ecocomDP

Dataset Design Pattern for Ecological Community Surveys

Environmental Data Initiative (EDI)

2018

Session Agenda

Introduction & justification

1) **Process**
2) **Results**

Progress

1) Model
2) What we've found in data
3) Code and tools

Your input, code demo, and/or play with data

Introduction

Goals

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

Thematic approach

Work with scientists synthesizing primary data:
“Metacommunities”, “Synchrony” - LTER working groups

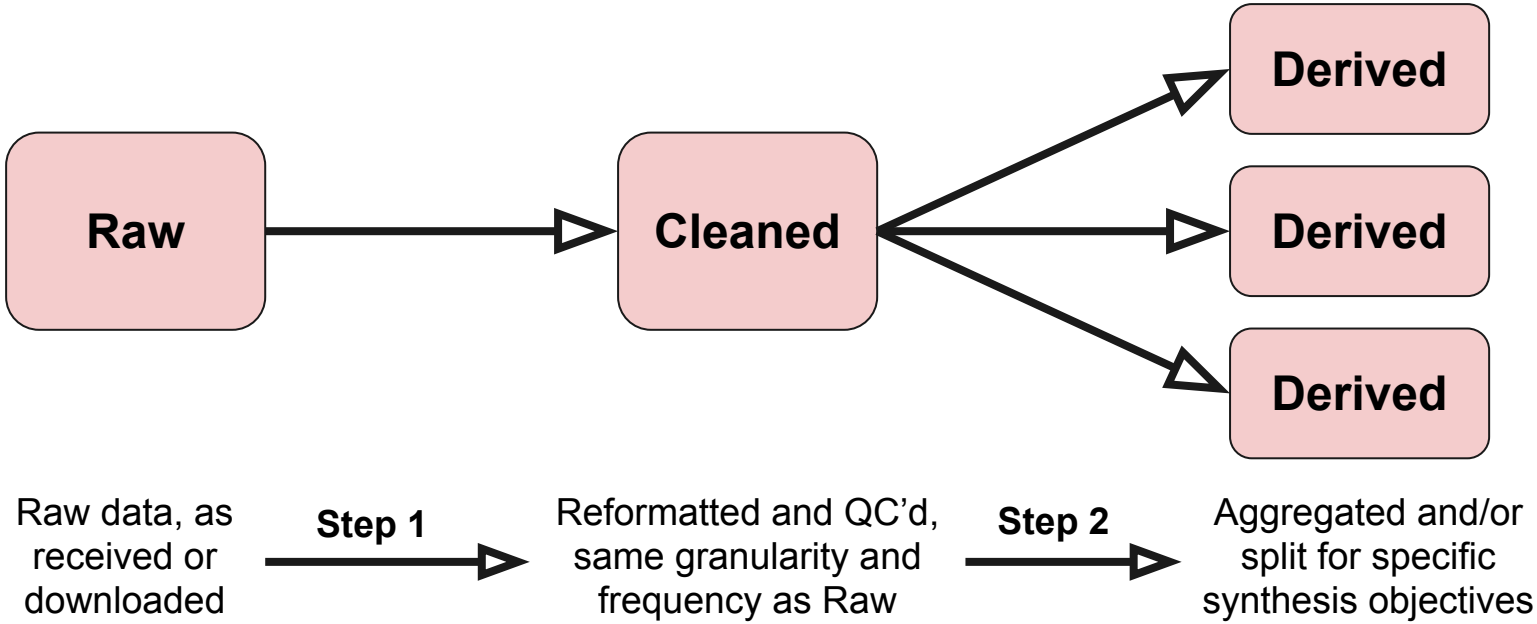
3) **Template for a process that can be reused**

Summer 2017, EDI workshop, Albuquerque

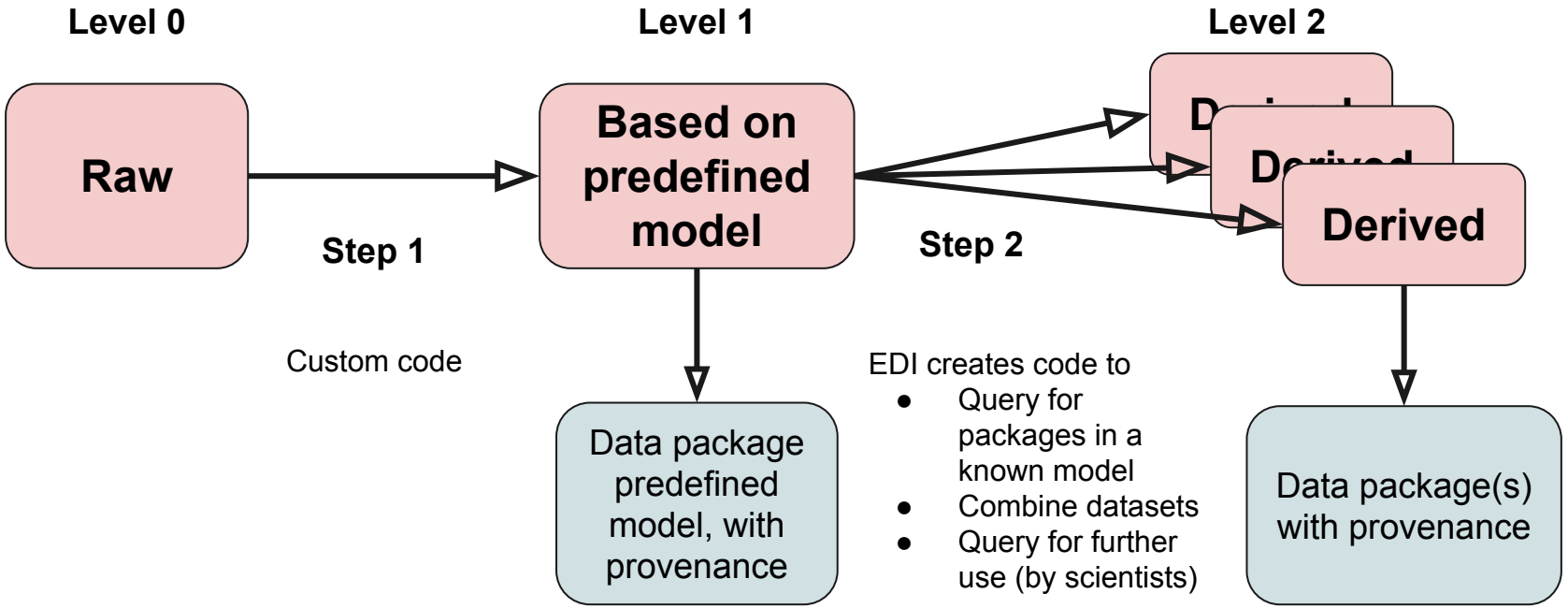
Background

	Popler	Darwin Core (Archive)	BioTIME
Authors	Miller, Compagnoni, Bibian, et al	Biodiversity community	Dornelas, et al
Support	NSF	GBIF/TDWG	ERC
Timeline	2015 (funded)	1998 (coined), 2009 (ratified)	2016 (data paper)
Description	Relational DB and associated R code	Vocabulary of terms and dataset format	Relational database with web interface
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 added as a dataset become independent; query infers a time series from a group of records	Optimized for assessing global biodiversity change Describes community level abundance global

Typical Synthesis Workflow



Ideal Synthesis Workflow



Objective - Design Pattern for Level 1 Dataset

Flexible format, for multiple types of measurements and synthesis projects

Metadata in EML

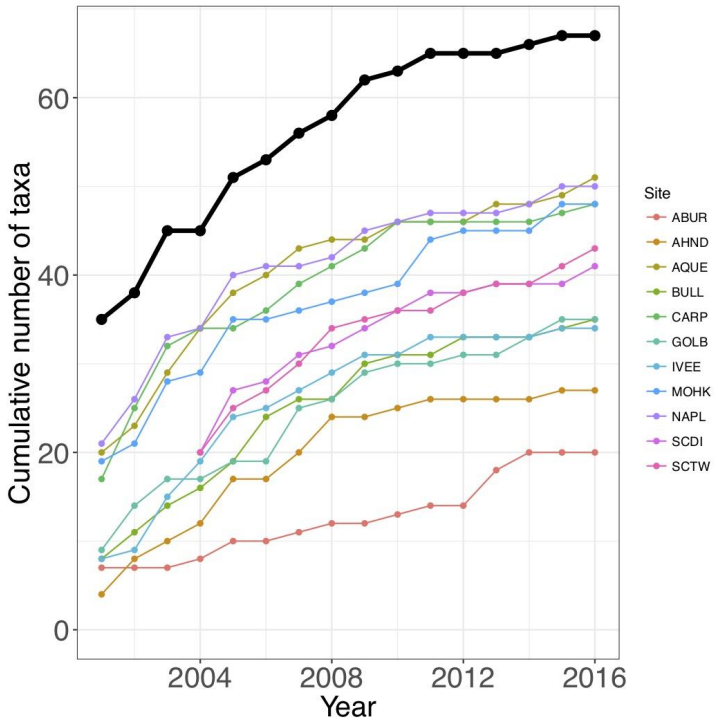
Reformat only, no calculations

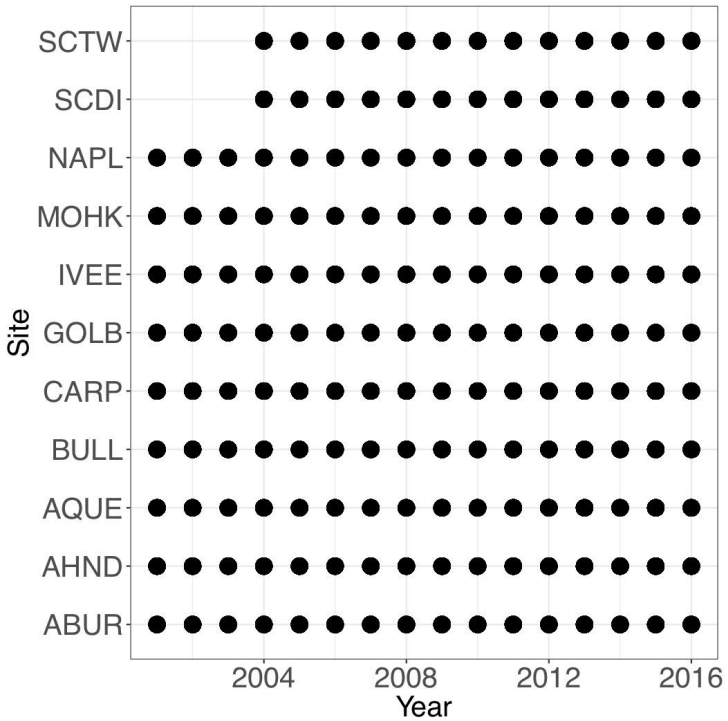
Original data referenced

Complete; original records can be recreated

Database-style linking between tables

Harmonized Format -> Harmonized Plots





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

Basic Process

Examine available models currently in use

Examine ad hoc cleaned (Level 1) data created by synthesis working groups

Find and describe patterns

Define common design pattern tables, typing

Test model against data of interest to WGs

Create utility scripts for QC, metadata

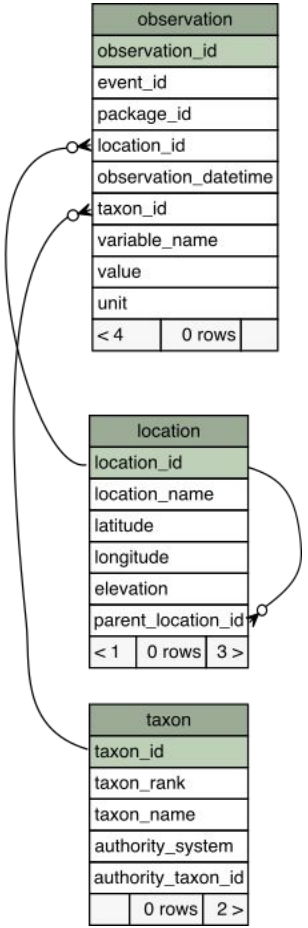
Model Overview

Observation table for data

- Count, biomass, abundance, density
- Primary organization
  - Entity, attribute, value, unit (EAV, U)

Essential tables

- Sampling location
- Organism



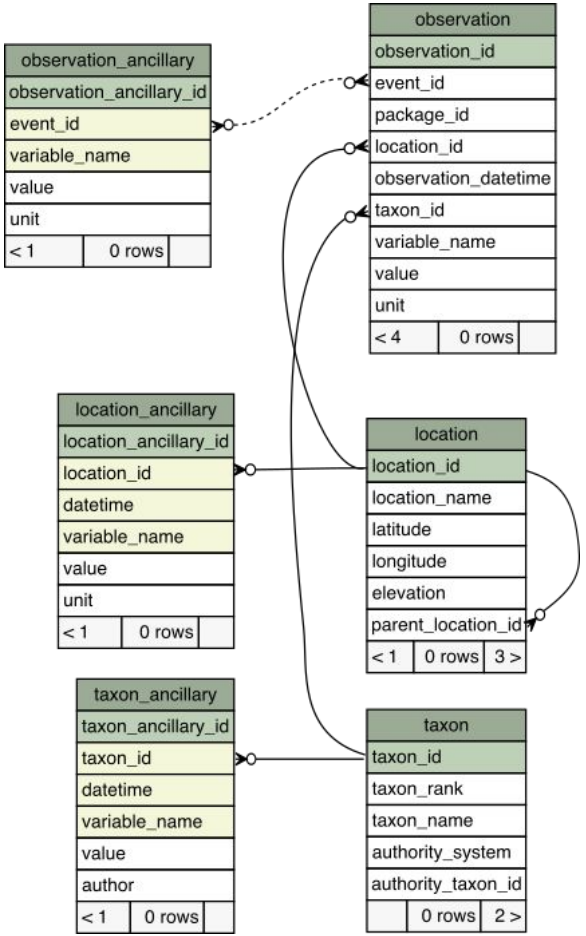
Model Overview

Ancillary tables

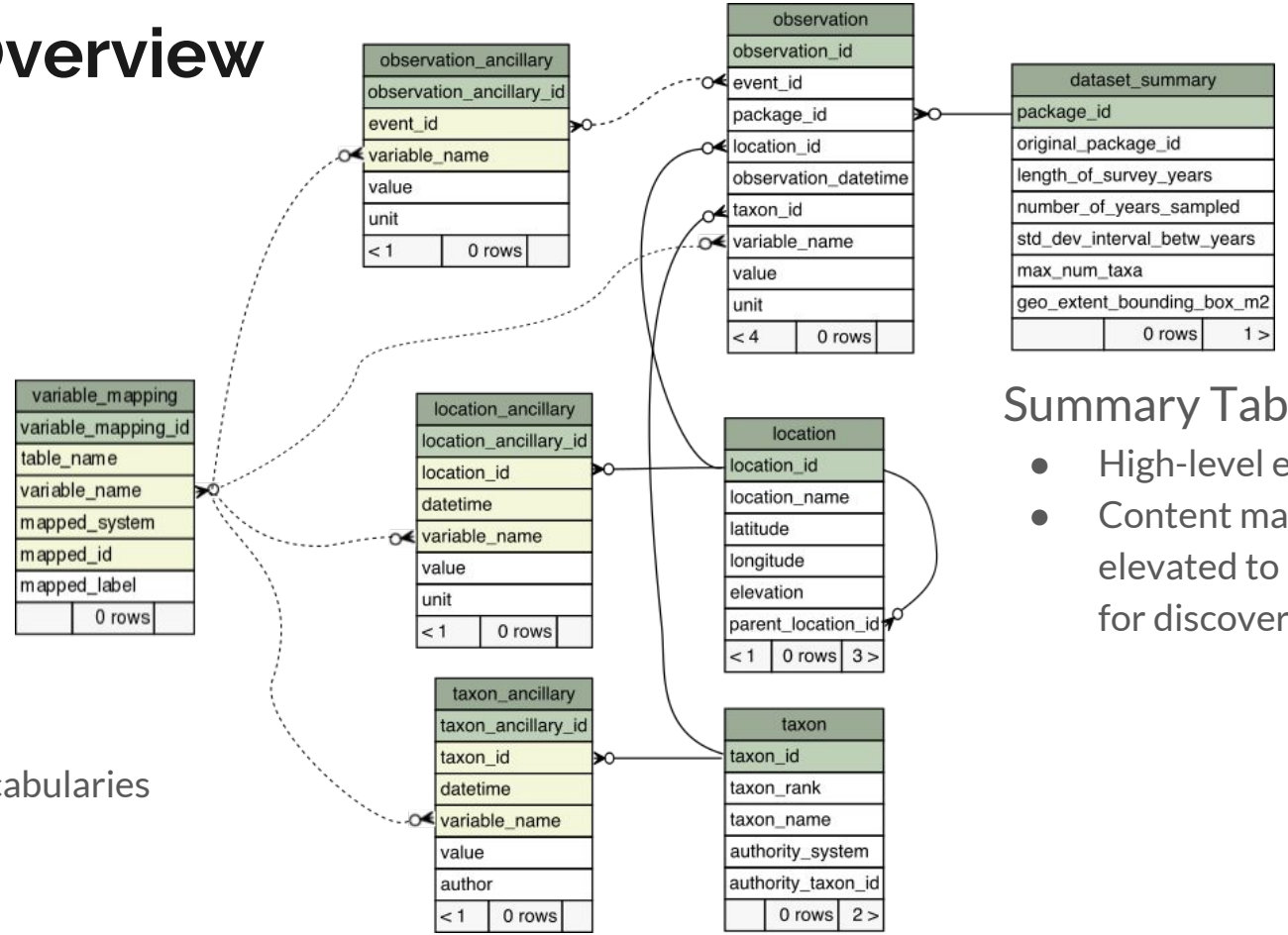
- Observation
- Location
- Organism

Primary organization

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



Model Overview



Summary Table

- High-level evaluation
- Content may be elevated to metadata for discovery

Variable Mappings

- Link to external vocabularies

Summary - Table Features

Table	arrangement	Req?	Unique constraint
Location	Long (“tidy”)	yes	location_id
Taxon	Long (“tidy”)	yes	taxon_id
Observation	Long, EAVU	yes	observation_id, event_id, package_id, sampling_location_id, observation_datetime, taxon_id, variable_name
Location_ancillary	Long, EAVU	no	location_id, datetime, variable_name
Taxon_ancillary	Long, EAVU	no	taxon_id, datetime, variable_name
Observation_ancillary	Long, EAVU	no	observation_id, variable_name
Variable_mapping	Long, EAV		table_name, variable_name, mapped_system, mapped_id
Summary	One line, generated	yes	summary_id

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 (measurements are long)
Approx size	10 datasets, 4 m rows	209 datasets (est), 6.6 m rows (total)	Unknown, >1 b GBIF occurrences
Data coverage	TBD (ostensibly complete)	Incomplete (time-limited)	Incomplete (contributor-limited)
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	Yes	Yes	Yes
Updates accepted	Yes, by anyone	unknown	Yes, by anyone

Key-Value Pairs

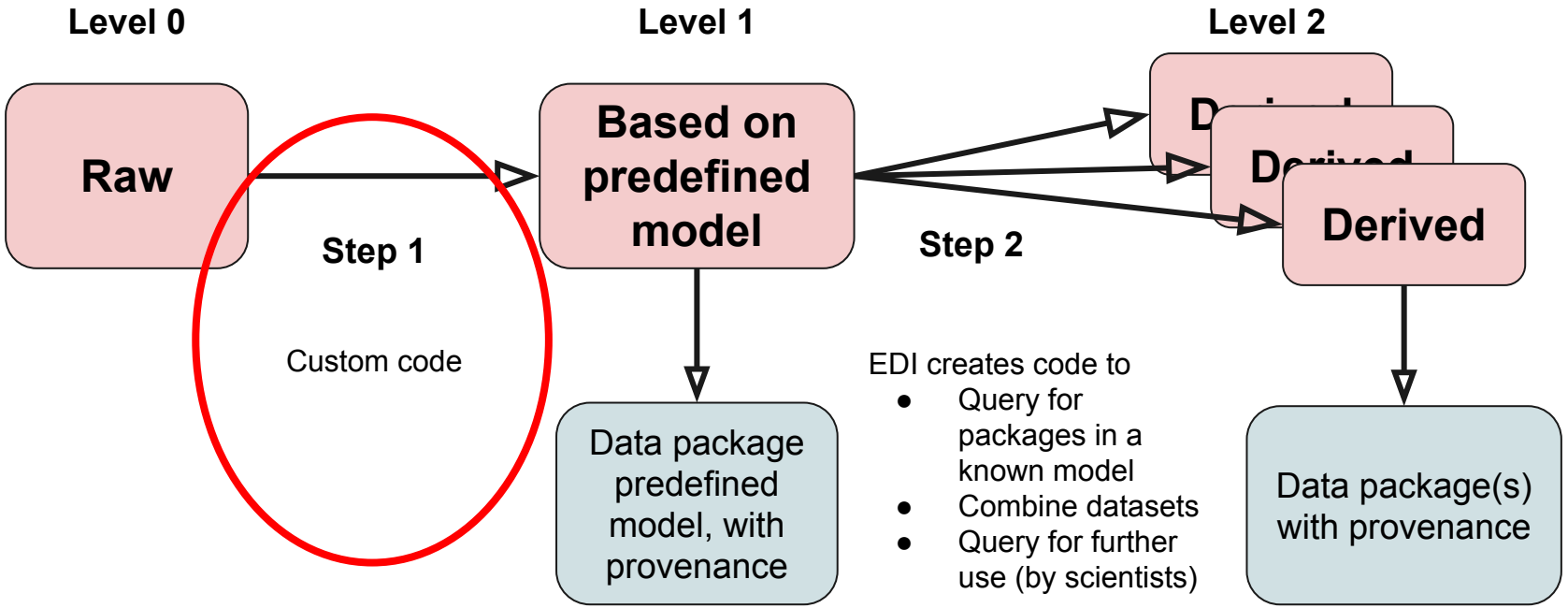
In general:

Values: lack typing

Keys: lack a vocabulary

	Key (variable_name)	Value typing	Unit
ecocomDP	Supported, vocabularies lacking	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

Ideal Synthesis Workflow



Utility Scripts - Dataset Conversion

Validate ecocomDP tables

- Referential integrity
- Unique constraints

Create EML metadata

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

Documentation

- Model description
- Script use
- Recommendations for practice (in progress)

https://github.com/EDIorg/ecocomDP

Metrics - Converted Datasets

	Required Tables			Ancillary Tables		
	Location	Taxon	Observation	Location	Taxon	Observation
N - table occurrences, ecocomDP packages	17	11[1]	17	10	9	13
N - Locations or Taxa (named tables)	7859	2030[2]	-	-	-	-
N - Variables (Observation, ancillary tables)	-	-	26	29[4]	37	160
Median (N/dataset)	124	118	1	4	3	6
N taxonomic DBs referenced	-	4[3]	-	-	-	-
N taxa with external DB identifier	-	707	-	-	-	-

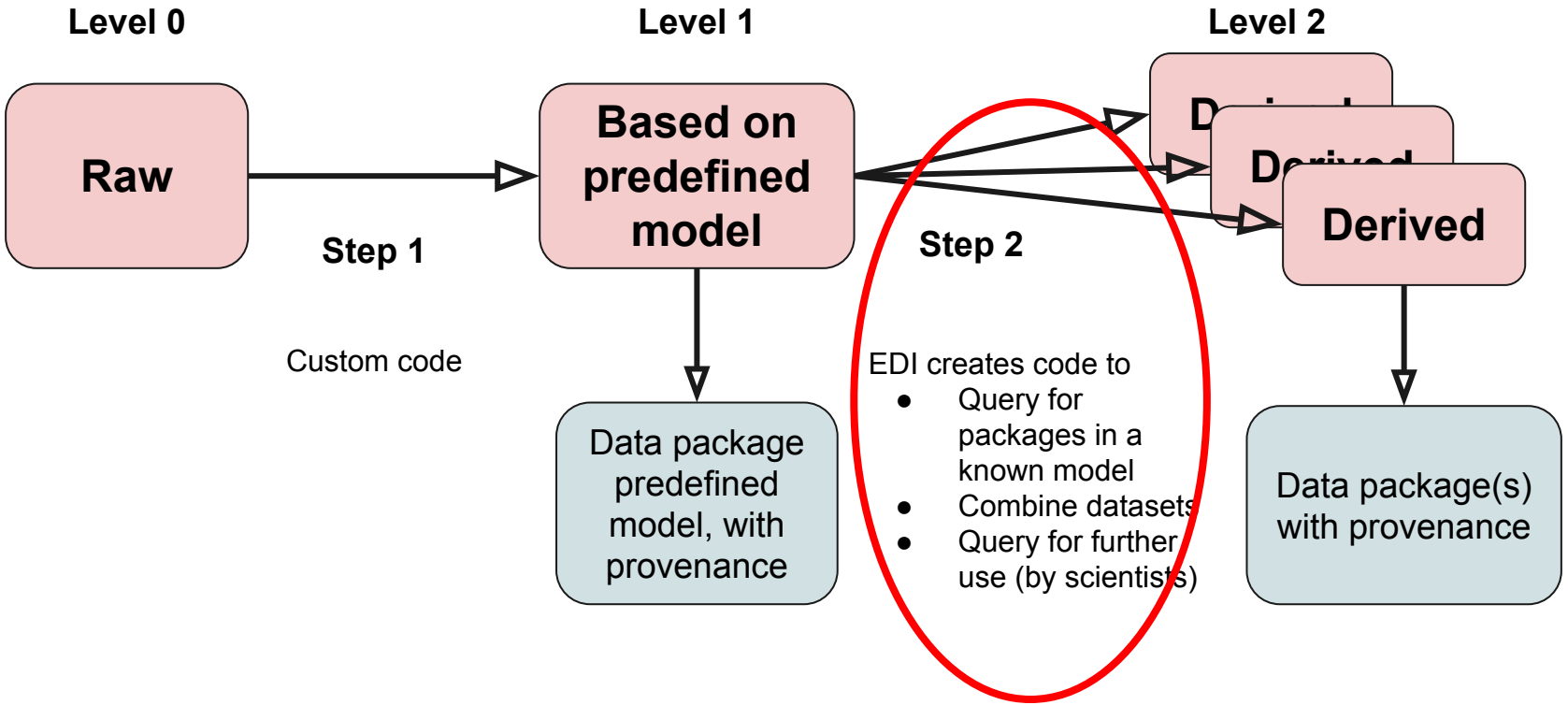
Legend:

Green background: required tables
Yellow border: tables with EAV-U design
"-" metric not appropriate to this table

Footnotes:

[1] *Six datasets have taxon information in progress*
[2] *6146 OTUs (1 dataset) omitted from total, included in median*
[3] *OTUs reference one paper, rather than an authoritative taxon system*
[4] *Total does not include 199 land use features (2 datasets)*

Ideal Synthesis Workflow



Utility Scripts - Aggregation

EDI creates R code to

- Query for packages in a known model
- Combine datasets
- Query for further use (by scientists)

NEON/EDI R code to

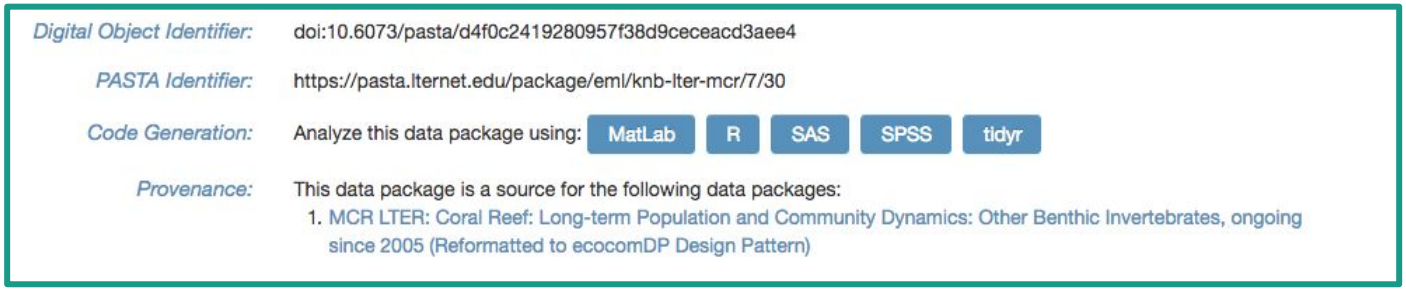
- Query NEON for macroinvertebrate data, export ecocomDP
- Filter by site

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

… documentation/examples/user_workflow.R

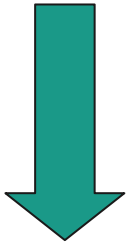
Provenance

https://portal.edirepository.org/nis/mapbrowse?scope=knb-lter-mcr&identifier=7



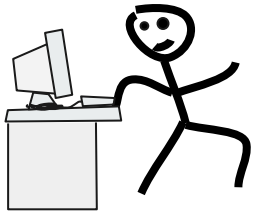
https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=194

Source Dataset

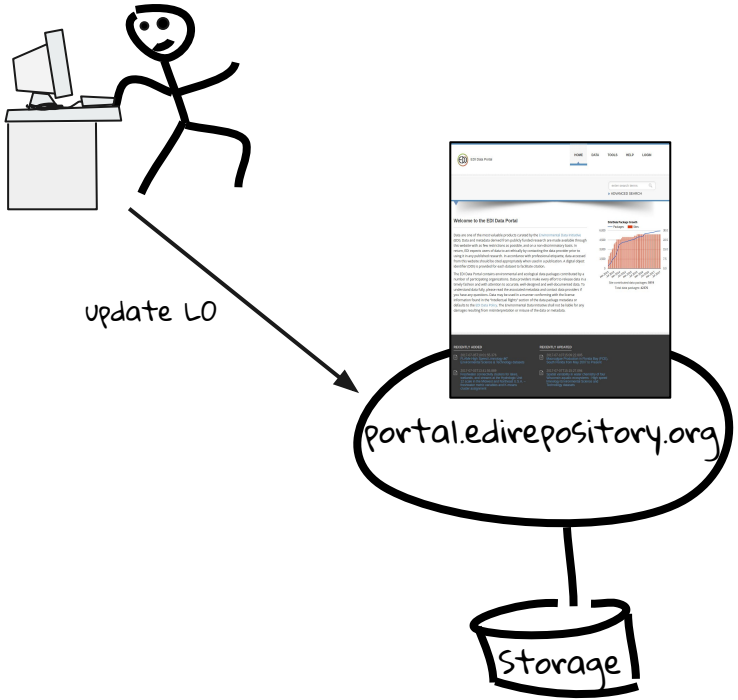


Derived Dataset

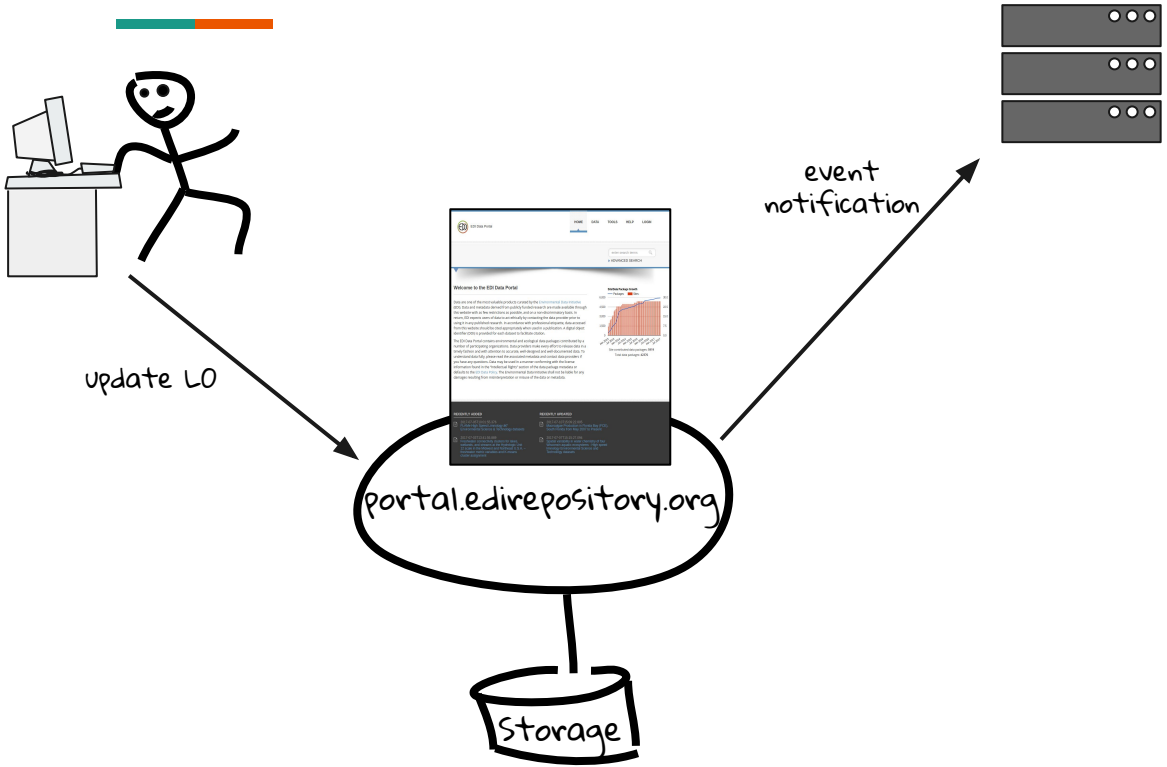
Automated Updates



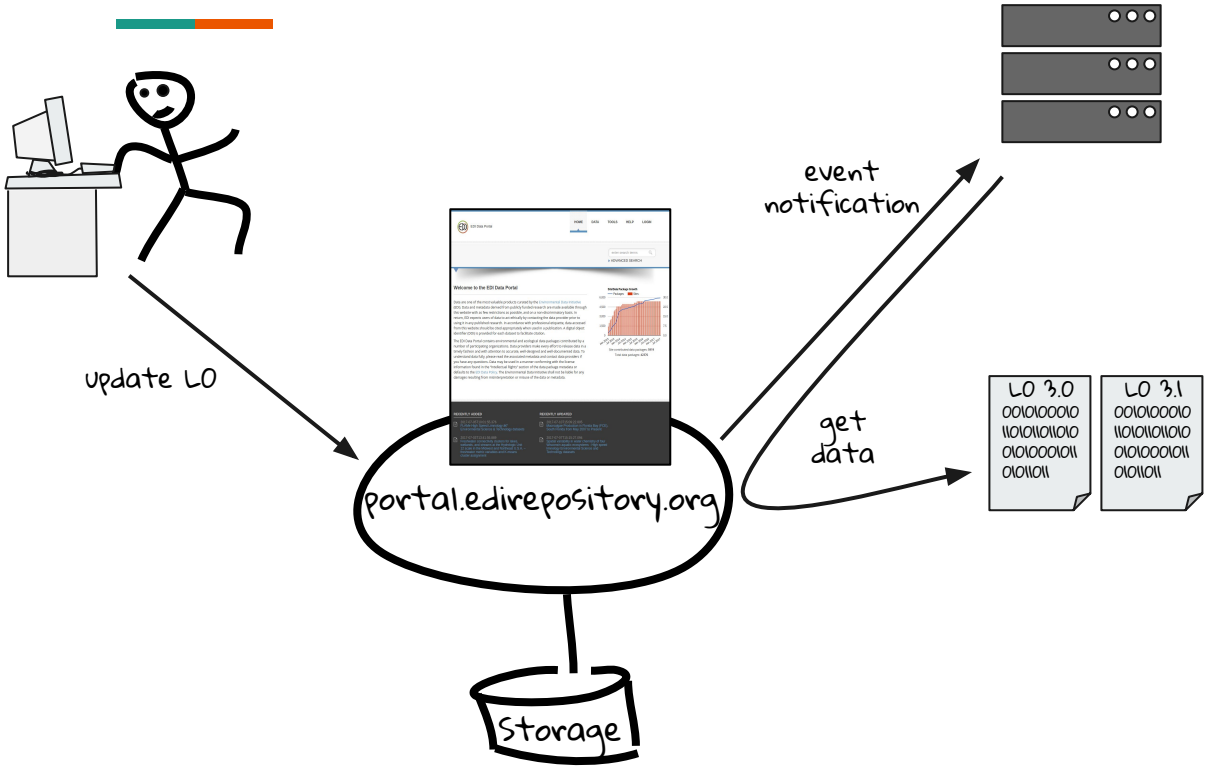
Automated Updates



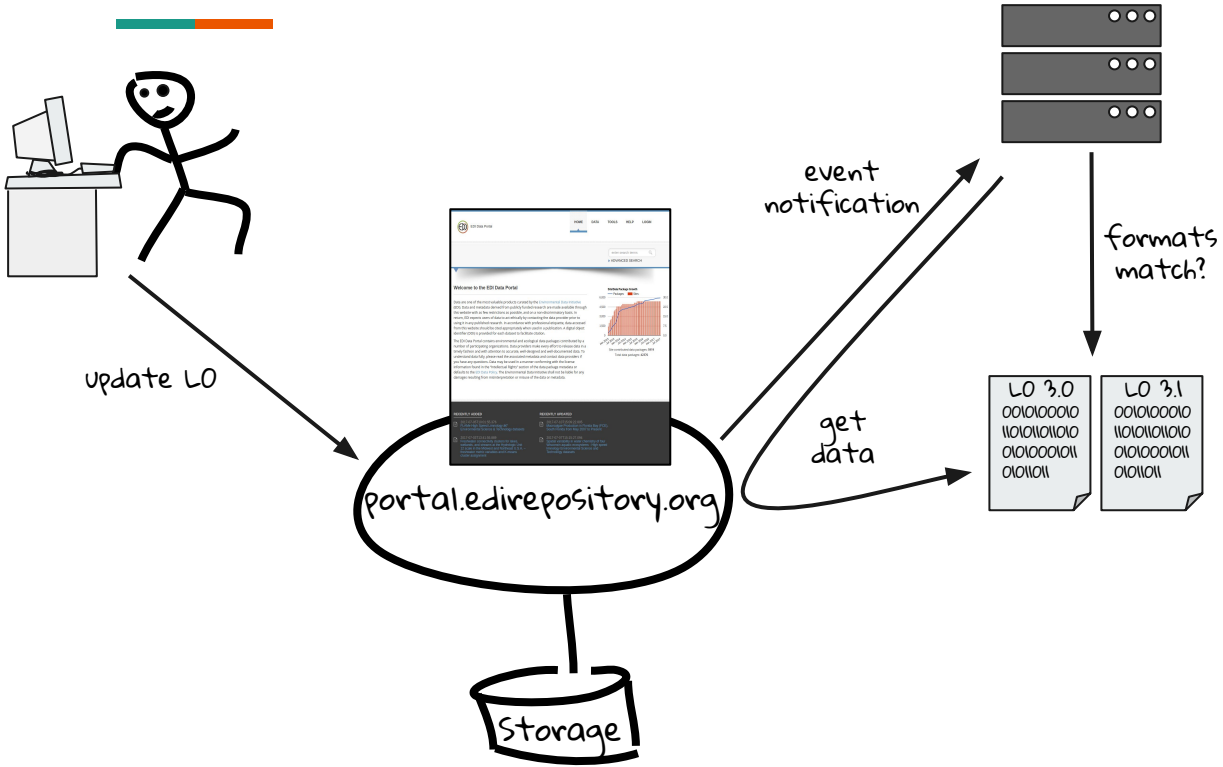
Automated Updates



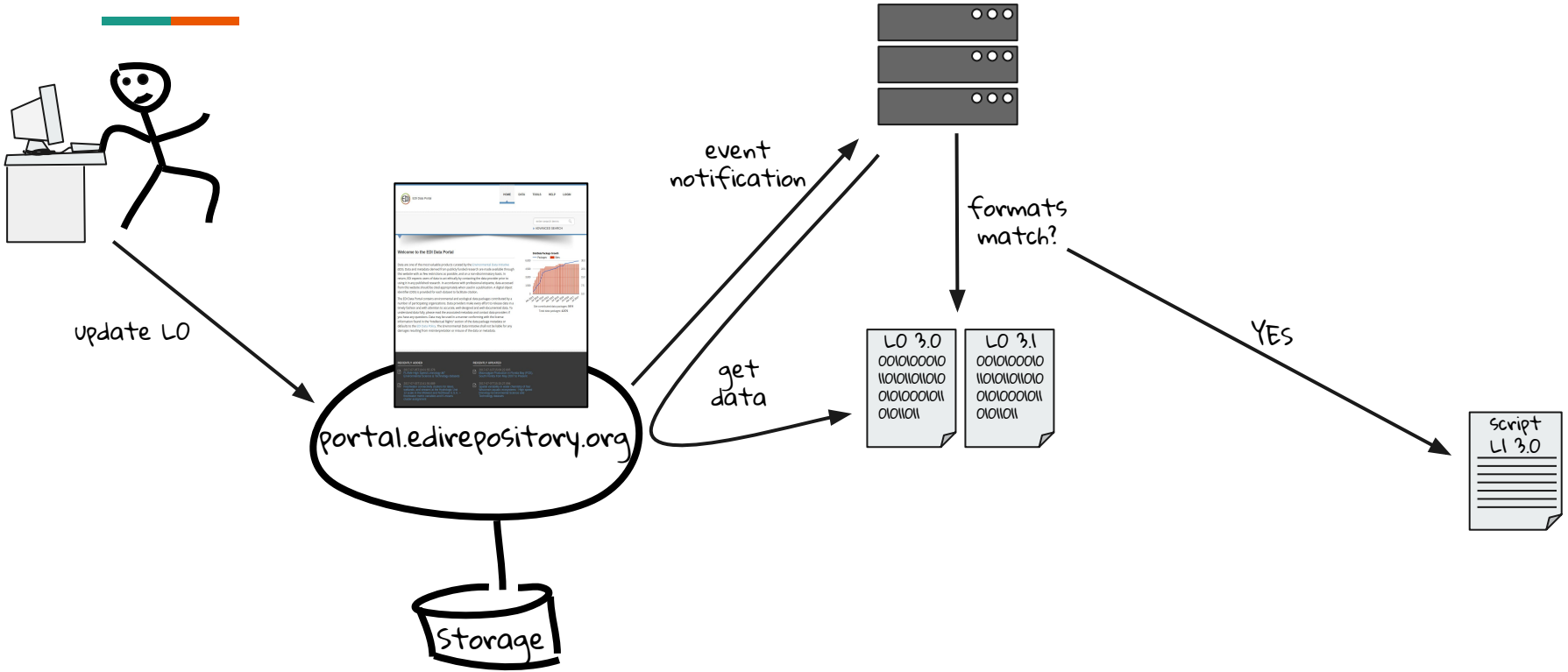
Automated Updates



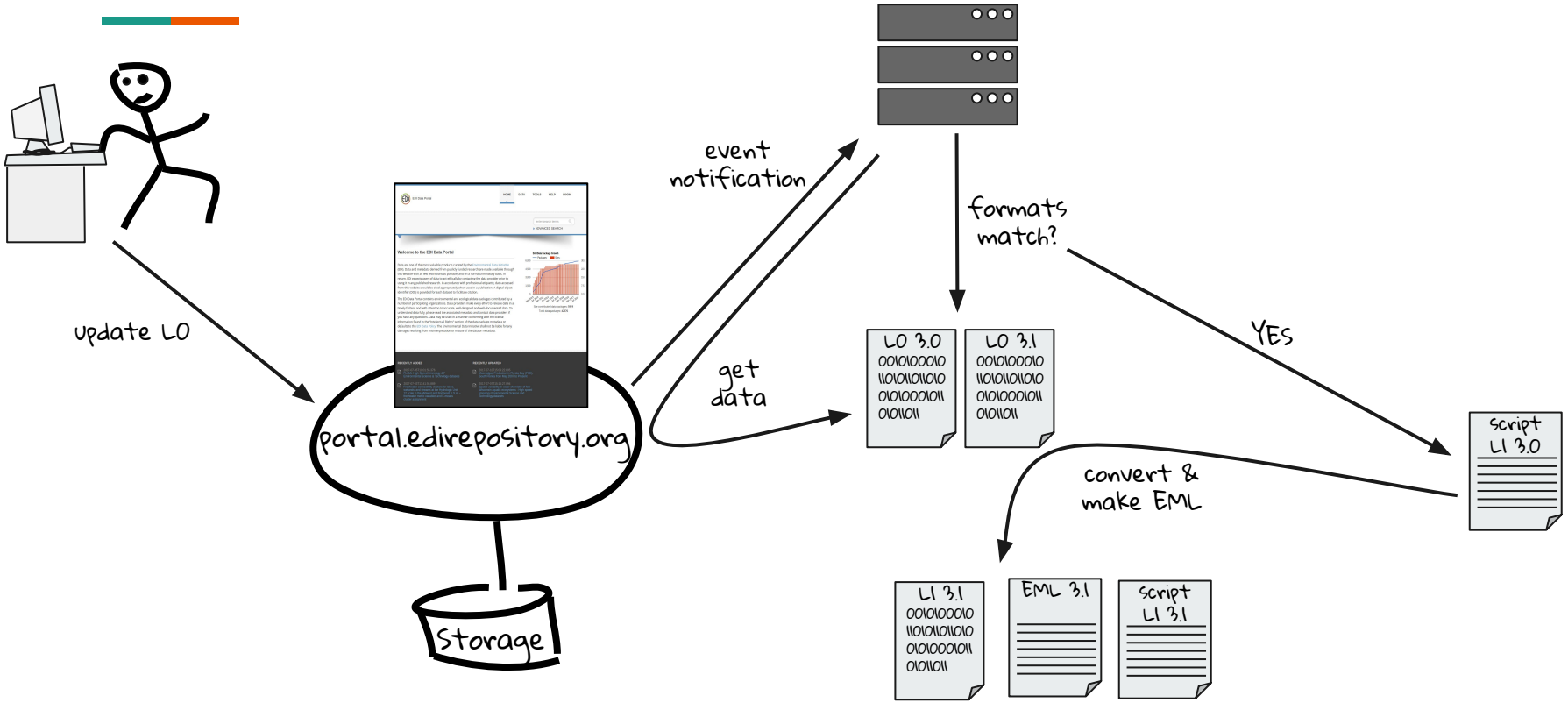
Automated Updates



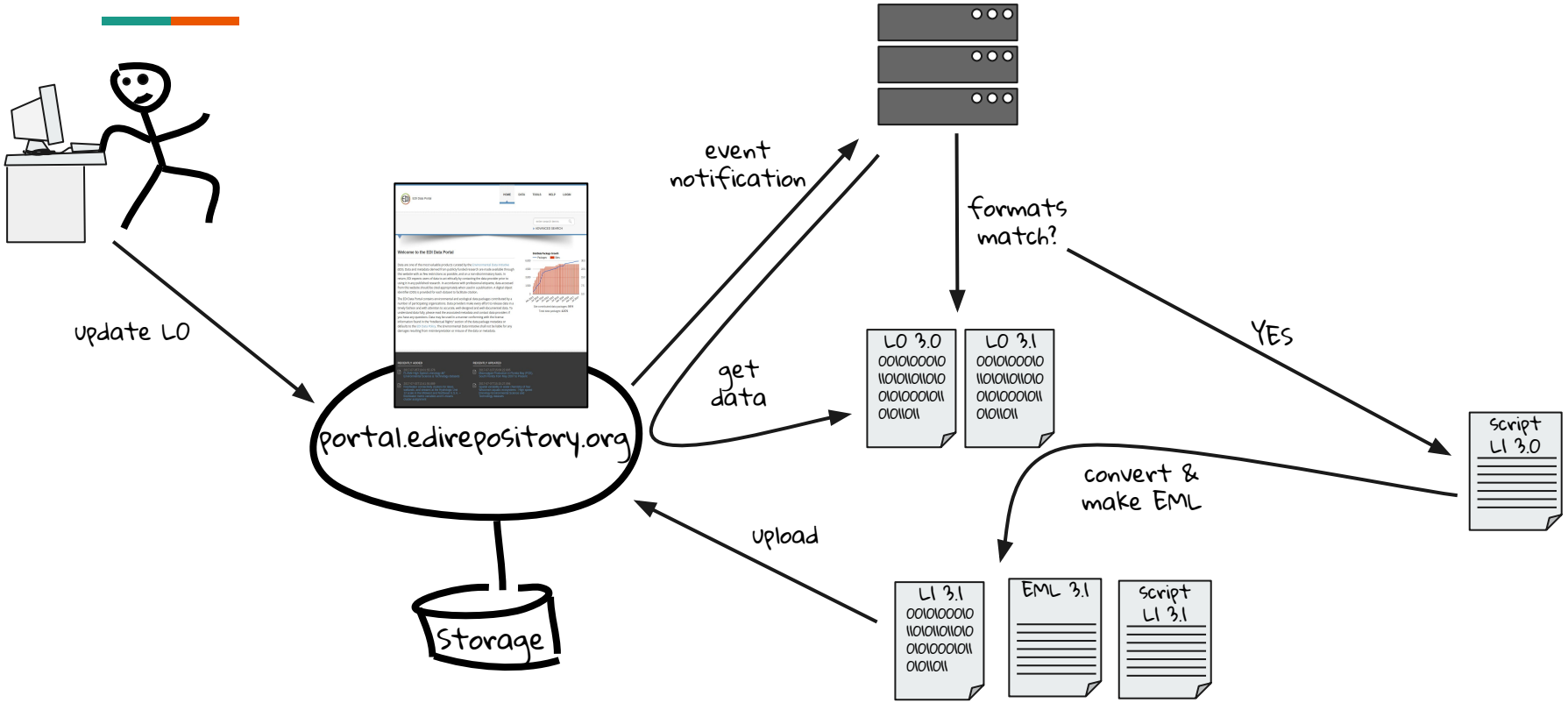
Automated Updates



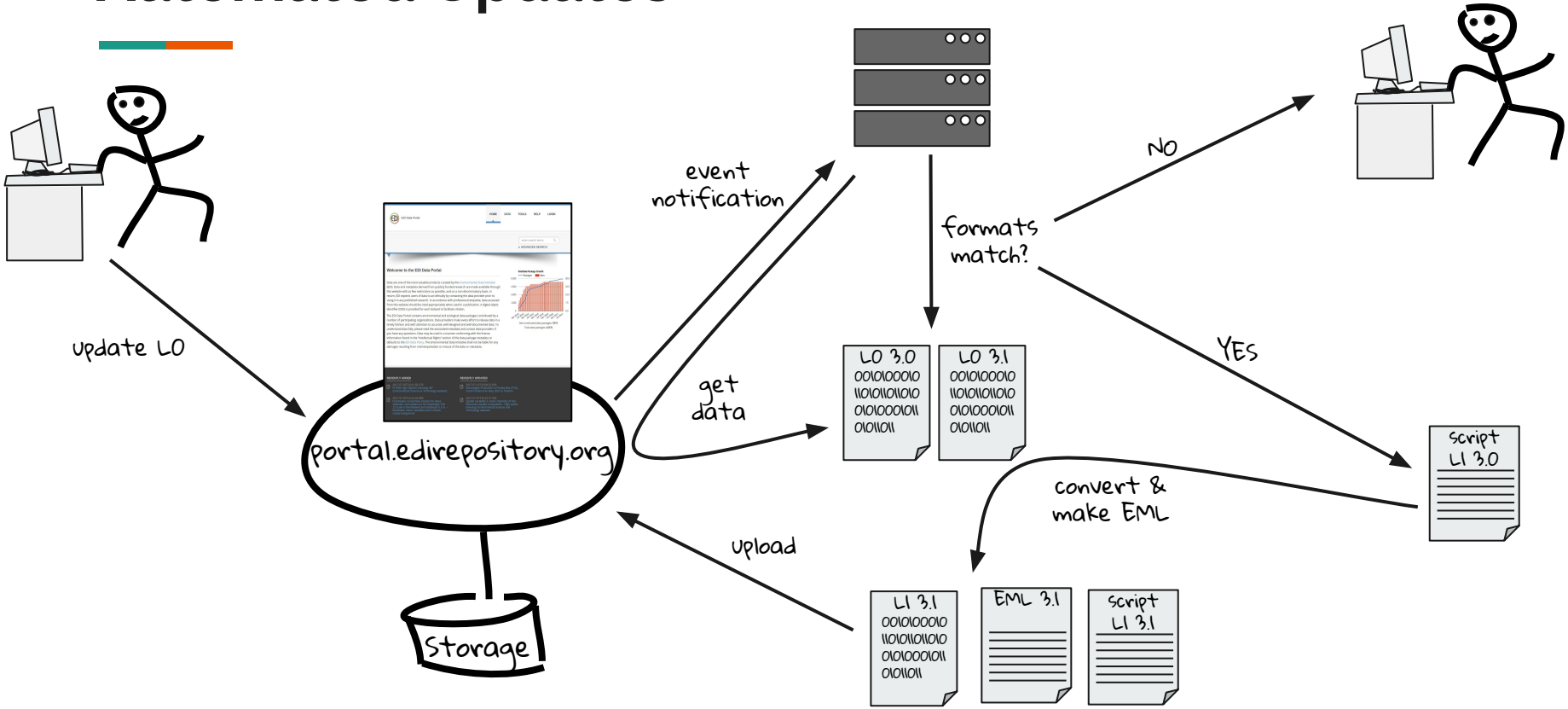
Automated Updates



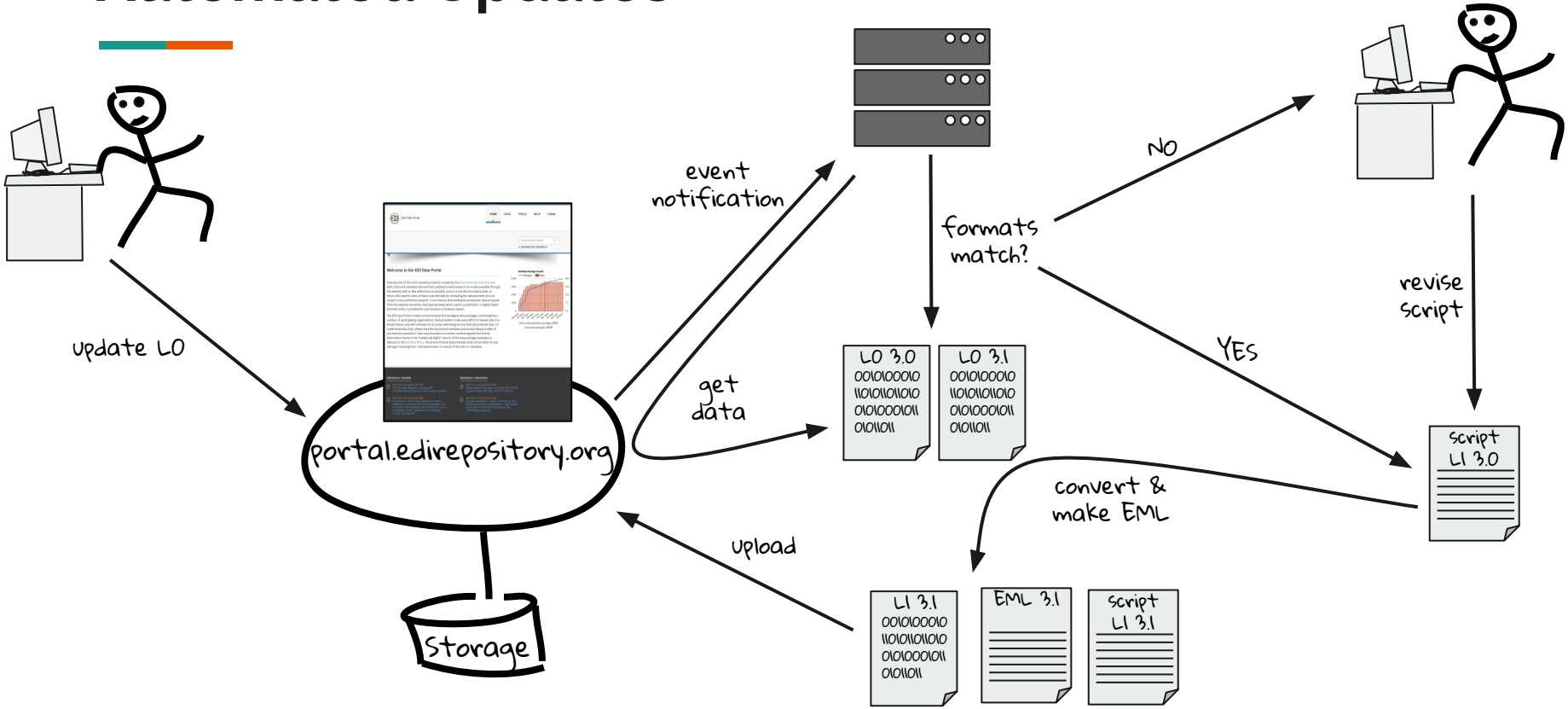
Automated Updates



Automated Updates



Automated Updates



Your input

Synthesis scientists:

- Would you use a common model instead of developing your own?
- Do you know of data we should add to the processing queue?
- Do you have synthesis projects in other scientific domains?

Data Contributors:

- Our conversions to L1 have identified important dataset features
- Can we inform you of these?
- How?

Important L0 Features

- Sampling site nesting can be understood
- Locations are complete (with latitude, longitude)
- Taxa can be resolved (e.g., species binomials)
- Work with EDI to build robust measurement vocabularies

Demo/discussion

Appendix, additional material

- Potential collaborators, follow-up projects
- Variables observed in L1 datasets to date
  - Observation table (few)
  - Ancillary tables (many)
- Taxonomic authorities observed in L1 datasets to date
- Misc references, incomplete

Potential Future Collaborations

Activity	**With**	**Preparation**	**Issues to resolve** [1]
Convert data from ecocomDP to other models	Popler GBIF Biotime	Stable source (ecocomDP)	> Revision management > Best use of features in destination-model
Structured vocabulary of variable descriptions	GBIF NCEAS ADC DataONE	Lists of expected measurements	> Integration with existing partial vocabularies and ontologies > Practices for contributions

Footnotes
[1] *Not included: funding issues*

Variables - Observation Table

Variable name	N	Unit	Unknown aspects
abundance	1	NA	Areal? Volumetric?
biomass	2	gram	Wet? Dry? Allometric? Single individual? A group?
count, number_of_plants, number_of_arthropods, number_size_class_*	14	NA, number	opportunities for QC here
CPUE	1	NA	Ratio of two measurements (catch, effort); QC steps
LOGCPUE	1	NA	Units of original measurements
relative abundance	1	NA	
cover_amount	1	NA	

Many aspects of "biological measurements" are not well described.

Data cannot be fully understood until nuances are described.

Variables - Ancillary Tables

Location

moose.cage
park_acreage
park_code
park_district
point_code
point_location
restored
treatment
urbanized
water

Taxon

behavior
biogeographic.affinity
Clade
Class
Coarse_Trophic
colony.size
common_name
feeding.preference
Fine_Trophic
Fish_length
hl
Kingdom
Lineage
nest.substrate
Order
Phylum
primary.habitat
rel
rll
secondary.habitat
seed.disperser
slavemaker.sp
source
Total_Length
Tribe

Observation

Accession_Number
air_temp_F
area
cloud
Cloud_Cover
Date
Diver
DO
End
Gear Type
height
Number of replicate samples
observer
pH
sample_subtype
Sea_State
Secchi Depth
Start
subproject
Surge
surveys_notes
surveys_observation_notes
Swell
Temperature
time_end
time_start
trap.num
trap.type
Visibility
wind
Wind_Velocity

Taxonomic Authorities - Taxon Table

Used to date	Coverage	Notes
ITIS		
Catalog of Life	> 100 expert taxonomic DBs	
WoRMS	Temperate marine	
GBIF Backbone Taxonomy		Aggregates several databases

For More Information

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