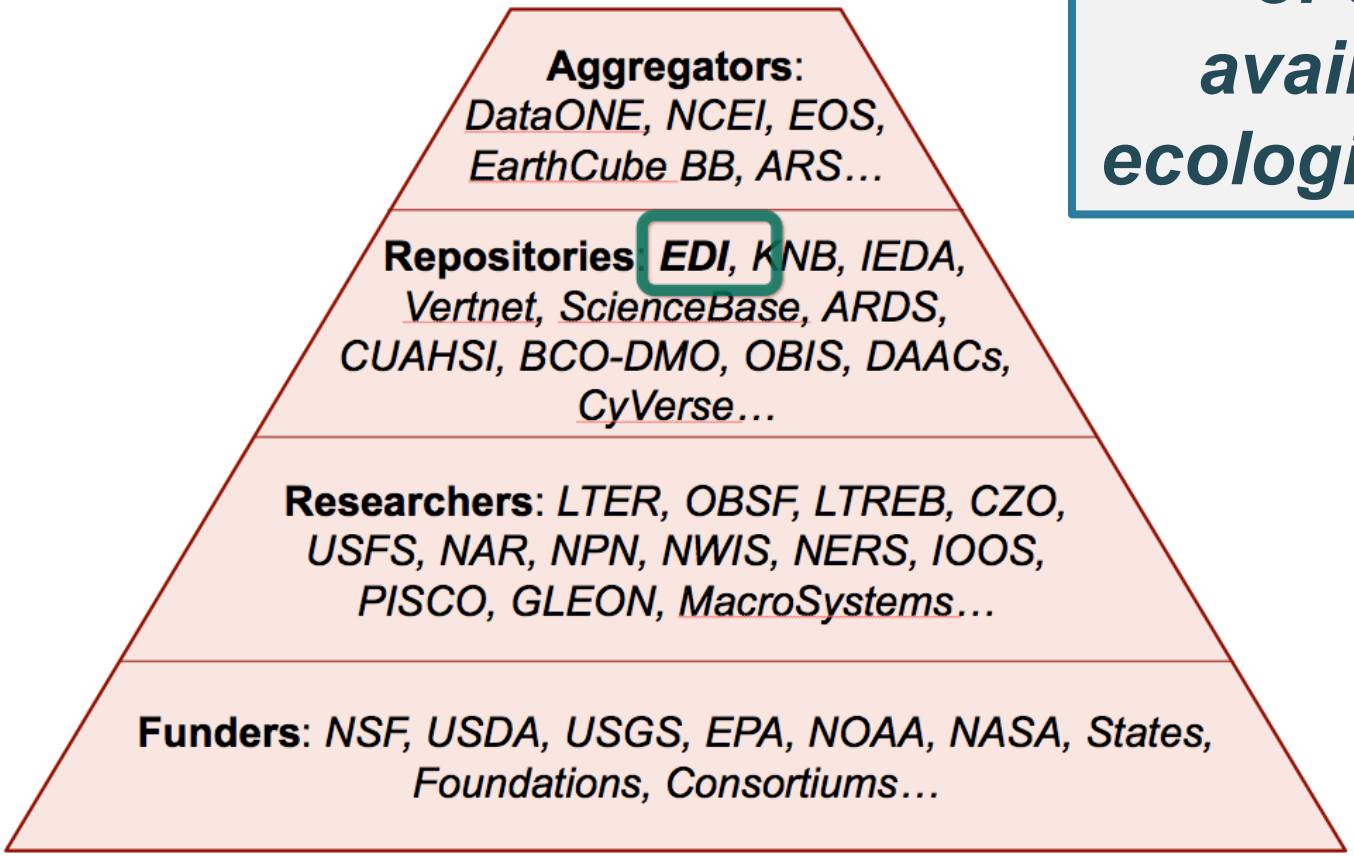
Environmental Data Initiative

Systematic curation and metadata standardization

Margaret O'Brien, ESIP Winter 2017
http://environmentaldatainitiative.org

Introduction and Goals

EDI in the Landscape of Environmental Data Management



Increase the number and quality of datasets available from ecological research

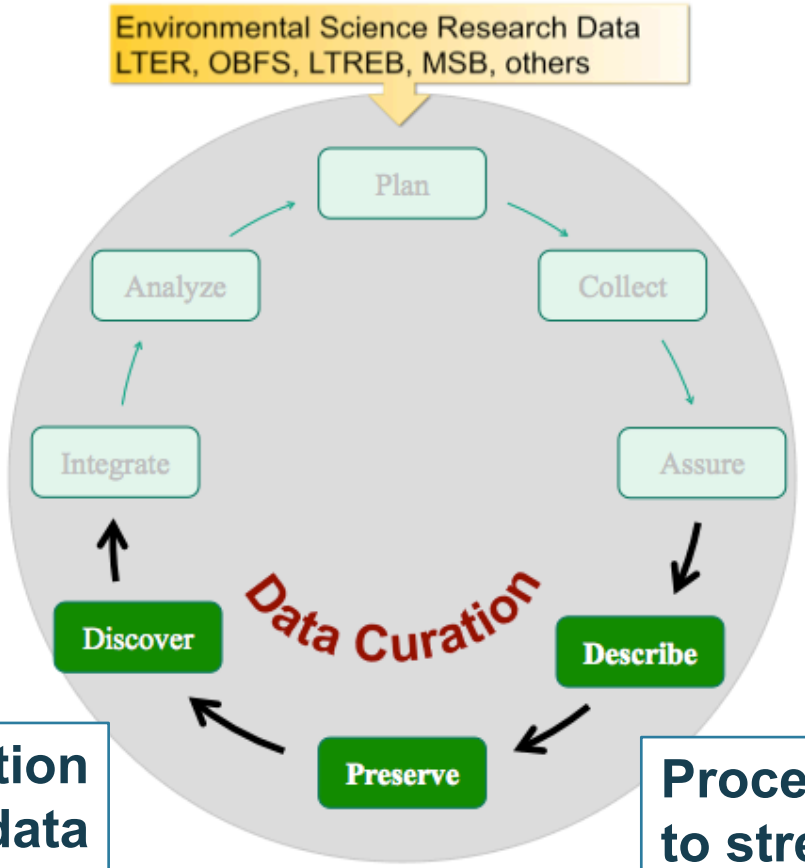
LTER
LTREB
MSB
OBFS

Background

- Evolved from LTER data management
  - Software
  - Practices
  - Data management style

- Builds on existing partnerships
  - NCEAS (data mgt, LTER communication)
  - ESIP (data mgt practices)
  - Agencies

Approach

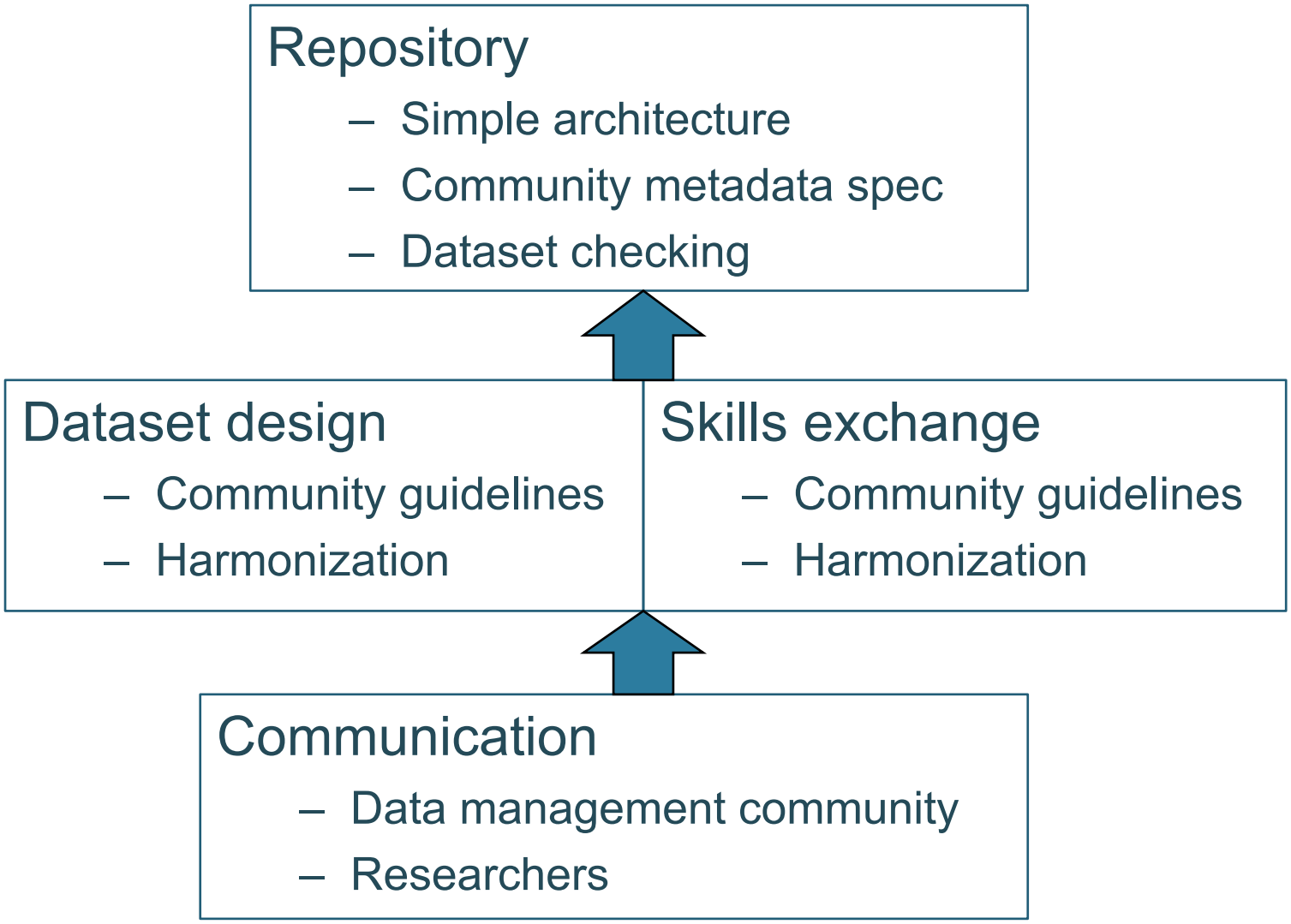


Content Standardization to enhance data DISCOVERY

Process Standardization to streamline data DESCRIPTION

Software Development to promote data PRESERVATION

Activities


Repository
– Simple architecture
– Community metadata spec
– Dataset checking
Dataset design
– Community guidelines
– Harmonization
Skills exchange
– Community guidelines
– Harmonization
Communication
– Data management community
– Researchers

Metadata - Structure

Feature	Ensures
Ecological Metadata Language	Uniform views
Best Practice Recommendations	Stable reference material
Metadata review system	Feedback to submitters
Entrance conditions	Minimum integrity

Metadata - Content

Type	Recommended content
Who	ORCID
Where	presume WGS84
When	ISO 8601
What	**
Why	Text metadata
How	**

Metadata checking system

- 33 checks implemented
- Based on community Best Practice
- Included with every dataset
- New checks proposed

O'Brien et al, doi: 10.016/ecoinf/2016/08/001

Spatial Coverage:

N: 34.87315 S: 32.8 E: -118.4 W: -120.6344833

Package ID: edi.5.1

Resources:

Metadata
Report
Data *

1. SBCMBON integrated fish *(600773962 bytes)*
2. SBCMBON site geolocation *(7078 bytes)*
3. SBCMBON fish density R code *(1459 bytes)*

Download Zip Archive

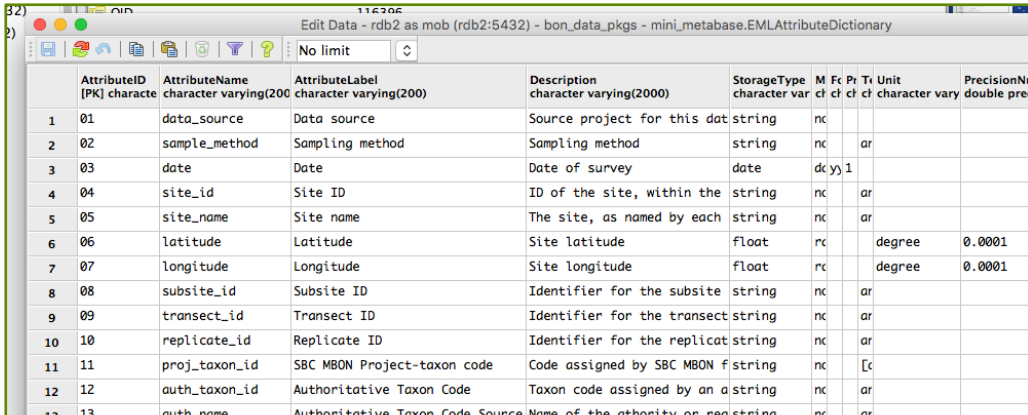
* *By downloading any data you implicitly acknowledge the LTER Data Policy*

Digital Object Identifier: doi:10.6073/pasta/ae7a51738a412dda3cc7ced221c5e90d

#	Identifier	Status	Quality Check	Name	Description	Expected	
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 'm' are integers and is one of an allowed values	
2	emlVersion	valid	Type: metadata System: lter On Failure: error	EML version 2.1.0 or beyond	Check the EML document declaration for version 2.1.0 or higher	eml://ecoinformatics 2.1.0 or eml://ecoinformatics 2.1.1	
3	schemaValid	valid	Type: metadata System: knb On Failure: error	Document is schema-valid EML	Check document schema validity	schema-valid	Document validated for namespace: 'eml://ecoinformatics.org/eml-2.1.1'
4	parserValid	valid	Type: metadata System: knb On Failure: error	Document is EML parser-valid	Check document using the EML IDs and references parser	Validates with the EML IDs and references parser	EML IDs and references parser succeeded
5	schemaValidDereferenced	valid	Type: metadata System: lter On Failure: error	Dereferenced document is schema-valid EML	References are dereferenced, and the resulting file validated	schema-valid	Dereferenced document validated for namespace: 'eml://ecoinformatics.org/eml-2.1.1'
6	keywordPresent	valid	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	4 '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.	1 'methods' element(s) found

Measurement-level Metadata

- Find existing vocabularies where
  - *Structure is detailed; on par with EML's*
  - *Maintenance plan*
  - *Community vetted*
- Contribute to ongoing efforts
  - *DataONE Ecosystem Ontology of Measurement*
- Local dictionaries



	AttributeID [PK] characte	AttributeName character varying(200	AttributeLabel character varying(200)	Description character varying(2000)	StorageType character var	M ch	Fo ch	Pr ch	Ta ch	Unit character vary	PrecisionN double pre
1	01	data_source	Data source	Source project for this dat	string	nc					
2	02	sample_method	Sampling method	Sampling method	string	nc			ar		
3	03	date	Date	Date of survey	date	dc	yy	1			
4	04	site_id	Site ID	ID of the site, within the	string	nc			ar		
5	05	site_name	Site name	The site, as named by each	string	nc			ar		
6	06	latitude	Latitude	Site latitude	float	rc				degree	0.0001
7	07	longitude	Longitude	Site longitude	float	rc				degree	0.0001
8	08	subsite_id	Subsite ID	Identifier for the subsite	string	nc			ar		
9	09	transect_id	Transect ID	Identifier for the transect	string	nc			ar		
10	10	replicate_id	Replicate ID	Identifier for the replicat	string	nc			ar		
11	11	proj_taxon_id	SBC MBON Project-taxon code	Code assigned by SBC MBON f	string	nc			[c		
12	12	auth_taxon_id	Authoritative Taxon Code	Taxon code assigned by an a	string	nc			ar		

Dataset Design Strategies

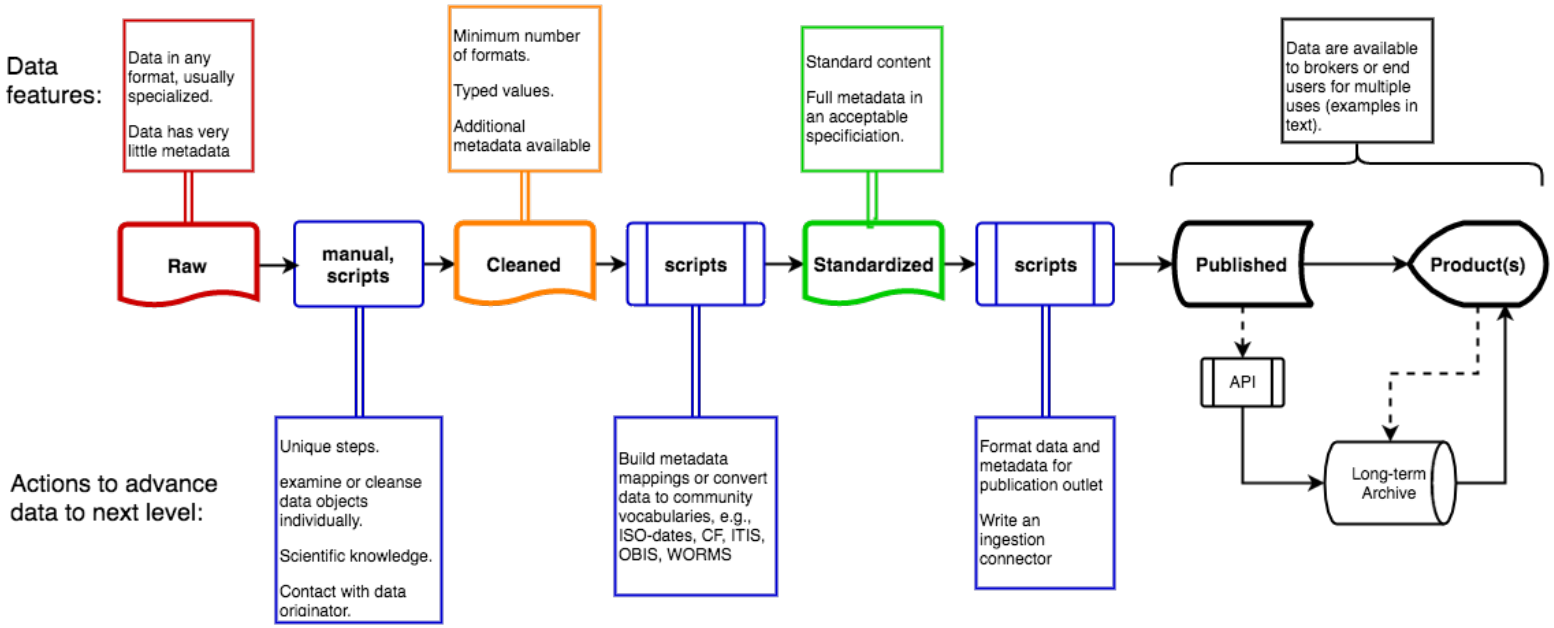
- Collect examples
- Participate in scientific working groups
- Understand how data might be used

Most often, the barriers are social

Data diversity - LTER

- Meteorology, Oceanography, Hydrology, …
- Population, community ecology
- Biogeochemistry, elemental fluxes, nutrients
- Human subjects, Archeology
- GIS
- Genomics
- Imagery, acoustics, telemetry
- Near real-time sensors
- QC’d sensors
- Model output

Example – Population biology



Example 1 – Population biology

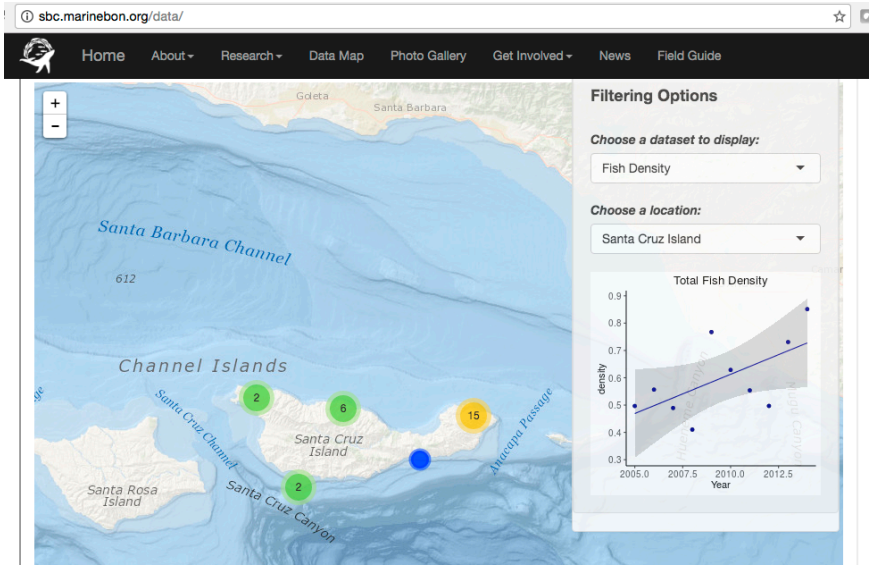
- Define use
  - Extraction of occurrence records (e.g., OBIS)

- Recommended minimum set of measurements (e.g. table columns)
    - Date
    - Latitude
    - Longitude
    - Taxon code
    - Taxon registry
    - Organism count

Example 2 – Population biology

- Define Use
  - Plot time-series of fish areal density

- Minimum set of measurements
  - Date
  - Latitude
  - Longitude
  - Taxon code
  - Taxon registry
  - Areal density



But wait, there might be more …

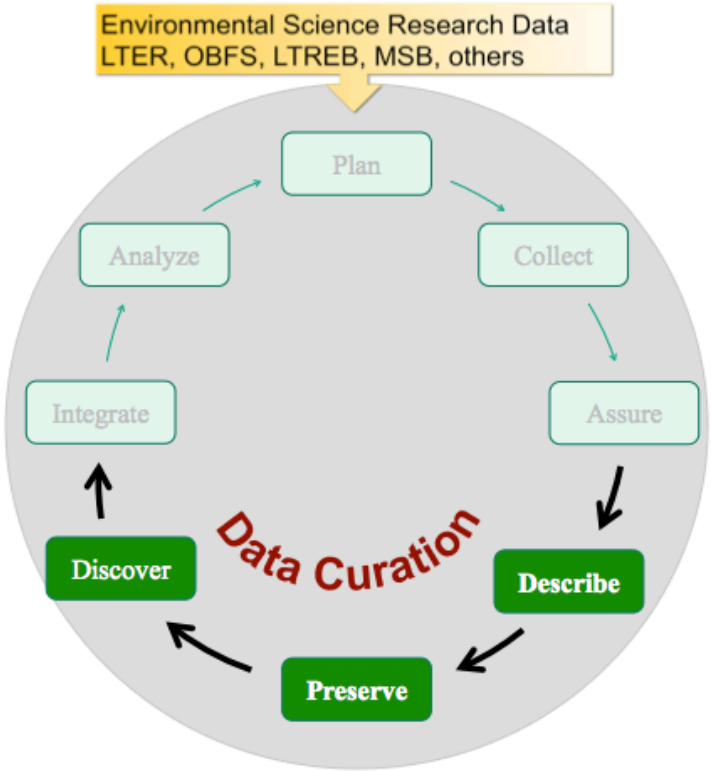
Example 2, cont.

Wait a sec… this is a "time-series plot?"

- Do you expect updates?
- How frequently?
- How will you know
  - New data are available?
  - Data still meet your needs?

Some datasets benefit from specialized metadata or subscription mechanisms

Thank you


Environmental Science Research Data
LTER, OBFS, LTREB, MSB, others
Plan
Collect
Assure
Describe
Preserve
Discover
Integrate
Analyze
Data Curation