Converting community survey data packages into the ecocomDP data model format

Environmental Data Initiative (EDI)
2019

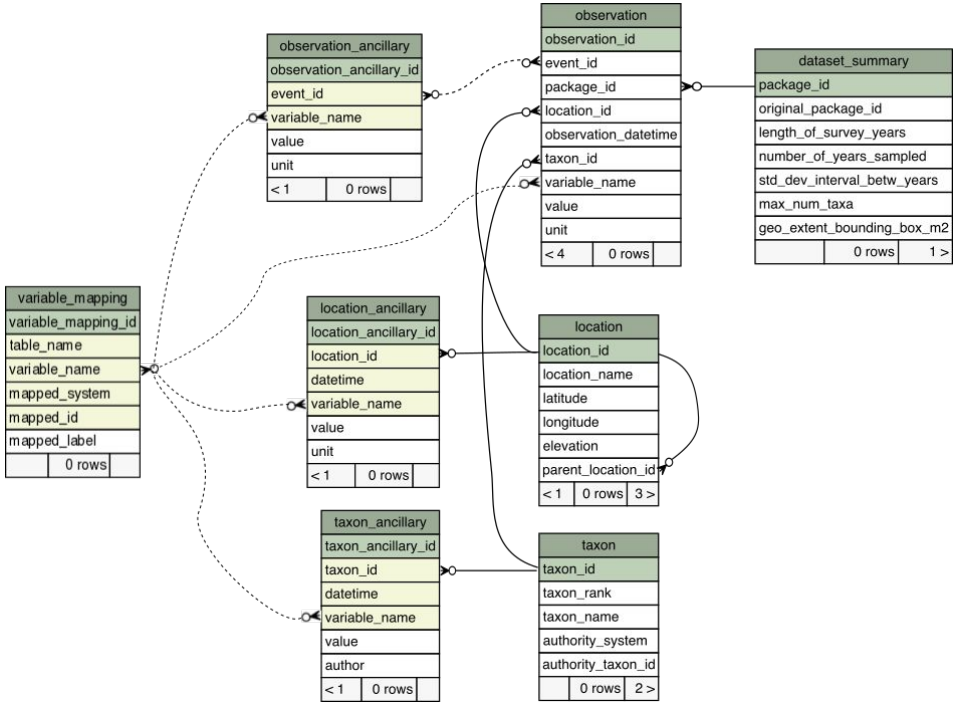
Agenda

Recap

Status

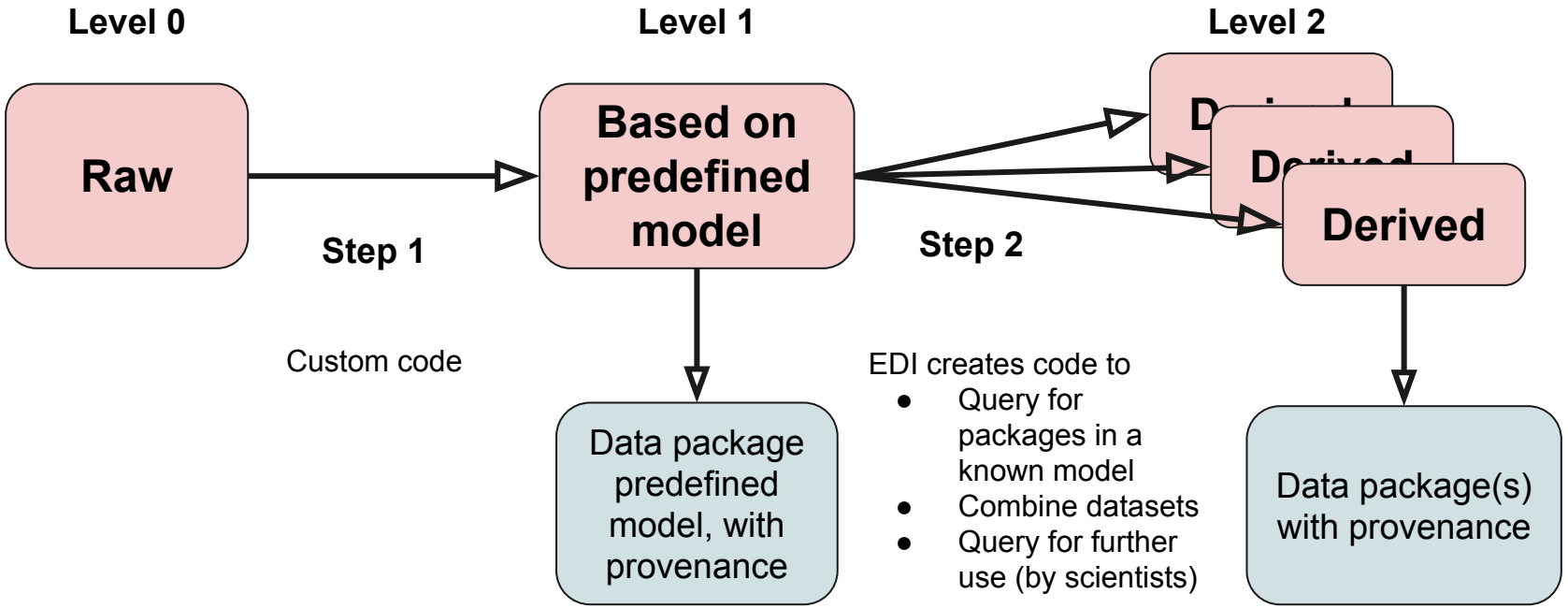
R-tools

Lessons learned

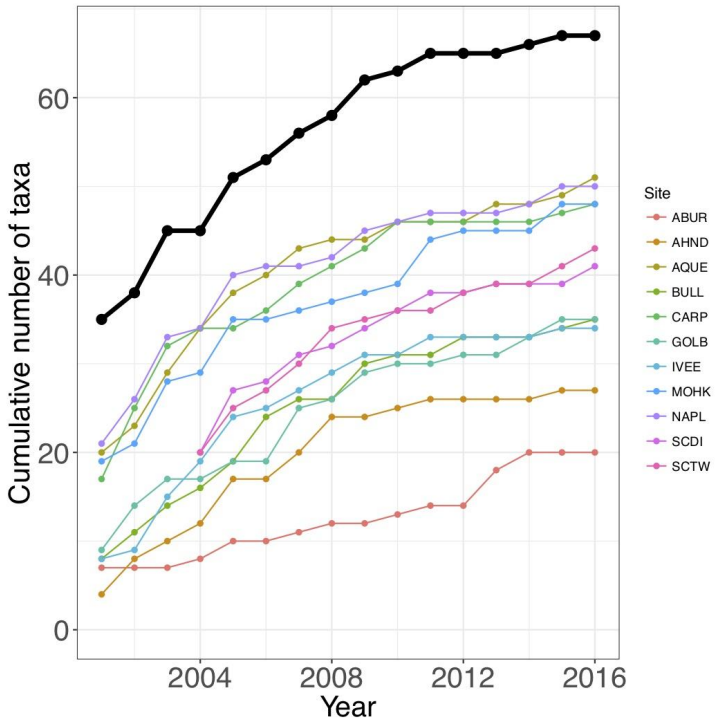


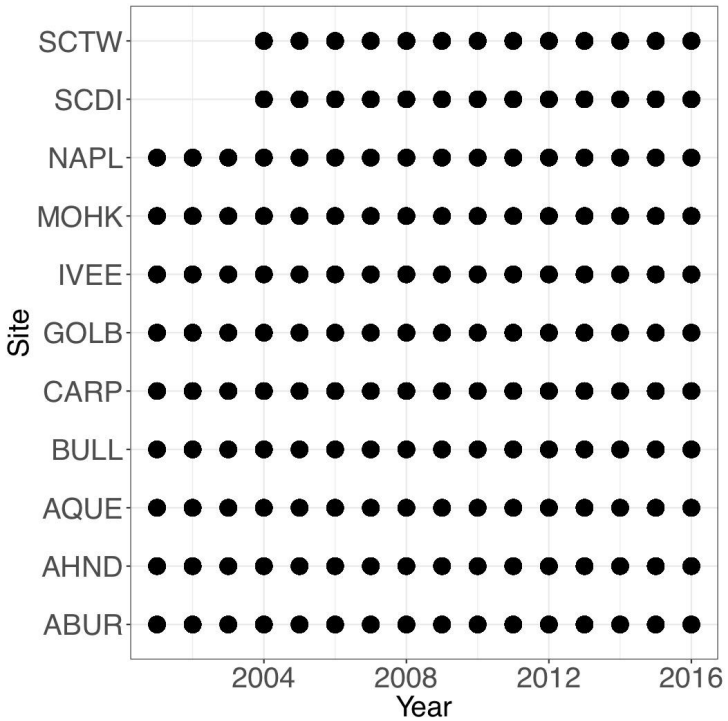
RECAP

Ideal Synthesis Workflow



Harmonized Format -> Harmonized Plots





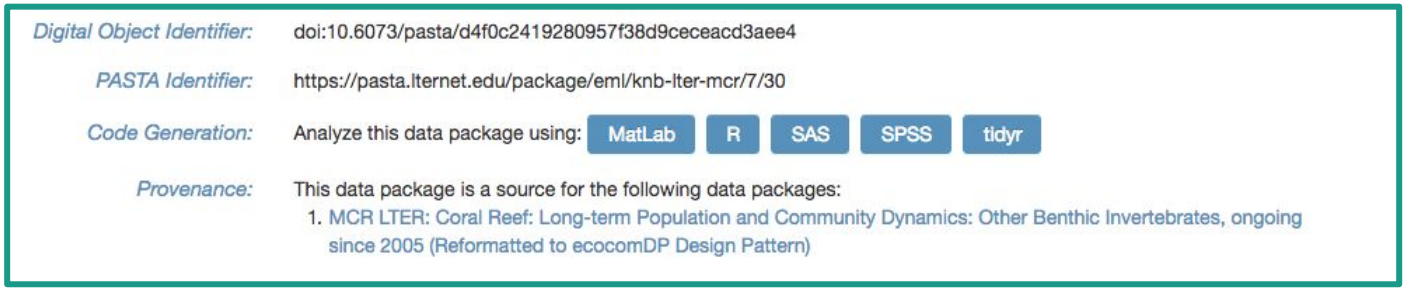
From: Lany et al, 2018. A reproducible workflow for synthesizing disparate LTER data (this meeting, Poster session)

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

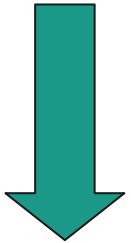
Provenance

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

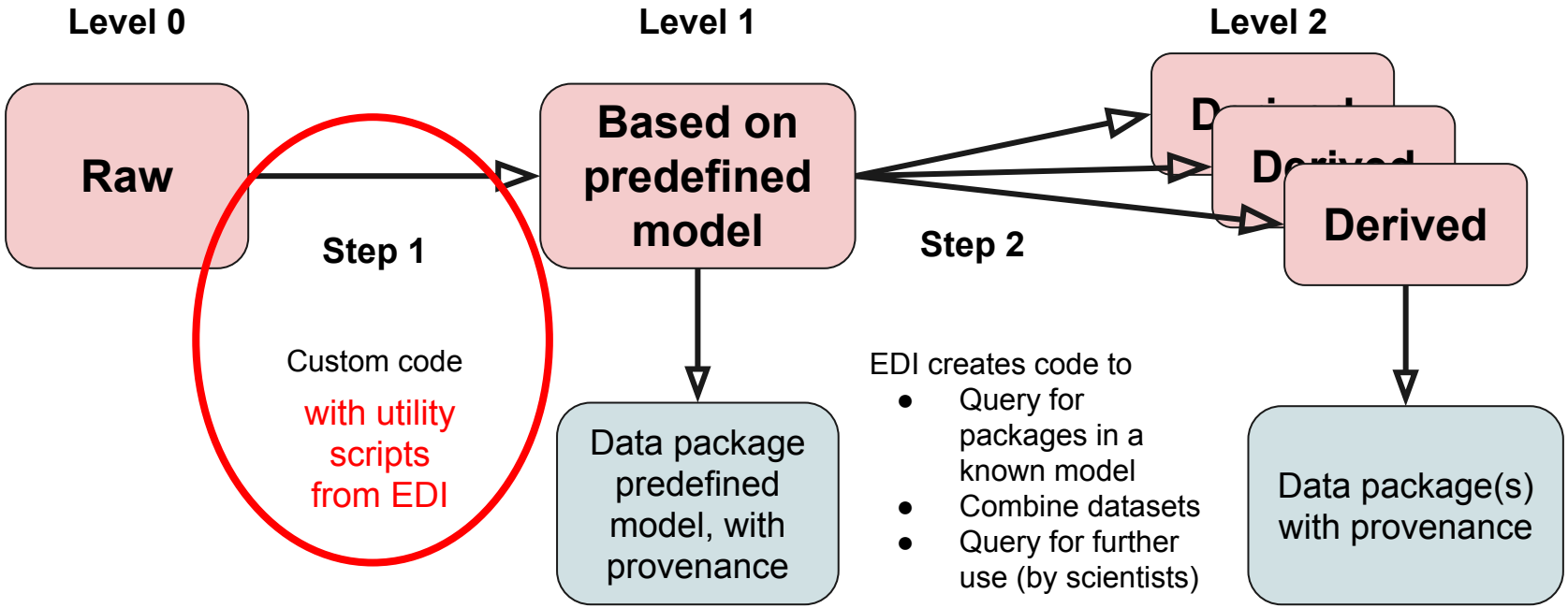
STATUS

Summary Metrics

Without NEON					NEON
	N	Min	Max	Median	
Number of datasets	28	-	-	-	1
Temporal coverage (years)	28	4	38	12	4
Temporal evenness (interval SD)	28	0	10.8	0.43	.93
Geographic coverage (km^2, > 0)	25	1368	3.9×10^8	9.9×10^5	NA
Taxonomic coverage (without OTUs)	27	1	1752	62	1066

R-TOOLS

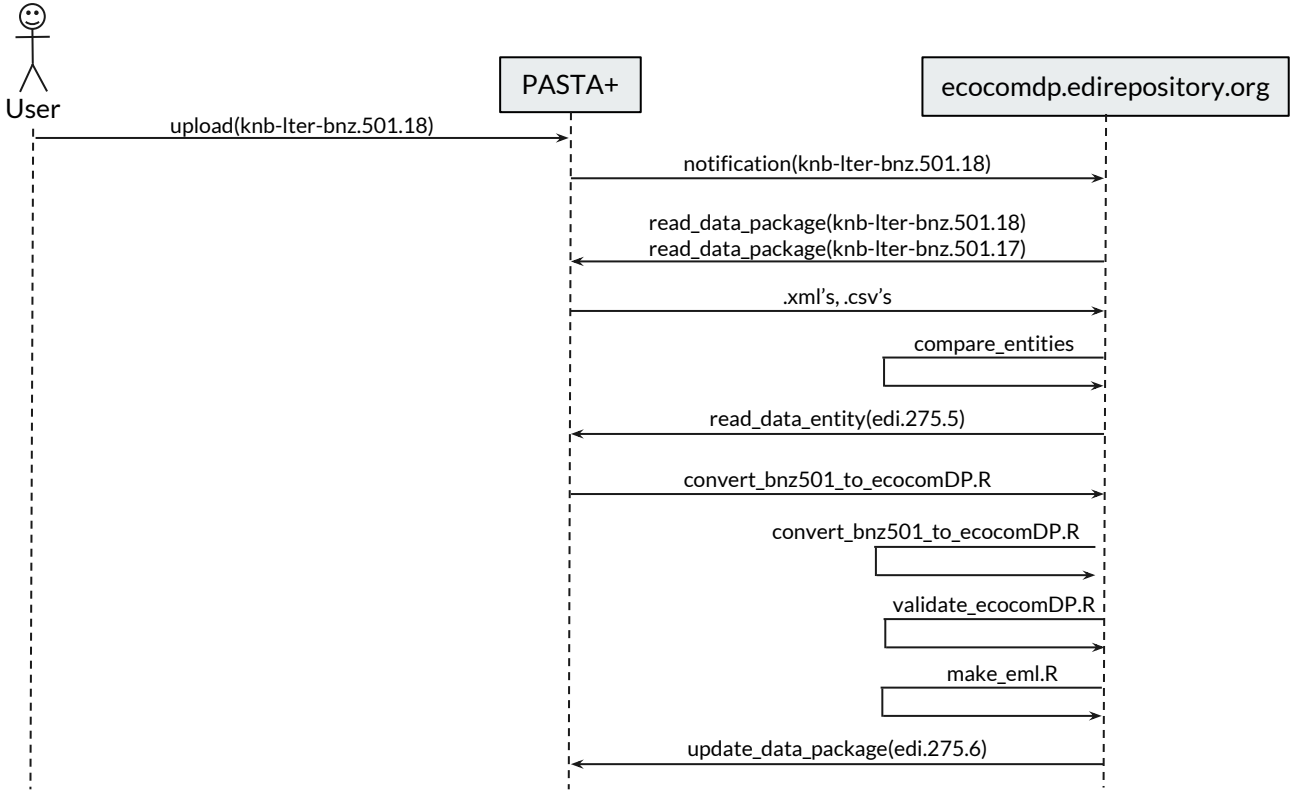
Ideal Synthesis Workflow



Utility Scripts - Dataset Conversion

https://github.com/EDIorg/ecocomDP

Maintenance



LESSONS LEARNED

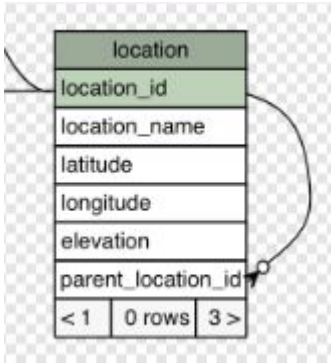
Important L0 Features - Locations

Locations are complete (with latitude, longitude)

- Best: digital lat/lon
  - https://portal.edirepository.org/nis/metadataviewer?packageid=edi.5.3
- OK (need processing):
  - In metadata only: https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-sbc.17.33
  - Deg-min-sec (strings)
  - Locations in second table
- Not usable: sites codes without lat/lon

Important L0 Features - Site Nesting

- Sampling site nesting can be understood:
  - Best: subsites labeled
    - https://portal.edirepository.org/nis/metadataviewer?packageid=edi.5.3
  - OK:
  - Not useable:



Important L0 Features - Taxa

- Taxa can be resolved
  - Best: Taxon codes assigned at source
    - https://portal.edirepository.org/nis/metadataviewer?packageid=edi.3.5
  - OK: species binomials
    - https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-sbc.17.33
  - Not useable: local codes only
    - https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-sbc.17.33
      (*if all they had included was the column called “sp_code”)

Important L0 Features - Variables

- Metadata can be matched to entity column
  - Best: attributeName exactly matches column header
    - https://portal.edirepository.org/nis/metadataviewer?packageid=edi.3.5
  - OK: can be matched by manual examination
    - https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-mcr.1039.9
  - Marginal: no header
    -

Important L0 Features - Date times

- Temporal sampling regime is consistent
  - Best: consistent dateTime format throughout
    - https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-mcr.6.56
  - OK: sampling regime changes over time (yyyy, vs yyyy-mm-dd)
    - YYYY, vs YYYY-MM-DD
  - Not useable: date and time columns are not typed in EML as dateTimes (i.e, typed as strings, as below)

10/8/10	15:25
10/28/10 - 10/29/10	22:00 - 6:00
10/26/10	9:34

Important Lo Features - Table linkages

- FK linkages
  - Best: EML constraint included, with referential integrity
    - https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-mcr.6.56
  - OK: FK detected manually, has referential integrity
    - url
  - Not Usable: FK detected manually, but no referential integrity
    - url

Questions for you

- What is best way to communicate issues of “usability”?
- Maintenance options - what are your preferences? E.g., for
  - EDI server space for conversion scripts vs your local
  - Repeatable workflows and event notifications

Still needed

- Work with EDI to build robust measurement vocabularies
- Recommended taxonomic authorities for your domain

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

Questions?