2023 LTER Information Management Committee Annual Meeting Report

---

held at DoubleTree Inn, Burlington VT, in conjunction with the 2023 ESIP Summer meeting

August 17, 2023

Participants (25, from top left): Jason Downing (BNZ), Emery Boose (HFR), Marina Franz (CCE), Kris Hall (SEV), Adam Sapp (GCE), Dan Bahauddin (CDR), Paul Hanson (EDI), Chris Turner (NGA), John Porter (VCE), Greg Maurer (JRN), Renée Brown (MCM), Margaret O'Brien (SBC, EDI), Stevan Earl (CAP), Kate Morkeski (NES), Stace Beaulieu (NES), Gabriel Kamener (FCE), Corinna Gries (NTL, EDI), Mary Martin (HBR), Nick Lyon (LNO), Risa McNellis (PIE), Miguel Leon (LUQ), Yang Xia (KNZ), Jim Laundre (ARC), Suzanne Remillard (AND), Li Kui (SBC), Not pictured: Sarah Elmendorf (NWT), Colin Smith (EDI), Sven Bohm (KBS).

Table of Contents

Agenda

Time	Morning Activity
8:30	Welcome
9:00	Working group highlights
9:45	IM Exec election (1 member)
9:55	Volunteer for next Databits editor
10:05	Break
10:35	LNO report
10:50	EDI report
11:05	ILTER report
11:15	IMKE Lightning Talks
12:15	Lunch
13:45	Group Photo
14:00	Units WG Activity
15:00	Break
15:30	EML Best Practices WG Activity
16:30	Adjourn
18:00	Dinner

Meeting materials

A community slide deck for the meeting is in IMC's Google Drive (2023 IMC Annual Meeting). Slides from the full meeting program and Units and EML Best Practices WG activities are reproduced in Appendices A, B, and C, respectively.

Working Group Updates

To begin the day, active IMC working groups provided brief updates on their activity over the past year.

HyMet

The HyMet working group has continued its work to create a modern replacement for the Clim/Hydro DB system. Work has been relatively subdued over the past year, but there are now seven HyMet-formatted datasets on EDI Staging generated with the hyMetDP R package (https://github.com/EDIorg/hymetDP), and the group has begun a pilot project to integrate these datasets into the Dendra platform (https://dendra.science/orgs/lter). Over the next year this working group will be seeking support for a joint effort involving Dendra, EDI, LTER, and potentially CUAHSI to push more LTER sensor network data to Dendra, and to develop integrations between Dendra and EDI.

Metabase

The Metabase working group provided a short update detailing improvements to handling of taxonomic coverages. When populating taxonomic coverages in Metabase, or resulting EML documents, the taxonomic authority identifier can now be included, generally as a URI.

EML Best Practices

This group is updating the EML Best Practices document that was last rewritten in 2017. They presented on their work during an afternoon workshop in the IMC Annual Meeting. See their workshop slides and report for more information.

Unit Dictionary

This working group is actively working towards an updated replacement for the LTER units list. After establishing the scope of the group and explicitly specifying the need for a new system for managing units of measurement, the group has adopted QUDT (https://qudt.org/), which began within NASA but is now an independent 501c3 corp. The working group held a workshop in the afternoon during the IMC Annual Meeting, and more on group progress and next steps are outlined in their workshop slides.

Zotero

Last year this group added a section to the Zotero best practices for LTER sites (https://bit.ly/ZOTERO_BP) describing how to add a dataset as a “document” item type and add information in Zotero's Extra field indicating that the item is a dataset, for example:

```
Type: dataset
Version: 3
DOI: 10.6073/PASTA/CA34BE7554DDC67C9FA0F8DEA01F375B
Previous Version: 2 DOI: 10.6073/PASTA/F7204A847A1D71FCE18ED880363E62F8
Previous Version: 1 DOI: 10.6073/pasta/8f9188936b08ef7cbe84ce3077487d6a
```

This year, the news is that Zotero has added a “dataset” item type. Datasets may be now entered in Zotero using this new type and will be listed accordingly in a Zotero-backed data catalog. The working group plans to integrate best practices for handling dataset items with Zotero into its documentation in the future.

Annotation/Keywording

This is a new working group tasked with revising the LTER Controlled Vocabulary (LTER CV) that is applied to most LTER datasets and generating improved dataset discovery tools at the EDI repository. The group is currently mapping LTER CV keywords found in EML to ontologies and controlled vocabularies that have become standards in the environmental and Earth sciences data communities. As this process takes place, the group is also designing a hierarchical search key for EDI datasets that will be the backend to a faceted search interface. Corinna reported on early progress and the overall vision for the working group during the morning session (refer to her slides for details).

Resources

Slides used for the main program of the IMC annual meeting are in Appendix A, with Working Group updates beginning here.

Business Meeting

There was one election for a position on the IM Executive committee (IM Exec) during the 2023 IMC meeting. Greg Maurer ended his term on IM Exec and there was one nominee to take the resulting open position. We also reviewed assignments to the DataBits editorial team and heard updates from three of our partner organizations: the LTER Network Office (LNO), EDI, and ILTER’s (International Long-Term Ecological Research) Information Management Committee.

Statements from IM Exec nominees

Li Kui (SBC):

> “I am thrilled to nominate myself for the Information Manager Executive position. As an experienced information management professional at SBCLTER, I have consistently enhanced our data management system, expediting data publication workflows, improving data quality for researchers and graduate students. With expertise in data governance, knowledge management systems, statistical analysis, and leadership, I deliver innovative solutions and engage effectively

> with researchers. Committed to staying up-to-date with emerging technologies, optimizing information management processes, and strengthening data security, I am ready to make my contribution after receiving valuable support from the IM committee over the past 5 years."

Election outcomes

Li Kui was elected to fill the open seat on the IM Executive committee..

Databits

Marina Frantz and Dan Bahauddin are continuing as the Databits editorial team until the next IMC Annual Meeting. There was general agreement that Databits editors can apply more pressure to members of IMC to get them to write articles.

LTER Network Office Report

Nick Lyon from the LTER Network Office (LNO) delivered an update about recent developments there. The LNO is supporting a large number of synthesis groups, and the newest addition to the program, the SPARC groups, seem to be off to a good start. Nick and Angel Chen serve as the scientific computing team that aids these synthesis groups in a variety of ways, including data preparation, QA/QC, programming, version control, and analysis. Nick reported that the data-focused support their team provides has come to be greatly valued by many of the synthesis groups. The scientific computing team has also created a set of workshops and training materials that may be valuable to IMs themselves or personnel at their LTER sites.

There are changes coming to the LNO-supported suite of LTER community tools. Discourse is being rolled out as an alternative to Slack and LNO would like IMC to help test and build out the use-cases for the platform. The community system called LTER Hub is also changing. Though the Salesforce based system will remain as the primary database for tracking LTER people and groups, LTER users will primarily interact with that system through Discourse, a demographic survey system (Qualtrics), and the LTER Network website. IMs can still update the database through the same Google Sheets based system.

In other news, the LNO grant will be up for renewal in 2024, so the team will be writing the proposal in fall 2023. Among the things that will come along with this renewal are a new Network-level DEIJ coordinator (which is already approved), and a proposed graduate-level course that focuses on synthesis and collaboration. Members of IMC are invited to contribute to the content of this course, and to participate as instructors and learners. Finally, Nick presented the new `LTERtools` R package, designed to be a library of useful functions for LTER data analysis and management. The LNO scientific computing team designed the package with a robust authorship system, and members of IMC are encouraged to contribute R functions that they think can be generally useful to the LTER community. Nick presented more on this in an IMKE Lightning Talk.

EDI Report

Paul Hanson and Mark Servilla from EDI delivered an update about recent developments at LTER's partner data repository. The most important news is that Corinna Gries is retiring soon and Paul Hanson has assumed the role of PI for the University of Wisconsin-based portion of the EDI grant. Fortunately for IMC, Corinna will be continuing to contribute to EDI and the LTER Network for the time being. In other personnel news, EDI has discontinued its summer fellowship program, but has begun a new graduate student internship program at University of Wisconsin that will train students in data management as they assist EDI in publishing incoming datasets.

EDI is also working steadily towards its sustainability goals with a number of new initiatives and partnerships. The repository is being written into collaborative research grants as a data management partner, which should provide funding for data publication assistance. These research grants include some to LTER sites, and unspecified others. In early 2023 EDI received Core Trust Seal accreditation for the first time, which will make it competitive as a repository partner with a greater number of academic journals. EDI has also been selected as a partner to the new TIP directorate at NSF to explore sustainable funding models for research data repositories. One new partnership that will lead to sustainable funding for EDI is a contract for environmental data management for the City of Seattle. This collaboration is currently in the process of setting up financial agreements between the city and University of Wisconsin.

Mark reported that infrastructure and software tooling at the EDI repository is rapidly advancing, with a host of new potential benefits to IMC. All servers underpinning PASTA+ and other repository systems have recently had operating system and core software upgrades to bring them up to date. EDI is also working on a proposal to further renew the repository infrastructure and enable key parts to operate in the cloud. The EDI Data Portal has added new functionality for tracking data package citations and providing better metadata to DataCite. The ezEML tool has gained a major new collaboration feature, as well as numerous smaller updates over the past 15 months. There have also been a host of feature updates to EDI systems, including the Data Explorer (DEx), EML parser/validator, and the EDI website, and further enhancements to repository systems and services are being planned.

ILTER Report

Renée Brown (MCM), who serves as the U.S. representative to, and co-chair of, the ILTER Information Management Committee, described what she has learned in the past six months of interaction with the ILTER Network. Renée described the scope, participation and governance of the Network at large, and then focused on initiatives within the ILTER IMC. Given that ILTER is a widely distributed and diverse network without oversight by a common funding agency, the organizational activity and priorities of the network are somewhat less defined than in the U.S. LTER Network, and this extends to ILTER's IMC. In fact, many ILTER sites are very small, with limited funding, and therefore have no data manager and do not publish any data. Nevertheless

there are certainly some opportunities for engagement and collaboration between the U.S. and ILTER IMCs.

The ILTER network maintains a data management platform called DEIMS (not directly related to the DEIMS CMS that some U.S. LTER sites use) where participating sites can list information about their research site, personnel, and research themes. LTER IMs should consider updating their U.S. LTER site information in this system, and Renée provided information about how to do this (see the slides in Appendix A). Another area for collaboration is around controlled vocabularies, ontologies, and other scientific metadata standards. Such standards are widely used and discussed in ILTER, and it makes sense for IMC to participate in creating and maintaining these standards alongside our international counterparts. Finally, the ILTER network has recently been assessing the FAIR status of network data, and this would be a natural area for collaboration between U.S. LTER and ILTER. Renée ended the update with a call for members of the U.S. IMC to reflect on and then implement new engagement and collaboration activities with ILTER.

Resources

Slides used for the main program of the IMC annual meeting are in Appendix A, with business meeting and partner updates beginning here.

IMKE Lightning Talks

Prior to the annual meeting, members of IMC were asked to submit ideas for a new segment of the IMC Annual Meeting - IMKE Lightning Talks. IMKE lightning talks are short presentations (~3 min) on a tool, technique, or trick for doing information management work at LTER sites. IMs were invited to prepare talks about software tools, web apps, file formats, metadata standards, database schemas, workflows, and other topics areas long as they could be short, to the point, and useful as a jumping-off point for interested IMs from another site. We received nine submitted talks, making this a relatively successful new feature of our annual meetings. All presenters submitted slides that are archived in **Appendix A**.

Units Working Group Activity

Resources

Slides for the IMC annual meeting update and workshop are in Appendix B .

EML Best Practices Working Group Activity

Resources

Slides for the IMC annual meeting update and workshop are in Appendix C.

Appendix A: Full-day meeting slides

Full meeting program slides are below (or Google Drive: 2023 IMC Annual Meeting). Slides from the Units and EML Best Practices WG activities are in Appendices B and C, respectively.

Welcome

Lets re-introduce ourselves

- My name is _____
- I'm at ___ LTER
- I'm excited about this ESIP session: _____
- Ask me about ____

____

Agenda

Time	Morning Activity
8:30	Welcome
9:00	Working group highlights
9:45	IM Exec election (1 member)
9:55	Volunteer for next Databits editor
10:05	Break
10:35	LNO report
10:50	EDI report
11:05	ILTER report
11:15	IMKE Lightning Talks
12:15	Lunch

Time	Afternoon Activity
13:45	Group Photo
14:00	Units WG Activity
15:00	Break
15:30	EML Best Practices WG Activity
16:30	Adjourn
18:00	Dinner

Working group updates

To review…

Working group charge: replace the ClimDB/HydroDB system with something modern & better

To do this we need:

1. A data format and metadata standard
   - *Ideally specific to sensor networks & timeseries data*
2. A platform to allow discovery & access of LTER met/hydro data
   - *Web based and well integrated with EDI, preferably*
3. Time and effort to standardize and upload LTER datasets
   - *And we're all pretty busy….*

So, where are we?

1. Data format and metadata standard

We initially chose ODM 1.1 CSV format

EDI wrote an R package to convert met/hydro data to ODM format

- Needs a little work, but a good starting point

However, ODM is losing support as a standard

We need to co-develop something that works for *LTER IMs, EDI, and Dendra*.

- ODM will be one option here…
- Dendra has some opinions

We will need your feedback to determine next steps

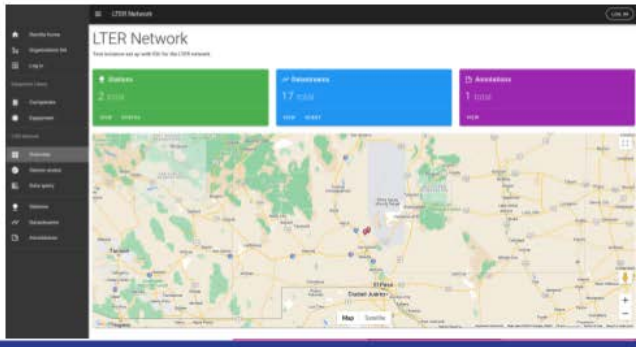
2. The discovery/access platform

Web-based that integrates with EDI

We found no surplus web developers in IMC or at EDI, but we found Dendra

- **EDI:** HyMet datasets published with DOI
- **Dendra:** Discovery & access
  - Other functions might include publication to EDI

Pilot at https://dendra.science/orgs/lter (JRN only)



3. Time and effort to standardize datasets

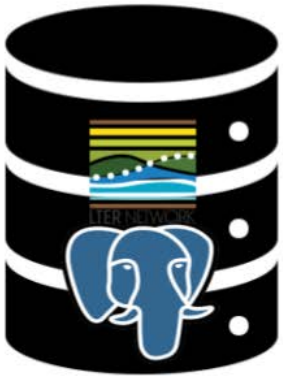
There are 7 HyMet datasets already on EDI

We need to support someone in this work

Dendra also needs support to integrate with EDI

We're looking for a funding opportunity to support this work (spring 2024 submission, meetings with NSF POs this fall)

LTER Core Metabase

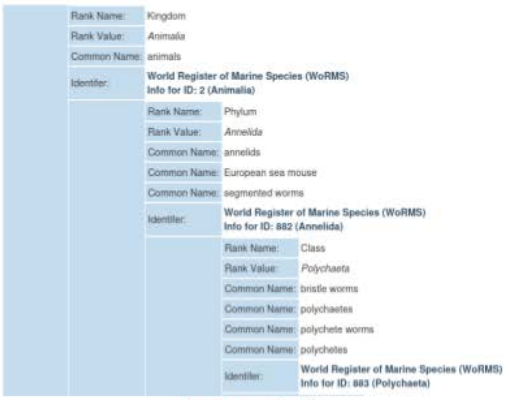
- 

WG members: Greg Maurer, An Nguyen, Tim Whiteaker, and anyone else using some iteration of metabase.

LTER Core Metabase

Not very much to report since last September!

- Taxonomic coverage and authority identifiers are now working better!
- Our R tooling has improved
  - See Greg's IMKE Lightning talk for a demo
- An may have other updates and we are still working on a VWC.
- We need more help… join us!



Rank Name:	Kingdom
Rank Value:	*Animalia*
Common Name:	animals
Identifier:	**World Register of Marine Species (WoRMS) Info for ID: 2 (Animalia)**

Rank Name:	Phylum
Rank Value:	*Annelida*
Common Name:	annelids
Common Name:	European sea mouse
Common Name:	segmented worms
Identifier:	**World Register of Marine Species (WoRMS) Info for ID: 882 (Annelida)**

Rank Name:	Class
Rank Value:	*Polychaeta*
Common Name:	bristle worms
Common Name:	polychaetes
Common Name:	polychete worms
Common Name:	polycheles
Identifier:	**World Register of Marine Species (WoRMS) Info for ID: 883 (Polychaeta)**

(dataset link)

EML Best Practices WG

- We're updating the EML Best Practices document
- You'll hear much more later today

Units!

We've made some substantial progress in developing resources and tools for annotation of attribute units, but we need your advice on tools and your help on improving our ability to link raw units to units in an existing ontology (QUDT).

SEE YOU AT OUR AFTERNOON SESSION!

Zotero for Data

Extra:

Type: dataset

Version: 2

DOI: 10.6073/PASTA/F7204A847A1D71FCE18ED880363E62F8

Previous Version: 1 DOI: 10.6073/pasta/8f9188936b08ef7cbe84ce3077487d6a

BLE example

BLE result

Plus see NES IMKE slide

Best practices at https://bit.ly/ZOTERO_BP

Annotation / Keywording Working Group

John H. Porter, Colin Smith, Margaret O'Brien, Christine Laney, Greg Maurer, Tim Whiteaker, Mary Martin, Renée F. Brown, Chris Turner, Eric Sokol, Nick Lyon, Marty Downs, Corinna Gries

Goals

I need data to predict blue-green algae blooms affecting beaches and drinking water supplies

Assign standardized terms/concepts that describe a dataset in several categories (ecosystem, process, substance, units etc.)

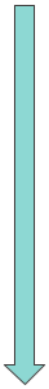
Faceted Search

Harmonize and simplify keywords to higher level categories

Graph query, reasoner

Map terms used to concepts in existing ontologies

Process



Gather keywords used to describe datasets

LTER controlled vocabulary
Analysis of metadata
Maps (ecosystems, vegetation, place names)
Taxonomic standards

Map to appropriate ontology terms

Map to higher level concepts

Organize higher level concepts

Assign standard terms to datasets

Search metadata and assign higher level keywords (NLP)
Geographic map analysis
Unit mapping

Apply NLP principles and general rules

Trophic would find heterotrophic - \Wtrophic\W

Vertebrate would find invertebrate - \Wvertebrate\W

Estuar - estuary, estuaries, estuarine

Exclude many place names, e.g., Harvard Forest is not all forest ecosystem

No 'tropical' ecosystem in Hubbard Brook

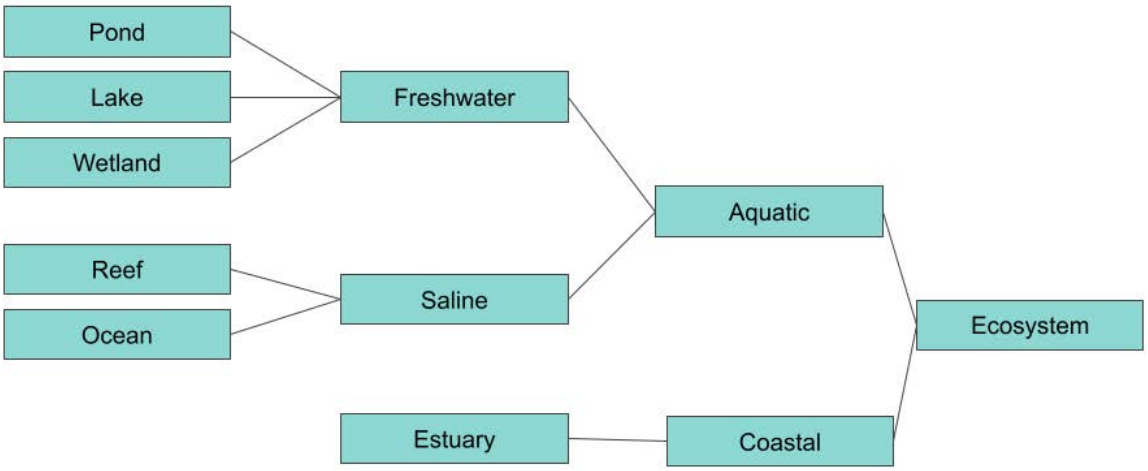
Harmonize spelling

estuaries	ENVO:00000045	estuary	http://vocabs.lter-europe.net/EnvThes/21807	estuary
estuarine	ENVO:00000045	estuary	http://vocabs.lter-europe.net/EnvThes/21807	estuary
estuary	ENVO:00000045	estuary	http://vocabs.lter-europe.net/EnvThes/21807	estuary

Map to higher level concept

bog	ENVO:00000044	peatland	http://vocabs.lter-europe.net/EnvThes/21798	bog	wetland
fen	ENVO:00000232	fen			wetland
marsh	ENVO:00000035	marsh	http://vocabs.lter-europe.net/EnvThes/21801	marsh	wetland
peatland	ENVO:00000044	peatland	http://vocabs.lter-europe.net/EnvThes/21803	peatland	wetland
wetland	ENVO:01001209	wetland ecosystem	http://vocabs.lter-europe.net/EnvThes/21797	wetland	wetland

Concept hierarchy

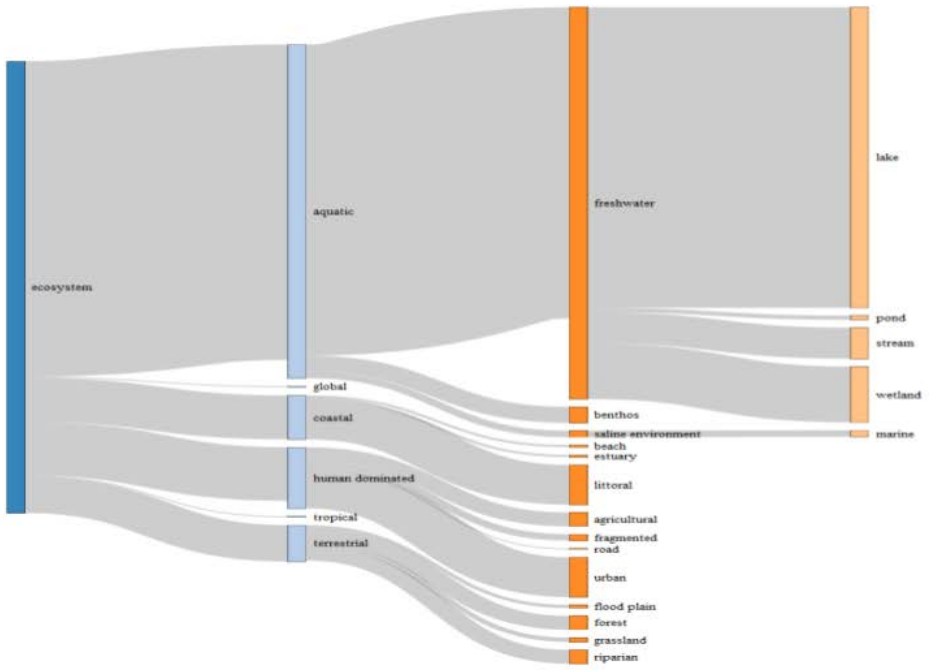


Initial Results

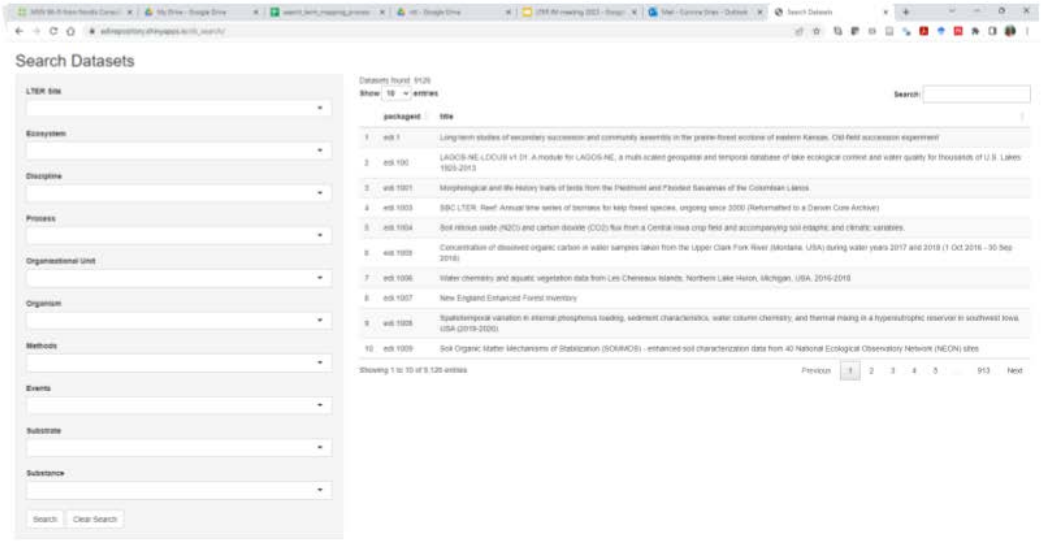
https://drive.google.com/drive/folders/1X6ucoozr_86aeGQyWzrVPDxTLOlmagWA?usp=sharing

discipline	ecosystem	events	methods	org_unit	organism	process	substance	substrate	title
aquatic ecology, biogeochemistry, species interaction, limnology, water biogeochemistry, plant cover	aquatic, freshwater, lake, wetland	NA	NA	NA	animal, vertebrate, fish	hydrologic process, ice	ions, nitrogen compounds, phosphorus compounds, inorganic matter, nutrient, organic matter, carbon, ammonia, ammonium, nitrate, nitrogen, phosphorus, silicon, inorganic nutrient, bicarbonate, silica	air, water, air, ice	North Temperate Lakes LTER: Chemical Limnology of Primary Study Lakes: Nutrients, pH and Carbon 1981 - current

¶¶¶



https://edirepository.shinyapps.io/ds_search/



ChatGPT

Soil Respiration at Forest Edges along an Urban to Rural Gradient in Massachusetts since 2018

geography, meteorology, soil science, human dominated, terrestrial, ecosphere event, experiment, biogeochemical process, environmental process, physiological process, ions, landuse, climatology, temperature, soil moisture, soil temperature, fragmented, urban, forest, disturbance, field manipulation, carbon, decomposition, disturbance, respiration, carbon, warming

Soil Respiration, Carbon, Carbon Dioxide, Decomposition, Land Use, Microbes, Soil Moisture, Soil Temperature, Urbanization, Forest Fragmentation

Any other working group updates?

A question before we move on

Do we need a spatial data & GIS working group?

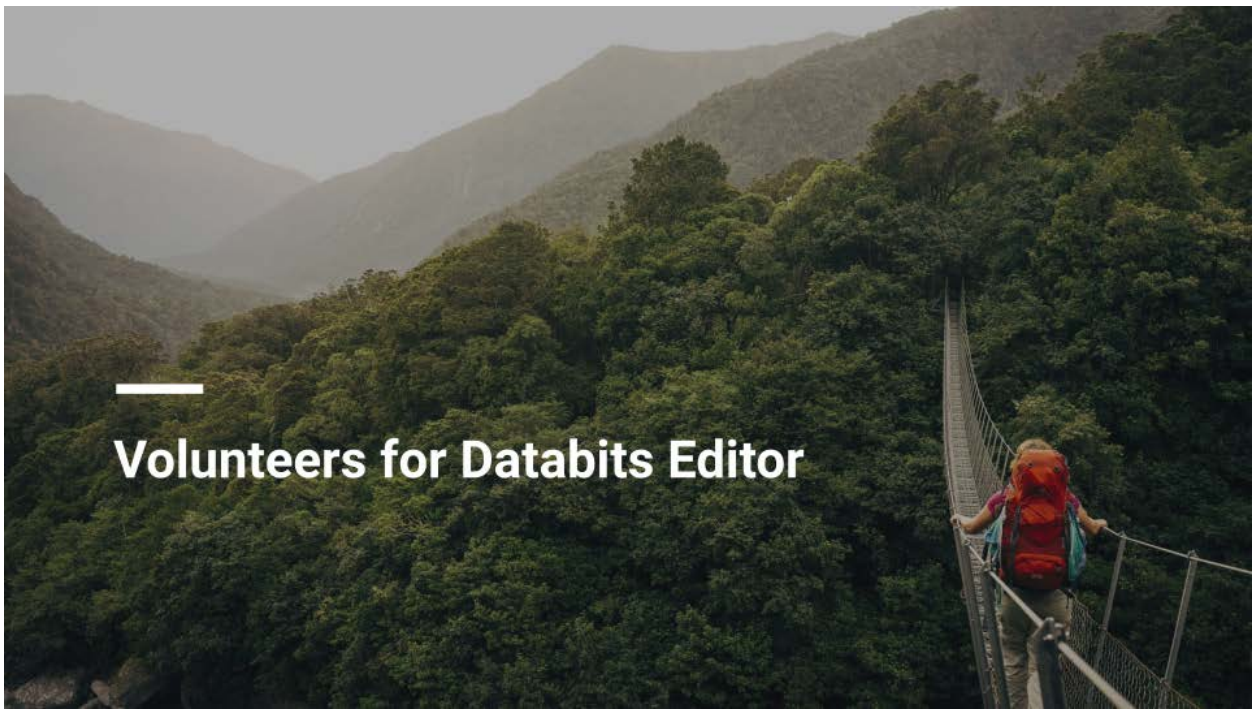
What about a "next generation tech" working group (AI methods, cloud infra, etc.)

Any other working group ideas?

Business meeting

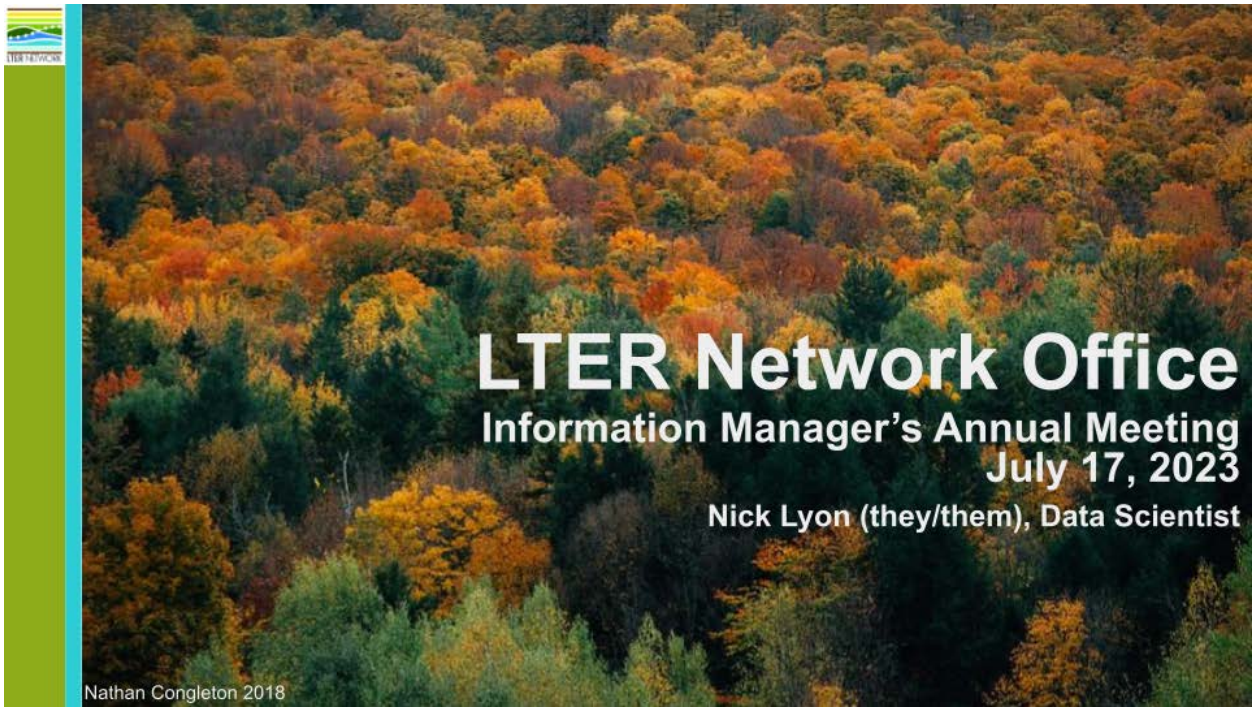
Election for IM Executive Committee member

- The job: plan VWCs and the annual IMC meeting, prepare reports, coordinate IMC activities
- Meet monthly via Zoom
- One opening, 3-year term (2023-2026)
- One nominee: **Li Kui**



Break Time!

Return in 30 minutes



Synthesis Group Activities

New Groups!

- 4 SPARC Groups (1 in-person meeting)
  - SPARC = Scientific Peers Advancing Research Collaborations
- 3 full working groups (~4 in-person meetings)
- 5 pre-existing working groups

In-Person Meetings

- 1 SPARC already met, 1 this week, 1 next week and final one in end of September
- Working groups all through October & December

Scientific Computing Support


Angel Chen


Nick Lyon

Data-related support for working groups

- Quality Control (QC) and data tidying
- Analytical workflows
- Publication-quality figure creation

Our process:

1. Build relationship with group
2. Assess possible needs
3. Share examples of previous work
4. Emphasize open communication lines
5. Perform any requested task quickly, transparently, and reproducibly

Wrangling Workflow Development


Angel Chen


Nick Lyon

SPARC Soil Phosphorus Control of C & N | github.com/lter/lter-sparc-soil-p

- Task: harmonize many idiosyncratic datasets
- Developed a “column key” approach
  - Flexibly identified synonymous columns
  - Minimal labor for group leads / participants
- Focus on maintaining a sustainable level of effort for maintenance

Controls on River Silica Exports | github.com/lter/lterwg-si-export

- Task: identify watershed shapes and extract various spatial / climatic data
- Created a “cookie cutter” workflow
  - Used global watershed sub-polygons to identify complete watershed shape for all sites
  - Leverage that shape to “cookie cut” any spatial data of interest

Scientific Computing Support


Angel Chen


Nick Lyon

Environmental Drivers of Plant Reproduction | github.com/lter/lterwg-plant-repro-synchrony

- Task: create and organize a standalone repository for a particular paper
  - Previous GitHub repository contained scripts spanning multiple papers
  - Needed only one paper’s code split off in preparation for sharing / publication
- Discussed group’s desire for balanced transparency and data privacy
- Provided suggestions and formatting tips for maximum efficacy

Offered 6 Workshops | nceas.github.io/scicomp.github.io/workshops

- Collaborative Coding with GitHub
  - Target audience: beginner – intermediate prior experience
- Coding with the `tidyverse`
  - Target audience: beginner

Collaboration & Community Update

Rolling out new platform: **Discourse** (lternet.discourse.group)

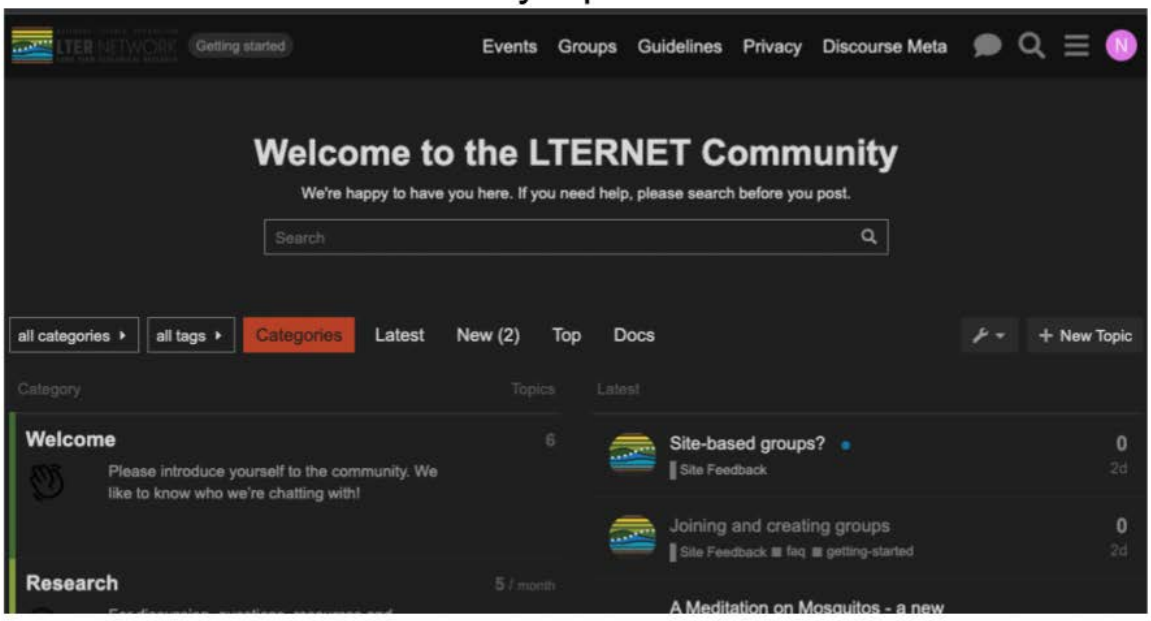
Discourse has many functionalities:

- Discussion forum with group and topic organization tags
- Centralized event calendar for LNO events
- Venue for announcements / position advertisements / training opportunities

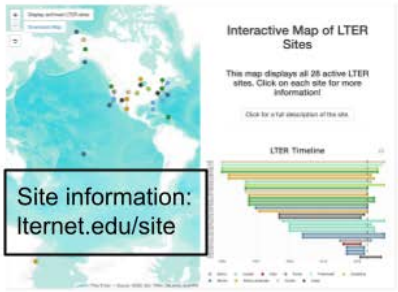
You should have received an invite link from Marty, please join us!

- Still early-days but we think its role will grow with the user base

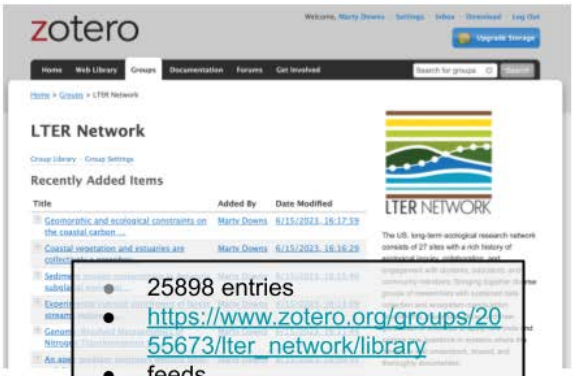
Collaboration & Community Update



Network Information



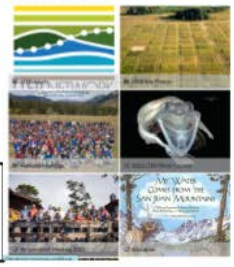
Site information: lternet.edu/site



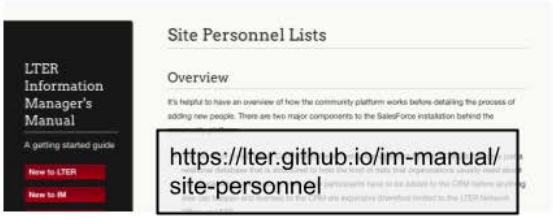
- 25898 entries
- https://www.zotero.org/groups/2055673/lter_network/library
- feeds https://lternet.edu/bibliography/

SmugMug

LTER Network

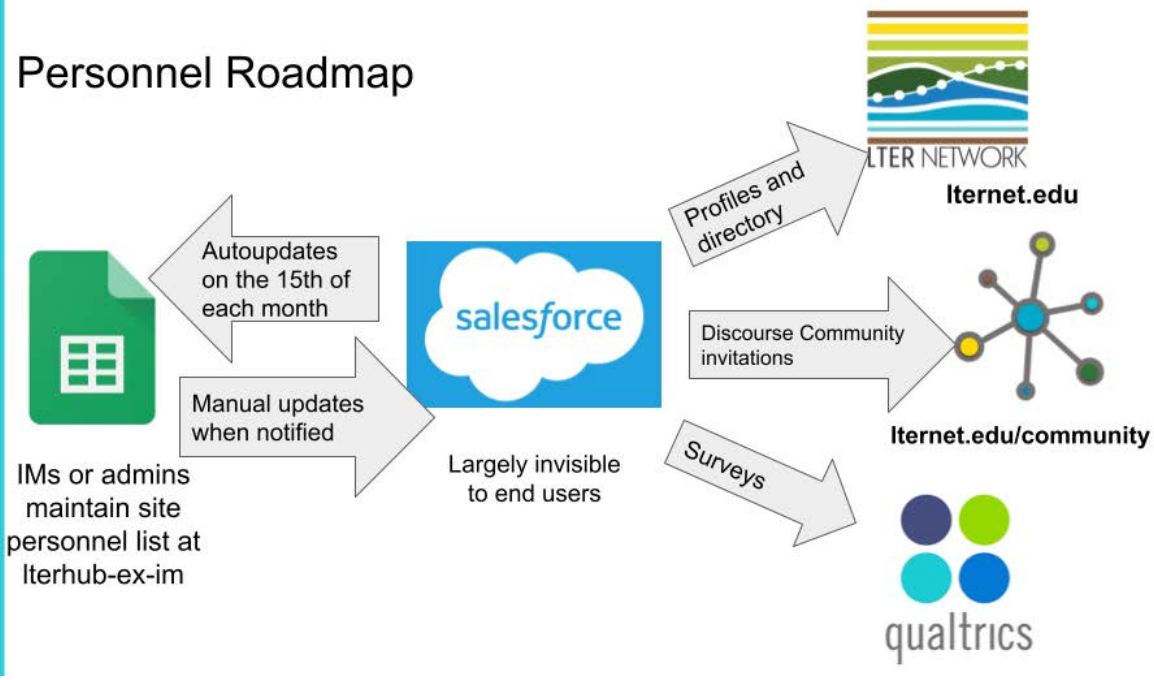


https://lternetwork.smugmug.com/



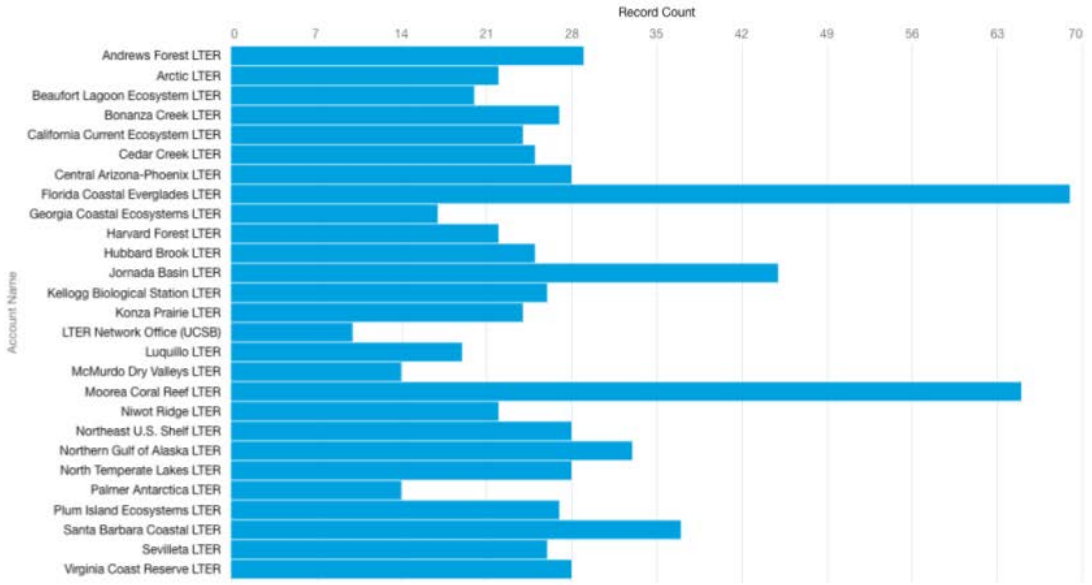
https://lter.github.io/im-manual/site-personnel

Personnel Roadmap

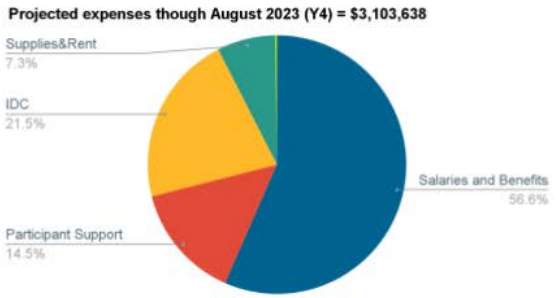


Demographic Responses

Demographic Response Rate



Budget / Upcoming Renewal



LNO Renewal in 2024

- Writing proposal in fall 2023

Noteworthy Elements

- New DEIJ Coordinator position approved
- Proposing new graduate student course
  - Focus on synthesis and collaboration
  - If IMs interested in participating, we're happy to include you in the planning stages!
    - Contribute modules
    - Refine content
    - Guest lecture
    - Attend as a learner

Happy to discuss more!

New R Package: `ltertools`

Hex Logo Under Construction!

Rationale

- We (likely) face similar data challenges
- Individual's custom functions are often useful to others
- Pre-established architecture of R package reduces sharing barriers

Robust authorship credit for contributing

- Author on package
- Name & website of choice built into function documentation
- See pkgdown for details: **lter.github.io/ltertools**

More details in IMKE Lightning Talk!

EDI Developments Supporting Site Data Management

EDI team

Personnel

- WI Principal investigator change
- Fellowship program change
- Graduate student internship experiment
- Collaborations with funded projects
- Analysis ready data initiatives

Scope growth

- Core Trust Seal approved Spring 2023
  - Relevance: Accreditation, some journals require
- EDI selected by NSF TIP (Technology, Innovation and Partnerships) to explore sustainability pathways
- Seattle collaboration

Infrastructure stability

- Ubuntu 18.04 LTS to 22.04 LTS (support to April 2027)
- Tomcat 8 to 9
- Python at version 3.11
- System clocks now set to UTC from Mountain/Denver Timezone

Data Portal & PASTA Feature Upgrades

- Journal Citations UI improved, along with Crossref integration to populate citation fields
- UI notifications for Crossref and DataCite outages
- Data package zip file downloads now stream (almost) immediately
- Search results can now be downloaded as CSVs
- IDs (ORCID, ROR, ISNI and GRID.AC) now added to DataCite metadata
- PASTA Gatekeeper service switched from Java to Python

Reminder - updates from fall 2022 report at ASM:

- Audit report UI improved with pagination
- Audit reports can now be downloaded as CSVs
- Text, DocBook XML, and Markdown are now rendered on landing page abstract and other elements

ezEML - New Collaboration Features - *No more sending files back and forth!*

- Peer to peer
  - Invite others to collaborate on a package
  - Peers edit the package as if it were in their own ezEML account
- With EDI Curators group
  - Submit Package to EDI sets up a collaboration with the EDI Curators group
  - Curators can review and edit the package in place before submitting it to the EDI repository

ezEML - Other features added in past 15 months

- Check Data Tables - quality checks on data tables' contents and fidelity to the metadata
- Enhancements to Check Metadata - check DateTime formats with recommended formats
- Load taxonomic coverage from CSV file - load your taxa in a bulk operation
- Structured Data Source info - for documenting data provenance
- Manage Data Packages page - see and manage all of your packages in a single page
- Clone Column Properties for tables with many columns - column selection streamlined
- Templates - to pre-populate metadata with values tailored to your site
- Fetch packages directly from EDI repository - fetch and open them in ezEML

Other Ongoing Upgrades

- Data Explorer (DeX) provides insights to tabular based data
- EML Validator/Parser for XML validation and attribute resolution (Python)
- New project website - https://edirepository.org
- Deprecation of environmentaldatainitiative.org website and email addresses

What we are discussing

- Proposal to renew infrastructure
- Implementation of annotations and other value adding information to be used by the system
  - Shadow EML
- Sustainable funding strategies

Always interested in discussing ways to help LTER IMs!

INTERNATIONAL LONG TERM ECOLOGICAL RESEARCH

ILTER Information Management

Renée F. Brown, McMurdo Dry Valleys LTER
Co-chair ILTER IMC

- Network of 39 member networks that together operate 750+ research sites
- Focus on long-term, site-based research and monitoring
- Relies on financial contributions from its members
- Governance (simplified version):
  - Executive Committee
  - Coordinating Committee
  - 2 Standing Committees: Science and Information Management
- Open Science Meeting
  - Next meeting tentatively scheduled for 14-19 October 2024 in Kunming, China

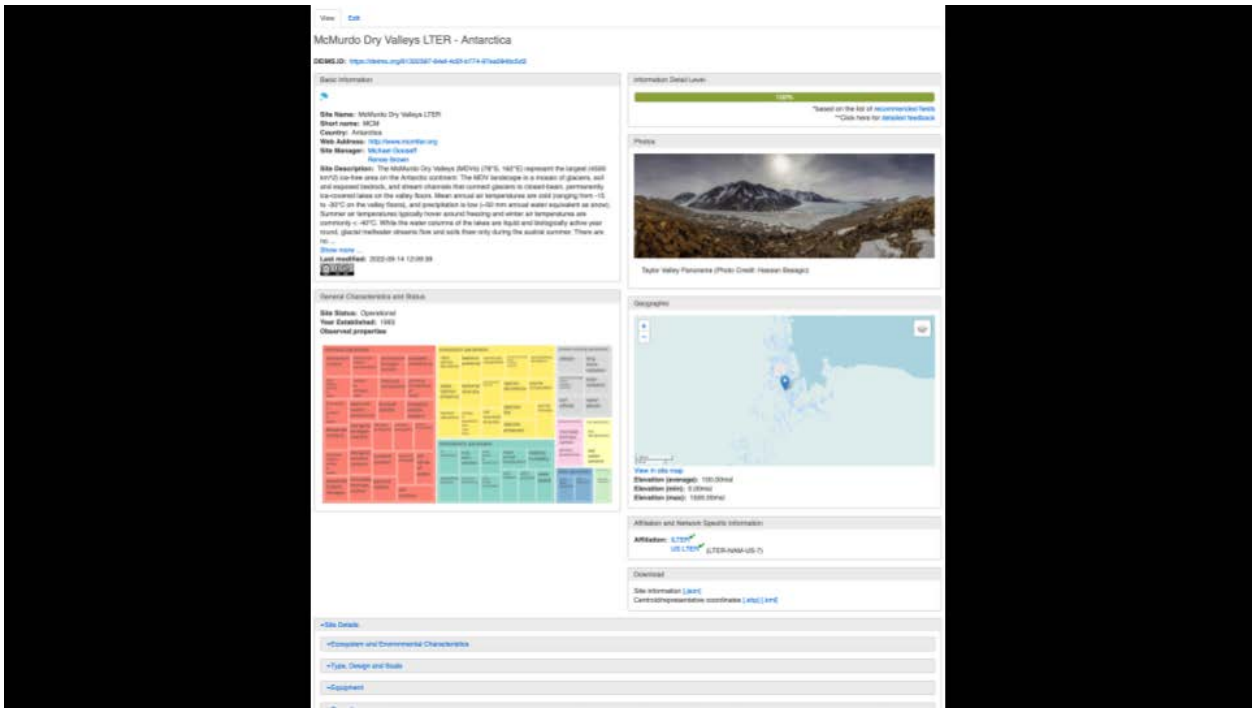
ILTER Information Management Committee Goals

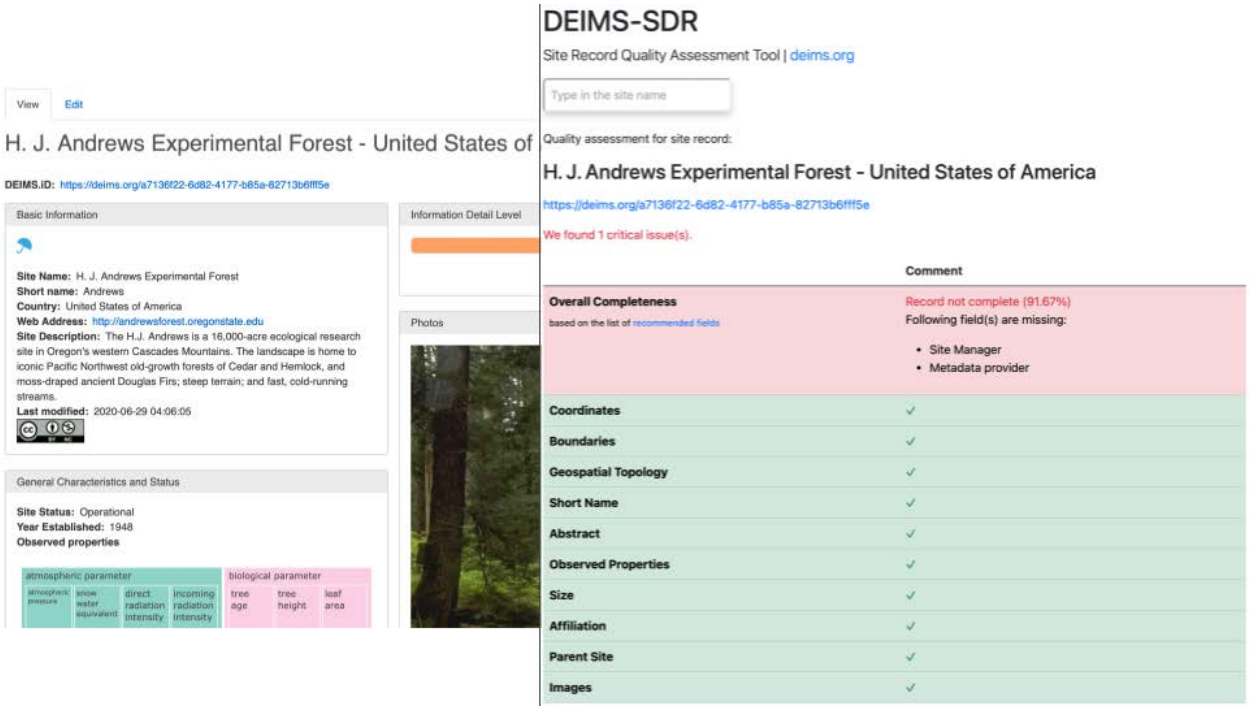
- Establish an ILTER Data Management Framework
  - *Define a federated shared infrastructure and the specifications and standards associated with it, along with policies and best practices*
- **Develop and maintain an open registry of ILTER research sites, variables under observation, and instrumentation (DEIMS-SDR)**
- **Survey members with respect of their readiness to contribute site and data catalogue information into a federated data structure**
- Link ILTER to other international initiatives such as GEO and RDA

- The **Dynamic Ecological Information Management System (DEIMS) – Site and Dataset Registry (SDR)** is a web portal where one can discover and register information about ILTER sites.
- **IM Responsibilities:**
  - Review & update your site's info annually
  - Recommend including IM name and contact info in both the *Site Manager* and *Metadata Provider* fields
  - Keep a record of this information in your site's IM documentation
- Update your site info at **https://deims.org/**
  - **Username: US**
  - **Password: ?US80**

DEIMS-SDR

Year Last Updated	Site Code
2022	JRN, KNZ, MCM, SEV
2021	*CWT**, FCE, LUQ
2020	AND, ARC, *BES**, BLE, BNZ, CAP, CCE, CDR, GCE, HBR, HFR, KBS, MCR, NES, NGA, NTL, NWT, PAL, PIE, SBC, VCR
Not verified	MSP





Inside ILTER – survey on network level data management capabilities

Johannes Peterseil[1]

johannes.peterseil@umweltbundesamt.at

Ivo Offenthaler[1]

christoph.wohner@umweltbundesamt.at

& Information Management Committee Team Members

[1] Environment Agency Austria

ILTER conference

Information Technology Workshop

12 October 2022 9:00-10:20

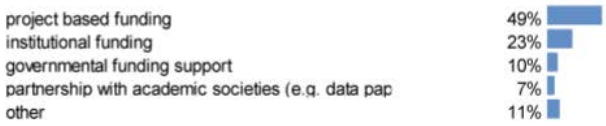
Novi Sad, Serbia

Organisational capabilities

- **Dedicated personnel** for data management is still scarce, e.g. done by the research staff themselves, or with a limited amount (less than 1 person/FTE)
- **Funding** is done mainly on project based funding or by in-kind contributions by the organisations
- □ funding of data management and curation is critical to the implementation and sustainability of ILTER as a data provider

Do you have dedicated personnel for data management, curation and development in your LTER network?

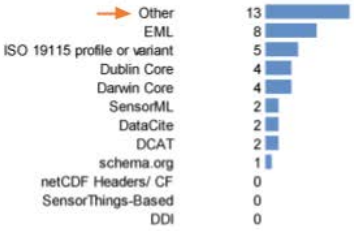
Do you have dedicated funding or resources for information / data management (incl. data publication) for your LTER network?



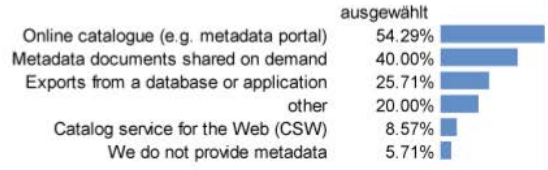
Comments: no funding, not planned, unfortunately none

Findability

Which metadata content standards does your LTER network support?

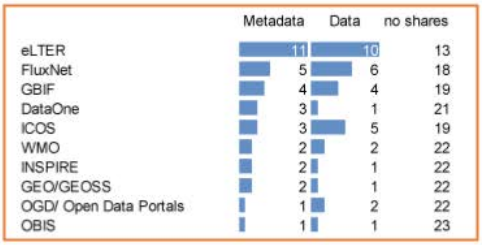


How does your LTER network provide metadata for your data?



Accessibility

Do you currently share metadata and/ or data with other regional or international data networks?

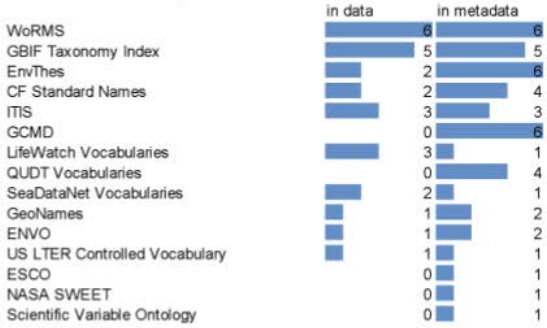


e.g. ICP Forests, ICP Waters, LifeWatch, Copernicus, Eco-Bank, EMODNET, Friend-ERB, google data search, OceanDataStandards.org, Interact, DIAS, Pangea, WGMS, etc.

- □ Data portal, download service, some machine-2-machine services
- □ Provision to networks through dedicated workflows

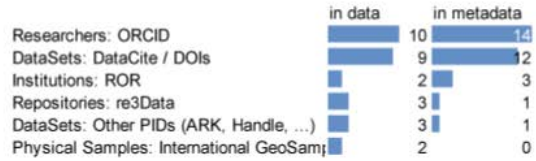
Interoperability

Which externally and publicly available domain vocabularies does your LTER network use?



Others: DEIMS, ECSO, http://www.oceandatastandards.org, https://dwc.tdwg.org/, https://www.bodc.ac.uk/resources/vocabularies/parameter_codes/, PANGEA, WGMS

Which Linked Open Data Services does your LTER network use?

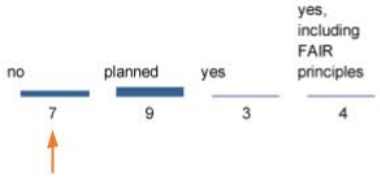


Others: e.g. CWB (https://opendata.cwb.gov.tw/devManual/insrtuction), LSID for scientific names (
Part of data stored in JaLTER data base has DOI if they are published as data paper on "Ecological Research".

☐ Use of persistent and resolveable identifiers to enable linking (e.g. provenance) and citation of datasets and information

Re-usability

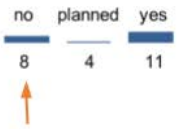
Does your LTER network have a dedicated Data Management Plan?



Does your LTER network provide a dedicated network level Data Policy?

Does each member site in your LTER network have a dedicated Data Management Plan?

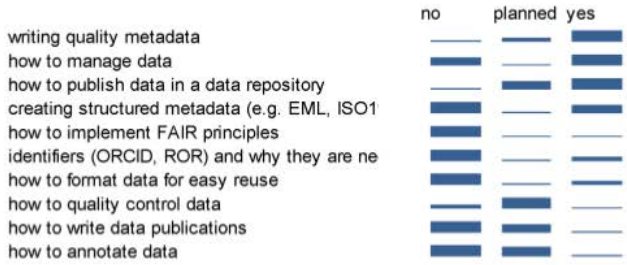
Does your LTER network provide a dedicated network level Data License?



ILTER Conference 2022 - 12 October, Novi Sad, RS

Training

Does your network offer training for member scientists about how to document and format data so that it can be published? If yes, which of these topics does your training cover?



Challenges and Opportunities

- How to encourage and incentivize good data management practices when many member sites have little or no resources?
- What roles can the US LTER IMC play with respect to ILTER efforts, and vice-versa?
- US LTER International Committee wants to establish an "Americas LTER Network" like the eLTER (European LTER Network) - Could US LTER IMC help set a good example from the start?
- Is there a role that EDI could play with respect to an Americas regional network and/or to under-resourced ILTER member sites?
- Ideas, questions, comments? Let's discuss!

IMKE Lightning Talks

IMKE Lightning Talks

IMKE: Using Zotero for NES-LTER Data Products Catalog

Find [link] on NES-LTER Data webpage https://nes-lter.whoi.edu/data/

Pros:

- Catalog data products in many different repositories
- No need to set-up or maintain infrastructure
- Widely used and easy-to-use app
- Using a bibliographic / reference manager could encourage citation

Pro (& Con):

- Multiple users can contribute

Con:

- Relying on external cloud service

Another puppy

NORTHEAST U.S. SHELF
Long-Term Ecological Research

Stace Beaulieu &
Kate Morkeski,
LTER Network IM Meeting,
July 2023

Miguel Leon (he/him/él)

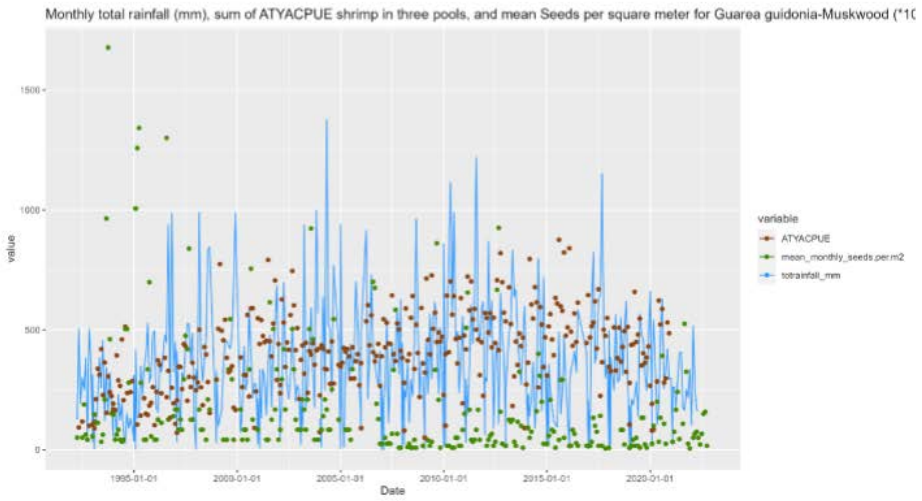
Luquillo LTER Information Manager (IM)

University of New Hampshire

LUQ LTER Signature dataset

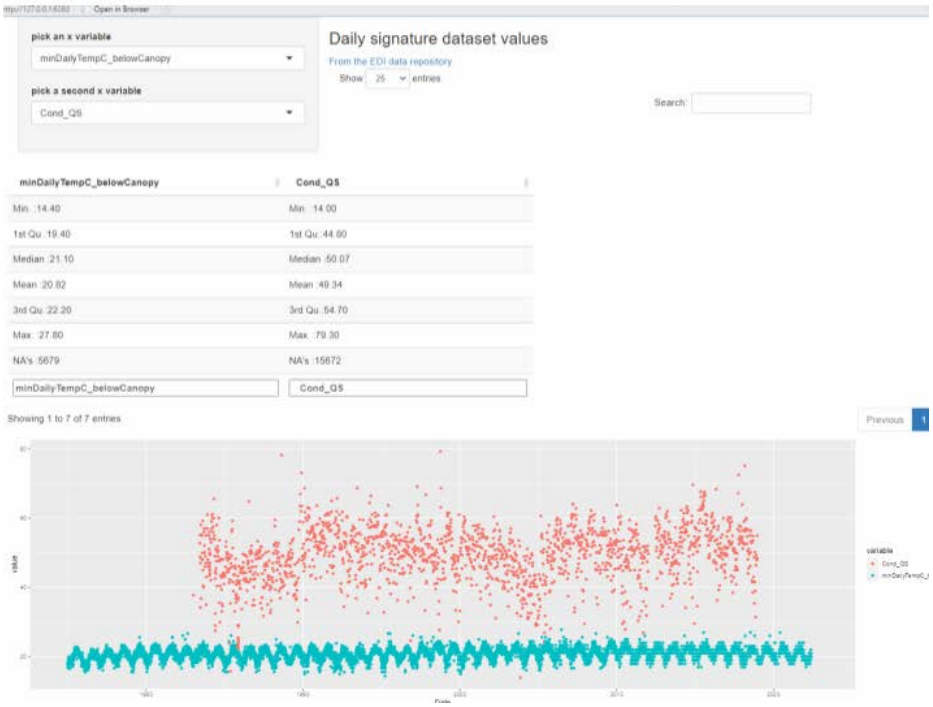
- Long term datasets, targeting datasets with ~30 years of data at El Verde
  - Aggregated to Daily and Monthly
- Precipitation, Max, Min temperature, manual measurements,
- Stream chemistry at Quebrada Sonadora and Quebrada Prieta
- Rainfall chemistry
- LFDP Bird counts
- Quebrada Prieta Shrimp population counts
- Phenology
- Coming – Gastropods, walking sticks, Litterfall

Signature dataset



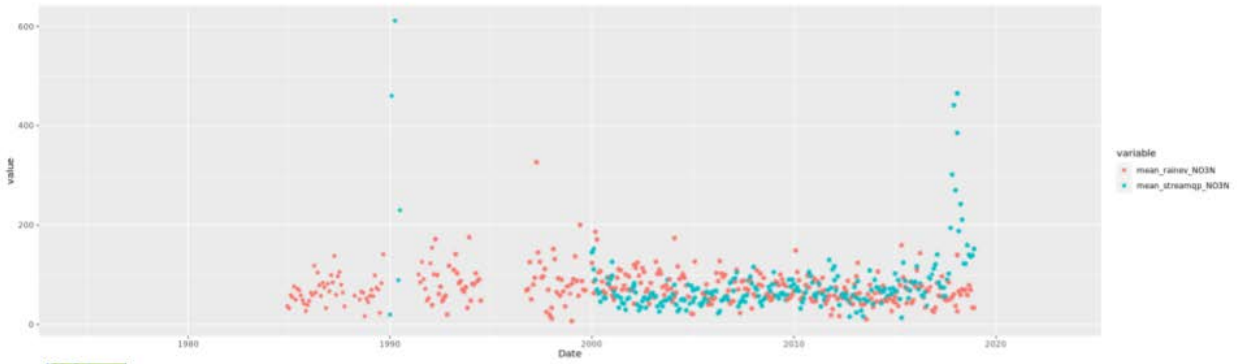
Signature dataset

- Shiny App -



Signature dataset

mean_streamqp_NO3N	mean_rainev_NO3N
Min. : 14.00	Min. : 7.0
1st Qu.: 65.75	1st Qu.: 44.0
Median : 80.00	Median : 69.5
Mean : 107.06	Mean : 75.4
3rd Qu.:126.00	3rd Qu.: 97.0
Max. :611.00	Max. :326.3



TEMPLATES

Adam Sapp

2023 IMC Annual Meeting

Excel Data Templates

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)
- Great way to do initial QA/QC and constrain data at collection

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)
- Great way to do initial QA/QC and constrain data at collection

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)
- Great way to do initial QA/QC and constrain data at collection

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)
- Great way to do initial QA/QC and constrain data at collection

Excel Data Templates

- Inspired (in part) by the Field Data Application VWC (April 2022)
- Great way to do initial QA/QC and constrain data at collection
- Experimented with web forms
  - Errors
  - Uncomfortable

Excel data template

Please DO NOT edit the format of this spreadsheet. If you need to make changes, contact the GCE IM office.

Monthly Sample Data Sheet

CruiseID: _____
Start Date: _____ End Date: _____
Sampler: _____
Dissolved samples: PO4, Nox, TDN/DON, DOC, PO4, TDP/DOP, Si, NH4
Particulate samples: PC/PN, Chl, TSS

			Sample Identification			Sample Processed	PC/PN		Chl		TSS		Dissolved (n=8)	pH
Station	GCE Site	Tide	Depth	Time (GMT)	Date	Time (GMT)	Petri Dish #	mLs	Tube #	mLs	Petri Dish #	mLs	Bottle #	mV
DB+03	GCE 6	H	0.5m											
AL+24	GCE 7	H	0.5m											
DB+03	GCE 6	L	0.5m											
-	Field Blank	-												Temp °C
-	Lab Blank	-												

Sample processing notes: MQ water is from UGAMI Lab 5 (Room 116) . Samples were processed on the same date they were taken in the field.

Calibration Standard Readouts (mV):

	mV
pH 4	
pH 7	
pH 10	

Excel data template

Please DO NOT edit the format of this spreadsheet. If you need to make changes, contact the GCE IM office.

Monthly Sample Data Sheet

CruiseID: LTER0723
Start Date: 7/17/2023 End Date: 7/17/2023
Sampler:
Dissolved samples: PO4, Nox, TDN/DON, DOC, PO4, TDP/DOP, Si, NH4
Particulate samples: PC/PN, Chl, TSS

			Sample Identification			Sample Processed	PC/PN		Chl		TSS		Dissolved (n=8)	pH
Station	GCE Site	Tide	Depth	Time (GMT)	Date	Time (GMT)	Petri Dish #	mLs	Tube #	mLs	Petri Dish #	mLs	Bottle #	mV
DB+03	GCE 6	H	0.5m				4929A		4929		4929B		2000	
							4930A		4930		4930B			
							4931A		4931		4931B			
AL+24	GCE 7	H	0.5m				4932A		4932		4932B		2001	
							4933A		4933		4933B			
							4934A		4934		4934B			
DB+03	GCE 6	L	0.5m				4935A		4935		4935B		2002	
							4936A		4936		4936B			
							4937A		4937		4937B			
-	Field Blank	-											2003	Temp °C
-	Lab Blank	-											2004	
														mV
Sample processing notes: MQ water is from UGAMI Lab 5 (Room 116) . Samples were processed on the same date they were taken in the field.									Calibration Standard Readouts (mV):				pH 4	
													pH 7	
													pH 10	

Steve's Nag Machine

Adam Sapp

2023 IMC Annual Meeting

Automated Workflow Tracking System

(Called 'Steve's Nag Machined' by everyone but Steve)

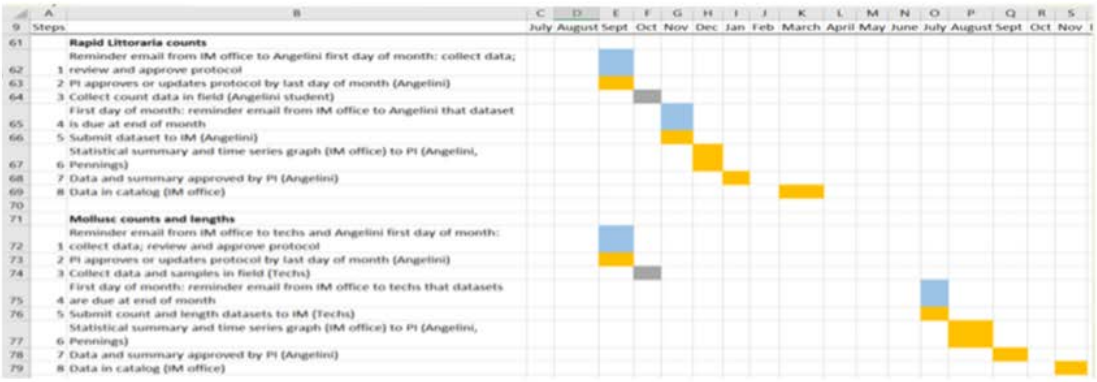
Tracks different parts of our fall monitoring study

Fall Monitoring each October - 10 marsh sites with ~20 plots each sampled for:

- Plant species, height and density
- Barnacle settlement
- Mollusc length and abundance
- Littoraria count
- Crab hole count
- Soil salinity and organic matter
- Grasshopper count

Each data set has a different timeline from collection to catalog.

Automated Workflow Tracking System



Automated Workflow Tracking System

(Called 'Steve's Nag Machined' by everyone but Steve)

Tracks different parts of our fall monitoring study

Fall Monitoring each October - 10 marsh sites with ~20 plots each sampled for:

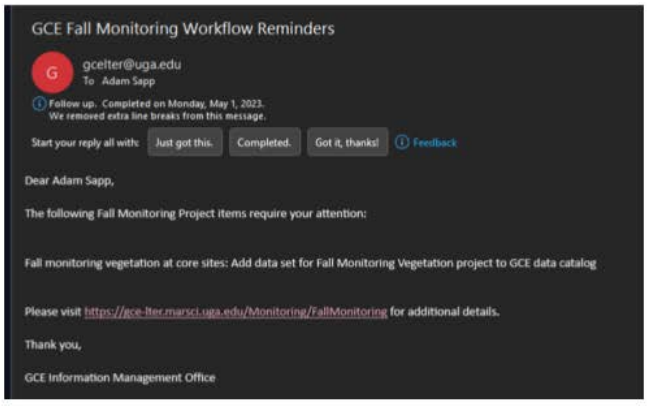
- Plant species, height and density
- Barnacle settlement
- Mollusc length and abundance
- Littoraria count
- Crab hole count
- Soil salinity and organic matter
- Grasshopper count

Each data set has a different timeline from collection to catalog.

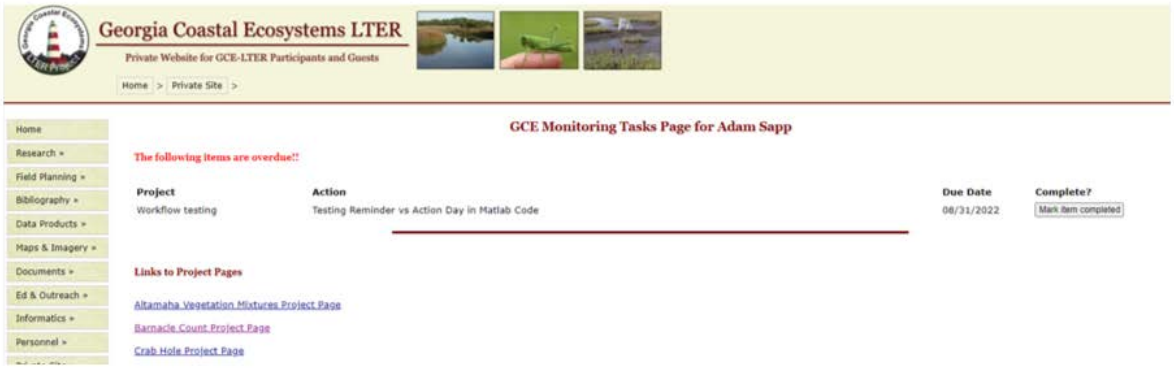
Ensures protocols are reviewed and updated regularly

Ensures monitoring data are submitted and reviewed in a timely manner

Automated Workflow Tracking System



Automated Workflow Tracking System



Rapid publishing of big datasets to EDI with metabase and some R packages

Greg Maurer, with support from LTER-core-metabase WG

Use cases

You manage metadata with metabase and need a way to create EML and publish datasets.

Sending large datasets to EDI (500MB limit on data entity uploads)

Building scripted (in R) data publication workflows that create/update datasets with a few commands



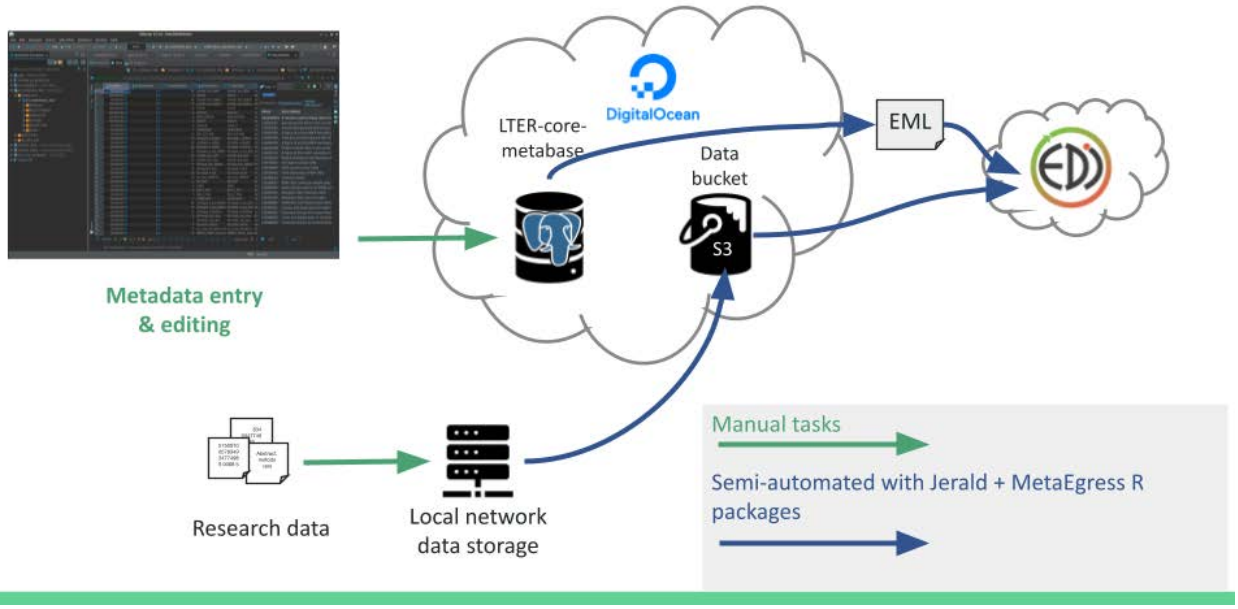
Requirements

LTER-core-metabase – relational database schema (PostgreSQL) for populating EML

MetaEgress R package – create EML documents from a metabase with R (An Nguyen, BLE)

jerald R package – R-based workflows to publish datasets to EDI using APIs & cloud infrastructure (Greg Maurer, JRN)

EML, **EDIutils**, **aws.s3** R packages are dependencies



Live demo (danger!)

Its just that easy!!!

TO DOs

- More options to specify data entity locations and upload method
- Check MD5 hashes to avoid unnecessary data entity transfers
- Import EML documents into a metabase
- Refactor MetaEgress & jerald into one big R package?

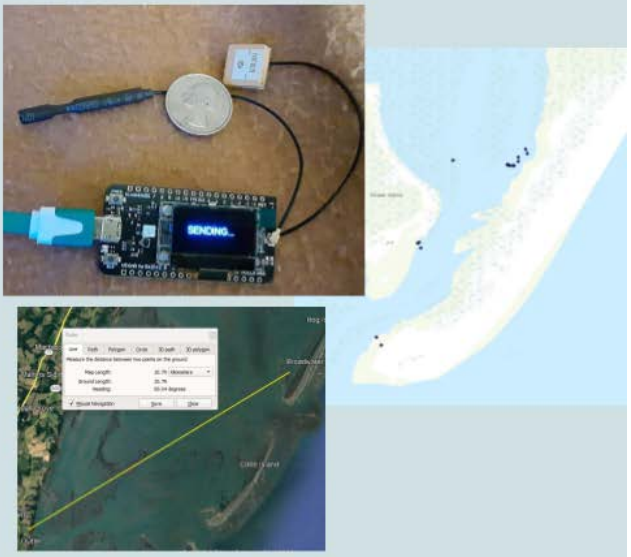
CALL NOW!

Getting Data from Isolated Sites with Minimal Power

-John Porter, VCRLTER

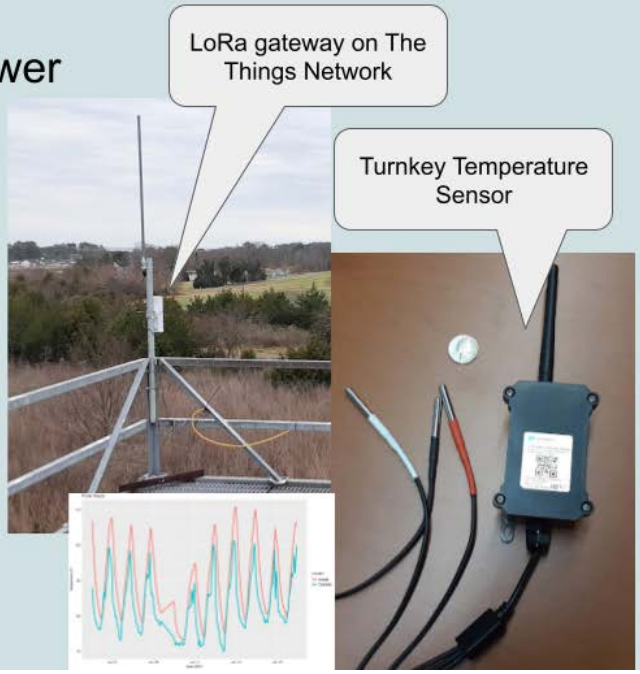
WiFi is great for retrieving data - when you have adequate power. Cell connections are great if you have lots of money (e.g,. $40/mo per station)

LoRa and Low-Cost Cellular may be helpful for getting data from the field into databases from isolated field sites



LoRa - LOngRAnge Low Power

- IOT technology
- Battery life measured in YEARS
  - For infrequent measurements (e.g., hourly+)
- Can be run as a mesh, or using established LoRa networks (e.g., TheThingsNetwork, Helium)
- Range of 1-16 km based on intervening landscape and antennas used
- Many turn-key sensors available

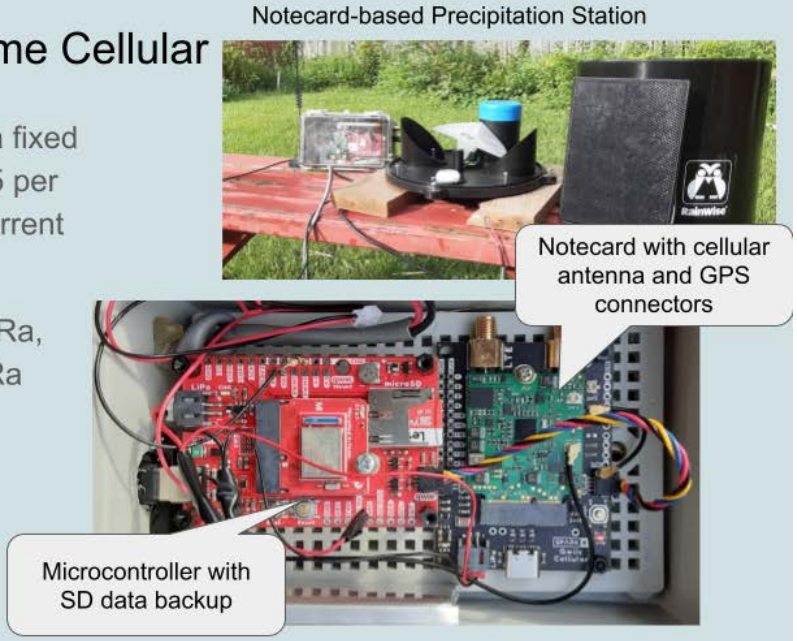


Low-Cost, Low-Volume Cellular

"Notecards" © are sold with a fixed amount of cellular data (~$85 per ½ GB) but there are NO recurrent charges. Good for 10 years.

Require more power than LoRa, but no need to be near a LoRa gateway (just LTE cellular coverage)

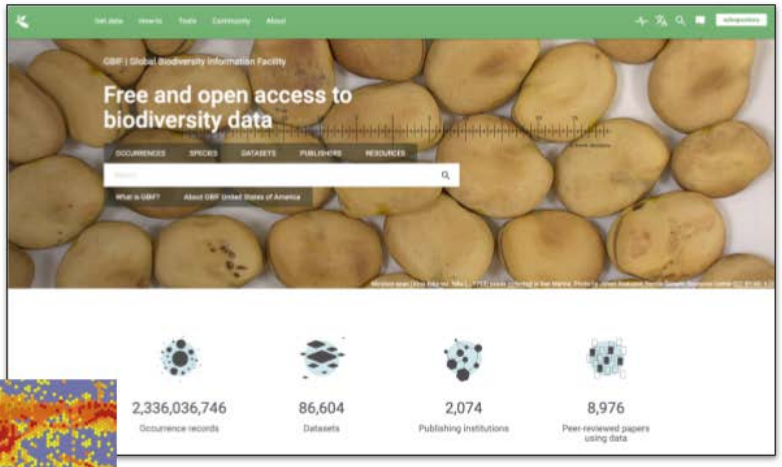
Session using Notecards tomorrow afternoon at ESIP

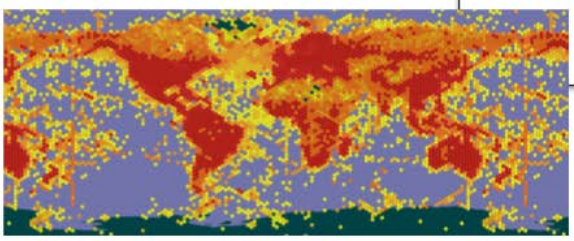


EDI Pipeline for Contributing Biodiversity Data to GBIF

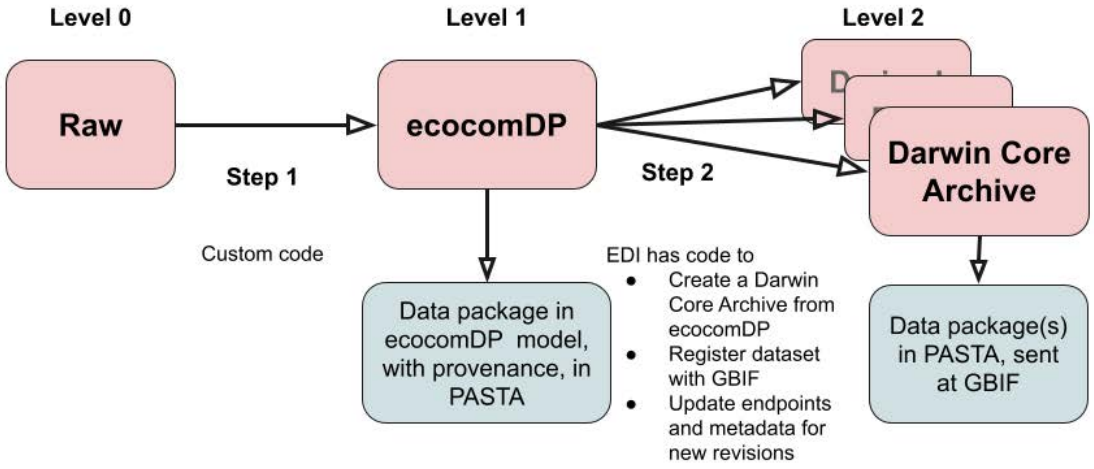
Colin Smith, Margaret O'Brien
2023-07-17, IMC

GBIF

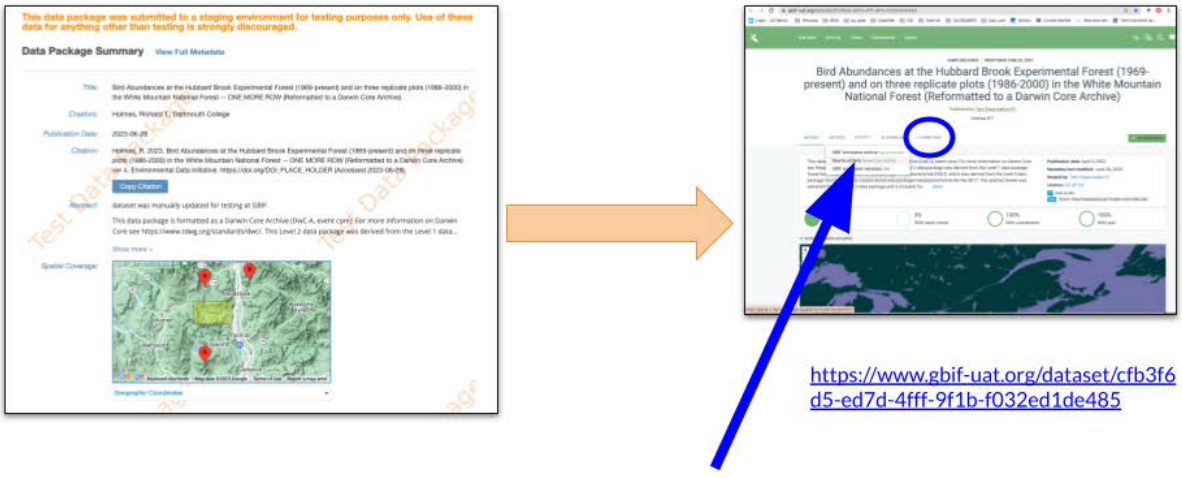




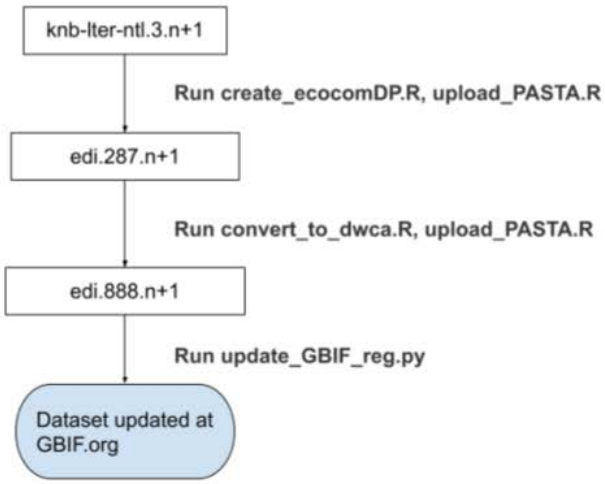
Workflow



Example



Goal: Automated Sequence



Setup

1. Create L0 > L1 conversion script
2. Register L0
3. Create PASTA event subscriptions

Next Steps

73 datasets queued (66 from LTER)

- PASTA search on "Darwin Core Archive"
- Prioritize, identify others

Formalize registration process

- Technical
- Social

EMLvp (validator and parser)

Mark Servilla, EDI

parserValid	error	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	Failed to parse IDs and references

parserValid	error	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	Failed to parse IDs and references

EMLvp

EMLvp is a Python 3 package to validate and parse Ecological Metadata Language XML documents for compliance with the EML metadata standard, including XML schema validation and ensuring that references resolve to existing ids.

EMLvp

EMLvp is a Python 3 package to validate and parse Ecological Metadata Language XML documents for compliance with the EML metadata standard, including XML schema validation and ensuring that references resolve to existing ids.

EMLvp

Compliance includes the following inspections:

- id attributes in all elements are unique,
- references elements for subject id,
- for circular references (references parent elements with id attributes),
- for system attribute consistency,
- customUnit for STMML definitions,
- parents of annotation elements for subject id (sans the annotations element),
- references attribute of annotation(s) for subject id, and
- additionalMetadata describes attribute for subject id.

EMLvp

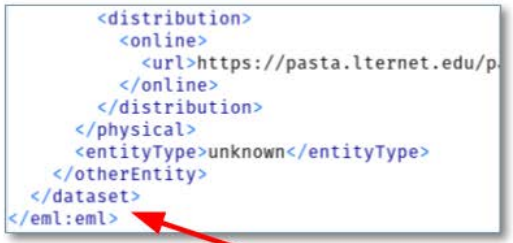
- Python3 package available through PyPI: https://pypi.org/project/emlvp
- Documentation: https://emlvp.readthedocs.io/en/latest
- Command-line interface application
- Import as Python package

```
❯ emlvp --help
Usage: emlvp [OPTIONS] [TARGET]...

  Performs validation of EML XML file(s)

      1. XML schema validation

      2. EML parsing for references/id resolution

      3. Dereference references/id into expanded EML XML and re-validate/parse

      TARGET: EML XML file or directory containing EML XML file(s) (may be repeated)

Options:
  -d, --dereference   Dereference EML XML file(s) (default is False).
  -f, --fail-fast     Exit on first exception encountered (default is False).
  -p, --pretty-print  Pretty print output for dereferenced EML XML (default is
                      False).
  -s, --statistics    Show post processing inspection statistics.
  -v, --verbose       Send output to standard out (-v or -vv or -vvv for
                      increasing output).
  --version           Output emlvp version and exit.
  -h, --help          Show this message and exit.
```

```
❯ emlvp  edi.628.1.xml
edi.628.1.xml
Missing custom unit id(s): ['gramPercentimeterSquared']
```

```
❯ python
Python 3.10.8 | packaged by conda-forge | (main, Nov 22 2022, 08:26:04) [GCC 10.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import emlvp.validator as validator
>>> from emlvp.validator import Validator
>>> with open("edi.628.1.xml", "r") as f:
...     xml = f.read()
...
>>> schema_path = validator.schema_path()
>>> v = Validator(schema_path + "/EML2.2.0/xsd/eml.xsd")
>>> v.validate(xml)
>>>
>>> from emlvp.parser import Parser
>>> p = Parser()
>>> p.parse(xml)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/home/servilla/git/EMLvp/src/emlvp/parser.py", line 199, in parse
    raise exceptions.ParseError(msg_queue.strip())
emlvp.exceptions.ParseError: Missing custom unit id(s): ['gramPercentimeterSquared']
>>>
```

```
<measurementScale>
  <ratio>
    <unit>
      <customUnit>gramPercentimeterSquared</customUnit>
    </unit>
    <numericDomain>
      <numberType>real</numberType>
      <bounds>
        <minimum exclusive="false">174</minimum>
        <maximum exclusive="false">820.2</maximum>
      </bounds>
    </numericDomain>
  </ratio>
</measurementScale>
```



<additionalMetadata>?????

NATIONAL SCIENCE FOUNDATION
LTER NETWORK
LONG TERM ECOLOGICAL RESEARCH

Hex Logo Under Construction!

Introducing `ltertools`

Tools Developed by the Long Term Ecological Research Community

github.com/lter/ltertools

Package Rationale & Theme

We likely face similar data challenges!

- By sharing our custom tools, we can help one another
- DRY – Don't Repeat Yourself

Hurdle to R package development is the required architecture

- That is already in place for `ltertools`
- New functions can be easily integrated into this existing skeleton

Function theme is "tools useful to LTER community"

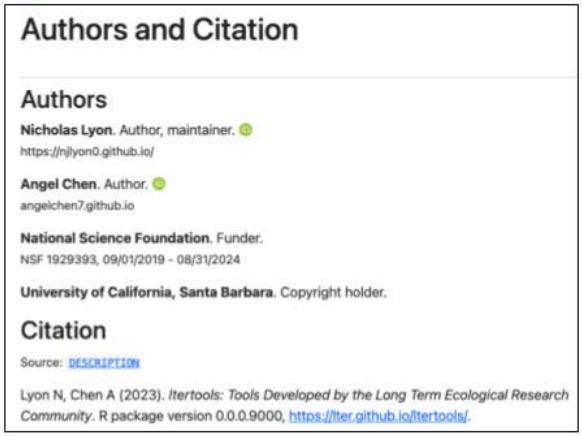
- Any of your functions fit under this umbrella

github.com/lter/ltertools

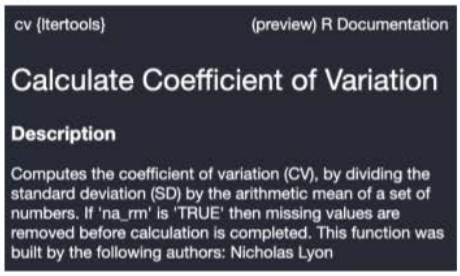
Contribution Credit

Contributed functions net you 3 modes of credit:

1 Package authorship



2 Named in function documentation



3 Named in GitHub README

github.com/lter/ltertools

“How do I Contribute?”

Function Idea

Open a GitHub Issue

Minimum information:

- Description of function purpose

Ideal information:

- Use-case of function
- Script + example data accomplish that operation

Function Script

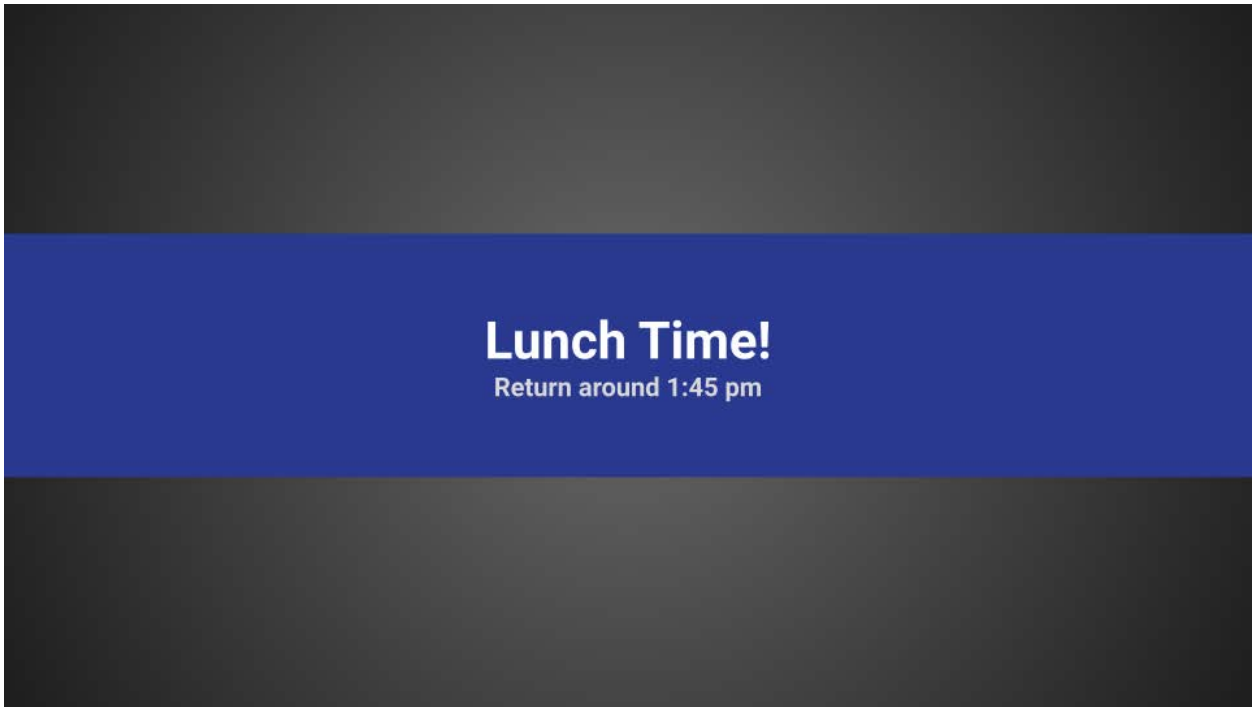
A) Open a GitHub Issue

- See “Function Idea” instructions

B) Fork `ltertools`

- Only if you want!
- Put all content in “dev” folder

github.com/lter/ltertools



Afternoon activities

Group Photo

Units WG activity

SLIDES AT:
https://docs.google.com/presentation/d/1AB1uWsNx_IBeL0C5MBkNhToEUt-Nz0xl5rdpnnocbX8/edit?usp=drive_link

Break Time!

Return in 30 minutes

EML Best Practices WG activity

Slides at:
https://docs.google.com/presentation/d/1k-sY_u1woW19jTWPx-55IT-PMns5HA6Eb65aBOMY3bo/edit?usp=sharing

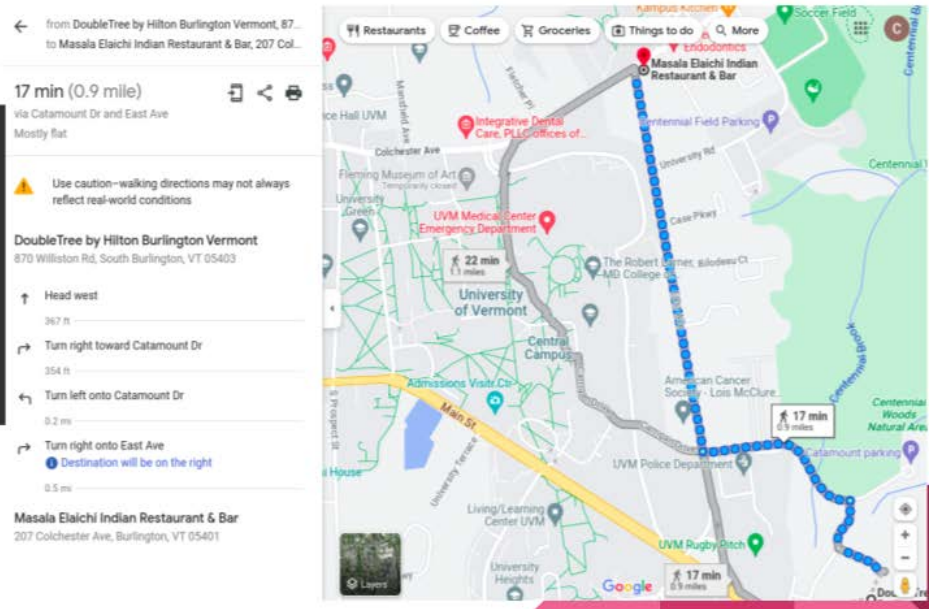
6pm: Dinner at Masala Elaichi

Then we'll see where the night takes us!

https://masalaelaichi.com/

THANK YOU GREG!

https://masalaelaichi.com/



Puppy Photos

Appendix B: Units WG Activity

Slides on Google Drive: UnitsWG_2023-07-17

Units Working Group

2023-07-17

LTER IMC annual meeting

Numbers without units are not data, they are just numbers!

Today's Agenda

Working Group Progress

Web services demo & playtime

Feedback, questions, discussion

Recap - - Working Group Goals

Find a replacement for an antiquated Unit Dictionary and Services

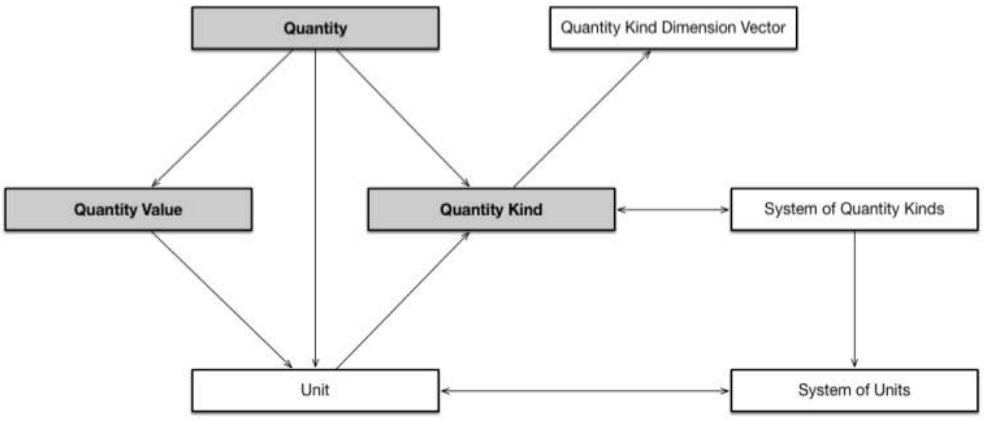
High-quality Units

Services that make units easy to incorporate into
- EML
- EML generation systems
    - ezEML
    - EML Assembly Line
    - Site-based data management systems

Advise dataset curators on best practices

Facilitate scientific linkages to other data

QUDT https://QUDT.org



Quantities, Units, Dimensions and Types Ontology

Units vs measurements (a clarification)

A unit is *part* of a measurement

Measurement: grams carbon per square meter of grassland

Unit: gram per meter squared

Issue: EML provides a specific "place" for the unit (<standardUnit>,<customUnit>), but other important parts of a measurement are in free text (<attributeName>, <attributeDescription>)

Possible solution: Use EML annotations to identify other measurement components in a machine-interpretable way. However …

WE ARE ADDRESSING UNITS HERE - NOT MEASUREMENTS

Typical unit description

URI:
http://qudt.org/vocab/unit/MilliGM-PER-L

QUDT records contain

- definitions and conversion factors
- IDs from other unit systems (e.g., ucumCode, uneceCommonCode)

https://ble-lter.github.io/QUDT_Browser/

Sample EML - Annotation with QUDT Unit

```
<attribute id="ds1-e2-att3">
  <attributeName>ca_conc</attributeName>
  <attributeLabel>Calcium concentration</attributeLabel>
  <attributeDefinition>Concentration of Ca ion in the water sample</attributeDefinition>
  <storageType>float</storageType>
  <measurementScale>
    <ratio>
      <unit>
        <customUnit>milligramPerLiter</customUnit>
      </unit>
      <precision>0.01</precision>
      <numericDomain>
        <numberType>real</numberType>
      </numericDomain>
    </ratio>
  </measurementScale>
  <annotation>
    <propertyURI label="has unit">http://qudt.org/schema/qudt/hasUnit</propertyURI>
    <valueURI label="MilliGM-PER-L">http://qudt.org/vocab/unit/MilliGM-PER-L</valueURI>
  </annotation>
</attribute>
```

Attribute ID: unique within the dataset

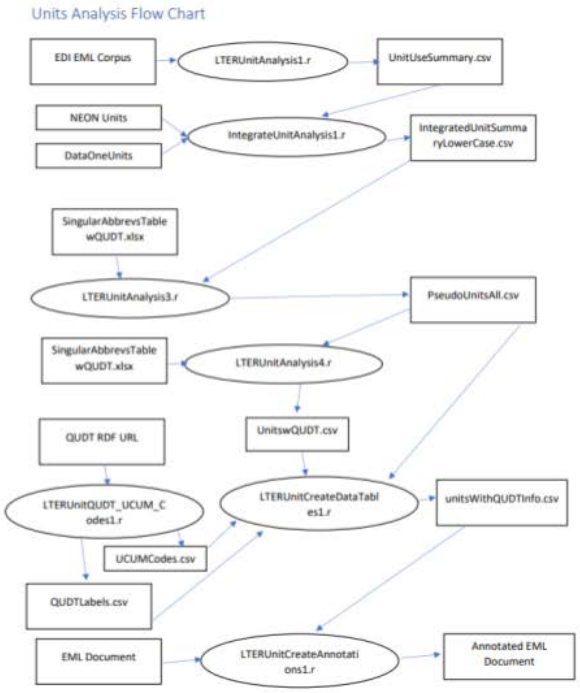
The **SUBJECT** in a semantic triple

ANNOTATION ELEMENT: contains URS for the QUDT unit

The **PREDICATE** and **OBJECT** in a semantic triple

Working Group Products

1. Existing EML units mapped to QUDT URIs
   a. From EDI, LTER, DataONE, NEON
2. New Units to be added to QUDT
3. Code and web service to add <annotation> elements (to EML 2.2) for QUDT Units



Compile, Integrate & Tally Unit Lists

Transform and match to QUDT

Tools for annotating EML

Mapping to QUDT

Distinct* units: **7,110**

Distinct QUDT units matched: **294**

355,057 individual instances (~60k datasets):

- 317,764 matched
- **89.5%** of all unit **uses** matched to QUDT

EDI: 90.5% of 176,722

NEON: 96.1% of 15,765

DataONE: 87.8% of 162,570

Lookup Table: 5,148 ad hoc + QUDT + UCUM units and their QUDT equivalents

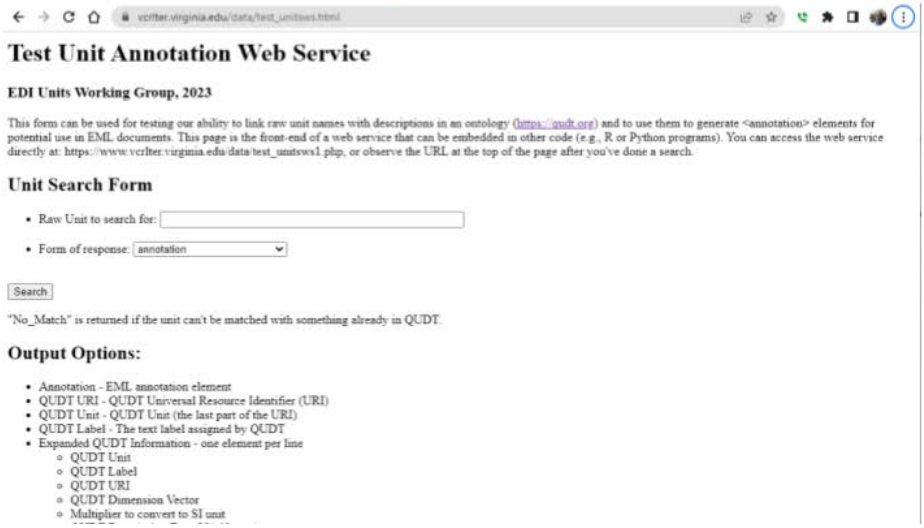
Working Group Products

1. ~~Existing EML units mapped to QUDT URIs~~
2. IN PROGRESS New Units to be added to QUDT
3. IN PROGRESS Code and web service to add <annotation> elements (EML 2.2) for QUDT Units

QUDT id	qudt:uomCode	rdfs:label	qudt:abbreviation				
unit	rdf:type	rdfs:label	qudt:conversionMultiplier	qudt:conversionOffset	qudt:hasDimensionVector	qudt:hasQuantityKind	qudt:plainTextDescription
NUM-PER-MilL	{#}.mL-1		number/ml	1		volumetricDe	number of particles or organisms per
NUM-PER-CentM2-HR	{#}.cm-2.hr-1		number/cm²	1		frequency	rate of change of areal density of a su
MilL-PER-L	mL.L-1		ml/L	0.001	0		milliters of solution per total volume
	microMOL{photons}.m-2.s-1		µEin*2/s	1	0	illuminance	micro Einsteins (1E-06 moles of photo
GM-PER-M2-HR	g.m-2.h-1		g/m^2/h	0.000019666	0		grams per meter square per hour
GM-PER-GM-HR	g.g-1.h-1		1/h			massPerMas	mass per mass per time
GM-PER-M2-HR	g.m-2.h-1		g/m2/h			arealMassDe	mass per area per time
KiloGM-PER-MOL	kg.mol-1		Kg/mol				kilogram per mol
microMOL-PER-CentM2	umol.cm-2		µmol/cm^2			arealDensity	µmoles per square centimeter
MilGM-PER-CentM2-DAY	mg.cm-2.d-1		mg/cm2/day		0	arealMassDe	Milligrams per square centimeter per
MilGM-PER-L-SEC	mg.L-1.s-1		mg/l/s			massDensity	mass per volume per time
MilGM-PER-MilGM-DAY	mg.mg-1.d-1		mg/mg/day		0	massPerMas	Milligrams per milligram per day
MOL-PER-KiloGM	mol.kg-1		mol/Kg			amountOfSul	moles per kilogram

Web Service Demo

https://www.vcrlter.virginia.edu/data/test_unitsws.html



vcrlter.virginia.edu/data/test_unitsws.html

Test Unit Annotation Web Service

EDI Units Working Group, 2023

This form can be used for testing our ability to link raw unit names with descriptions in an ontology (https://qudt.org) and to use them to generate <annotation> elements for potential use in EML documents. This page is the front-end of a web service that can be embedded in other code (e.g., R or Python programs). You can access the web service directly at: https://www.vcrlter.virginia.edu/data/test_unitsws1.php, or observe the URL at the top of the page after you've done a search.

Unit Search Form

- Raw Unit to search for:
- Form of response: annotation

Search

"No_Match" is returned if the unit can't be matched with something already in QUDT.

Output Options:

- Annotation - EML annotation element
- QUDT URI - QUDT Universal Resource Identifier (URI)
- QUDT Unit - QUDT Unit (the last part of the URI)
- QUDT Label - The text label assigned by QUDT
- Expanded QUDT Information - one element per line
  - QUDT Unit
  - QUDT Label
  - QUDT URI
  - QUDT Dimension Vector
  - Multiplier to convert to SI unit

Web Services

Input	Output	Issues
EML document	EML document with unit annotations and attribute IDs added (if not already there)	Options? - update revision number - location of <annotation>
Single unit string*	QUDT URI	
Single unit string*	populated <annotation> element	
List of units*	list of QUDT URIs	
List of units*	group of populated <annotation> snippets	
attribute IDs and associated unit string*	<annotation> element	
QUDT ID	EML <annotation> element	Option: stmml/unit

DIY, with R-code & your configuration

Input	Output
EML document	EML document with unit annotations and attribute IDs added (if not present)
Directory of EML documents	new directory of EML docs with unit annotations and attribute IDs (if not present)
QUDT ID	EML <annotation> element

Improving Mapping results -LTER

View your Site's Unit uses and how they map Google Sheets (one sheet/site):

https://tinyurl.com/ViewUnits23 *Note uppercase U*

Add Unit Mappings & Propose New QUDT Units

https://tinyurl.com/Addunits23 *Note NO uppercase U*

Browse QUDT for other units:

https://ble-lter.github.io/QUDT_Browser/

Your TASKS:

1. Provide input on preferred annotation approaches (web service, DIY tools)
2. View your site's units and identify:
   a. Any units that should be linked to an existing QUDT unit
   b. Any units that are missing from QUDT, but should be added
   c. Add those to the online sheet

Resources:

- **View** your Site's Unit uses: https://tinyurl.com/ViewUnits23
- **Add** Unit Mappings & Propose New QUDT Units: https://tinyurl.com/Addunits23
- Web Services Test Page: https://www.vcrlter.virginia.edu/data/test_unitsws.html
- Browse QUDT for other units: https://ble-lter.github.io/QUDT_Browser/
- To see which dataset a unit came from, see the raw units list: https://github.com/EDIorg/Units-WG/blob/main/RCode_JP/DataFiles4R/RawUnitList.csv

Additional questions for discussion

How can units annotations be leveraged (e.g., automate STMML production, automated unit conversion tools)?

What is the best process for adding units to the mapping list used by the web services?

What if the EML we want to add QUDT units to is not version 2.2 yet?

Can I put the annotations in before I submit the dataset to PASTA?

Where do I learn more about semantic annotation in EML?
https://eml.ecoinformatics.org/semantic-annotation-primer.html

Thank You

Margaret O'Brien

John Porter

Christine Laney

Stevan Earl

Mary Martin

Marina Franz

Old slides below here

Old and in the way!

Automated Mapping to QUDT URIs

Step 1: Use string substitutions to convert existing “raw” units into consistent “pseudounits”.

RAW
microGramsPerLiter
microgramsperliter
microgramperliter
micrograms per liter
microgram/liter
micrograms / liter
ug/l
micrograms/liter
micrograms/l
microgram per l

“Pseudounit”

microgramperliter

Automated Mapping to QUDT URIs

Step 2: Use string substitutions to convert “psuedounit” into QUDT URIs

microgramperliter - base pseudounit

↓

MicroGMperliter - substitute for “microgram”

↓

MicroGMperL - substitute L for liter

↓

MicroGM-PER-L - replace “per” with -PER-

↓

http://qudt.org/vocab/unit/MicroGM-PER-L - Draft URI

Automated Mapping to QUDT URIs

Step 3:

Verify draft URI with QUDT

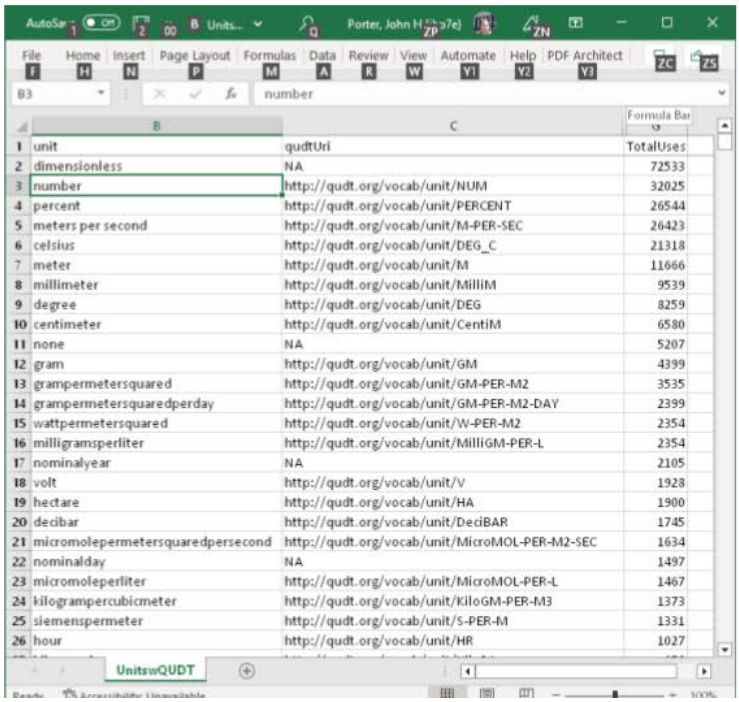
Draft URIs that match those in QUDT (late 2022 version) are kept

Results

Table:

- original unit (lowercase)
- QUDT URI

Sorted by frequency of use



unit	qudtUri	TotalUses
dimensionless	NA	72533
number	http://qudt.org/vocab/unit/NUM	32025
percent	http://qudt.org/vocab/unit/PERCENT	26544
meters per second	http://qudt.org/vocab/unit/M-PER-SEC	26423
celsius	http://qudt.org/vocab/unit/DEG_C	21318
meter	http://qudt.org/vocab/unit/M	11666
millimeter	http://qudt.org/vocab/unit/MilliM	9539
degree	http://qudt.org/vocab/unit/DEG	8259
centimeter	http://qudt.org/vocab/unit/CentiM	6580
none	NA	5207
gram	http://qudt.org/vocab/unit/GM	4399
grampermetersquared	http://qudt.org/vocab/unit/GM-PER-M2	3535
grampermetersquaredperday	http://qudt.org/vocab/unit/GM-PER-M2-DAY	2399
wattpermetersquared	http://qudt.org/vocab/unit/W-PER-M2	2354
milligramsperliter	http://qudt.org/vocab/unit/MilliGM-PER-L	2354
nominalyear	NA	2105
volt	http://qudt.org/vocab/unit/V	1928
hectare	http://qudt.org/vocab/unit/HA	1900
decibar	http://qudt.org/vocab/unit/DeciBAR	1745
micromolepermetersquaredpersecond	http://qudt.org/vocab/unit/MicroMOL-PER-M2-SEC	1634
nominalday	NA	1497
micromoleperliter	http://qudt.org/vocab/unit/MicroMOL-PER-L	1467
kilogrampercubicmeter	http://qudt.org/vocab/unit/KiloGM-PER-M3	1373
siemenspermeter	http://qudt.org/vocab/unit/S-PER-M	1331
hour	http://qudt.org/vocab/unit/HR	1027

"Missing URIs" in Automated mapping

Four causes for failure to get a QUDT match:

1) List of substitutions is incomplete
   a) Units we don't understand yet, possibly domain specific (phi)
2) Units with underlying issues
   a) Uncaught misspellings (e.g., miligram, precent, hectorpascal, celcius)
   b) Units that mix units with a substance (e.g., kg/dry soil, birds/ha, "number of stems counted")
   c) Entries that are not units at all (e.g., 999, 1970s, diveroperatedpushcore, feather, yoypollock)
3) Units that aren't in QUDT, but could be
   a) Nominal months, days (as compared to number of months or days)
4) Entries that aren't units but indicate a lack of units
   a) "Dimensionless", alone, accounts for 72,533 uses
   b) "None" 5,207 additional uses

Unmatched units

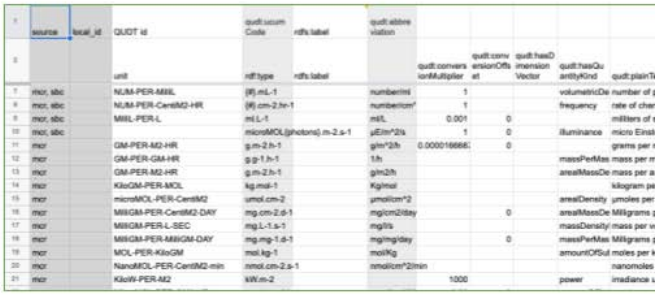
Four causes for failure to get a QUDT match:

1) List of substitutions is incomplete
   a) Units we don't understand yet, possibly domain specific (phi)
2) Units with underlying issues
   a) Uncaught misspellings (e.g., miligram, precent, hectorpascal, celcius)
   b) Units that mix units with a substance (e.g., kg/dry soil, birds/ha, "number of stems counted")
   c) Entries that are not units at all (e.g., 999, 1970s, diveroperatedpushcore, feather, yoypollock)
3) Units that aren't in QUDT, but could be
   a) Nominal months, days (as compared to number of months or days)
4) Entries that aren't units but indicate a lack of unit
   a) "Dimensionless", alone, accounts for 72,533 uses
   b) "None" 5,207 additional uses

Working Group Progress with QUDT

1. Collating units
   a. EML in EDI (~9k datasets)
   b. EML from DataONE (~50k datasets)
   c. NEON Datasets (76 Units, total)
2. Analyzing these, automating mapping to QUDT
3. Manual mapping to QUDT
   a. Using datasets from 4 (NTL,CAP,MCR,SBC) LTER sites, NEON

WG Next Steps



	source	local_id	QUDT id	qudt:ucumCode	rdfs:label	qudt:abbreviation					
			unit	rdf:type	rdfs:label		qudt:conversionMultiplier	qudt:conversionOffset	qudt:hasDimensionVector	qudt:hasQuantityKind	qudt:plainText
7	mcr, sbc		NUM-PER-MIL	{#}.mL-1		numberl/ml	1			volumetricDe	number of p
8	mcr, sbc		NUM-PER-CentiM2-HR	{#}.cm-2.hr-1		number/cm^	1			frequency	rate of chan
9	mcr, sbc		MIL-PER-L	mL-1		mil/L	0.001	0			milliters of
10	mcr, sbc			microMOL{photons}.m-2.s-1		µEin/m^2/s	1	0		illuminance	micro Einst
11	mcr		GM-PER-M2-HR	g.m-2.h-1		g/m^2/h	0.000016666	0			grams per s
12	mcr		GM-PER-GM-HR	g.g-1.h-1		1/h				massPerMas	mass per m
13	mcr		GM-PER-M2-HR	g.m-2.h-1		g/m2/h				arealMassDe	mass per a
14	mcr		KiloGM-PER-MOL	kg.mol-1		Kg/mol					kilogram pe
15	mcr		microMOL-PER-CentiM2	umol.cm-2		µmol/cm^2				arealDensity	µmoles per
16	mcr		MilliGM-PER-CentiM2-DAY	mg.cm-2.d-1		mg/cm2/day		0		arealMassDe	Milligrams p
17	mcr		MilliGM-PER-L-SEC	mg.L-1.s-1		mg/l/s				massDensity	mass per v
18	mcr		MilliGM-PER-MilliGM-DAY	mg.mg-1.d-1		mg/mg/day		0		massPerMas	Milligrams p
19	mcr		MOL-PER-KiloGM	mol.kg-1		mol/Kg				amountOfSul	moles per k
20	mcr		NanoMOL-PER-CentiM2-min	nmol.cm-2.s-1		nmol/cm^2/min					nanomoles
21	mcr		KiloW-PER-M2	kW.m-2			1000			power	irradiance u

1. Assemble corpus of new units to be added to QUDT
2. Define programmatic services needed by the community
   a. ezEML
   b. EML AssemblyLine
3. Document practices for EML metadata
   a. Annotation - measurement level
   b. STMML

QUDT to STMML

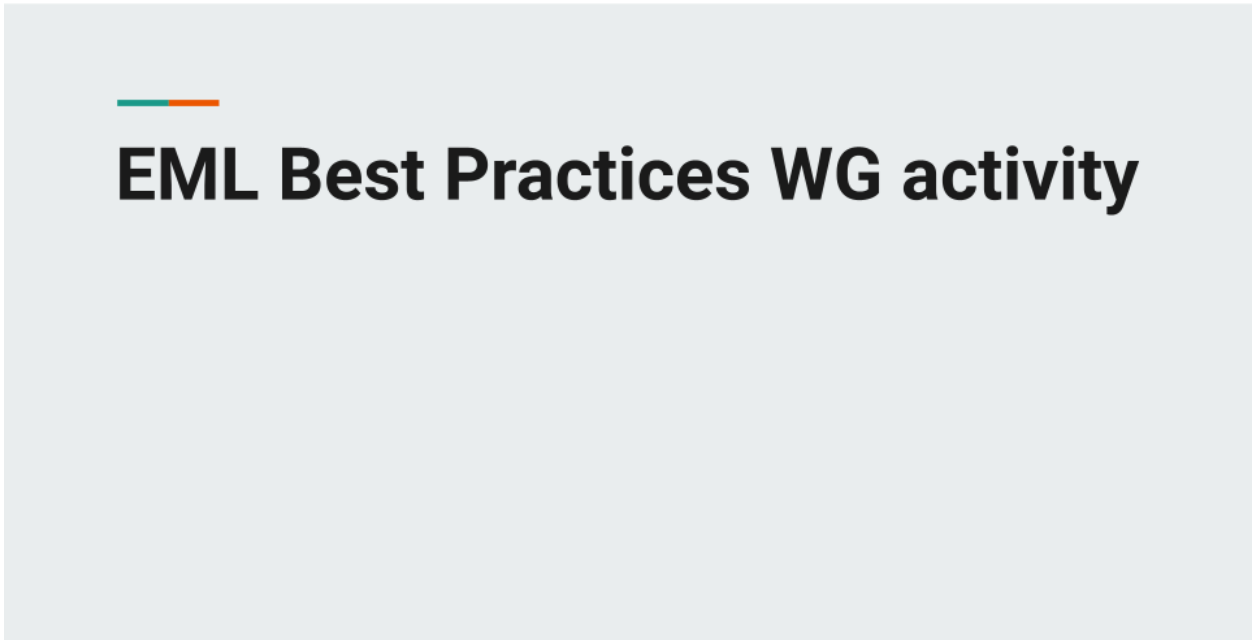
Breakout group questions

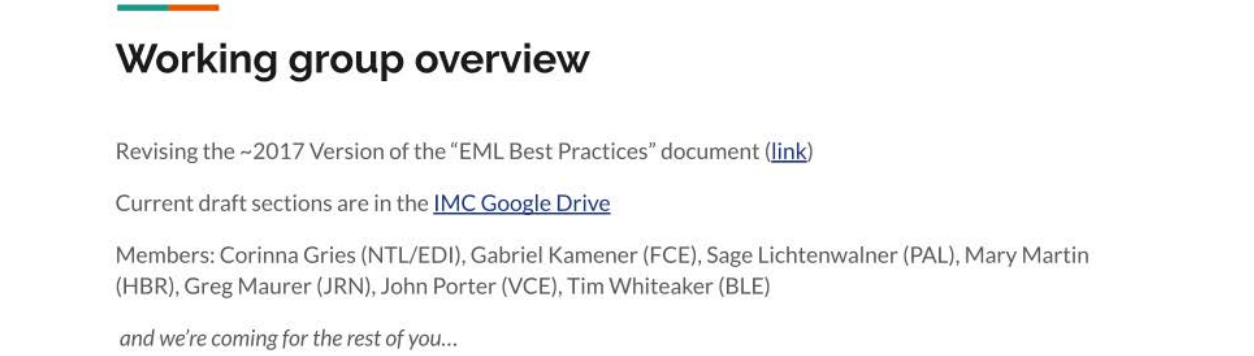
1. What do you need to incorporate QUDT URIs?
    a. As <annotation> elements
    b. As <customUnit> with STMML
2. How should “new” units be identified and linked to QUDT URIs?
    a. Editable Google Docs spreadsheet?
    b. Web form?
    c. Requests sent to a committee?
3. How should unit annotations be used for searching and integrating data?
    a. What changes to search systems might be required?
4. Anything else related to units?

Notes doc:
https://docs.google.com/document/d/1FDKWkJCNLXX9xiQBXUav67HBY6pUMUH_oXbr4xXrfWc/edit#

Appendix C: EML Best Practices WG Activity

Slides on Google Drive: EML Best Practices WG activity





Our agenda

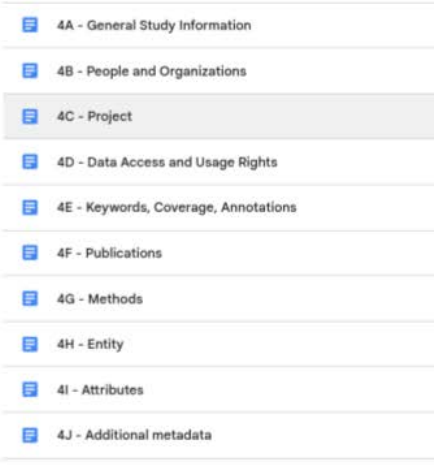
1. Quick overview of document structure
2. Review the major changes in recommendations so far
3. Feedback time
    a. What do you think about these changes?
    b. What do you do at your site that should be a best practice?
    c. What is still confusing and should be addressed?

Document structure

Four Chapters

1. Introduction (What is EML, history, conventions, resources)
2. Metadata and dataset design principles
3. XML, schemas, and EML (the underlying tech and application for EML)
4. Best practices for each element in EML
    a. 10 subsections so far

Some sections are incomplete



We need your feedback

We need your input on Chapter 4!

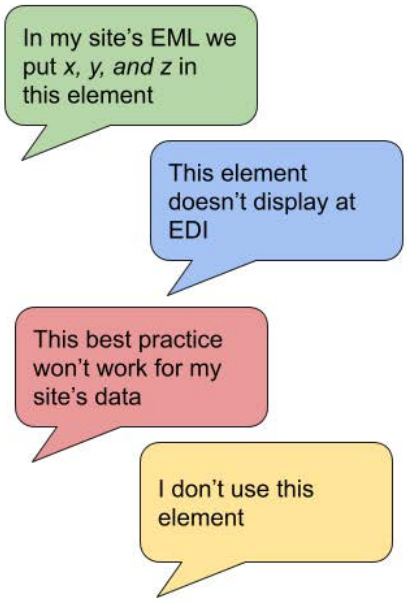
EML 2.2 introduced new EML elements

Some EML elements are controversial

Some best practices are unclear and hard to decide

Different sites have different use-cases

One size of EML does not fit all



Revision 1: Semantic annotation

New section - currently taken from the EML schema documentation's "What's new in EML 2.2" section

- It only describes how and where to enter terms from an ontology, but **not why or when...**
- We don't have many use cases from LTER IMs

Questions for IMC

- Why and when do you use annotations?
- Which ontologies?
- How do you use annotations once a dataset is published?

Revision 2: EML 2.2 Citation elements

New recommendation - There are new citation elements in EML 2.2. Use them when possible.

- <referencePublication> & <usageCitation> are new citation types (<citation> pre-dates EML 2.2)
- <literatureCited> (also new) is a list containing multiple <citation> or <bibtex> elements

Questions for IMC

- Which of these do you use and where do you put them (child of <dataset>, <methodStep>, etc)?
- Is the purpose of each clear?
- What do you expect from EDI for displaying citation elements?

Purple
Name/site optional

Revision 3: Abstract

New recommendation - we are recommending including "why" in addition to "what, where, when"

- Adding "why" is new to the BP doc, but many IMs may do this already.

Questions for IMC

- Should all published datasets include "why" in the abstract?

Yellow
Name/site optional

Revision 4: Geographic coverage

New recommendation - we are recommending bounding boxes over individual sample points

- It is common to use points but how useful is this in practice?
- Multiple bounding boxes are also common

Questions for IMC

- How often to you use points?
- How often and why do you use multiple bounding boxes?
- What is a reasonable and useful level of detail in <geographicCoverage>?

Pink
Name/site optional

Revision 5: spatialRaster & spatialVector entities

Question- Should we deprecate these entity types in the EML BP doc?

- These are still used, but not consistently
- Many/most IMs include spatial data files (ESRI shapefile, GeoTIFF, etc.) as otherEntities

Questions for IMC

- Are there use-cases for metadata included in spatialRaster and spatialVector entities?
- How many sites regularly create and use these entity types?
- Is using otherEntity for spatial files (with good metadata included) better?
- Could a working group help with this question?

Blue
Name/site optional

Revision 6: Maintenance and versioning

New recommendation - we will make new recommendations here, but not sure what

- Ongoing/complete distinction seems necessary commonly used in <description>
- There is a lot of other detail available in this element (<maintenanceUpdateFrequency>, <changeHistory>, <changeDate>, <changeScope>)

Questions for IMC

- How do you use the <maintenance> element and its child elements?
- Do you have a use-case for data or metadata versioning in EML?

What do you think about these changes?

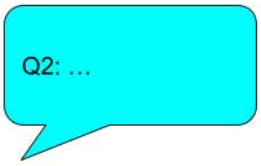
General comments and discussion - raise your hand.

Question 1: What do you do with your site's EML that should be a best practice?

Raise hand, or any sticky note with Q1 on it

Question 2: What parts of EML are still confusing and should be addressed?

Raise hand, or any sticky note



More revisions to come

- **<project> element:** use <funding>, <award>, and <relatedProject> within projects, but do what works for your site. Should we be more specific for LTER datasets?
- **Keywords (and annotations):** We are waiting to collaborate with the Keywording/Annotation WG
- **Attributes:** Waiting to collaborate with the Units WG.

Comment on our draft document anytime

Come help us discuss these at meetings every other Tuesday through September.