Thematic harmonization of environmental data: Facilitating interoperability of data within and among repositories in support of data reuse and scientific synthesis

Margaret O'Brien[1], Colin Smith[2], and Corinna Gries[2]

[1]University of California, Santa Barbara
[2]University of Wisconsin, Madison

November 22, 2022

Abstract

Data repositories and research networks worldwide are publishing a diverse array of long-term and experimental data for meaningful reuse, repurpose, and integration. However, in synthesis research the largest time investment is still in discovering, cleaning and combining primary datasets until all are completely understood and converted to a usable format. To accelerate this process, we have developed an approach to define flexible domain specific data models and convert primary data to these models using a light-weight and distributed workflow framework. The approach is based on extensive experience in synthesis research workflows, takes into account the distributed nature of original data curation, satisfies the requirement for regular additions to the original data, and is not determined by a single synthesis research question. Furthermore, all data describing the sampling context are preserved and the harmonization may be performed by data scientists that are not specialists in each specific research domain. Our harmonization process is 3-phased. First, a Design Phase captures essential attributes, considers already existing standardization efforts, and external vocabularies that disambiguate meaning. Second, an Implementation Phase publishes the data model and best practice guides for reference, followed by conversion of relevant repository contents by data managers, and creation of software for data discovery and exploration. Third, a Maintenance Phase implements programmatic workflows that run automatically when parent data are revisioned using event notification services.In this presentation we demonstrate the harmonization process for ecological community survey data and highlight the unique challenges and lessons learned. Additionally, we demonstrate the maintenance workflow and data exploration and aggregation tools that plug in to this data model

American Geophysical Union, Fall Meeting 2019 IN51F-0699

Data Harmonization Facilitates Interoperability and Reuse

Colin Smith[1], Margaret O'Brien[2], Corinna Gries[1]

[1]University of Wisconsin, Madison, [2]University of California, Santa Barbara

Abstract

Data repositories and research networks publish a diverse array of primary data for meaningful synthesis, integration and future reuse. However, in synthesis research the largest time investment is in cleaning and combining primary datasets until all are completely understood and converted to a usable format. To accelerate this process, EDI defines flexible domain-specific data models, and converts primary data to these models using a lightweight and distributed workflow framework.

Advantages of a Harmonization Workflow

- *Original data description and curation practices are maintained*
- *Workflow framework is repeatable (essential for ongoing datasets)*
- *Intermediate format is not determined by a single synthesis research question*
- *Most harmonization steps can be performed by non-specialists*

3-Phase Process

Design

- *Capture essential attributes from the community*
- *Consider existing standardization efforts*
- *Evaluate external vocabularies to disambiguate meaning.*

Participants in harmonization workshops. Above: community survey data (2017) Right: meteorology and hydrology data (2019)

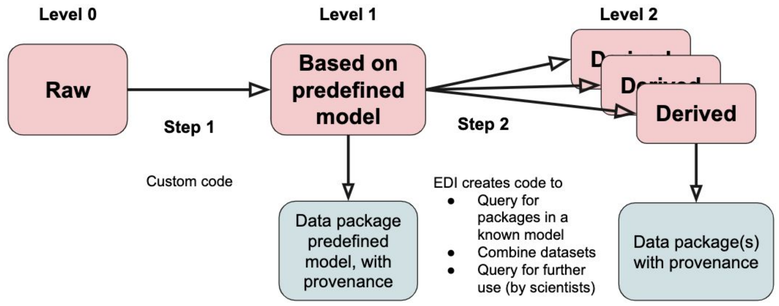
Implementation

- *Distribute the data model; convert relevant datasets*
- *Create*
  - *templates for building data packages*
  - *best practice guides for reference*
  - *software for discovery, exploration, denormalization*

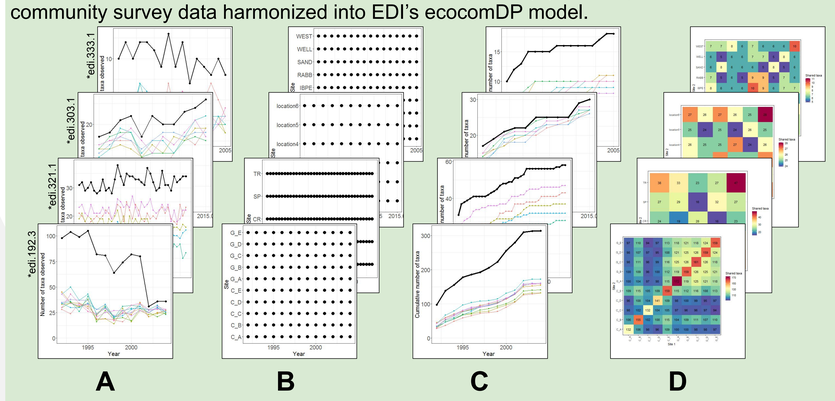
Maintenance

- *Build workflows to run when source data are updated*
- *Automate with event notification services*

Harmonization Framework



Harmonized data (Level 1) allows common visualization of many datasets to evaluate features such as number of taxa over time (A), spatio-temporal sampling effort (B), species accumulation curves (C) and species shared among sites (D). These plots are from community survey data harmonized into EDI's ecocomDP model.



*All datasets can be accessed with prefix "https://portal.edirepository.org/nis/mapbrowse?packageid="

Principles

Level 0:

- Well-described data, by those close to the research
- Repository allows access via API
- Metadata in EML, text-based tables

Level 1:

- Reformatted only (no aggregation or processing)
- Metadata in EML, text-based tables
- Archive in a repository aids discovery

Level 2:

- Researchers use harmonized inputs and code created for the harmonized model to speed their analyses
- Repository archive for Level 2 is recommended

Use cases

Research synthesis

Up to 80% of synthesis work can be related to cleaning, interpreting and reformatting input data. A harmonization workflow reduces that effort considerably.

Interoperability

A Level 1 data model facilitates integration.

- *NEON collaborates with EDI to include NEON biodiversity data in the discovery functionality of ecocomDP's R tools*
- *EDI is exploring integration with the Popler model (https://github.com/ropensci/popler)*

Targeted Applications

Harmonization is a necessary step in facilitating contributions to external applications, including

- *GBIF, for biodiversity data: https://www.gbif.org*
- *CUAHSI for hydrology data: https://www.cuahsi.org*

Results

Ecological community surveys

*Design and Implementation in 2017 & 2018. Today, thousands of records are available as **Level 1**, in EDI design pattern for ecological community data, "**ecocomDP**". 2019: Maintenance phase.*

Metrics of datasets converted to date

	non-NEON (N= 70 datasets)				NEON (N = 1)
	N	Min	Max	Median	
Temporal coverage (years)	70	4	70	14	4
Temporal evenness (interval SD)	69	0	10.8	0.05	.93
Geographic coverage (km^2, > 0)	70	1368	1.3×10^{14}	1.9×10^{8}	NA
Taxonomic coverage (without OTUs)	69	1	1752	48	1066

Meteorology & hydrology data

Design phase: 2018/2019

- *Working group recommends **CUAHSI ODM 1.1** model*

Implementation phase: 2019

- *Sample datasets converted*
- *Reference guides curated*

Maintenance phase: 2020 (planned)

- ***LTER & USFS** early adopters*

http://environmentaldatainitiative.org

This work is supported by the US National Science Foundation, grants 1565103, 162923, 1931174, 1931143