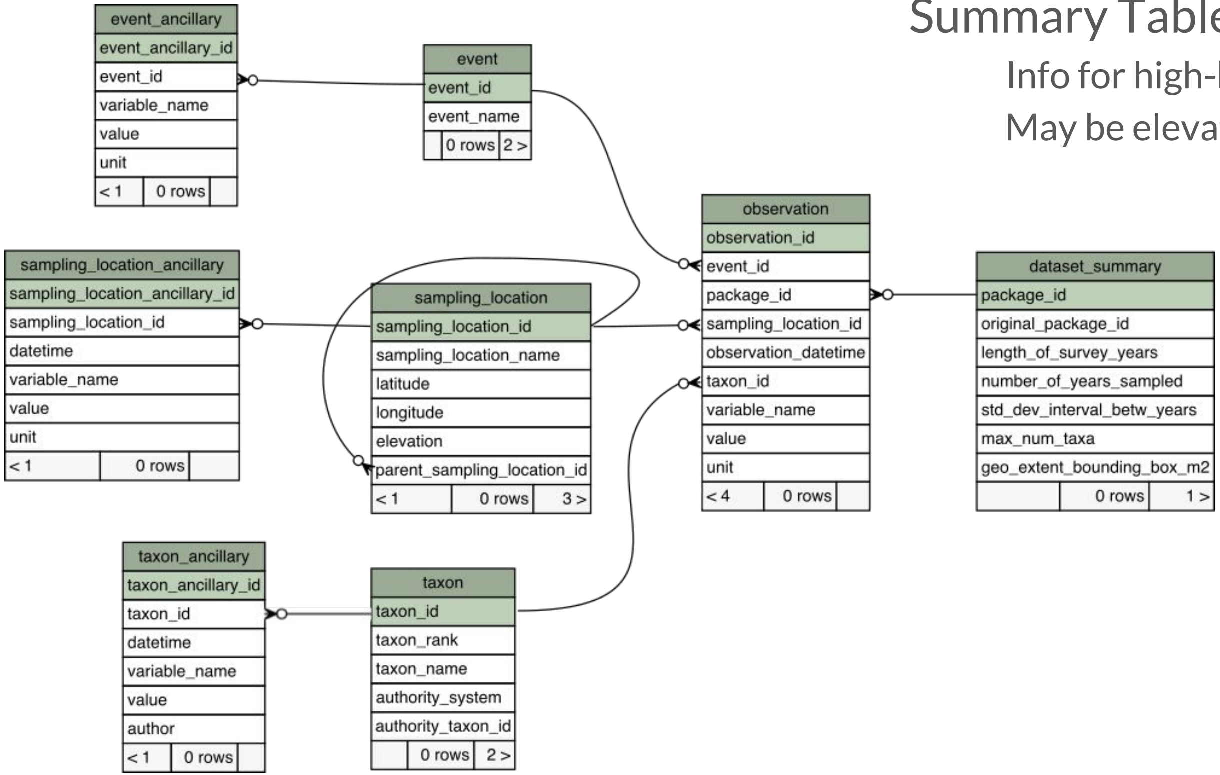
taxon
taxon_id
taxon_rank
taxon_name
authority_system
authority_taxon_id
0 rows | 2 >

Model Overview

Summary Table

Info for high-level evaluation

May be elevated to metadata



event_ancillary
event_ancillary_id
event_id
variable_name
value
unit
< 1 0 rows

event
event_id
event_name
0 rows 2 >

sampling_location_ancillary
sampling_location_ancillary_id
sampling_location_id
datetime
variable_name
value
unit
< 1 0 rows

sampling_location
sampling_location_id
sampling_location_name
latitude
longitude
elevation
parent_sampling_location_id
< 1 0 rows 3 >

observation
observation_id
event_id
package_id
sampling_location_id
observation_datetime
taxon_id
variable_name
value
unit
< 4 0 rows

dataset_summary
package_id
original_package_id
length_of_survey_years
number_of_years_sampled
std_dev_interval_betw_years
max_num_taxa
geo_extent_bounding_box_m2
0 rows 1 >

taxon_ancillary
taxon_ancillary_id
taxon_id
datetime
variable_name
value
author
< 1 0 rows

taxon
taxon_id
taxon_rank
taxon_name
authority_system
authority_taxon_id
0 rows 2 >

Summary - Table Features

Table	arrangement	Typing (value col)	Req?	Unique constraint
Location	Long (“tidy”)	-	yes	sampling_location_id
Taxon	Long (“tidy”)	-	yes	taxon_id
Event	Long (“tidy”)	-	no	event_id
Observation	Long, EAVU	numeric	yes	observation_id, event_id, package_id, sampling_location_id, observation_datetime, taxon_id, variable_name
Location_ancillary	Long, EAVU	character	no	sampling_location_id, variable_name
Taxon_ancillary	Long, EAVU	character	no	taxon_id, variable_name
Event_ancillary	Long, EAVU	character	no	event_id, variable_name
Summary	One line, generated	numeric	yes	summary_id

Progress - Datasets

Description	L0 ID	Total L0 Values	L1 ID	observation	location	taxon	event	location_ ancillary	taxon_ ancillary	event_ ancillary
NTL LTER Microbial Observatory, Bogs	NA	NA	lter-knb-ntl.344.2	1 var	9 sites	6208 taxa	1387 events		8 vars	86 vars
Wisconsin Lakes fish sizes 1944 - 2012	NA	NA	lter-knb-ntl.345.1	1 var	3148 sites	19 taxa	55 k events			1 var
Wisconsin Lakes fish abundance 1944 - 2012	NA	NA	lter-knb-ntl.346.2	2 vars	2594 sites	9 taxa	18 k events			2 vars
Santa Barbara Channel, integrated fish density	edi.5.2	73 m	tbd	1 var	3718 sites	390 taxa	tbd	2 vars		3 vars
Moorea fish size and abundance	lter-knb-mcr.6	1.7 m	tbd	3 vars	241 sites	388 taxa	792 events		3 vars	11 vars
Ctl AZ- Phoenix bird abundance and diversity	lter-knb-cap.46.14	4.2 m	tbd	x	x	x				
Ant assemblages during a Hemlock removal experiment	lter-knb-hfr.118.28	40 k	tbd	x	x	x				

Progress - Utility Scripts

Validate ecocomDP tables

- Referential integrity
- Unique constraints

Create EML metadata

- Using EML R library
- Metadata templates (entities, attributes, keywords)
- Summary table

Documentation

https://github.com/EDIorg/ecocomDP

Model Comparison

	ecocomDP	Popler	Darwin Core Archive (DwC-A)
Description	Design pattern for text tables that together comprise a data package	RDB with R libraries written to access/analyze content	Star schema, with vocabulary and text dataset for upload to GBIF
Table format	long	wide	wide
Approx size	4 datasets, 4 m rows	209 datasets (est), 6.6 m rows (total)	Unknown, 800 m (GBIF occurrences)
Data coverage	Ostensibly, complete	incomplete	incomplete
Source traceable	yes	Yes	Left to contributor
Spatial	Infinite nesting; spatial characteristics with location_ancillary	5 levels (labeled cols); 1 other characteristic (extent)	Left to contributor
Taxonomy	tree not present, retrieve from referenced authority	Entire tree included, with controlled levels (zoology)	Authority ID required, tree not required
R access	(planned)	Yes	Yes
Updates accepted	Yes, by anyone	unknown	Yes, by anyone

References

ecocomDP
Schema (postgres implementation): http://sbc.lternet.edu/~mob/EDI/schemaSpy/ecocom_dp/
GitHub: https://github.com/EDIorg/ecocomDP

Popler
Schema ERD: http://sbc.lternet.edu/~mob/EDI/schemaSpy/popler
GitHub (R package): https://github.com/AldoCompagnoni/popler
GitHub (database): https://github.com/bibsian/database-development

DwC Archive:
Homepage: http://www.tdwg.org/standards
GitHub: https://github.com/tdwg/dwc

Potential Issues

Key-Value pairs:

- Values: lack typing
- Keys: lack a vocabulary

	Key (variable_name)	Value typing	Unit
ecocomDP	tbd	numeric	Required field
Popler	unknown (possibly by table name)	numeric	Unknown (possibly via metadata key)
DwC-A	vocabularies suggested, not required	No typing (char)	Required field

Next Steps - ecocomDP

Conversion/creation resources

Mapping/planning template, “Best Practices”,

Summary table creation

Continue with QC and validation

Manipulations with `gather()`, `spread()` from the **tidyr** package

Aggregation scripts

Require model to be stable, with example datasets converted

Model enhancement

Linkages to measurement vocabularies (following example in Taxon)

Renaming (suggested: “Taxon” > “Organism”)

Discussion

EDI-Popler Collaboration?

Collaborators could...	But first...
Write code to convert data in ecocomDP to popler (and reverse)	Both need to be stable.
Drop ecocomDP, use Popler	Understand, identify and handle certain Popler limitations: ‘LTER data’, usefulness for other types of synthesis
Merge ecodomdp and popler, in a data package implementation	Suggest changes to Popler. eg, merge the 5 community obs tables. Add cols for external meas vocab.
Develop a vocab for variables	create lists of expected measurements

Popler questions

How can new data be added

process, formats, restrictions

How does popler handle ancillary observations in original data?

Eg, depth in a water column, size classes of taxa,