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

**Table of Contents**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agenda</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>Meeting materials</strong></td>
<td>3</td>
</tr>
<tr>
<td>Working Group Updates</td>
<td></td>
</tr>
<tr>
<td>HyMet</td>
<td>4</td>
</tr>
<tr>
<td>Metabase</td>
<td>4</td>
</tr>
<tr>
<td>EML Best Practices</td>
<td>4</td>
</tr>
<tr>
<td>Unit Dictionary</td>
<td>4</td>
</tr>
<tr>
<td>Zotero</td>
<td>4</td>
</tr>
<tr>
<td>Annotation/Keywording</td>
<td>5</td>
</tr>
<tr>
<td>Resources</td>
<td>5</td>
</tr>
<tr>
<td><strong>Business Meeting</strong></td>
<td>5</td>
</tr>
<tr>
<td>Statements from IM Exec nominees</td>
<td>5</td>
</tr>
<tr>
<td>Election outcomes</td>
<td>6</td>
</tr>
<tr>
<td>Databits</td>
<td>6</td>
</tr>
<tr>
<td>LTER Network Office Report</td>
<td>6</td>
</tr>
<tr>
<td>EDI Report</td>
<td>7</td>
</tr>
<tr>
<td>ILTER Report</td>
<td>7</td>
</tr>
<tr>
<td>Resources</td>
<td>8</td>
</tr>
<tr>
<td><strong>IMKE Lightning Talks</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>Units Working Group Activity</strong></td>
<td>8</td>
</tr>
<tr>
<td>Resources</td>
<td>8</td>
</tr>
<tr>
<td><strong>EML Best Practices Working Group Activity</strong></td>
<td>9</td>
</tr>
<tr>
<td>Resources</td>
<td>9</td>
</tr>
<tr>
<td><strong>Appendix A: Full-day meeting slides</strong></td>
<td>10</td>
</tr>
<tr>
<td>Working group updates</td>
<td>12</td>
</tr>
<tr>
<td>Business meeting</td>
<td>24</td>
</tr>
<tr>
<td>IMKE Lightning Talks</td>
<td>43</td>
</tr>
<tr>
<td>Afternoon activities</td>
<td>71</td>
</tr>
<tr>
<td><strong>Appendix B: Units WG Activity</strong></td>
<td>75</td>
</tr>
<tr>
<td><strong>Appendix C: EML Best Practices WG Activity</strong></td>
<td>89</td>
</tr>
</tbody>
</table>
Agenda

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

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:
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**

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

<table>
<thead>
<tr>
<th>Time</th>
<th>Afternoon Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:45</td>
<td>Group Photo</td>
</tr>
<tr>
<td>14:00</td>
<td>Units WG Activity</td>
</tr>
<tr>
<td>15:00</td>
<td>Break</td>
</tr>
<tr>
<td>15:30</td>
<td>EML Best Practices WG Activity</td>
</tr>
<tr>
<td>16:30</td>
<td>Adjourn</td>
</tr>
<tr>
<td>18:00</td>
<td>Dinner</td>
</tr>
</tbody>
</table>
Working group updates

HyMet (Clim/Hydro DB 2.0)

WG members: Corinna Gries, Miguel Leon, Greg Maurer, Mark Servilla, Colin Smith (plus Collin Bode & Scott Smith from Dendra)
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](https://dendra.science/orgs/lter) (JRN only)

3. Time and effort to standardize datasets

There are [7 HyMet datasets](https://dendra.science/orgs/lter) 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)
LGTER Core Metabase

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

---

**LTRER 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!
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


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 - \( \text{Trophic} \)
Vertebrate would find invertebrate - \( \text{Vertebrate} \)
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

<table>
<thead>
<tr>
<th>estuaries</th>
<th>ENVO:00000045</th>
<th>estuary</th>
<th><a href="http://vocab.biosphere.net/EnvThes/21807">http://vocab.biosphere.net/EnvThes/21807</a> estuary</th>
</tr>
</thead>
<tbody>
<tr>
<td>estuarine</td>
<td>ENVO:00000045</td>
<td>estuary</td>
<td><a href="http://vocab.biosphere.net/EnvThes/21807">http://vocab.biosphere.net/EnvThes/21807</a> estuary</td>
</tr>
<tr>
<td>estuary</td>
<td>ENVO:00000045</td>
<td>estuary</td>
<td><a href="http://vocab.biosphere.net/EnvThes/21807">http://vocab.biosphere.net/EnvThes/21807</a> estuary</td>
</tr>
</tbody>
</table>
### Map to higher level concept

<table>
<thead>
<tr>
<th>Term</th>
<th>ID</th>
<th>Type</th>
<th>Link</th>
<th>Parent Term</th>
<th>Parent ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>bog</td>
<td>ENVO:00000044</td>
<td>peatland</td>
<td><a href="http://vocabs.iter-europe.net/EnvThes/21798">http://vocabs.iter-europe.net/EnvThes/21798</a></td>
<td>wetland</td>
<td></td>
</tr>
<tr>
<td>fen</td>
<td>ENVO:0000232</td>
<td>fen</td>
<td></td>
<td>wetland</td>
<td></td>
</tr>
<tr>
<td>marsh</td>
<td>ENVO:0000035</td>
<td>marsh</td>
<td><a href="http://vocabs.iter-europe.net/EnvThes/21801">http://vocabs.iter-europe.net/EnvThes/21801</a></td>
<td>wetland</td>
<td></td>
</tr>
<tr>
<td>peatland</td>
<td>ENVO:0000044</td>
<td>peatland</td>
<td><a href="http://vocabs.iter-europe.net/EnvThes/21803">http://vocabs.iter-europe.net/EnvThes/21803</a></td>
<td>wetland</td>
<td></td>
</tr>
<tr>
<td>wetland</td>
<td>ENVO:01001209</td>
<td>wetland</td>
<td><a href="http://vocabs.iter-europe.net/EnvThes/21797">http://vocabs.iter-europe.net/EnvThes/21797</a></td>
<td>wetland</td>
<td></td>
</tr>
</tbody>
</table>

### Concept hierarchy

- Pond
- Lake
- Wetland
- Reef
- Ocean
- Freshwater
- Aquatic
- Saline
- Ecosystem
- Estuary
- Coastal
### Initial Results

https://drive.google.com/drive/folders/1X6ucoozr_86aeGQyWzrVPDxTLQImgWA?usp=sharing

<table>
<thead>
<tr>
<th>discipline</th>
<th>ecosystem</th>
<th>events</th>
<th>methods</th>
<th>org_unit</th>
<th>organism</th>
<th>process</th>
<th>substance</th>
<th>substrate</th>
<th>site</th>
</tr>
</thead>
<tbody>
<tr>
<td>aquatic ecology, biogeochemistry,</td>
<td>aquatic, freshwater,</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>animal, vertebrate,</td>
<td>hydrologic process, ice</td>
<td>ions, nitrogen,</td>
<td>air, water, air</td>
<td>North Temperate</td>
</tr>
<tr>
<td>interaction, limnology, water</td>
<td>lake, wetland</td>
<td></td>
<td></td>
<td></td>
<td>fish</td>
<td></td>
<td>compounds, phosphorus, inorganic matter</td>
<td>ice</td>
<td>Lakes LTER: Chemical Limnology of Primary Study Lakes: Nutrients, pH and Carbon 1981 - current</td>
</tr>
</tbody>
</table>
Soil Respiration at Forest Edges along an Urban to Rural Gradient in Massachusetts since 2018

geochemistry, 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

Volunteers for Databits Editor
Break Time!
Return in 30 minutes

LTER Network Office
Information Manager’s Annual Meeting
July 17, 2023
Nick Lyon (they/them), Data Scientist
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

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

SPARC Soil Phosphorus Control of C & N | github.com/ltter/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/ltter/wg-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

Environmental Drivers of Plant Reproduction | github.com/ltter/wg-plan-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 (Lнтернет.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
- SmugMug: https://lternetwork.smugmug.com/
- Zotero:
  - 25898 entries
  - https://www.zotero.org/groups/2055673/lter_network/library
  - feeds
    - https://lternet.edu/bibliography/
  - Site Personnel Lists
    - Overview

Personnel Roadmap

- Autoupdates on the 15th of each month
- Manual updates when notified
- IMs or admins maintain site personnel list at lterhub-ex-im
- Largely invisible to end users
- Profiles and directory
- Discourse Community invitations
- Surveys
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: `Itertools`

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: Iter.github.io/Itertools

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](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!

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/](https://deims.org/)
  - Username: US
  - Password: ?US80

<table>
<thead>
<tr>
<th>Year</th>
<th>Last Updated</th>
<th>Site Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022</td>
<td>JRN, KNZ, MCM, SEY</td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>CWT*, FCE, LUQ</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>AND, ARG, BES, BLE, BNZ, GAP, CCE, CDR, GCE, HBR, HFR, KBS, MCR, NES, NGA, NTL, NWT, PAL, PIE, SBC, VCR</td>
<td></td>
</tr>
<tr>
<td>Not verified</td>
<td>MSP</td>
<td></td>
</tr>
</tbody>
</table>
Inside ILTER – survey on network level data management capabilities

Johannes Peterseil
johannes.peterseil@umweltbundesamt.at

Ivo Offenthaler
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 ILTER network?

<table>
<thead>
<tr>
<th>IT support (infrastructure, helpdesk, etc.)</th>
<th>none</th>
<th>planned</th>
<th>&lt;1</th>
<th>1-2</th>
<th>2-5</th>
<th>&gt;5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data management and curation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software development / engineering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- project based funding: 49%
- institutional funding: 23%
- governmental funding support: 10%
- partnership with academic societies (e.g. data paper): 7%
- other: 11%

Comments: no funding, not planned, unfortunately none

Findability

Which metadata content standards does your ILTER network support?

<table>
<thead>
<tr>
<th>Standard</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>13</td>
</tr>
<tr>
<td>EML</td>
<td>8</td>
</tr>
<tr>
<td>ISO 19115 profile or variant</td>
<td>5</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>4</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>4</td>
</tr>
<tr>
<td>SensorML</td>
<td>2</td>
</tr>
<tr>
<td>DataGrid</td>
<td>2</td>
</tr>
<tr>
<td>DCAT</td>
<td>2</td>
</tr>
<tr>
<td>schemas.org</td>
<td>1</td>
</tr>
<tr>
<td>netCDF Headers/ CF</td>
<td>1</td>
</tr>
<tr>
<td>SensorThings-Based</td>
<td>0</td>
</tr>
<tr>
<td>DOI</td>
<td>0</td>
</tr>
</tbody>
</table>

Accessibility

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

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Data</th>
<th>no shares</th>
</tr>
</thead>
<tbody>
<tr>
<td>sILTER</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>FluxNet</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>GBIF</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DataOne</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ICOS</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>WMO</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>INSPIRE</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GEOGEOSS</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>OSO/ Open Data Portals</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>OBS</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

How does your ILTER network provide metadata for your data?

- Online catalogue (e.g. metadata portal): 54.29%
- Metadata documents shared on demand: 40.00%
- Exports from a database or application: 26.71%
- Other: 20.00%
- Catalog service for the Web (CSW): 8.57%
- We do not provide metadata: 5.71%

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

e.g. ICP Forests, ICP Waters, LifeWatch, Copernicus, Eco-Bank, EMODNET, Friend-E, google data search, OceanDataStandards.org, Interact, OASIS, Pangra, WGMS, etc.
Interoperability

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

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>in data</th>
<th>in metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>WoRMS</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>GBIF Taxonomy Index</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>EnvThes</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>CF Standard Names</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>ITS</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GCMD</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LifeWatch Vocabularies</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>QUDT Vocabularies</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SeaDataNet Vocabularies</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>GeoNames</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ENVCO</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>US LTER Controlled Vocabulary</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ESCO</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>NASA SWEET</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Scientific Variable Ontology</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>


Which Linked Open Data Services does your LTER network use?

<table>
<thead>
<tr>
<th>Service</th>
<th>in data</th>
<th>in metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers: ORCID</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>DataSets: DataCite / DOIs</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Institutions: ROR</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Repositories: re3Data</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>DataSets: Other PIDs (ARK, Handle, ...)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Physical Samples: International GeoSamp</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Others: e.g. CWB (https://opendata.cwb.gov.tw/DevManual/Instructio), 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?

<table>
<thead>
<tr>
<th></th>
<th>no</th>
<th>planned</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>none</th>
<th>no site DMP</th>
<th>most sites</th>
<th>most or all</th>
<th>yes all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>no</th>
<th>planned</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>no</th>
<th>planned</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>4</td>
<td>11</td>
</tr>
</tbody>
</table>
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?

- writing quality metadata
- how to manage data
- how to publish data in a data repository
- creating structured metadata (e.g., EML, ISO 19115)
- how to implement FAIR principles
- identifiers (ORCID, ROR) and why they are needed
- how to format data for easy reuse
- how to quality control data
- how to write data publications
- how to annotate data

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

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

Monthly total rainfall (mm), sum of AYACPJUE shrimp in three pools, and mean Seeds per square meter for Guarea guianensis-Susliwood (1°C)

Signature dataset

- Shiny App -
Signature dataset

TEMPLATES

Adam Sapp
2023 IMC Annual Meeting
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

- Experimented with web forms
  - Errors
  - Uncomfortable

Excel data template
Excel data template

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

(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

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

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

As seen on TV!
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

**LTTER-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, EDlutils, 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
Workflow

Level 0
Raw

Step 1
Custom code

Level 1
ecocomDP

Data package in ecocomDP model, with provenance, in PASTA

Step 2
EDI has code to
- Create a Darwin Core Archive from ecocomDP
- Register dataset with GBIF
- Update endpoints and metadata for new revisions

Level 2
Darwin Core Archive

Data package(s) in PASTA, sent at GBIF

Example

https://www.gbif-uit.org/dataset/cfb3f6d5-ed7c-44ff-9f1b-4032ed1de485
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
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 STXML 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 ed1.628.1.xml
ed1.628.1.xml

Missing custom unit id(s): ['gramPercentimeterSquared']
```python
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)

p = Parser()
p.parse(xml)
```

Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/home/servilla/git/EMLPy/src/emlen/parser.py" line 199, in parse
raise exceptions.ParseException(msg_queue.strip())
emlvp.exceptions.ParseException: Missing custom unit id(s): ['gramPercentimeterSquared']
```
Introducing **ltertools**

Tools Developed by the Long Term Ecological Research Community
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

Contribution Credit

Contributed functions net you 3 modes of credit:

1. Package authorship

Authors and Citation

Authors
Nicholas Lyon. Author, maintainer. 🌐
https://github.com/nlyon
Angel Chen. Author. 🌐
angelchen@ucla.edu
National Science Foundation. Funder.
NSF 1929393, 08/01/2019 - 08/31/2024.
University of California, Santa Barbara. Copyright holder.

Citation
Source: DESCRIPTION
Lyon N, Chen A (2023). ltertools: Tools Developed by the Long Term Ecological Research
Community. R package version 0.0.0.9000, https://github.com/nlyon/ltertools

2. Named in function documentation

Calculate Coefficient of Variation

Description
Computes the coefficient of variation (CV), by dividing the
standard deviation (SD) by the arithmetic mean of a set of
numbers. If `na.rm` is TRUE then missing values are
removed before calculation is completed. This function was
built by the following authors: Nicholas Lyon

3. Named in GitHub README

README.md
- `cv` – Calculate coefficient of variation (CV) from a vector of numbers. Function written
  by Nick Lyon
“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

---

**Lunch Time!**

Return around 1:45 pm
Afternoon activities

Group Photo

Units WG activity

SLIDES AT: https://docs.google.com/presentation/d/1AB1uWsNx_IBELOC5MBkNhToEUt-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_u1woW19iTWPx-55IT-PMns5HAD6Eb65aBOMY3bo/edit?usp=sharing
6pm: Dinner at Masala Elaichi
*Then we’ll see where the night takes us!*

[Image: Masala Elaichi logo]

THANK YOU GREG!

[Website: https://masalaelaichi.com/]
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

[Diagram: Quantities, Units, Dimensions and Types Ontology]

https://QUDT.org
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
(\texttt{<standardUnit>,<customUnit>}), but other important parts of a measurement are in free text \texttt{(attributeName, attributeDescription)}

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

\textit{WE ARE ADDRESSING UNITS HERE - NOT MEASUREMENTS}

TYPICAL UNIT DESCRIPTION

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

QUOTD records contain
\begin{itemize}
  \item definitions and conversion factors
  \item IDs from other unit systems (e.g., ucumCode, uneceCommonCode)
\end{itemize}

https://ble-iter.github.io/QUOTD_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>
      <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
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

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

EDI: 90.5%
of 176,722
NEON: 96.1%
of 15,765
DataONE: 87.8%
of 162,570
**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

---

**Web Service Demo**

[https://www.vcrlter.virginia.edu/data/test_unitsws.html](https://www.vcrlter.virginia.edu/data/test_unitsws.html)
### Web Services

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

### DIY, with R-code & your configuration

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>EML document</td>
<td>EML document with unit annotations and attribute IDs added (if not present)</td>
</tr>
<tr>
<td>Directory of EML documents</td>
<td>new directory of EML docs with unit annotations and attribute IDs (if not present)</td>
</tr>
<tr>
<td>QUDT ID</td>
<td>EML <code>&lt;annotation&gt;</code> element</td>
</tr>
</tbody>
</table>
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:
ADDITIONAL QUESTIONS FOR DISCUSSION

How can units annotations be leveraged (e.g., automate STMMML 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 QUODT 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
ug/l
micrograms/liter
micrograms/l
microgram per l

"Pseudounit"

microgramperliter
Automated Mapping to QUDT URIs

- microgram/permiliter - base pseudounit
  - MicroGM/permiliter - substitute for “microgram”
    - MicroGM/permiliter - substitute L for liter
      - MicroGM-PER-L - replace “per” with -PER-
        - http://qudt.org/vocab/unit/MicroGM-PER-L - Draft URI

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

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

“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, celsius)
   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” 5,297 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) Misspelled units (e.g., miligram, precent, hectorpascal, celsius)
   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” 5,297 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

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. STMMI
**QUDT TO STMML**

Breakout group questions

1. What do you need to incorporate QUODT URIs?
   a. As `<annotation>` elements
   b. As `<customUnit>` with STMML

2. How should “new” units be identified and linked to QUODT 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/1FDKwK3CNLXX9x+iQBXUqv67HBY6pUMUH_oXbr4xXrFWc/edit#
Appendix C: EML Best Practices WG Activity

Slides on Google Drive: EML Best Practices WG activity

---

EML Best Practices WG activity

---

Working group overview

Revising the ~2017 Version of the "EML Best Practices" document (link)

Current draft sections are in the IMC Google Drive

Members: Corinna Gries (NTL/EDI), Gabriel Kamener (FCE), Sage Lichtenwalner (PAL), Mary Martin (HBR), Greg Maurer (JRN), John Porter (VCE), Tim Whiteaker (BLE)

and we're coming for the rest of you...
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?

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

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?
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.