The Environmental Data Initiative (EDI)

Supporting Curation and Archiving of Environmental Data

EDI is funded by the NSF DEB

Overview

EDI’s mission and approaches - Corinna Gries

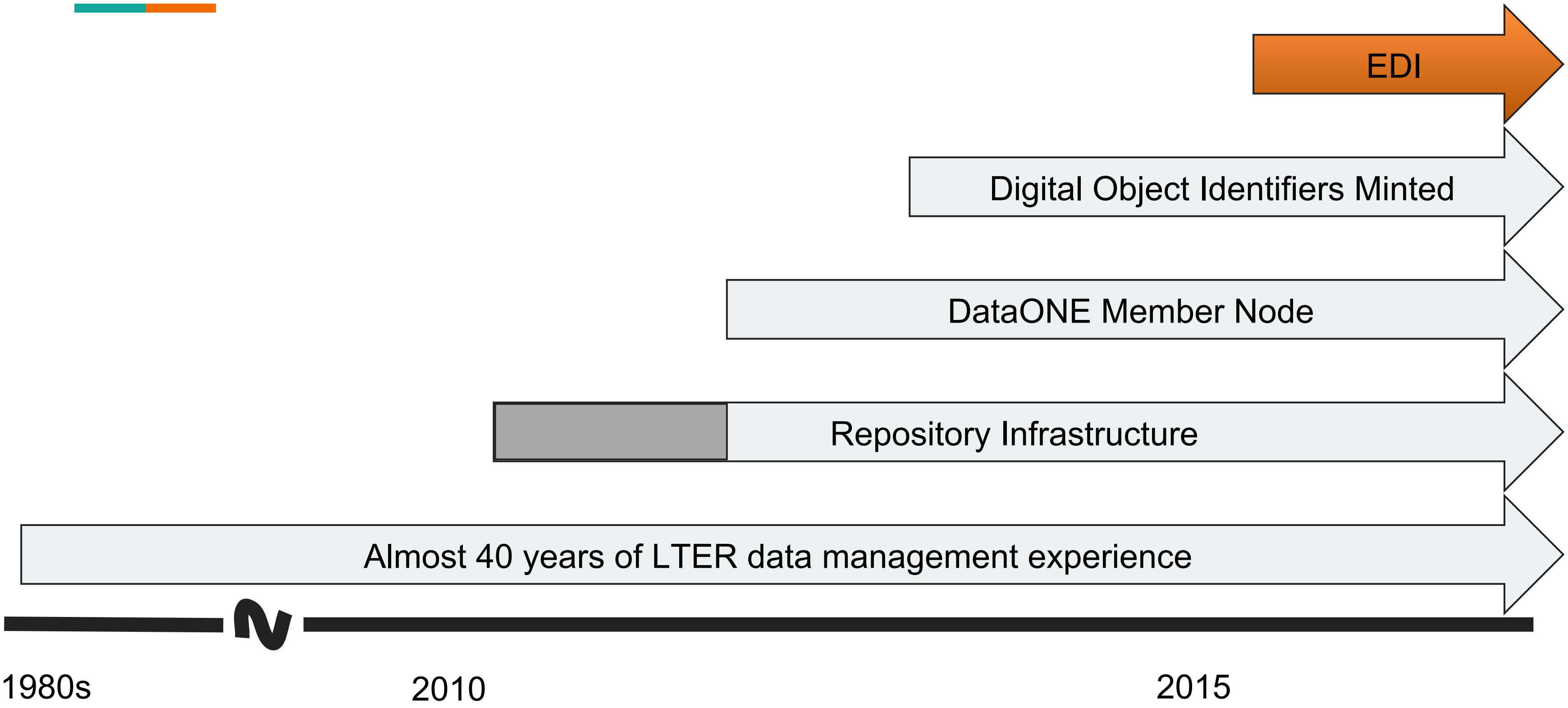
Data repository and publishing - Duane Costa

Data publication workflow support - Colin Smith

Outreach and training - Kristin Vanderbilt

Data integration support - Margaret O’Brien

History - we are standing on the shoulders of giants


EDI
Digital Object Identifiers Minted
DataONE Member Node
Repository Infrastructure
Almost 40 years of LTER data management experience
1980s
2010
2015

Mission and Goals

Accelerate the curation, archiving, and dissemination of environmental data

Ensure that environmental data are:

- Deposited into a data repository for long-term preservation and data integrity
- Easily discoverable and seamlessly accessible
- Documented with rich science metadata to enable reuse and integration

Curation

Data curation support for data providers

- Experienced data managers on staff
- Consultation
- Training
- Data Curation Tool Development

Archiving

EDI’s repository builds on cyber infrastructure developed for LTER

Certified and registered as trustworthy repository (re3data, Nature, ESA, others)

Data are documented in Ecological Metadata Language standard

High quality standards enforced through automated congruence and completeness checking

Dissemination

DataONE member node

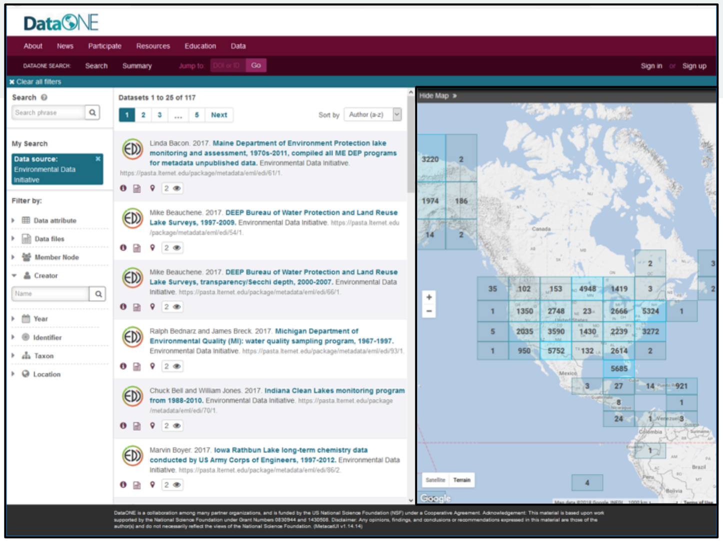
Local search interface

Digital Object Identifier (DOI) through DataCite

Collaborations with FAIR project, journal publishers

Linking of publications and data sets

Documentation of data provenance



Datasets 1 to 25 of 117

Linda Bacon. 2017. **Maine Department of Environment Protection lake monitoring and assessment, 1970s-2011, compiled all ME DEP programs for metadata unpublished data.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/61/1.

Mike Beauchene. 2017. **DEEP Bureau of Water Protection and Land Reuse Lake Surveys, 1997-2009.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/54/1.

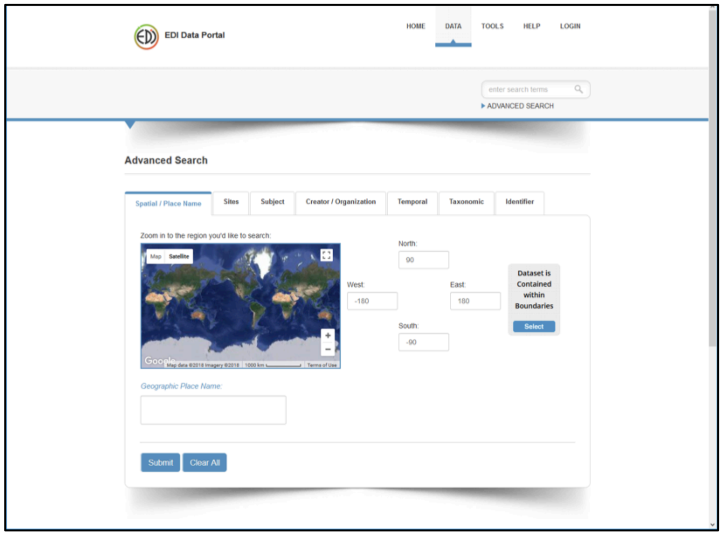
Mike Beauchene. 2017. **DEEP Bureau of Water Protection and Land Reuse Lake Surveys, transparency/Secchi depth, 2000-2007.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/66/1.

Ralph Bednarz and James Breck. 2017. **Michigan Department of Environmental Quality (MI): water quality sampling program, 1967-1997.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/93/1.

Chuck Bell and William Jones. 2017. **Indiana Clean Lakes monitoring program from 1988-2010.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/70/1.

Marvin Boyer. 2017. **Iowa Rathbun Lake long-term chemistry data conducted by US Army Corps of Engineers, 1997-2012.** Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/86/2.

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. (MetacatUI v1.14.14)



Advanced Search

Zoom in to the region you'd like to search:

North: 90

West: -180

East: 180

South: -90

Dataset is Contained within Boundaries

Geographic Place Name:

Overview

EDI’s mission and approaches - Corinna Gries

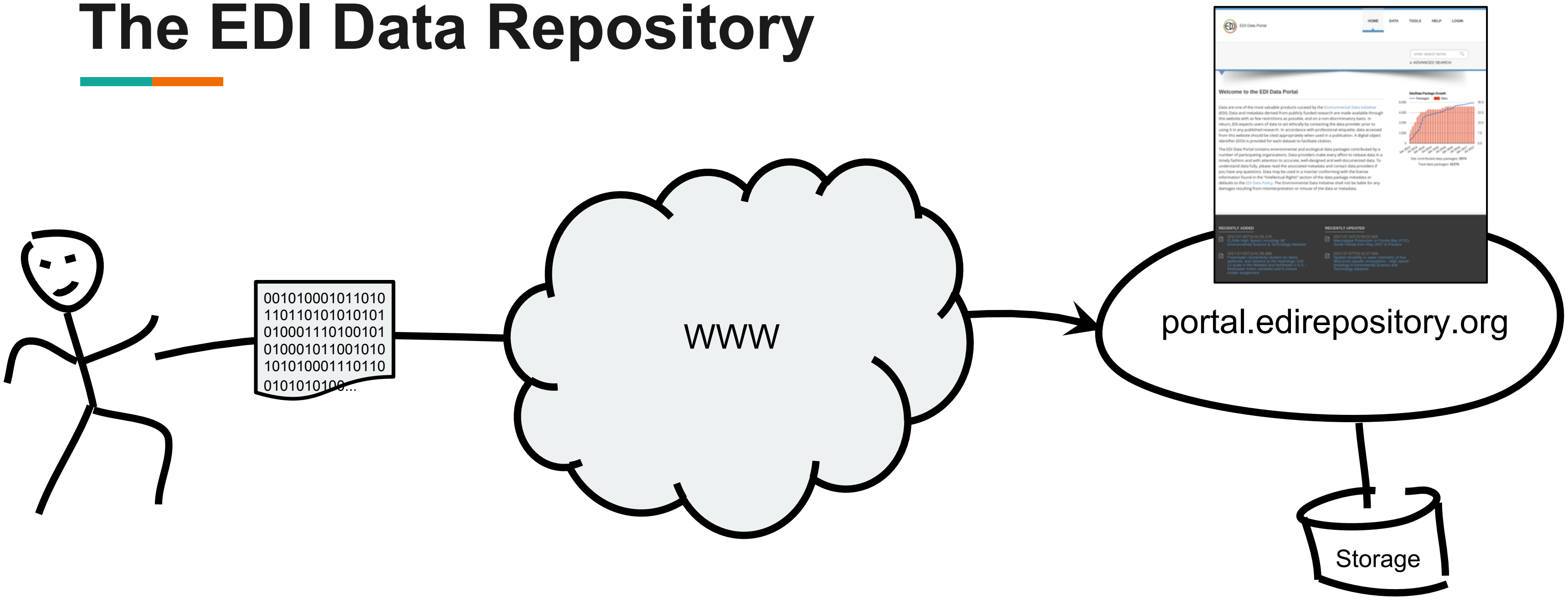
Data repository and publishing - Duane Costa

Data publication workflow support - Colin Smith

Outreach and training - Kristin Vanderbilt

Data integration support - Margaret O’Brien

The EDI Data Repository



Welcome to the EDI Data Portal

Data are one of the most valuable products curated by the Environmental Data Initiative (EDI). Data and metadata derived from publicly funded research are made available through this website with as few restrictions as possible, and on a non-discriminatory basis. In return, EDI expects users of data to act ethically by contacting the data provider prior to using it in any published research. In accordance with professional etiquette, data accessed from this website should be cited appropriately when used in a publication. A digital object identifier (DOI) is provided for each dataset to facilitate citation.

The EDI Data Portal contains environmental and ecological data packages contributed by a number of participating organizations. Data providers make every effort to release data in a timely fashion and with attention to accurate, well-designed and well-documented data. To understand data fully, please read the associated metadata and contact data providers if you have any questions. Data may be used in a manner conforming with the license information found in the "Intellectual Rights" section of the data package metadata or defaults to the EDI Data Policy. The Environmental Data Initiative shall not be liable for any damages resulting from misinterpretation or misuse of the data or metadata.

Site contributed data packages: 5919
Total data packages: 42376

RECENTLY ADDED

2017-07-05T10:01:55.376
FLAMe High Speed Limnology â€" Environmental Science & Technology datasets

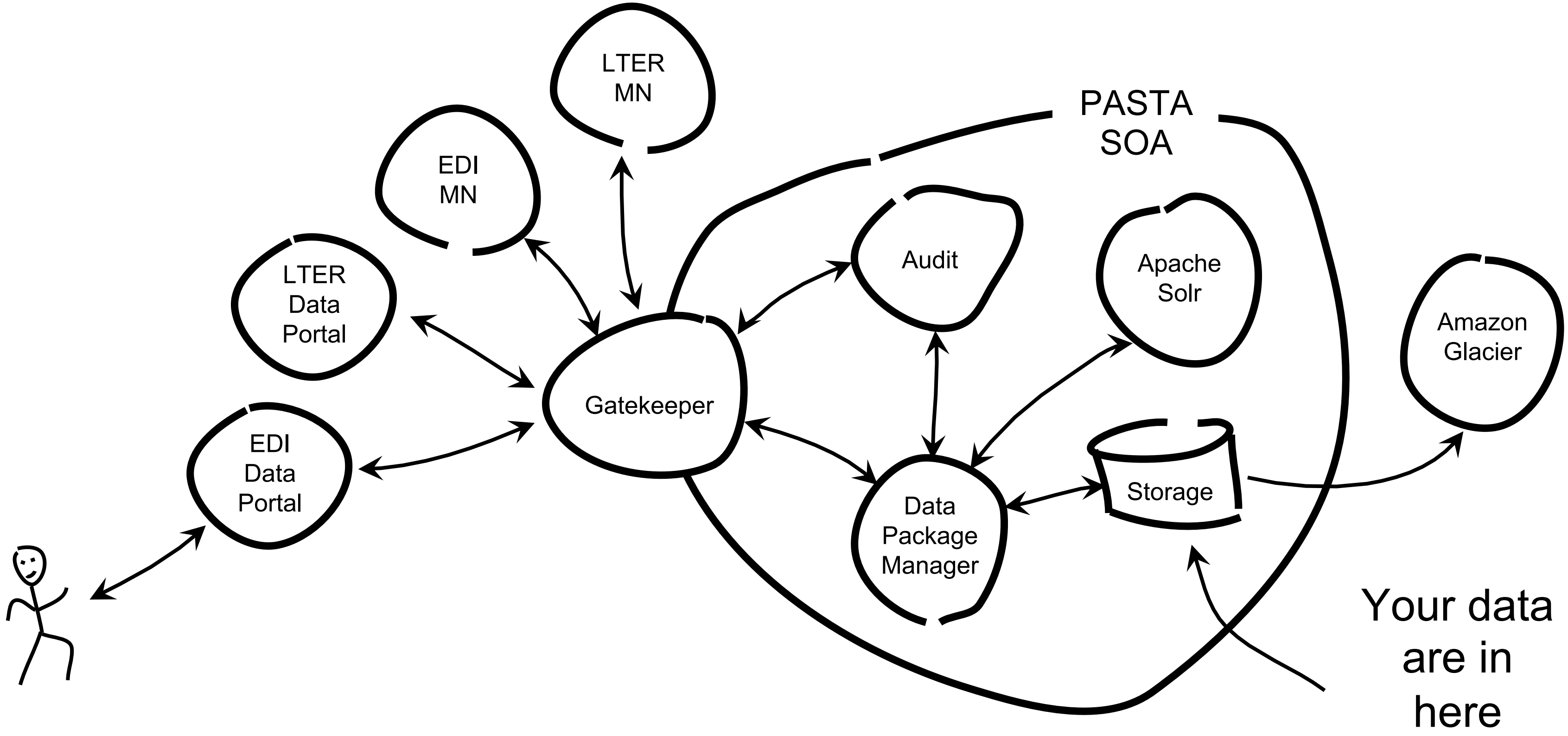
2017-07-03T13:41:55.889
Freshwater connectivity clusters for lakes, wetlands, and streams at the Hydrologic Unit 12 scale in the Midwest and Northeast U.S.A. – freshwater metric variables and K-means cluster assignment

RECENTLY UPDATED

2017-07-10T15:09:22.805
Macroalgae Production in Florida Bay (FCE), South Florida from May 2007 to Present

2017-07-07T15:15:27.094
Spatial variability in water chemistry of four Wisconsin aquatic ecosystems - High speed limnology Environmental Science and Technology datasets

Components of the EDI Data Repository



About the EDI Data Repository (1 of 2)

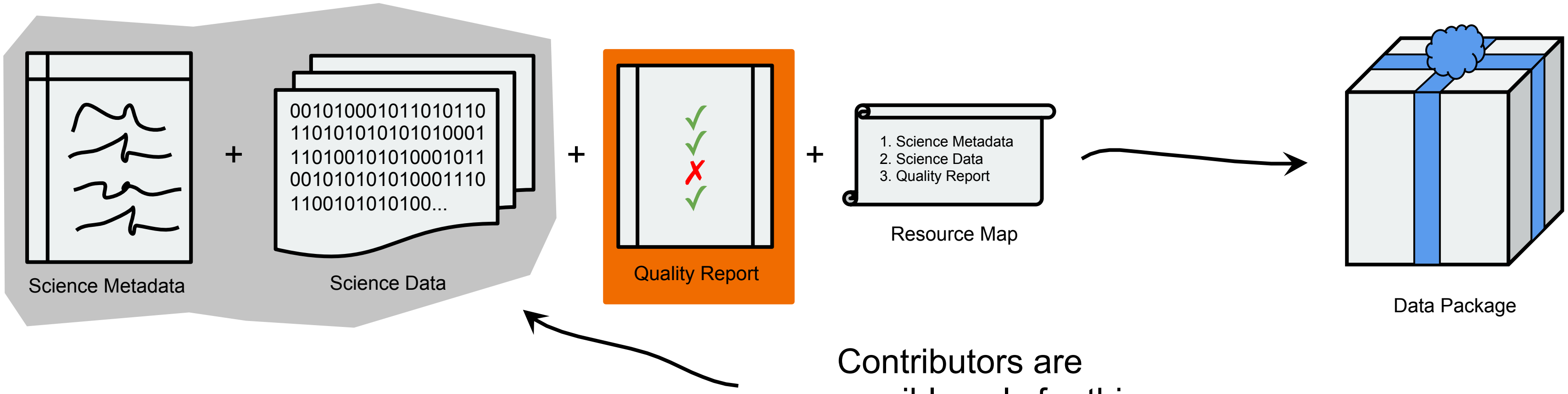
- Community (LTER Network) designed with transparency in mind
- In continuous production since January 2013
- An Internet accessible, open data repository, open source project (GitHub)
- The EDI Data Repository is an *implementation instance* of a PASTA data repository. To put it another way, the EDI Data Repository is *powered by* PASTA software.
    - Metadata-driven (Ecological Metadata Language)
    - Java SE/EE implementation
    - Service Oriented Architecture (SOA) with a RESTful web service API
    - PASTA itself is *not* a website designed for direct human interaction
    - EDI Data Portal provides a GUI interface to PASTA web services

About the EDI Data Repository (2 of 2)

- Content summary
  - 42,000+ data packages
  - ~9TB data volume
- DataONE member node operator for the EDI MN and LTER MN (GMN)
- DOIs minted for all public data packages through DataCite
  - **10.6073/pasta/c868deb66af9c0f72c9672f65b484995**
  - Assigned at the data package level, not assigned to its individual components
- Ensures data integrity through checksums and multiple backup strategies (AWS Glacier)
- Supports upload notification to drive workflows (e.g. Twitter bot)
- Funded by US National Science Foundation - DEB/LTER/ABI and ARRA

What’s A Data Package?

Data Package (noun): an assemblage of science metadata (e.g., EML) and one or more science data objects; PASTA data packages include a “quality report” object and are described by package metadata called a “resource map” (i.e., manifest)



Contributors are responsible only for this

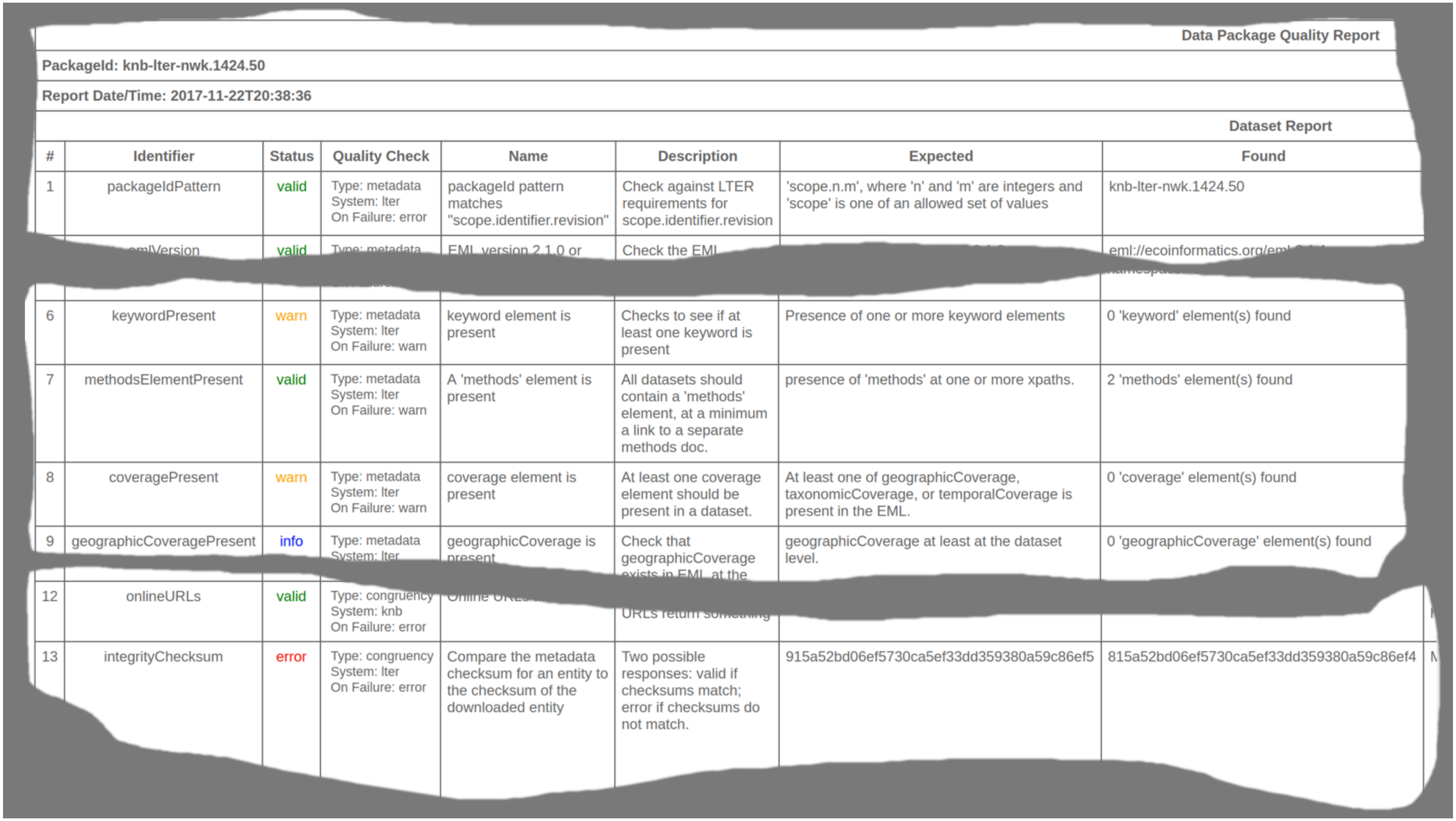
Identifiers in the EDI Data Repository

- PASTA data package identifier
  - edi.12.1 (internal)
  - https://pasta.lternet.edu/package/eml/edi/12/1 (external, public facing)
- DOI
  - doi:10.6073/pasta/43e9a619cbebb98da0011ada25ad5c12
  - https://dx.doi.org/10.6073/pasta/43e9a619cbebb98da0011ada25ad5c12
- DataONE identifier
  - doi:10.6073/pasta/43e9a619cbebb98da0011ada25ad5c12 (data package)
- Identifiers support
  - immutability
  - strong versioning
  - reliable access

PASTA Congruence/Quality Checking

- Are all required/recommended metadata fields present?
- Do metadata values comply with best practices?
- Is the data available for download?
- Does metadata accurately describe the data, i.e. are the two congruent?

info, valid, warn, error



Data Package Quality Report

PackageId: knb-lter-nwk.1424.50

Report Date/Time: 2017-11-22T20:38:36

Dataset Report

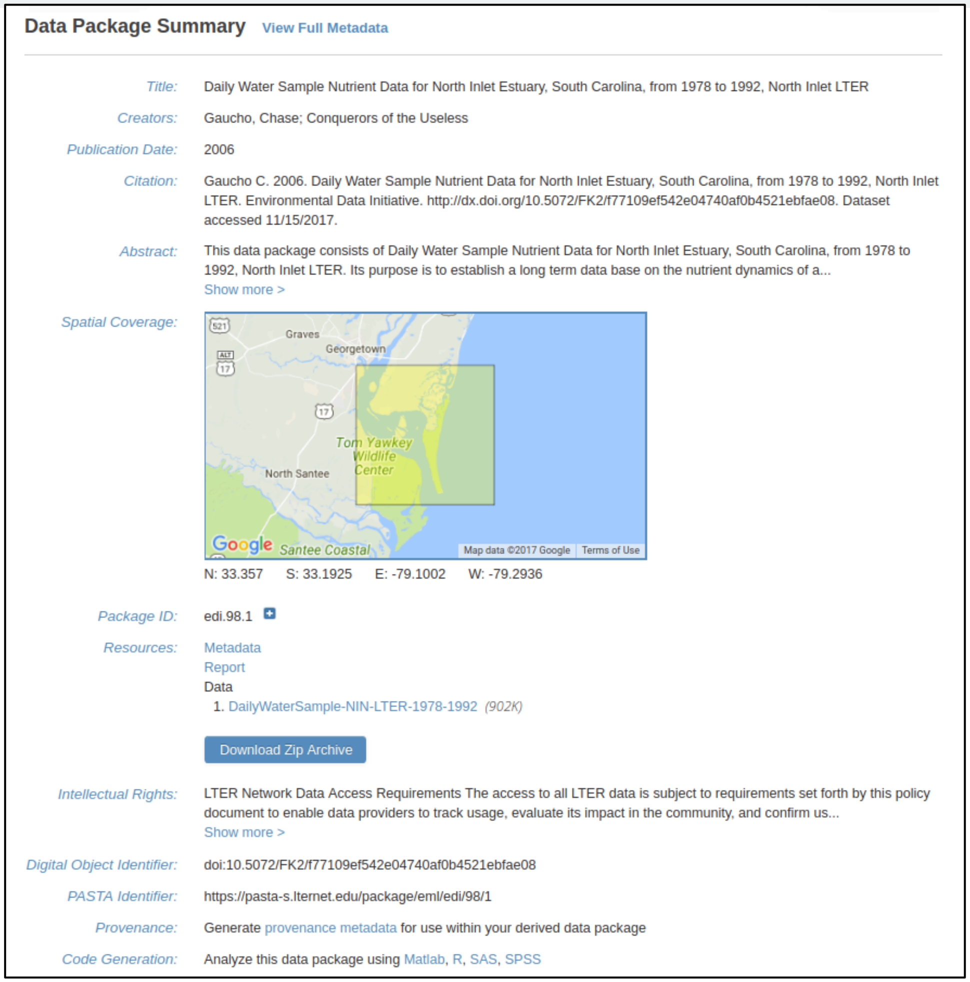
#	Identifier	Status	Quality Check	Name	Description	Expected	Found	
1	packageIdPattern	valid	Type: metadata System: lter On Failure: error	packageId pattern matches "scope.identifier.revision"	Check against LTER requirements for scope.identifier.revision	'scope.n.m', where 'n' and 'm' are integers and 'scope' is one of an allowed set of values	knb-lter-nwk.1424.50	
	emlVersion	valid	Type: metadata	EML version 2.1.0 or	Check the EML		eml://ecoinformatics.org/eml-2.1.1	
6	keywordPresent	warn	Type: metadata System: lter On Failure: warn	keyword element is present	Checks to see if at least one keyword is present	Presence of one or more keyword elements	0 'keyword' element(s) found	
7	methodsElementPresent	valid	Type: metadata System: lter On Failure: warn	A 'methods' element is present	All datasets should contain a 'methods' element, at a minimum a link to a separate methods doc.	presence of 'methods' at one or more xpaths.	2 'methods' element(s) found	
8	coveragePresent	warn	Type: metadata System: lter On Failure: warn	coverage element is present	At least one coverage element should be present in a dataset.	At least one of geographicCoverage, taxonomicCoverage, or temporalCoverage is present in the EML.	0 'coverage' element(s) found	
9	geographicCoveragePresent	info	Type: metadata System: lter	geographicCoverage is present	Check that geographicCoverage exists in EML at the	geographicCoverage at least at the dataset level.	0 'geographicCoverage' element(s) found	
12	onlineURLs	valid	Type: congruency System: knb On Failure: error	Online URLs	URLs return something			
13	integrityChecksum	error	Type: congruency System: lter On Failure: error	Compare the metadata checksum for an entity to the checksum of the downloaded entity	Two possible responses: valid if checksums match; error if checksums do not match.	915a52bd06ef5730ca5ef33dd359380a59c86ef5	815a52bd06ef5730ca5ef33dd359380a59c86ef4	M

Data Package Landing Page

Citation suggestion

Provenance information

Code generation



Data Package Summary

View Full Metadata

Title: Daily Water Sample Nutrient Data for North Inlet Estuary, South Carolina, from 1978 to 1992, North Inlet LTER

Creators: Gaucho, Chase; Conquerors of the Useless

Publication Date: 2006

Citation: Gaucho C. 2006. Daily Water Sample Nutrient Data for North Inlet Estuary, South Carolina, from 1978 to 1992, North Inlet LTER. Environmental Data Initiative. http://dx.doi.org/10.5072/FK2/f77109ef542e04740af0b4521ebfae08. Dataset accessed 11/15/2017.

Abstract: This data package consists of Daily Water Sample Nutrient Data for North Inlet Estuary, South Carolina, from 1978 to 1992, North Inlet LTER. Its purpose is to establish a long term data base on the nutrient dynamics of a...
Show more >

Spatial Coverage:

N: 33.357 S: 33.1925 E: -79.1002 W: -79.2936

Package ID: edi.98.1

Resources: Metadata
Report
Data
1. DailyWaterSample-NIN-LTER-1978-1992 *(902K)*

Download Zip Archive

Intellectual Rights: LTER Network Data Access Requirements The access to all LTER data is subject to requirements set forth by this policy document to enable data providers to track usage, evaluate its impact in the community, and confirm us...
Show more >

Digital Object Identifier: doi:10.5072/FK2/f77109ef542e04740af0b4521ebfae08

PASTA Identifier: https://pasta-s.lternet.edu/package/eml/edi/98/1

Provenance: Generate provenance metadata for use within your derived data package

Code Generation: Analyze this data package using Matlab, R, SAS, SPSS

Overview

EDI’s mission and approaches - Corinna Gries

Data repository and publishing - Duane Costa

Data publication workflow support - Colin Smith

Outreach and training - Kristin Vanderbilt

Data integration support - Margaret O’Brien

Data submitted to EDI

Long term data sets - regular updating (e.g. LTER, LTREB, OBFS)

Data associated with one project (e.g., MSB, other short term funding)

Data associated with a publication (may be subsets, usually are well curated)

Data policies in EDI

Data have to be submitted at the same time as metadata

- Except human subject data
- Special access to sensitive data (e.g., endangered species location) can be arranged, but data need to be on EDI server

Embargo time during paper review can be arranged

Data are licensed to be in the public domain or require attribution

Data publication workflow support

Best practices for organizing, cleaning, and documenting data

Software to help organize, clean, and document data

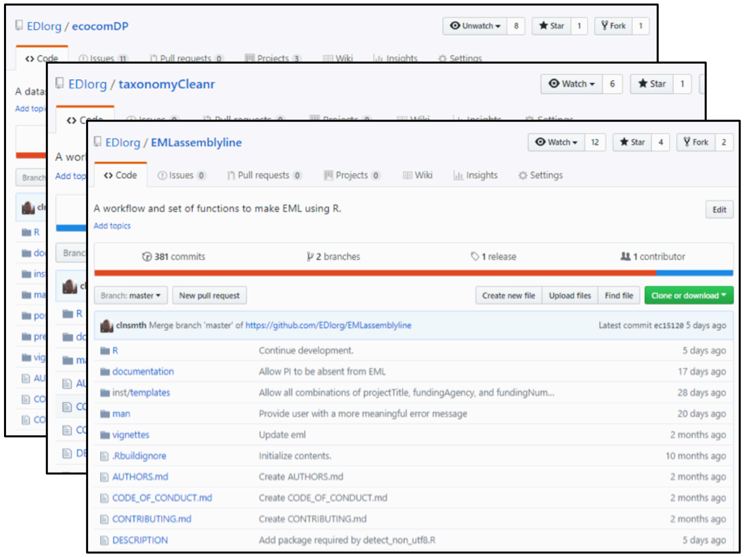
One-on-one support for any step of this process

Software to format, clean, and document data

Data formatting tools to help reformat data into a standardized structure

Data cleaning tools

Data documentation tools to help create high quality EML metadata



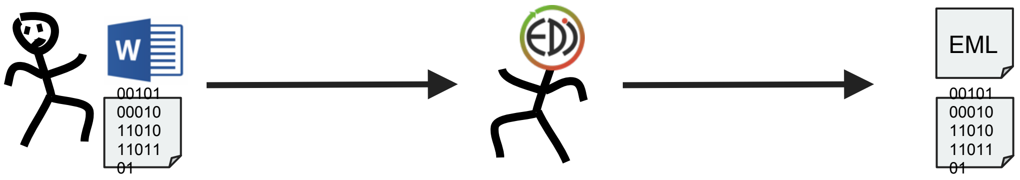
A workflow and set of functions to make EML using R.

clnsmth Merge branch 'master' of https://github.com/EDIorg/EMLassemblyline — Latest commit ec15120 5 days ago

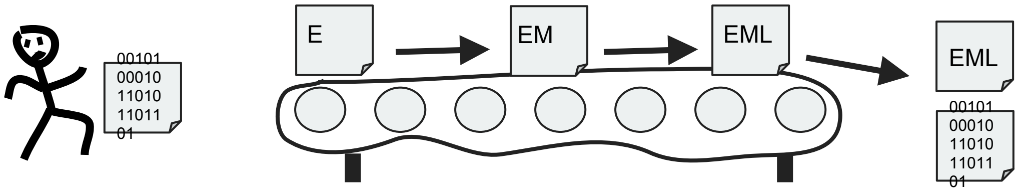
R	Continue development.	5 days ago
documentation	Allow PI to be absent from EML	17 days ago
inst/templates	Allow all combinations of projectTitle, fundingAgency, and fundingNum...	28 days ago
man	Provide user with a more meaningful error message	20 days ago
vignettes	Update eml	2 months ago
.Rbuildignore	Initialize contents.	10 months ago
AUTHORS.md	Create AUTHORS.md	2 months ago
CODE_OF_CONDUCT.md	Create CODE_OF_CONDUCT.md	2 months ago
CONTRIBUTING.md	Create CONTRIBUTING.md	2 months ago
DESCRIPTION	Add package required by detect_non_utf8.R	5 days ago

Metadata generation

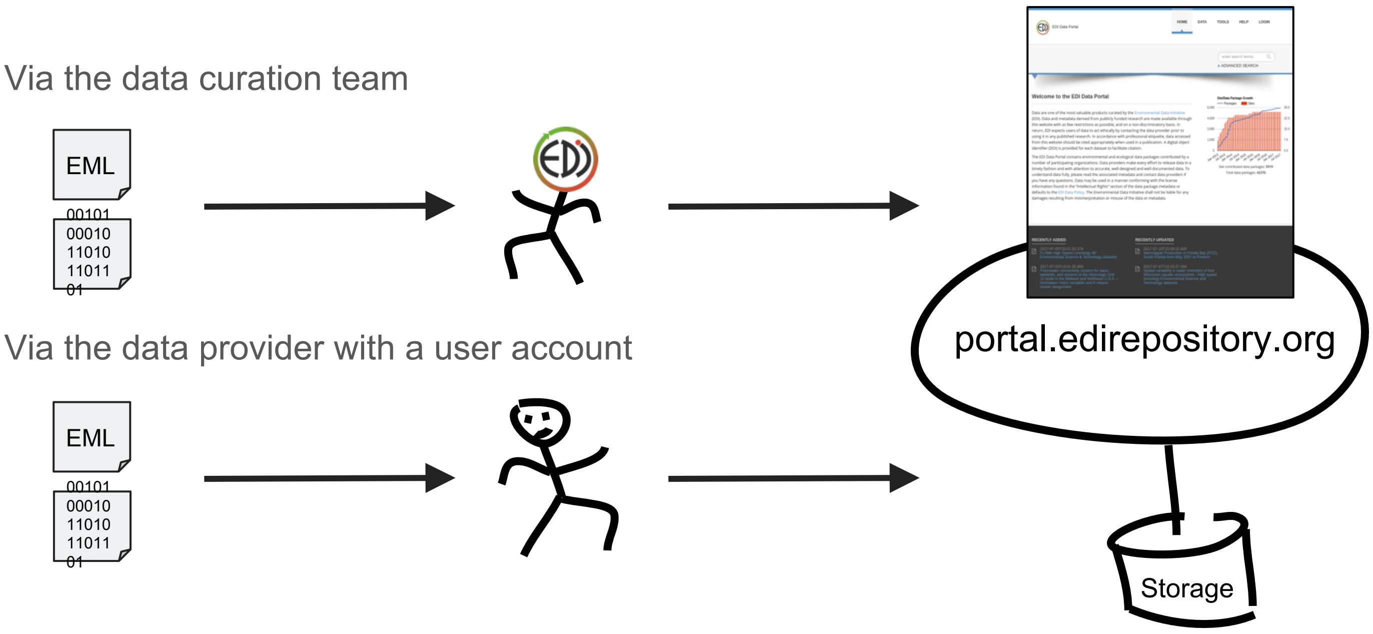
Metadata template, a Microsoft Word document completed by the data provider and converted to EML by EDI's data curation team



EML Assembly Line, a user friendly R library for data providers to generate EML metadata on their own



Data and metadata upload



Via the data curation team

Via the data provider with a user account

EML

00101
00010
11010
11011
01

EML

00101
00010
11010
11011
01

portal.edirepository.org

Storage

Overview

EDI’s mission and approaches - Corinna Gries

Data repository and publishing - Duane Costa

Data publication workflow support - Colin Smith

Outreach and training - Kristin Vanderbilt

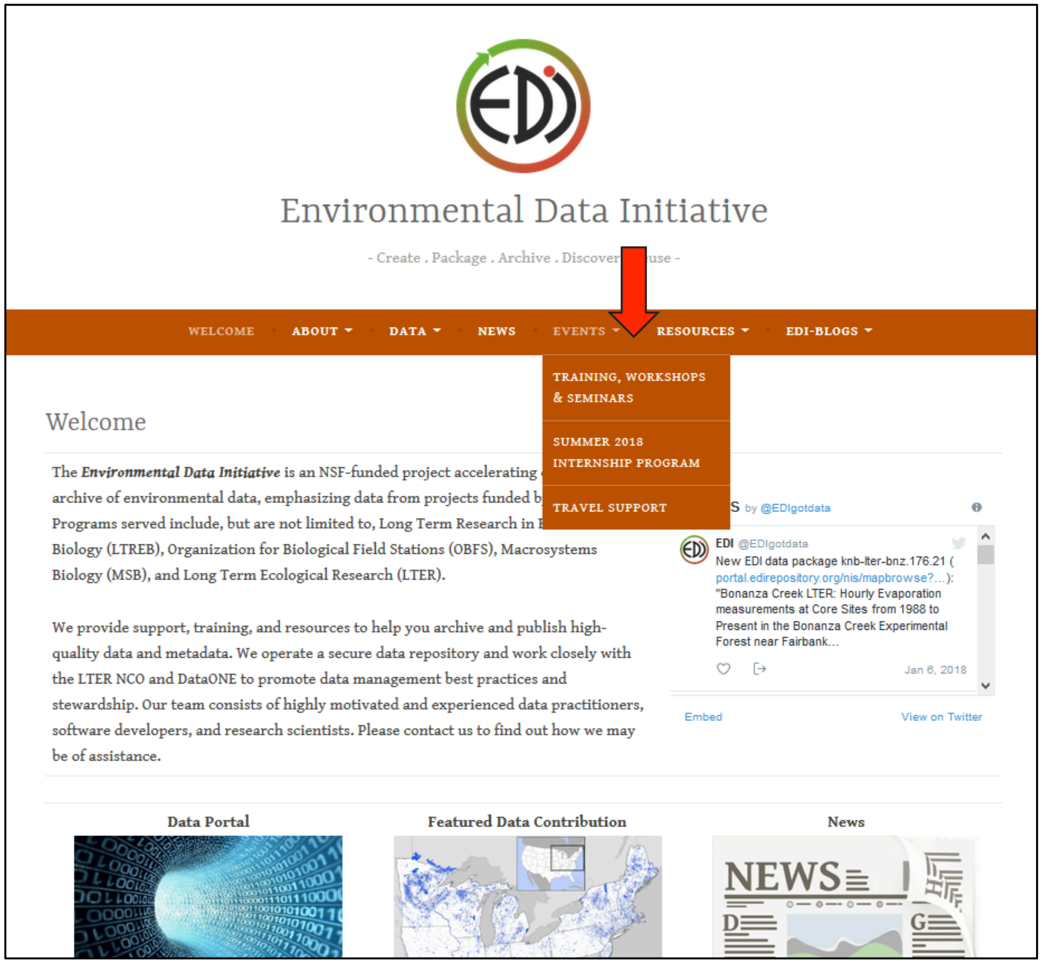
Data integration support - Margaret O’Brien

Outreach:

Website:

https://environmentaldatainitiative.org/

Twitter: @EDIgotdata


Environmental Data Initiative
- Create . Package . Archive . Discover use -
WELCOME
ABOUT
DATA
NEWS
EVENTS
RESOURCES
EDI-BLOGS
TRAINING, WORKSHOPS & SEMINARS
SUMMER 2018 INTERNSHIP PROGRAM
TRAVEL SUPPORT
S by @EDIgotdata
Embed
View on Twitter
Data Portal
Featured Data Contribution
News
NEWS

Welcome

The ***Environmental Data Initiative*** is an NSF-funded project accelerating
archive of environmental data, emphasizing data from projects funded b
Programs served include, but are not limited to, Long Term Research in E
Biology (LTREB), Organization for Biological Field Stations (OBFS), Macrosystems Biology (MSB), and Long Term Ecological Research (LTER).

We provide support, training, and resources to help you archive and publish high-quality data and metadata. We operate a secure data repository and work closely with the LTER NCO and DataONE to promote data management best practices and stewardship. Our team consists of highly motivated and experienced data practitioners, software developers, and research scientists. Please contact us to find out how we may be of assistance.

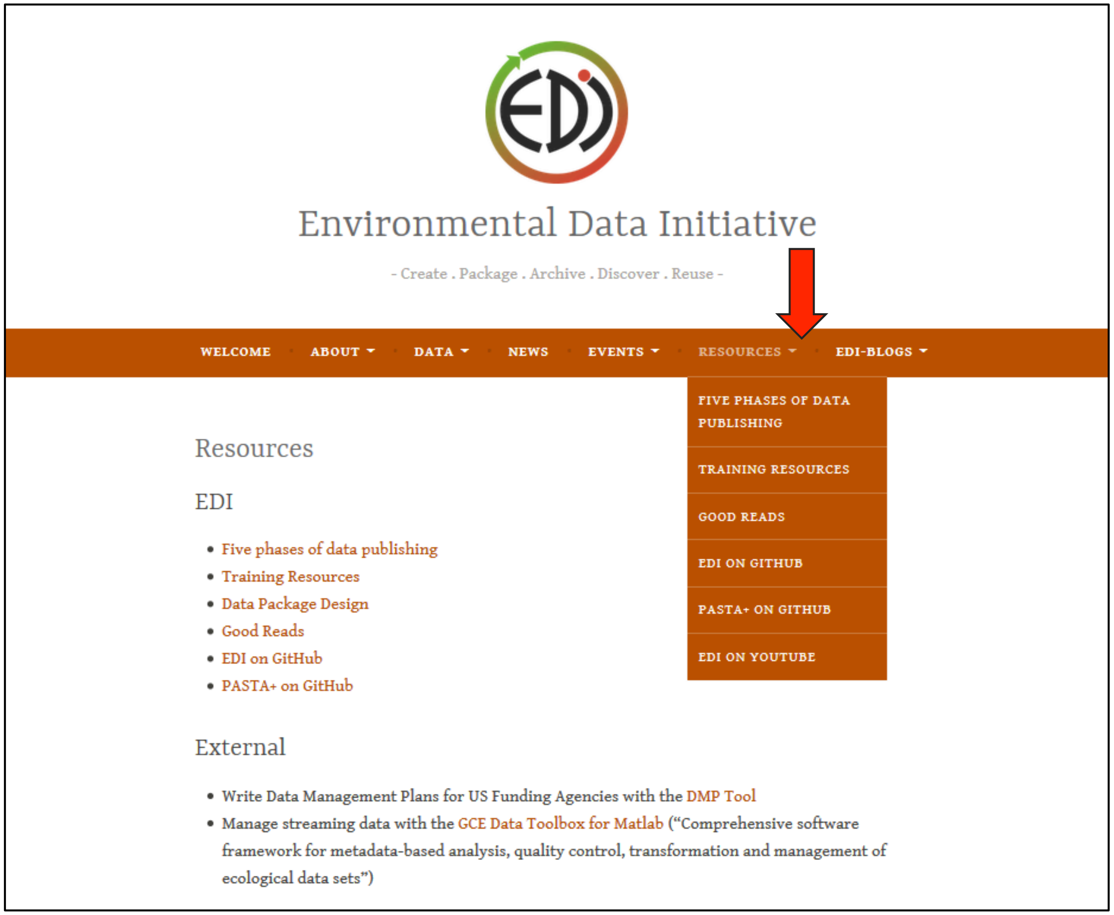
EDI @EDIgotdata
New EDI data package knb-lter-bnz.176.21 (portal.edirepository.org/nis/mapbrowse?...): "Bonanza Creek LTER: Hourly Evaporation measurements at Core Sites from 1988 to Present in the Bonanza Creek Experimental Forest near Fairbank...
Jan 6, 2018

Outreach:

Website:

https://environmentaldatainitiative.org/

Twitter: @EDIgotdata



Environmental Data Initiative

- Create . Package . Archive . Discover . Reuse -

Resources

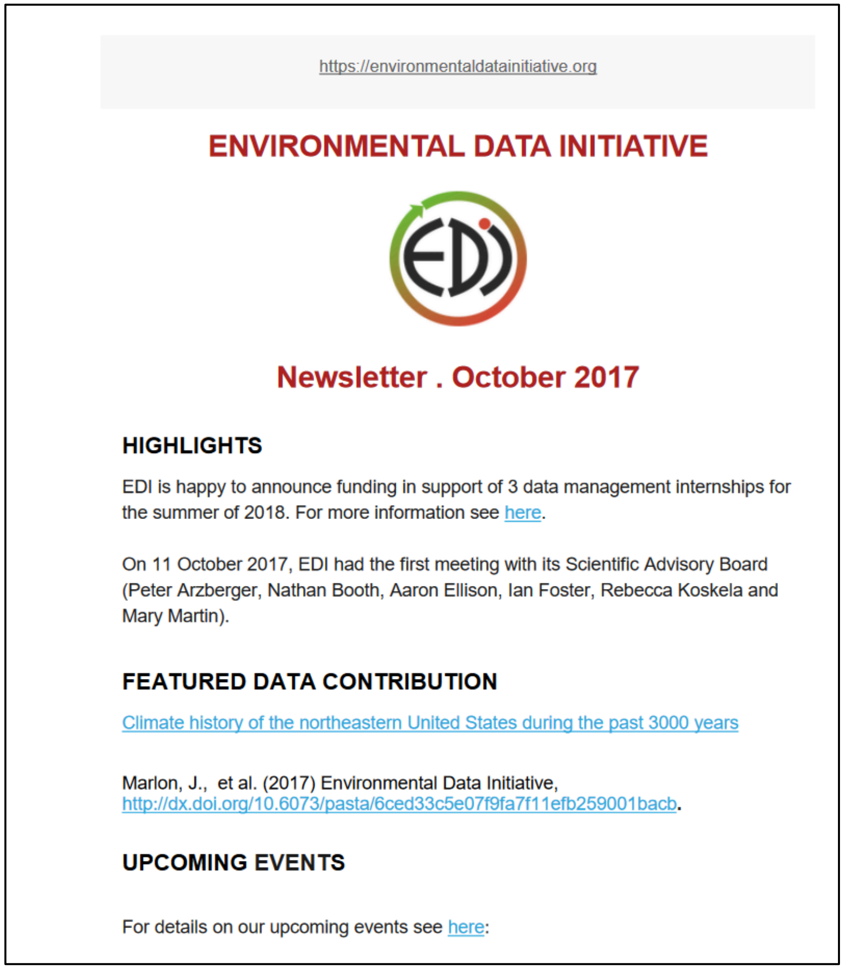
EDI

- Five phases of data publishing
- Training Resources
- Data Package Design
- Good Reads
- EDI on GitHub
- PASTA+ on GitHub

External

- Write Data Management Plans for US Funding Agencies with the DMP Tool
- Manage streaming data with the GCE Data Toolbox for Matlab ("Comprehensive software framework for metadata-based analysis, quality control, transformation and management of ecological data sets")

Newsletter



https://environmentaldatainitiative.org

ENVIRONMENTAL DATA INITIATIVE

Newsletter . October 2017

HIGHLIGHTS

EDI is happy to announce funding in support of 3 data management internships for the summer of 2018. For more information see here.

On 11 October 2017, EDI had the first meeting with its Scientific Advisory Board (Peter Arzberger, Nathan Booth, Aaron Ellison, Ian Foster, Rebecca Koskela and Mary Martin).

FEATURED DATA CONTRIBUTION

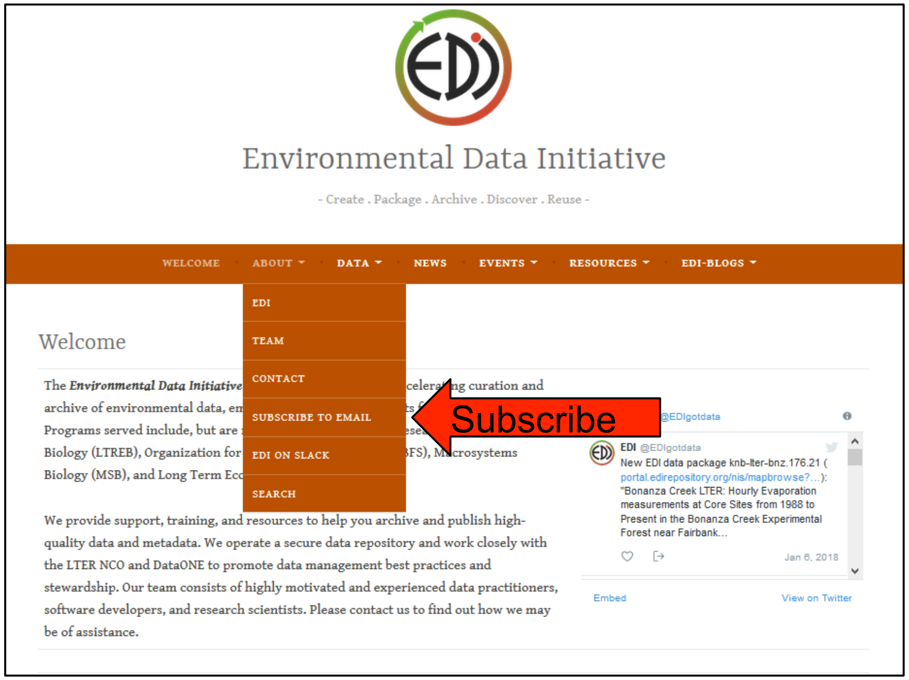
Climate history of the northeastern United States during the past 3000 years

Marlon, J., et al. (2017) Environmental Data Initiative, http://dx.doi.org/10.6073/pasta/6ced33c5e07f9fa7f11efb259001bacb.

UPCOMING EVENTS

For details on our upcoming events see here:

Newsletter

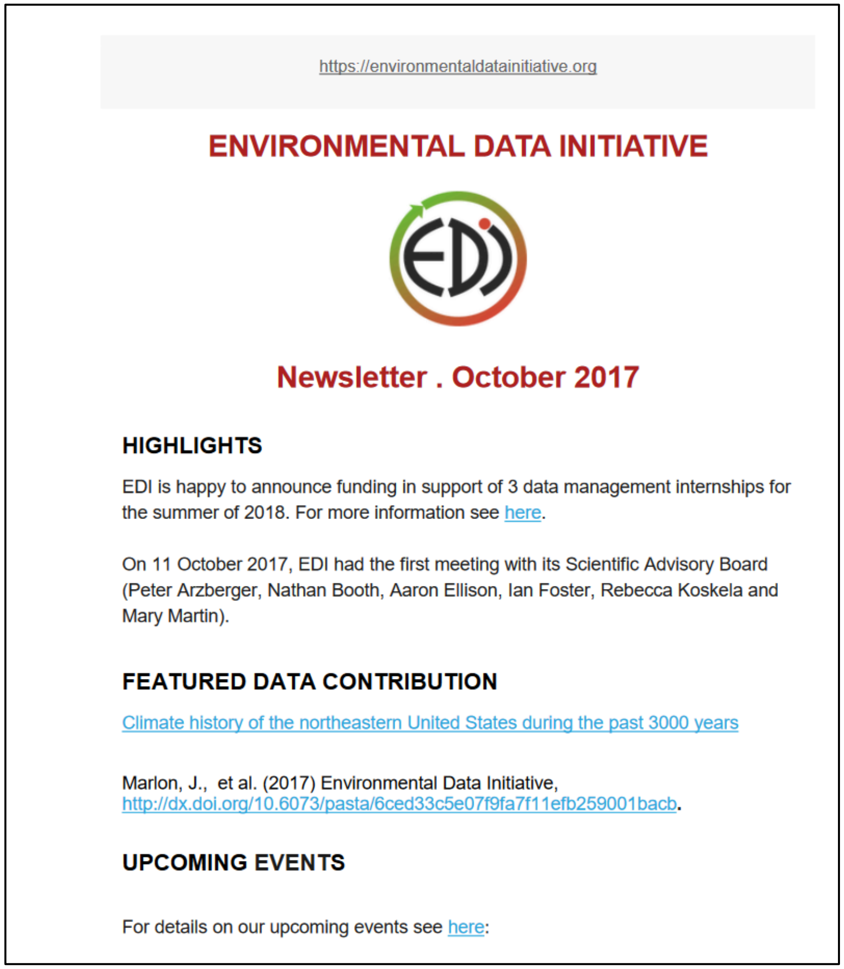


Welcome

The *Environmental Data Initiative* celerating curation and archive of environmental data, em ts f Programs served include, but are ese Biology (LTREB), Organization for FS), Microsystems Biology (MSB), and Long Term Ec

We provide support, training, and resources to help you archive and publish high-quality data and metadata. We operate a secure data repository and work closely with the LTER NCO and DataONE to promote data management best practices and stewardship. Our team consists of highly motivated and experienced data practitioners, software developers, and research scientists. Please contact us to find out how we may be of assistance.

EDI @EDIgotdata
New EDI data package knb-lter-bnz.176.21 (portal.edirepository.org/nis/mapbrowse?...): "Bonanza Creek LTER: Hourly Evaporation measurements at Core Sites from 1988 to Present in the Bonanza Creek Experimental Forest near Fairbank...
Jan 6, 2018



https://environmentaldatainitiative.org

ENVIRONMENTAL DATA INITIATIVE

Newsletter . October 2017

HIGHLIGHTS

EDI is happy to announce funding in support of 3 data management internships for the summer of 2018. For more information see here.

On 11 October 2017, EDI had the first meeting with its Scientific Advisory Board (Peter Arzberger, Nathan Booth, Aaron Ellison, Ian Foster, Rebecca Koskela and Mary Martin).

FEATURED DATA CONTRIBUTION

Climate history of the northeastern United States during the past 3000 years

Marlon, J., et al. (2017) Environmental Data Initiative, http://dx.doi.org/10.6073/pasta/6ced33c5e07f9fa7f11efb259001bacb.

UPCOMING EVENTS

For details on our upcoming events see here:

Student Internships

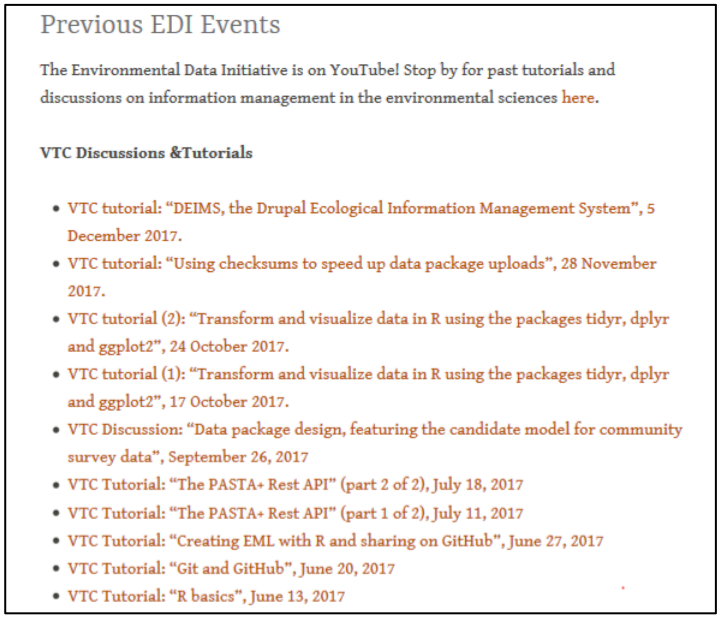
- Project leaders apply to host a student to work with particular data sets
- Students will
  - Learn to create metadata, quality control data and archive data packages in the EDI repository
  - Be trained and mentored by EDI
  - Conduct a small data analysis project
- $5000 scholarship for the summer

Information Management Training: In-Person

- Technical
  - IMs from OBFS, LTREB, LTER (2017)
  - EDI interns/OBFS/LTREB/LTER (2018)
  - Early Career faculty and their lab members (2018)
- Overview
  - GLEON scientists (2017)
  - OBFS site managers (2018)

Info Management Training: Online

- VTC (Git, EML Creation Using R, PASTA+ REST API)
  - Slides and YouTube videos
- Upcoming: Five Phases of Data Publishing
  - January 30: “What are Clean Data?”
- EDI data managers available to help!



Previous EDI Events

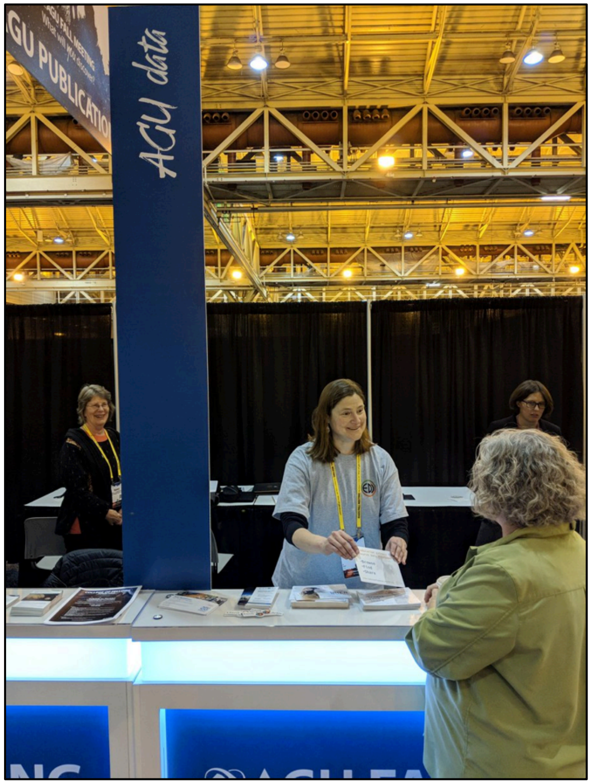
The Environmental Data Initiative is on YouTube! Stop by for past tutorials and discussions on information management in the environmental sciences here.

VTC Discussions &Tutorials

- VTC tutorial: “DEIMS, the Drupal Ecological Information Management System”, 5 December 2017.
- VTC tutorial: “Using checksums to speed up data package uploads”, 28 November 2017.
- VTC tutorial (2): “Transform and visualize data in R using the packages tidyr, dplyr and ggplot2”, 24 October 2017.
- VTC tutorial (1): “Transform and visualize data in R using the packages tidyr, dplyr and ggplot2”, 17 October 2017.
- VTC Discussion: “Data package design, featuring the candidate model for community survey data”, September 26, 2017
- VTC Tutorial: “The PASTA+ Rest API” (part 2 of 2), July 18, 2017
- VTC Tutorial: “The PASTA+ Rest API” (part 1 of 2), July 11, 2017
- VTC Tutorial: “Creating EML with R and sharing on GitHub”, June 27, 2017
- VTC Tutorial: “Git and GitHub”, June 20, 2017
- VTC Tutorial: “R basics”, June 13, 2017

Other Outreach:

- National and International Meetings (OBFS, ESA, AGU, ILTER, ESIP)
- Targeted seminars
- Code Repository



Overview

EDI’s mission and approaches - Corinna Gries

Data repository and publishing - Duane Costa

Data publication workflow support - Colin Smith

Outreach and training - Kristin Vanderbilt

Data integration support - Margaret O’Brien

https://environmentaldatainitiative.org/ https://portal.edirepository.org @EDIgotdata edi-got-data

Data Integration Support

Rationale

As archived, primary data are often difficult to use in synthesis - *even with complete metadata*

EDI Provides

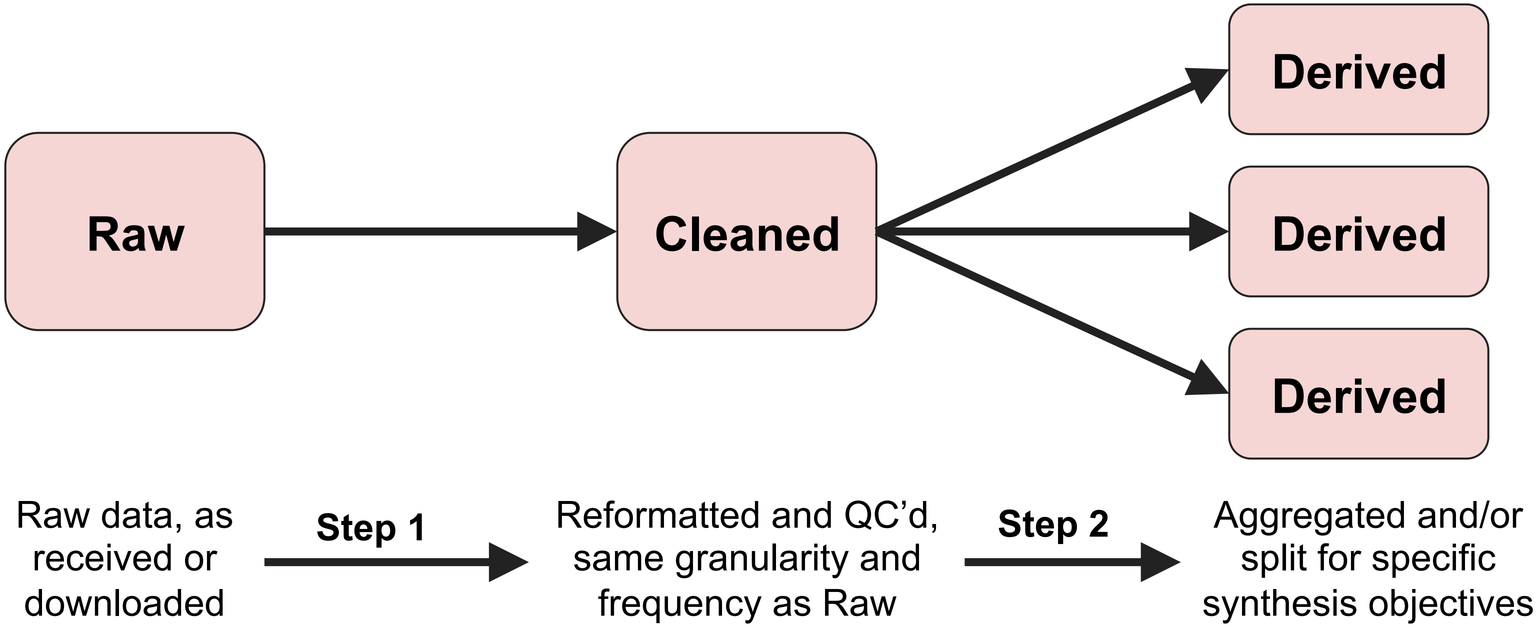
Expertise from data managers to define best practices for data package design

Approach

1) Recommend format and content so analyses can be streamlined
    a) Mechanism for preparers to know
        i) Data elements that are most important
        ii) Layouts that are easiest to use
2) Work with scientists currently engaged in synthesis of primary data

Template for a process that can be reused in many scientific domains

Typical Synthesis Workflow



Raw

Cleaned

Derived

Derived

Derived

Raw data, as received or downloaded

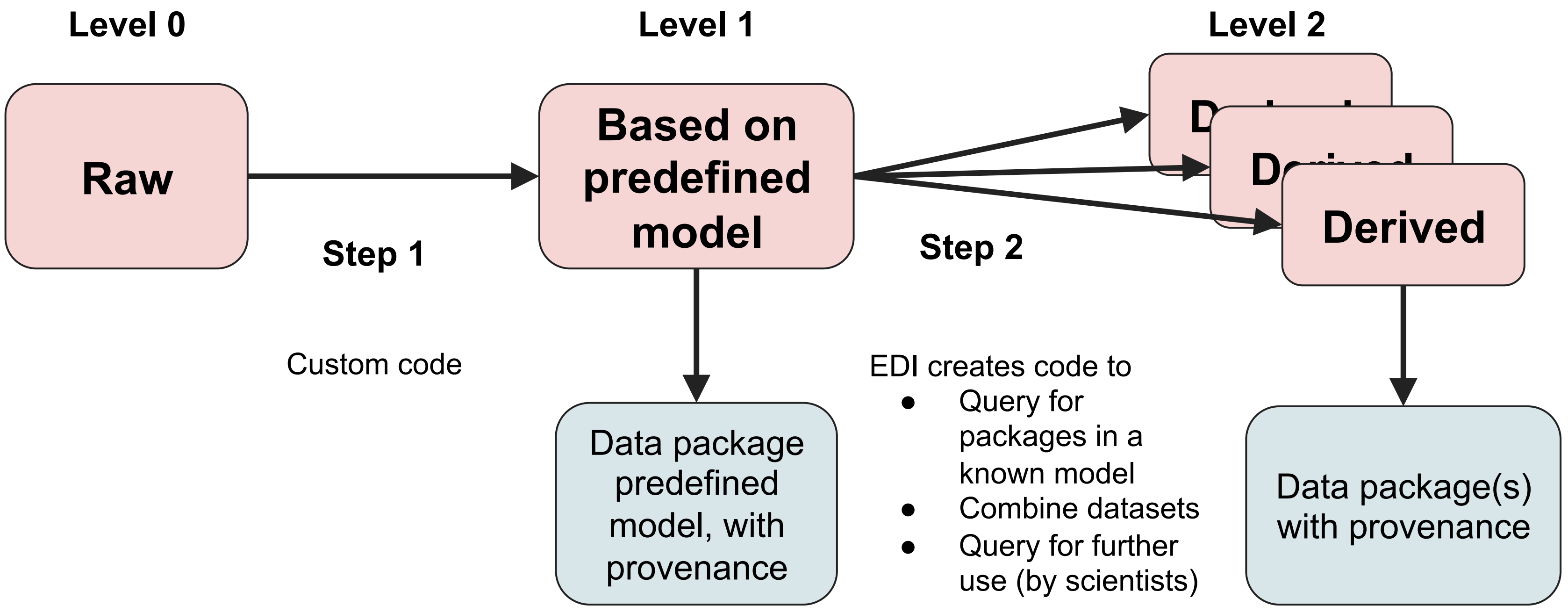
Step 1

Reformatted and QC'd, same granularity and frequency as Raw

Step 2

Aggregated and/or split for specific synthesis objectives

Ideal Synthesis Workflow



Level 0

Raw

Step 1

Custom code

Level 1

Based on predefined model

Data package predefined model, with provenance

Step 2

EDI creates code to

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

Level 2

D

D

Derived

Data package(s) with provenance

Objectives - Design Pattern for Level 1 Dataset

Flexible format for a particular domain of data, but usable for multiple types of measurements and synthesis projects

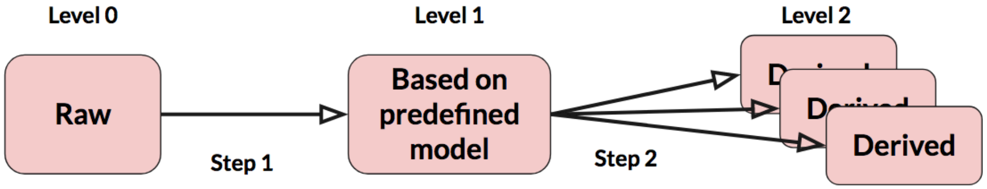
Metadata in EML

Reformat only, no calculations, aggregations or other steps requiring ecological judgement

Original data referenced, one to one relationship with reformatted Level 1 data

Complete; original records can be recreated

Distributed, not complex database application for maintenance

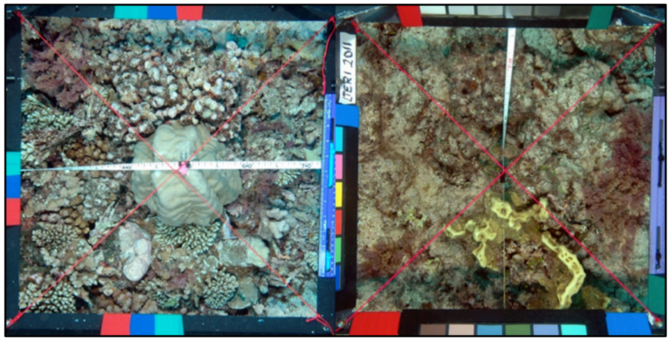


Basic Process

1. Examine available models currently in use
2. Examine *ad hoc* cleaned (Level 1) data created by a synthesis working group
3. Describe patterns
4. Define design pattern tables and value typing
5. Test model against data of interest
6. Create utility scripts for QC, metadata generation

Example - Design Pattern for Community Surveys

1. Examine available models currently in use: ***Popler, Darwin Core, CUAHSI ODM2***
2. Examine *ad hoc* cleaned (Level 1) data created by a synthesis working group: ***LTER Synthesis WG: Synchrony, Metacommunities***
3. Describe patterns: ***quite similar to DC-A but with many custom fields, especially for spatial info***
4. Define design pattern tables, typing - “***ecocomDP***”
5. Test model against data of interest
6. Create utility scripts for QC, metadata generation



Benthic cover at a Moorea Coral Reef (credit, http://lternet.edu)

ecocomDP

Observation table for data related to

Count, biomass, abundance, density

Primary organization

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

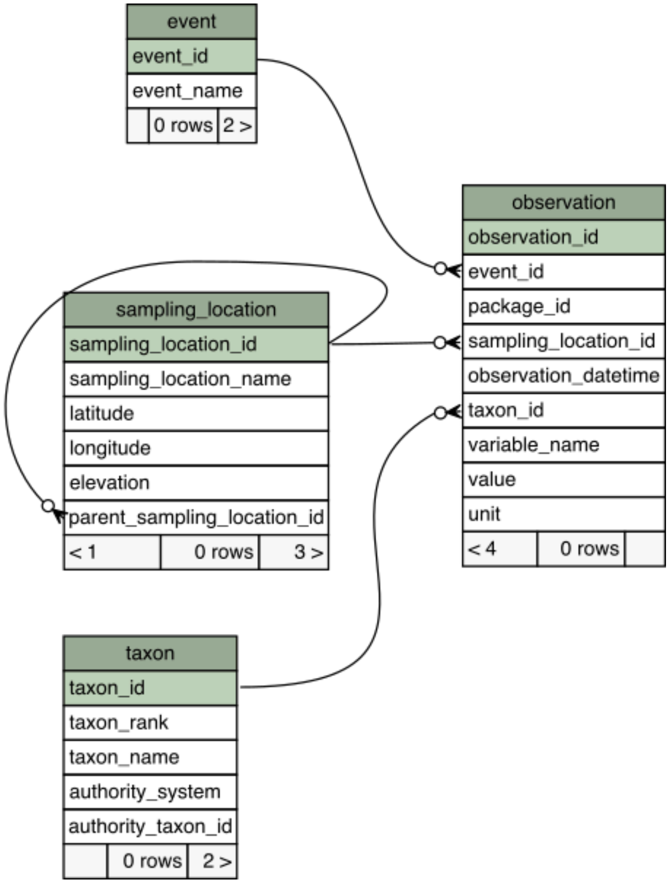
Linked Tables for

Sampling location

Organism

Event

Model includes ancillary tables and summary (https://github.com/EDIorg/ecocomDP)



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 >

Progress

Creating Level 1
- Usability
- Metadata generation

Using Level 1
- Query scripts

Woody encroachment at Konza Prarie (credit, http://lternet.edu)

Thank you

- Contact - info@environmentaldatainitiative.org
- Website - https://environmentaldatainitiative.org/
- Data portal - https://portal.edirepository.org
- Twitter - @EDIgotdata
- Slack - edi-got-data
- GitHub - https://github.com/EDIorg https://github.com/PASTAplus/PASTA
- PASTA+ User/Developer Documentation - http://pastaplus-core.readthedocs.io/en/latest/
- Data Package Manager Web Service API - https://pasta.lternet.edu/package/docs/api
- Audit Manager Web Service API - https://pasta.lternet.edu/audit/docs/api