Going beyond the spreadsheet - developing Best Practices in 'long-tail' environmental data curation and publishing

Corinna Gries[1], Renée Brown[2], Mary Gastil-Buhl[3], Sarah Elmendorf[4], Hap Garritt[5], Mary Martin[6], Greg Maurer[7], An Nguyen[8], John Porter[9], and Timothy Whiteaker[10]

[1]University of Wisconsin Madison
[2]University of New Mexico Main Campus
[3]Moorea Coral Reef LTER
[4]University of Colorado Boulder
[5]Woods Hole Marine Biological Lab
[6]University of New Hampshire Main Campus
[7]New Mexico State University
[8]University of Texas at Austin
[9]University of Virginia
[10]CRWR

November 26, 2022

Abstract

The research data repository of the Environmental Data Initiative (EDI) is a signatory of the FAIR Data Principles. Building on over 30 years of data curation research and experience in the NSF-funded US Long-Term Ecological Research program (LTER), it provides mature functionalities, well established workflows, and support for 'long-tail' environmental data publication. High quality scientific metadata are enforced through automatic checks against community developed rules and the Ecological Metadata Language (EML) standard. Although the EDI repository is far along the continuum of making its data FAIR, representatives from EDI and the LTER Information Management community have recently been developing best practices for the edge cases in environmental data publishing. Here we discuss and seek feedback on how to best handle the publication of these 'long-tail' data when extensive additional data are available along with e.g., genomics data, physical specimens, or flux tower data. While these latter data are better handled in other discipline-specific repositories such as NCBI, iDigBio, and AmeriFlux, they are frequently associated with other data collected at the same time and location, or even from the same samples. This is particularly relevant across the LTER Network, where sites represent integrative research projects. Questions we address (and seek community input from) include: How to archive documents and images when they are data, e.g., field notebooks, or time-lapse photographs of plant phenology? How to deal with data from Unmanned Vehicle (e.g., drones and underwater gliders), acoustic data, or model outputs, which may be several terabytes in size? How should processing scripts or modeling code be associated with data? Overall, these best practices address issues of Findability and Accessibility of data as well as greater transparency of the research process.

Going Beyond the Spreadsheet

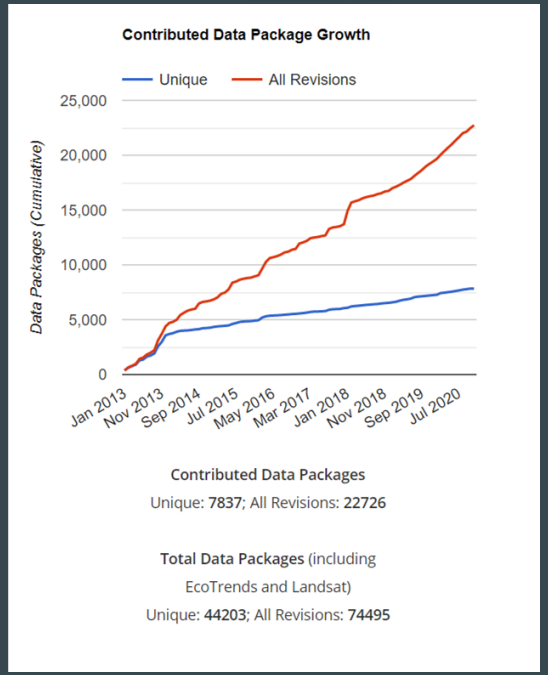
Developing Best Practices in 'long-tail' environmental data curation and publishing

C. Gries, R.F. Brown, G.Gastil-Buhl, S. Elmendorf, H. Garritt,
M. Martin, G. Maurer, A. Nguyen, J.H. Porter, T. Whiteaker

Introduction

Environmental Data Initiative repository

- FAIR data repository
- Ecological Metadata Language
- Data package: data table(s), metadata
- Metadata generation tools

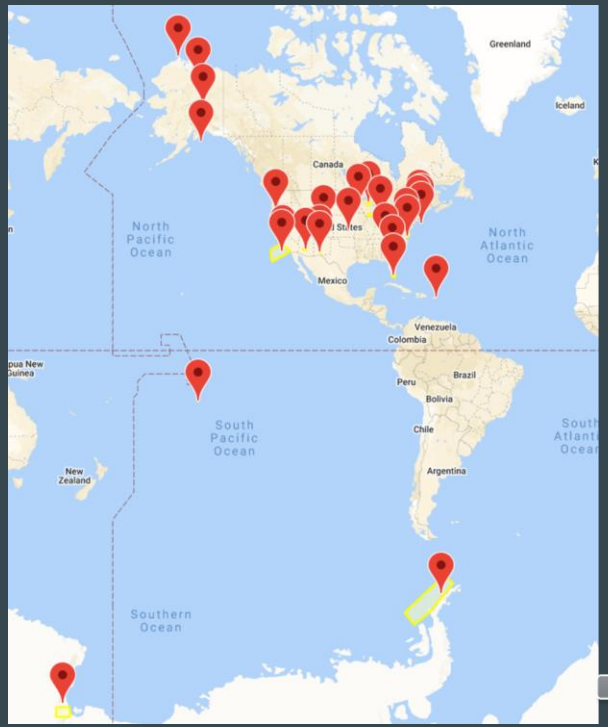


Introduction

Long-Term Ecological Research network

40 years of data management

Professional data managers at each site



Archiving other data types

Modeled Data

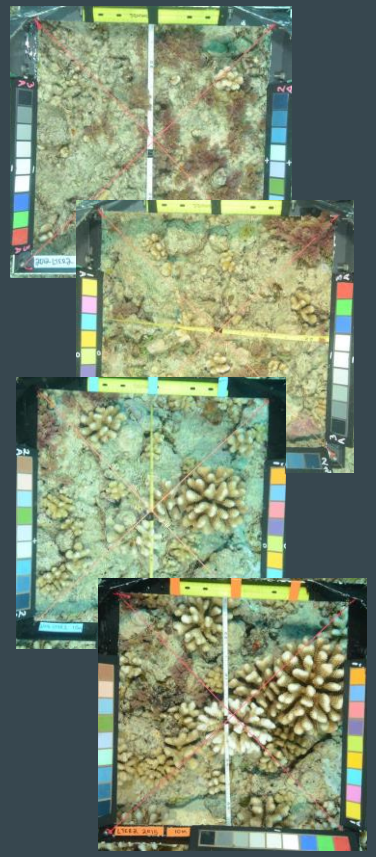
Code and Scripts

Data in More than one Repository

Images and Documents as Data

Spatial Data

Data from Small, Moving Platforms



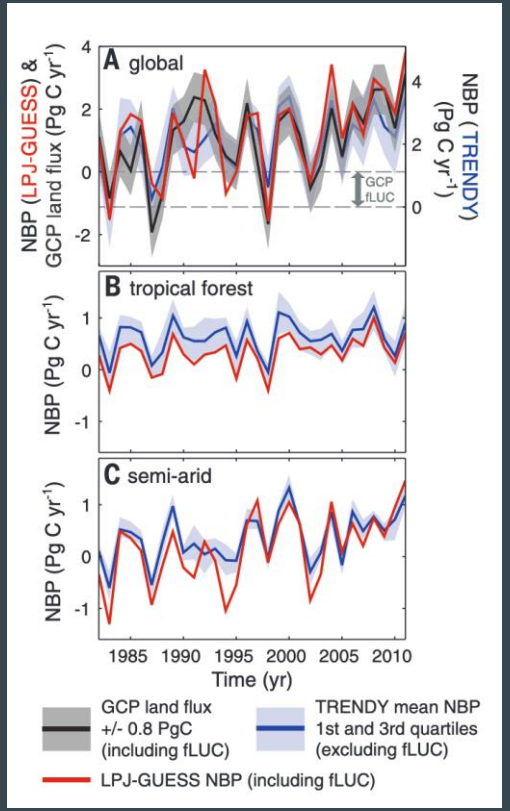
Modeled Data

Transparency / Reproducibility

Parameter settings, input data, output data

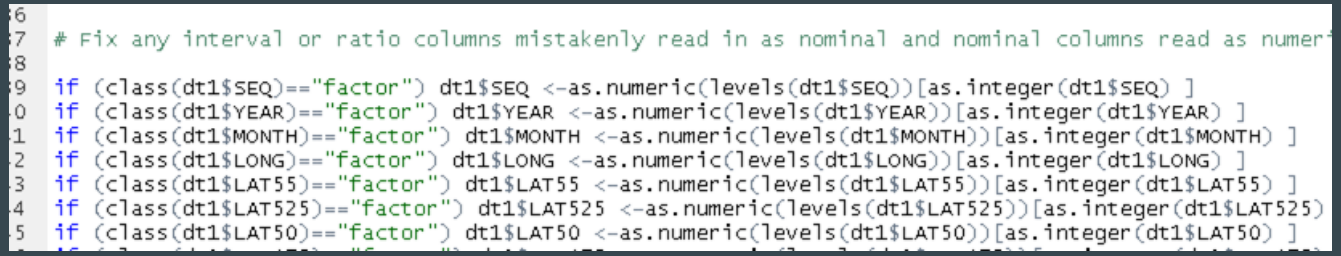
Model code

Size



From Ahstrom et al 2015

Code and Scripts



```
# Fix any interval or ratio columns mistakenly read in as nominal and nominal columns read as numeri

if (class(dt1$SEQ)=="factor") dt1$SEQ <-as.numeric(levels(dt1$SEQ))[as.integer(dt1$SEQ) ]
if (class(dt1$YEAR)=="factor") dt1$YEAR <-as.numeric(levels(dt1$YEAR))[as.integer(dt1$YEAR) ]
if (class(dt1$MONTH)=="factor") dt1$MONTH <-as.numeric(levels(dt1$MONTH))[as.integer(dt1$MONTH) ]
if (class(dt1$LONG)=="factor") dt1$LONG <-as.numeric(levels(dt1$LONG))[as.integer(dt1$LONG) ]
if (class(dt1$LAT55)=="factor") dt1$LAT55 <-as.numeric(levels(dt1$LAT55))[as.integer(dt1$LAT55) ]
if (class(dt1$LAT525)=="factor") dt1$LAT525 <-as.numeric(levels(dt1$LAT525))[as.integer(dt1$LAT525)
if (class(dt1$LAT50)=="factor") dt1$LAT50 <-as.numeric(levels(dt1$LAT50))[as.integer(dt1$LAT50) ]
```

Model code, data manipulation scripts (not binary software packages)

In EDI / code repository

Inside data package / referenced

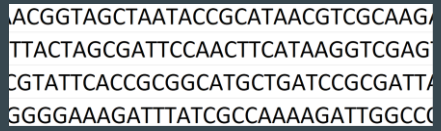
CodeMeta file

Data in More than one Repository

Related datasets managed in different repositories

E.g., GenBank, FluxNet, IDigBio

Data inventory table (accession #)



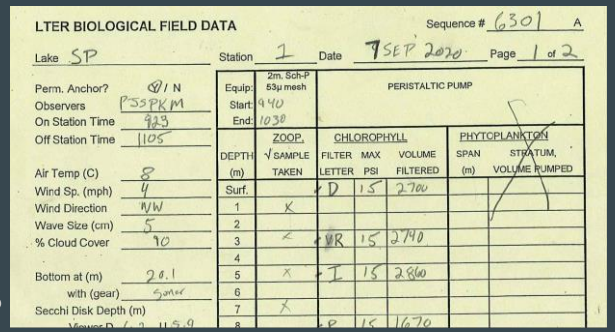
Images and Documents as Data

Images, digital field sheets, lab notebooks, protocols

Reanalysis, transparency, documentation

Design data packages for usability

Data inventory table


LTER BIOLOGICAL FIELD DATA

Sequence # 6301 A

Lake SP | Station 1 | Date 7 SEP 2020 | Page 1 of 2

Perm. Anchor? Ø / N
Observers PSSPKM
On Station Time 923
Off Station Time 1105

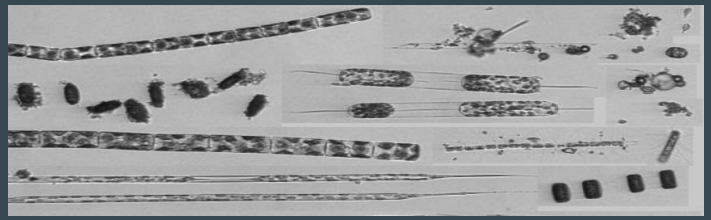
Air Temp (C) 8
Wind Sp. (mph) 4
Wind Direction NW
Wave Size (cm) 5
% Cloud Cover 90

Bottom at (m) 20.1
with (gear) Sonar
Secchi Disk Depth (m)

Equip: 2m, Sch-P 53µ mesh
Start: 940
End: 1038

PERISTALTIC PUMP

DEPTH (m)	ZOOP. √ SAMPLE TAKEN	CHLOROPHYLL FILTER LETTER	MAX PSI	VOLUME FILTERED	PHYTOPLANKTON SPAN (m)	STRATUM, VOLUME PUMPED
Surf.		D	15	2700		
1	X					
2						
3		R	15	2740		
4						
5	X	I	15	2860		
6						
7	X					
8		P	15	1670		

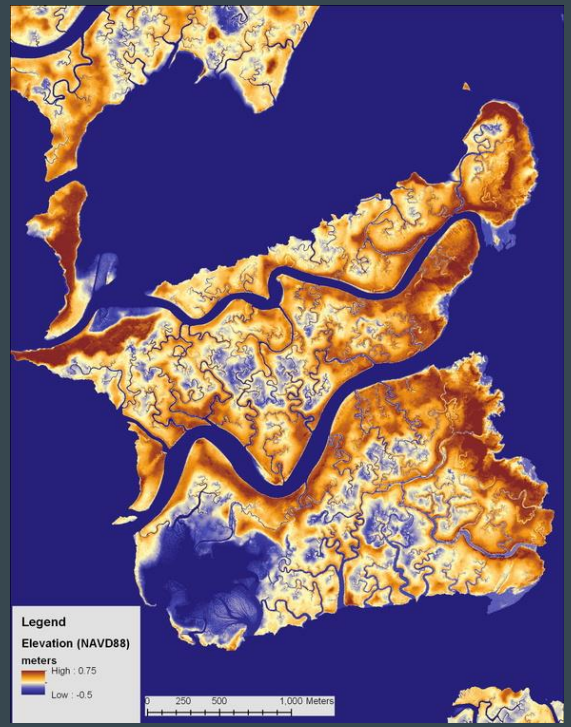


Spatial Data

Widely readable format (shapefile, geoTIFF)

Include metadata files in other formats

EML specific recommendations



Metadata Geography Table

Road GIS Layer for the Eastern Shore of Virginia 2013

Type Shapefile

Tags transportation, Roads, Street Centerline

Summary

The TIGER/Line shapefiles and related database files (.dbf) are an extract of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database (MTDB). The MTDB represents a seamless national file with no overlaps or gaps between parts, however, each TIGER/Line shapefile is designed to stand alone as an independent data set, or they can be combined to cover the entire nation. The All Roads Shapefile includes all features within the MTDB Super Class "Road/Path Features" distinguished where the MAF/TIGER Feature Classification Code (MTFCC) for the feature in MTDB that begins with "S". This includes all primary, secondary, local neighborhood, and rural roads, city streets, vehicular trails (4wd), ramps,

Data from Small, Moving Platforms

Sensors mounted on uncrewed aerial or underwater systems

Data Inventory Table

Publish raw data, processing code, derived data

What's new

Data inventory table

with additional metadata and accession numbers/GUIDs

Publish raw data and processing code

File with metadata in other formats

as part of the data package