



## LTER Databits

Information Management Newsletter of the Long Term Ecological Research Network

2018 *Spring Issue*

***Welcome to the Spring 2018 Issue of Databits!***

**Editorial Notes:** This issue features articles that present a wide set of themes including historical facts, a vision for collaboration strategies given new partnerships and challenges, descriptions of systems that produce EML packages to deposit data in data repositories, individual experiences of special events like the 2017 solar eclipse and a tropical hurricane, and information management post-LTER following termination of an LTER site from the former information manager.

We wanted to acknowledge the fact that the US LTER is not the only community of LTER scientists and information managers in the world. We invited the ILTER community of IMs to share their stories in this issue, so we can learn from each other and foster future collaborations. The reader will learn of the complexity of the ILTER community as a network of networks, as well as independent countries that want to participate in the effort of documenting and sharing data (see Blankman's article on ILTER). Readers will also learn about their challenges, goals, and solutions for establishing a common data repository including the need to develop common guidelines and governance to deal with local issues related to data sharing (see Peterseil's article on the eLTER Information System). Readers will learn about the similarities between ILTER and US LTER goals. Methods may vary, but both have adopted EML as a universal metadata language and have used the LTER Controlled Vocabulary as a keyword thesaurus. There are important differences between these communities that the reader can extract from these articles. It is important to keep in mind the difference in their definition of a "site". While a US LTER site represents all administrative and scientific resources to achieve scientific research, an ILTER site is the place where the data are gathered. Another common term with completely different meanings is "EDI". In the eLTER information managers' (IMs) world, the acronym "EDI" represents a metadata editor client, while in the US world it is an organization. Another interesting yet subtle difference is the reason for developing governance plans (see Earl et al.'s IMEXEC Message and Peterseil et al.'s article on an eLTER Information Management System).

It is evident from these articles that the IM community is a typical LTER community: it encompasses a wide variety of people with different perspectives on how to perform a common task; hence they produce a wide variety of solutions, especially in the US case. All members of our community (US and International collaborators) continue to make efforts to develop ways to produce EML packages with the goal of publishing the data in a data repository of their choice. In the wide spectrum of perspectives on how to achieve this common task, we see that all IMs are working to find the best way to produce good metadata to foster data sharing, synthesis, and sustainability of their system. Most sites have developed systems that best suit their site's data type and resources (see Walsh's article on BES Metadata Management Facility). The US LTER's new partner, EDI, has designed a system to provide scientists with do-it-yourself tools (see Smith's EML Assembly Line"). The International IM community has developed hybrid systems, including DEIMS (also used as the Information Management System by 6 of the US LTER sites). Two sites (see Kui and OBrien's article on Postgres) combine the use of R scripts and a relational database to produce EML files.

The US IM community is now facing changes in the way we operate, which brings new challenges. We need to develop plans to collaborate with our new partners (the NCO and EDI) who in turn are delineating strategies for conducting collaborations among members of the scientific community, including IMs, to develop synthesis projects. We want to further develop governance methods to continue our leadership in the field of informatics (see Vanderbilt and Gries' article on the EDI Initiative and Earl et.al.'s Message from the IMEXEC).

Also featured are two articles that assess the duration of US datasets for the LTER Network as a whole, and our present resources to suggest which future endeavors we should target (See Porter's articles on Visions of the LTER and Durations of LTER Datasets).

It is interesting to notice that early members of the US LTER IMC recommended annual meetings for this group (see Henshaw's IMC meeting history). It is my opinion that this proved to be the key to success, since it has fostered the development of standards (like EML), use of common vocabularies (data set keywords) which enhance data discovery, and the use of a standard set of units which ultimately facilitates data synthesis.

At present, we have 3 IMC members who have been the site data manager (now called information manager) since the 1980s, and many others that have stayed for more than 10 years (see Henshaw's article on the History of IMC meetings). The range of years of experience is wide and, in the past few years, the vision of the role of IMs has varied. All this accounts for the richness in solutions as well as conflicting perspectives on the role of an information manager in the LTER world. As one of the 3 "oldest" information managers of this group, I trust that whatever we decide to accomplish in the future as a group will be decided in an environment full of mutual trust and respect, as it has been throughout the past 35 years.

Editor: Eda C. Meléndez-Colom (LUQ); Co-Editors: Donald Henshaw (AND) and Hope Humphries (NWT)

## Featured Articles

### *International Section*

[The International Long-Term Ecological Research \(ILTER\) Global Network](#)

[eLTER Information System - a European contribution to share scientific data from long term ecosystem research in Europe](#)

### *USA Guest Section*

[The red thread of long-term networked information management re-imagined for LTAR](#)

### *USA Section*

[A history of LTER Information Management Committee \(IMC\) Meetings: venues and participation](#)

[Baltimore Ecosystem Study Metadata Management Facility](#)

[Hello from the BLE LTER Information Manager](#)

[Postgres, EML and R in a data management workflow](#)

[The EML Assembly Line: A Metadata Generation Tool for Data Providers in the Ecological Sciences](#)

[The Environmental Data Initiative - the first 1.5 Years Supporting LTER Information Managers](#)

## Commentaries

[A Message from IMEXEC: where we have been and where we are going](#)

[A survival kit for the shadows of a natural disaster](#)

[Duration of LTER Datasets](#)

[Total Eclipse of the Sun at the Andrews Forest](#)

[Visions of LTER IM: A Discussion at the 2017 Meeting](#)

## Good Reads

[The FAIR Guiding Principles for scientific data management and stewardship](#)

## Featured Articles – International Authors

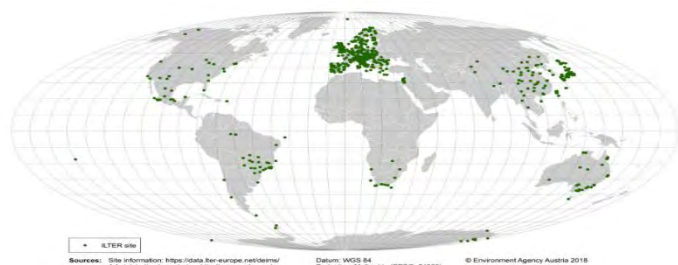
# The International Long-Term Ecological Research (ILTER) Global Network

**David Blankman**

**Chair, ILTER Information Management Committee, Director, Information Management, Israel LTER**

[ILTER](#) is a global network (Fig. 1) of long-term environmental observation sites, organized into regional and national member networks, and allowing affiliate membership for sites that do not have an appropriate country member to join. ILTER is organized into 4 regional networks (Americas, Africa, Europe, and East Asia), and at present comprise 44 members.

FIGURE 1 ILTER GLOBAL SITES



Sources: Site information: <https://data.ter-europe.net/ter/>  
Administrative boundaries: <http://www.gadm.org/>  
Datum: WGS 84  
Projection: Mollweide (EPSG: 54009)  
© Environment Agency Austria 2018

Country networks vary significantly in size and sophistication, and this is one of the challenges facing ILTER in the development of global e-infrastructure in support of its members' activities.

Affiliate membership usually entails individual sites joining a network in a neighbor country.

FIGURE 2 DEIMS-SDR METADATA INFRASTRUCTURE



Until now the focus of ILTER information management has been on gathering and dissemination of site information metadata, based on a system developed by the US ILTER (Fig. 2). This infrastructure, called [DEIMS-SDR](https://data.ilter-europe.net/deims) (<https://data.ilter-europe.net/deims>), requires approximately 40 site descriptive elements and is [currently relatively well populated](#) by contributing members. To date, [eILTER](#) (a European Union funded project) has provided funding for the development of DEIMS-SDR. While DEIMS-SDR started out as an ILTER only facility, other entities are now using it as well, including the [Critical Zones network](#).

ILTER objectives require us to develop a vision for a standards-driven, federated e-infrastructure serving the needs of scientists, institutions and networks, the funding community, and global initiatives. ILTER Information Management Committee is currently in the process of developing a strategic plan to focus on delivery of such a vision. The plan will need to focus on several questions and challenges, including:

1. How should ILTER interact with other global initiatives such as [GEO](#)/GEOSS[1] ([GEOBON](#)), [Future Earth](#), [Belmont Forum](#), and [RDA](#)?
2. In an era of [Open Science](#) and Open Data, what range of data policies are acceptable to member networks, and how can one limit the number of resulting [licenses](#) applicable to data?
3. How do we integrate networks of varying sophistication into a global e-infrastructure, while providing services and resources to less sophisticated participants?
4. How does [governance](#), [funding](#), and [participation](#) work in a federated, distributed infrastructure? [Can we certify our member networks and contributing data centers as trusted repositories?](#)
5. Which data and metadata standards are applicable, and how does one allow for flexibility while achieving desired levels of interoperability?
6. Furthermore, can one use the thrust to standardize variables and observation protocols across ILTER to inform and improve semantic interoperability in our data and metadata?

In the process, one needs to re-use and integrate existing investment to their maximum possible extent and allow incremental extension and improvement to the value offered by ILTER.

## **eLTER Information System - a European contribution to share scientific data from long term ecosystem research in Europe**

**Johannes Peterseil<sup>1</sup>, John Watkins<sup>2</sup>, Vladan Minic<sup>3</sup>, Ralf Kunkel<sup>4</sup>, Alessandro Oggioni<sup>5</sup>, Christoph Wohner<sup>1</sup>, Barbara Magagna<sup>1</sup>, David Ciar<sup>2</sup>, Jürgen Sorg<sup>4</sup>, Michael Mirtl<sup>1</sup> & Vladimir Crnojevic<sup>3</sup>**

<sup>1</sup> Umweltbundesamt GmbH (EAA), Spittelauer Lände 5, 1090 Vienna, Austria

<sup>2</sup> Centre of Ecology and Hydrology (CEH), Bailrigg, Lancaster, LA1 4AP, UK

<sup>3</sup> BioSense Institute (BSI), Zorana Đinđića 1, 21101 Novi Sad, Serbia

<sup>4</sup> National Research Council of Italy IREA (CNR-IREA), Via Bassini 15, 20133 Milano, Italy

<sup>5</sup> Forschungszentrum Jülich (FZJ), Wilhelm-Johnen-Straße, 52425 Jülich

### **Abstract**

Providing quality controlled and reliable data as the basis for scientific analysis and as input for environmental policies is one of the major aims of long term ecosystem monitoring and research. In order to foster information exchange and sharing, data must be discoverable and at least the metadata accessible. This requires proper documentation of data and services as well as the existence of infrastructure allowing the discovery and access of data in a web-based environment.

The eLTER Information System is a toolset developed to share and publish information resulting from the LTER network in Europe. This focuses on (a) adopting standards for documentation of research objects (observation facilities and datasets), (b) the use of controlled vocabularies, (b) the provision of time series data in standardised form, and (c) a catalogue of datasets across the different data resources.

### **Introduction**

Providing quality controlled and reliable data as the basis for scientific analysis and as input into the construction of new and evaluation of existing environmental policies is one of the major aims of long-term ecosystem monitoring and research not only in Europe (Mirtl 2010) but also on a global scale (Mirtl et al. in prep). In order to foster information exchange and sharing, data must be discoverable and at least the metadata accessible (Michener et al. 1997). This requires proper documentation of data and services as well as the existence of infrastructure allowing the discovery and access of data in a web-based environment.

The regional Long-term Ecosystem Research in Europe (LTER-Europe<sup>1</sup>) is a collaboration of 25 national long-term ecosystem research networks comprising 479 LTER sites and Long Term Socio-economic and Ecosystem Research (LTSER) platforms. The site network is an important component in the European landscape for ecosystem focused research infrastructures (Mirtl 2010). LTER-Europe membership consists of national networks and is part of the global International LTER (ILTER) network.

One of the goals of LTER is to improve comparability of long-term ecological data and facilitate exchange and preservation of these data (Vanderbilt et al. 2015). Currently, funding and organisation

---

<sup>1</sup> See <http://www.lter-europe.net/lter-europe>



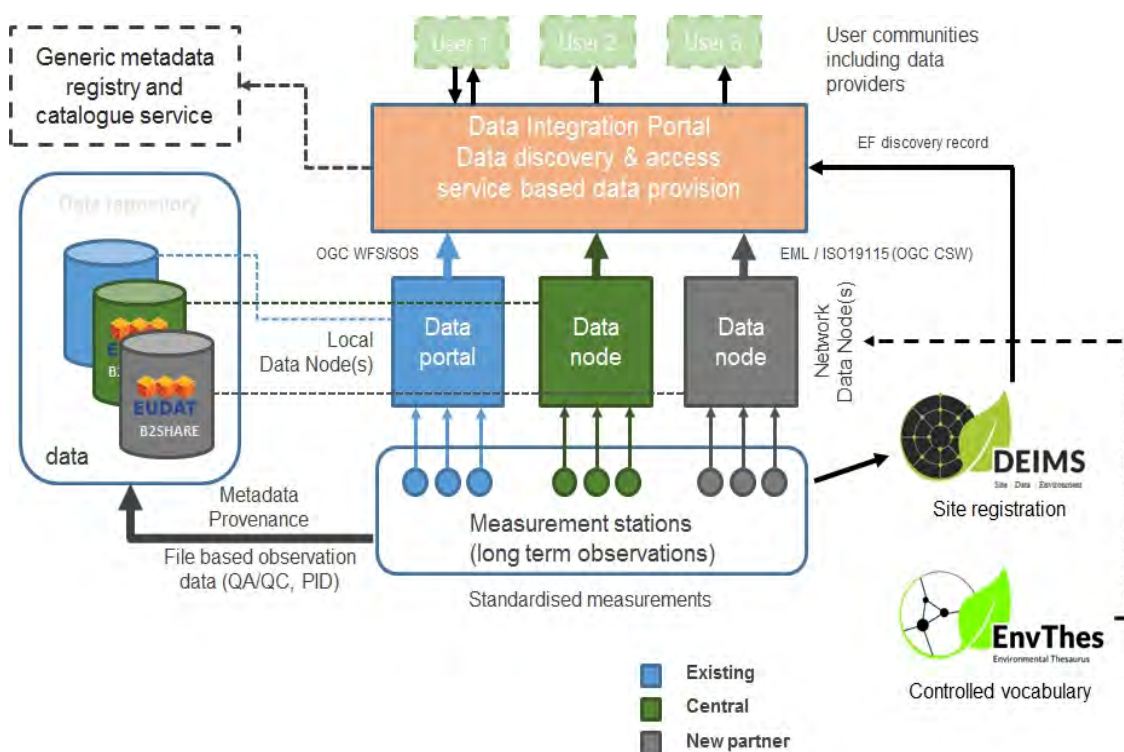
of the different components of the LTER network is strongly related to national funding opportunities thus leading to a diversity of data strategies and data management procedures. In order to overcome these challenges the European Union funded project eLTER (see <http://www.lter-europe.net/elter/about>, Grant number 65359), which aims to support this process and helps to develop and maintain infrastructure components for the publishing and sharing data sources from the LTER-Europe network.

This focuses on (a) adopting standards for documentation of research objects (observation facilities and datasets), (b) fostering the use of controlled vocabularies, (c) providing time series data in standardised form, and (d) providing a catalogue of datasets across the different data resources.

### Overview of the eLTER Information System

LTER is characterised by high inter-network as well as intra-network heterogeneity dealing with a variety of technical information management capabilities. Thus, a network of distributed (meta-)data sources linking to a common discovery portal is a way to unify and ease the access to data and information for end users. To implement this approach, we need to focus on a) a common architecture for the integration of distributed information and b) the use of common standards.

FIGURE 1. CONCEPTUAL ARCHITECTURE OF THE ELTER INFORMATION SYSTEM



### Conceptual architecture

Within the eLTER (H2020) project, LTER-Europe works towards the implementation of this vision. Figure 1 shows the conceptual architecture linking local and regional data nodes by the means of a common description of sites, datasets and data services including the following components:

- Site registration (DEIMS-SDR), providing harmonised and standardised documentation of long term observation facilities

- Data nodes (DN), providing metadata and access to data (including the link to data repositories, if data are stored in external trusted repositories)
- Data Integration Portal (DIP), providing tools for the discovery and access to data sources provided through the data nodes
- Common controlled vocabulary (EnvThes), providing a semantic backbone for keyword tagging and discovery

All different components of the eLTER Information system are interlinked using references in the metadata and standard web services.

The integration and discovery layer LTER-Europe uses GeoNetwork<sup>2</sup> and B2FIND as underlying technology applying ISO19115/19139 as basic dataset metadata schemata. In addition, access to and visualisations of time series data services based on OGC SOS is provided.

## Common standards

### Site metadata

Sufficient and standardised documentation of data is needed in order to ensure the sharing and reuse of data. This not only applies to the description of a single data object but also to the context of the observation, e.g., the research facility or infrastructure. For place-based observations information on the observation facilities (e.g., the research site) is an intrinsic and important asset for the discovery and reuse of data and expertise.

For the documentation of LTER sites a set of required fields was defined in order to allow proper accreditation of sites within the LTER network. Full documentation of the metadata model for the research site (Version 1.11) can be found on the DEIMS Site and Dataset Registry (see <https://data.lter-europe.net/deims/documentation/site>). LTER-Europe is also working on a common exchange format for site information adopting the INSPIRE Environmental Monitoring Facility (EF) application schema<sup>3</sup> and the provision of a cross research infrastructure site identification service (DEOS-ID) in order to reduce redundancies and improve discoverability of context information.

### Dataset metadata

Two main metadata schemas are supported for datasets. The Ecological Metadata Language (EML), a metadata specification for data resulting from the ecological domain (Michener et al. 1997), was adopted by ILTER and LTER-Europe as the main supported metadata standard. In addition, LTER-Europe also recommends the use of the INSPIRE metadata specification<sup>4</sup>, which is based on a European Community Directive. The INSPIRE directive defines the guidelines for the establishment of a spatial data infrastructure in Europe in order to support the community environmental policies, and policies or activities that may have an impact on the environment. The descriptive metadata are based on ISO19115/19139 as defined in the INSPIRE Metadata regulation (2008)<sup>5</sup>.

In order to ease the barrier of metadata provision a community metadata profile was defined selecting necessary required metadata elements to ensure discovery and reuse of data (Kliment & Oggioni 2011). This includes a mapping of the metadata elements implemented in DEIMS-SDR to both EML (Version 2.1.1) and ISO19115/139 (INSPIRE Profile).

<sup>2</sup> See <https://geonetwork-opensource.org/>

<sup>3</sup> See [http://inspire.ec.europa.eu/documents/Data\\_Specifications/INSPIRE\\_DataSpecification\\_EF\\_v3.0rc3.pdf](http://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf)

<sup>4</sup> see <http://inspire.ec.europa.eu/document-tags/metadata>

<sup>5</sup> see <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008R1205>

## Common semantics - EnvThes

The Environmental Thesaurus (EnvThes) was developed as a semantic backbone for data resulting from long-term ecosystem research and monitoring (Schentz et al. 2011, 2013). It is the core vocabulary used in the DEIMS Site and Dataset Registry to annotate keywords, research topics, observed parameters and various other metadata elements. It is a free and open thesaurus for the domain of long-term ecosystem research and monitoring including all related domains such as biodiversity, agriculture, forestry, etc. (see <http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn>).

Built on the US LTER Controlled Vocabulary (Porter 2010) as a primary source it partly incorporates and links to other relevant vocabularies including EUROVOC<sup>6</sup>, GEMET<sup>7</sup>, the INSPIRE spatial data themes<sup>8</sup>, and AGROVOC<sup>9</sup>. The vocabulary is based on current semantic web standards (SKOS and SPARQL) and supports multilinguality. Initial tests on the use of EnvThes as a multilingual thesaurus for annotation and discovery have been made (Vanderbilt et al. 2010, Vanderbilt et al. 2017).

## Implementation

### Metadata - DEIMS-SDR

The DEIMS Site and Dataset Registry<sup>10</sup> (Dynamic Ecological Information Management System, DEIMS-SDR) provides inter alia a common and standardised catalogue for the distinct identification of observation facilities (e.g., sites, stations, sensors) that is also used by ILTER. DEIMS-SDR is based on Drupal 7 and the current version of DEIMS (Version 2), is a branch of the DEIMS system developed by US LTER (Gries et al. 2010).

In addition to being the central metadata editor for LTER-Europe, DEIMS-SDR aims to provide a viable option to fulfil metadata requirements for research projects and national networks if they lack their own system.

In addition to the documentation of datasets provided by the DEIMS core branch, DEIMS-SDR also includes the documentation of research sites, networks and persons. For each of the research sites a landing page is provided (see Figure 2) containing information on the research sites, as well as related information (e.g., datasets and data products).

---

<sup>6</sup> See <http://eurovoc.europa.eu/drupal/>

<sup>7</sup> See <http://www.eionet.europa.eu/gemet/en/themes/>

<sup>8</sup> See <https://www.eionet.europa.eu/gemet/en/inspire-themes/>

<sup>9</sup> See <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>


<sup>10</sup> See <http://data.lter-europe.net/deims/>



**FIGURE 2. SITE DOCUMENTATION OF THE LTER SITE NP KALKALPEN (AT) (DEIMS-SDR LANDING PAGE FOR SITES)**

## Kalkalpen National Park - Austria

Basic information



Site Name: Kalkalpen National Park  
Site Code: LTER\_EU\_AT\_008  
Web Address: [Homepage](#)


[Map](#)

Country (Site Location): Austria  
LTER Member Network: Austria (LTER-Austria)  
Parent Site Name: LTER Platform Ebenwurz (EW) - Austria  
Contact: Site Manager: Franziska Popperl  
Christian Furjaeger


Keywords originating from EnvEurope Thesaurus: *ecopotential*

General Site Description:  
Kalkalpen National Park is made up of two mountain ranges \* The Reichraminger Hintergebirge is one of Austria's largest distinct forest areas - a sea of forest, which has not yet been dissected by public transportation routes and human habitation. Here, you will also find one of the longest intact stream systems of the Eastern Alps. Old shelters and overgrown trails remind us today of how wood was used and harvested in earlier times. \* The Sengengebirge is a northern outpost of the Limestone Alps. The ca. 20 km long main ridge reaches its highest point at the Hoher Nock (1,963 m). The name  
... [Show more](#)

UUID: 49515dda-1198-4013-8f43-c33e107af081

 Environmental monitoring facilities

Photos



© NP Kalkalpen/E. Mayrhofer

General Characteristics, Purpose, History


Metadata provider: Franziska Popperl  
Christian Furjaeger  
Forschung Kalkalpen

Site Status: existing  
Year Site Established: 1993  
Size : 20 850.00ha  
Purpose of Site :  
National parks provide enduring protection to unique natural landscapes for the benefit of future generations According to the definition by the International Union for the Protection of Nature and Natural Objects, national parks are natural areas on water or land, which are designated to protect the integrity of one or several ecosystems and to preserve them for current and future generations. They are intended to prevent exploitation and other activities that may cause damage to the area. They are also meant to provide a basis for spirituality, research, education, recreation, and sightseeing  
... [Show more](#)

Research Topics:  
*biology conservation ecology environmental science geography geology history hydrography hydrology limnology management meteorology*

Parameters:  
*ecosystem measure biological measure atmospheric measure experimental measure landscape measure soil measure water measure*

Geographic



Coordinates:  
Latitude: 47.772360000000  
Longitude: 14.392820000000  
Site Boundaries:  
[{"type": "Polygon", "coordinates": [[[14.167513979004, 47.856595940183], [14.554408519037, 47.856595940183], [14.554408519037, 47.67898892585], [14.167513979004, 47.67898892585], [14.167513979004, 47.856595940183]]]]]  
[Download boundaries \(.shp\)](#)

DEIMS-SDR implements the LTER dataset community profile (Kliment & Oggioni 2011) and allows the export to different metadata formats for datasets (e.g., EML 2.1.1, BDP, ISO19115, ISO19139). Additionally, DEIMS-SDR includes an interface to directly upload datasets to the open eScience data sharing platform B2SHARE (Ardestani et al. 2015).

## Data provision - Get-IT Data Node (DN)

The creation of a data node should enable the user a) to publish data, both geospatial and observations, through standard web services; b) to create spatial data repositories; and c) to facilitate the entry and maintenance of research and sensor data and metadata. GET-IT<sup>11</sup>, developed by a joint research group of CNR IREA and CNR ISMAR under the flagship project RITMARE<sup>12</sup>, is a software suite that aims to enable researchers to setup and operate an interoperable SDI following relevant standards from the OGC (WMS, WFS, WCS, CSW, SensorML, and SOS). Services, with entered data and metadata, are hosted by virtual machines that can be installed in server or in hosting sites.

Within eLTER GET-IT is further customised and updated to the newest version of the underlying software stack. GET-IT consists of a virtual machine, based on the Ubuntu operating system. The basic software component used in GET-IT is GeoNode<sup>13</sup>. While GeoNode does not typically include Sensor Web Enablement (SWE) and semantic enhancement, developments to overcome this shortcoming have been made. In particular, the new software implementations include:

<sup>11</sup> See <http://www.get-it.it>

<sup>12</sup> See <http://www.ritmare.it>

<sup>13</sup> See <http://geonode.org>

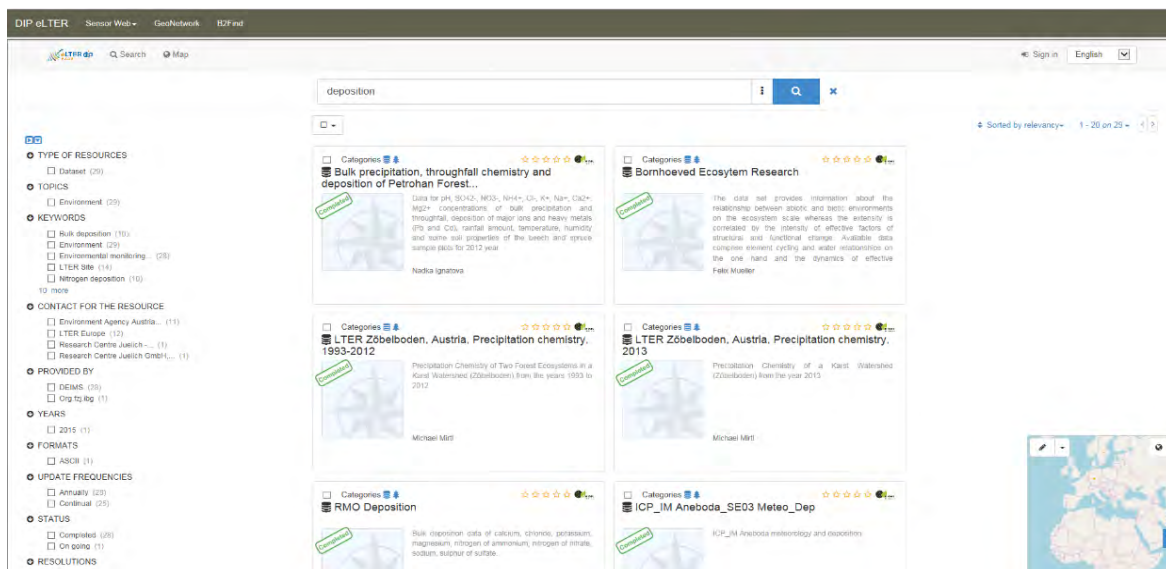
- A metadata editor client, named EDI, which allows the creation and validation of metadata in accordance with different profiles or templates. EDI allows plugging in external data sources that are made available as SPARQL endpoints;
- A SOS manager (52°North SOS<sup>14</sup>) that allows the registration of new sensors edited in Sensor Metadata Language (SensorML) metadata profile by EDI metadata editor;
- An insert observation interface allows uploading observations through copy and paste actions for a GET-IT registered sensor;
- A geographic data manager (GeoServer<sup>15</sup>) that allows sharing geographic data;
- An SOS client that allows viewing the information of registered sensors and data recording in a web map.

GET-IT is free and open source and is used within eLTER in setting up the central data node as well as contributing partner nodes (see Figure 1).

### Data discovery – Data Integration Portal (DIP)

Within the eLTER (H2020) project LTER-Europe works on the development of a Data Integration Portal<sup>16</sup> (DIP) based on GeoNetwork and B2FIND (see Figure 3). This portal should allow, in addition to the discovery of metadata, direct access to time series data services based on OGC SOS (see Figure 4). The implementation is based on a set of standards implemented through compliant software modules provided by GeoNetwork<sup>17</sup> and 52°North SOS. The metadata integration layer, based on this software stack, collates information from the different eLTER data nodes in the network and enables users to browse the metadata records and spatial information that describe the data available from the eLTER site network.

FIGURE 3. ELTER DIP DISCOVERY OF METADATA



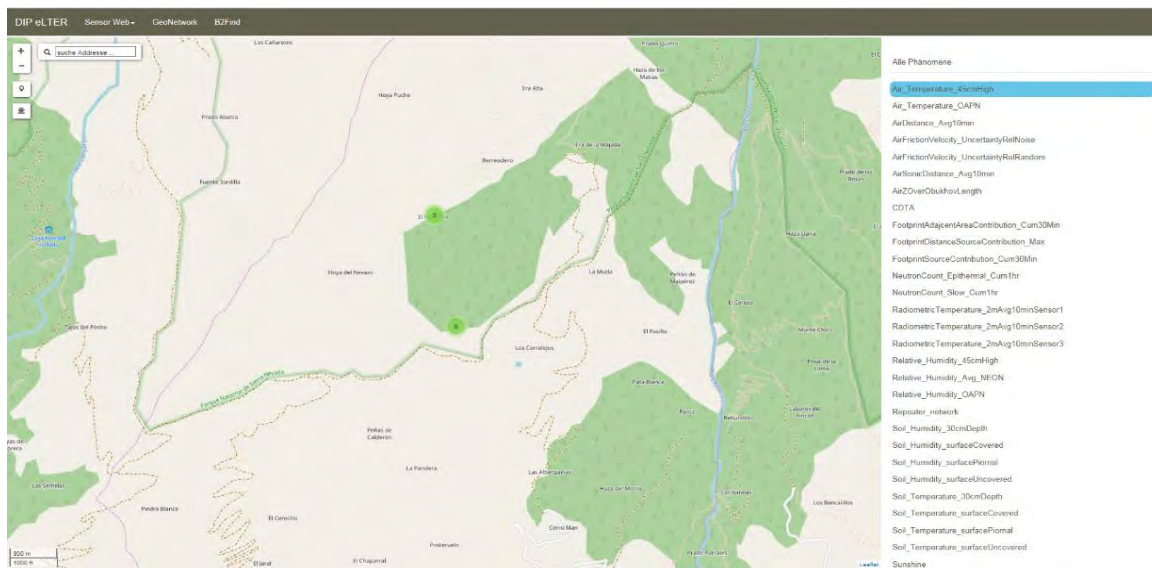
<sup>14</sup> See <http://52north.org/downloads>

<sup>15</sup> See <http://geoserver.org>

<sup>16</sup> See <http://dip.biosense.rs/>

<sup>17</sup> See <https://geonetwork-opensource.org/downloads.html>

**FIGURE 4. OVERVIEW MAP ON LTER SITES AND SOS PROVIDERS**



The harvesting of metadata records from the eLTER data nodes is based on OGC CSW services providing ISO 19139 standard dataset and service descriptions. These records can be browsed in the MD Discovery Layer.

Data search can be done using map services and time series graphing options through the Spatial Data Access Layer and the Time Series Access Layer. Sensor data can be accessed using the OGC SOS standard services between clients in the DIP and services in the eLTER data nodes. A visualization module will be integrated with the EnvThes vocabulary server in order to achieve unification of phenomena names that are currently not standardised across the network.

## Governance

Development work on the eLTER Information System satisfies important steps toward the implementation of tools and services enabling researchers and users to easily document and share the data. Nevertheless, the cultural and social aspects of data sharing need to be taken into account (Vanderbilt et al. 2015, Vanderbilt & Gaiser 2017). While agreeing on open data in principle on the global scale the implementation of common data sharing on the local level is still an issue in many of the member networks. To enable this vision eLTER is not only addressing the technological aspects of data publishing and sharing but also the social aspects. Common guidelines and governance will be developed in order to ensure the sustainability of data provision and updating.

## Summary

The eLTER Information System aims to provide a framework for the integration and provision of observation data from the different components of the LTER-Europe network. Building on standards and service interfaces it aims to implement the FAIR principles for open data sharing. The framework is an endeavour of LTER-Europe which will be undertaken with upcoming eLTER Research Infrastructure.

Nevertheless, the different components of the eLTER Information System are still in development. In addition to the development, linking to global data infrastructures as GEOSS and DataOne is tested and prepared.

Besides the different components in the eLTER Information System, the site documentation in DEIMS-SDR provides a global service which can also be used by other observation networks beyond



LTER-Europe and ILTER. Standard information exchange based on INSPIRE EF data specification and OGC services provides a valuable tool.

## Acknowledgement

The work on the eLTER Information System is funded by the eLTER (H2020) project. The DEIMS-SDR development is supported by the European H2020 projects eLTER, ECOPOTENTIAL, and EUDAT2020. All work is done in line with the information management activities of LTER-Europe and ILTER. The work was also supported by David Blankman and Inigo San Gil.

## References

- Ardestani, S.B., Blommesteijn, D., Dima, E., Hakansson, C.J., Laure, E., Livenson, I., Stranák, P., & Sanden, M.V. (2015). B2SHARE: An Open eScience Data Sharing Platform. e-Science. (link: <https://www.semanticscholar.org/paper/B2SHARE-An-Open-eScience-Data-Sharing-Platform-Ardestani-Hakansson/dd55f56ae4fe6c34fe0ed361aae9dfc7600c7b52>)
- Gries, C., San Gil, I., Vanderbilt, C. & Garrit, H. (2010) Drupal developments in the LTER network. (link: <http://databits.lternet.edu/spring-2010/drupal-semanticdevelopments-lter-network>)
- Kliment, T. & Oggioni, A. (2011) Metadatabase: EnvEurope Metadata specification for Dataset Level. EnvEurope (LIFE08 ENV/IT/000339) Project Report PD.A1.1.4 87pp. (link: [http://www.enveurope.eu/misc/PD\\_1\\_1\\_4\\_Kliment\\_Metadatabase\\_201112\\_final\\_v1.0.pdf](http://www.enveurope.eu/misc/PD_1_1_4_Kliment_Metadatabase_201112_final_v1.0.pdf))
- Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, Th.B., Stafford, S.G. (1997) Nongeospatial metadata for the Ecological Sciences. *Ecological Applications* 7:330-342
- Mirtl et.al (in prep). Genesis, Goals and Achievements of Long-Term Ecological Research at the global scale: A critical review of ILTER and future implications. *Science of the total environment*.
- Mirtl M. (2010) Introducing the Next Generation of Ecosystem Research in Europe: LTER-Europe's Multi-Functional and Multi-Scale Approach. In: Müller F., Baessler C., Schubert H., Klotz S. (eds) *Long-Term Ecological Research*. Springer, Dordrecht
- Porter, J.H. (2010) A controlled vocabulary for LTER datasets. (2010) (link: <http://databits.lternet.edu/spring-2010/controlled-vocabulary-lter-datasets>)
- Schentz, H., Peterseil, J. & Bertrand, N. (2013) EnvThes- interlinked thesaurus for long term ecological research, monitoring, and experiments. *Proceedings EnviroInfo 2013: Environmental Informatics and Renewable Energies*. Shaker Verlag, Aachen.
- Schentz, H., Peterseil, J., Magagna, B. & Mirtl, M. (2011) Semantics in Ecosystem Research and Monitoring. *Proceedings EnviroInfo 2011: Innovations in Sharing Environmental Observation and Information*. Shaker Verlag, Aachen.
- Vanderbilt, K., and E. Gaiser. 2017. The International Long Term Ecological Research Network: a platform for collaboration. *Ecosphere* 8(2):e01697. 10.1002/ecs2.1697
- Vanderbilt, K., John H. Porter, Sheng-Shan Lu, Nic Bertrand, David Blankman, Xuebing Guo, Honglin He, Don Henshaw, Karpjoo Jeong, Eun-Shik Kim, Chau-Chin Lin, Margaret O'Brien, Takeshi Osawa, Éamonn Ó Tuama, Wen Su, Haibo Yang (2017) A prototype system for multilingual data discovery of International Long-Term Ecological Research (ILTER) Network data. *Ecological Informatics* 40:93-101, <http://dx.doi.org/10.1016/j.ecoinf.2016.11.011>.
- Vanderbilt, K., Lin, Ch.Ch., Lu, Sh-Sh., Kassim, A.R., He., H., Guo, X., San Gil, I., Blankman, D. & Porter, J. (2015) Forstering ecological data sharing: collaborations in the international Long Term Ecological Research Network. *Ecosphere*(10) Article 204 18pp.

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Lin, Ch-Ch., Lu, S.-S., Ogawa, A., Ó Tuama, É., Schentz, H., Su, W. (2010) A multilingual metadata catalog for theILTER: Issues and approaches, *Ecological Informatics* 5:187-193.

## Featured Articles – USA Guest

# The red thread of long-term networked information management re-imagined for Long-Term Argoecosystem Research

**Nicole Kaplan**

**Computational biologist/information manager for the new Long-Term Argoecosystem Research Network site at the Central Plains Experimental Range (CPER), Co-leader of the LTAR Information Management Task Force (IMTF), and former Information Manager of the Shortgrass Steppe (SGS)**

I served as field crew manager for the SGS-LTER from 1998-2004 and information manager from 2000-2014. The overlap of my years working in the field and managing the data within the local and LTER network cyberinfrastructure gave me an appreciation for how crucial it is to understand and document the placed-based context in which the data were collected. In 2010, when the National Science Foundation terminated funding for the shortgrass steppe LTER project, my work shifted into decommissioning, which involved packaging, transforming and migrating SGS-LTER data from the local data management system into the Colorado State University (CSU) institutional repository (Kaplan et al. 2014). Despite having no formal guidelines available from the LTER program, we had a community of practice of information managers, and a smaller group of colleagues concerned particularly with information management tasks upon project termination. With input from the group and new partners at the CSU Libraries, we formed a collection of data that captures digital outcomes from the SGS-LTER era. Local infrastructure was created, that is 'infrastructuring ecology' (Baker and Millerand, 2010), including materials that provide a more detailed story as to why, how and where data were collected (e.g. proposals, protocols, images) and what products those data support (e.g. dissertations, papers, presentations). The collection was positioned as a local asset within the CSU research infrastructure and as interoperable with the LTER Network Information System (i.e. PASTA).

In the following years, as I attended celebratory milestone events for the SGS-LTER site manager's kids, whom I watched grow up over the course of the SGS-LTER project, I realized how time really does fly. Reflecting upon the termination and decommissioning work, I was confident that we created a rich legacy of data and artifacts that would be discoverable and accessible to researchers, especially those still working on the Central Plains Experimental Range (CPER), which had hosted SGS-LTER's core studies. The CPER has a >75-year history of research under the USDA Agricultural Research Service (USDA ARS), continues to support various projects and partnerships and data continue to be collected there. Data are collected on vegetation production and plant species composition, livestock weight gains, movement and health, climatological conditions, soil carbon, carbon and trace gas fluxes, precipitation variability, drought, elevated atmospheric CO<sub>2</sub> and increased temperatures, as well as remotely sensed images with high spatial (<10 mm) and spectral (>400 band) resolution. The CPER has many successful partnerships, both scientific and producer/customer related, all based in part on expectations of being able to share information and data. The Crow Valley Livestock Cooperative, Inc. is the oldest grazing association in the US and



has been a collaborative partner since the formation of CPER in 1937, and the Forest Service is also a key partner with collaborations in research that extend to the adjacent Pawnee National Grassland. New collaborations exist with the National Ecological Observatory Network (NEON) through CPER being the core site for Domain 10 and new networks, Greenhouse Gas Reduction through Agricultural Carbon Enhancement network (GRACEnet), and Nutrient Network (NutNet). Traditionally, USDA ARS research units have not created positions for information management within their organizations. With expectations of collaborators to share data in an age of open access (Krishnamurthy and Awazu 2016), a US federally funded research unit needs a dedicated information manager. In 2012 the need for information managers was formally identified when USDA established the Long-term Agroecosystem Research (LTAR) Network across 18 research sites in the US (Robertson et al. 2008).

My membership in the LTER Network ended, but my connection to the CPER as a place of research remained. In 2015, I was hired as computational biologist/information manager for the new LTAR Network site at CPER. In this role, I am able to re-engage in research data stewardship for CPER-LTAR. The LTAR network seeks to determine ways to ensure sustainability and enhance food production (and quality) and ecosystem services at broad regional scales. Research teams of scientists, support staff, university students and post-docs are conducting common experiments across the LTAR network to compare traditional production strategies with aspirational strategies, which include novel technologies and collaborations with farmers and ranchers. Success within LTAR research teams and for the network, requires data and metadata be provided in findable, accessible, useable, well documented and standardized formats so that information from sites can be applied to answering complex questions across regions.

I also co-lead the LTAR Information Management Task Force (IMTF) and find myself re-imagining information management for a new long-term collaborative research network and frequently refer to lessons learned from LTER. The LTAR IMTF has advised research teams to include an information manager as a participant in their scientific working groups, which has been formed to answer questions at broader scales that require data to be mobilized and integrated from across sites, an approach we discussed for LTER synthesis groups, but that I never fully experienced. The notion of an embedded information manager allows for co-design of data management solutions in parallel with the scientific process and facilitates trust in data stewardship. The LTAR is making published data collections findable via Ag Data Commons, the USDA institutional repository hosted by the National Agricultural Library (Waide et al. 2017). Data from three LTAR sites, CPER, JRN and KBS, which have also produced LTER data, have been made discoverable within the Ag Data Commons via web services for PASTA developed by the Environmental Data Initiative.

The future has opportunities as well as challenges, but I maintain an open mind regarding new technologies including, R for data wrangling and analysis, and ArcGIS Online for story maps to visualize data for a variety of audiences. Learning to manage new types of data and significantly larger data files generated from sensors has been fun and I have been fortunate to learn new tools from the next generation of scientists and generous colleagues in LTAR as well as LTER.

Lastly, in traveling with my family recently, we visited Oman, a country with rich reserves of oil that was able to use income from their natural resources to supplement incomes of those less fortunate. This is not a model of winners and losers but rather one that cultivates a shared community goal to leverage resources bringing the entire country forward with modern infrastructure for health, safety and quality of life. From my work with data and the long-term perspective, I see parallels in their approach with stewardship models for collaboration and partnership in sharing data and building cyberinfrastructure for the broader research and information management community. As we

steward data collections interoperating with centralized data centers, we insure the next generations can benefit from the wealth of knowledge that was created to address new challenges.

## References

Baker, K., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. . In J. N. Parker, N. Vermeulen, & B. Penders (Eds.), *In Collaboration in the New Life Sciences*. Ashgate.

Kaplan, N. E., Baker, K. S., Draper, D. C., & Swauger, S. (2014). Packaging, Transforming, and Migrating Data from A Scientific Research Project to an Institutional Repository: The SGS LTER Collection. Retrieved from Colorado State University, Fort Collins, Colorado. Retrieved from <http://hdl.handle.net/10217/87239>.

Krishnamurthy, Rashmi, and Yukika Awazu. "Liberating data for public value: The case of Data. gov." *International Journal of Information Management* 36.4 (2016): 668-672.

G. Philip Robertson, Vivien G. Allen, George Boody, Emery R. Boose, Nancy G. Creamer, Laurie E. Drinkwater, James R. Gosz, Lori Lynch, John L. Havlin, Louise E. Jackson, Steward T. A. Pickett, Louis Pitelka, Alan Randall, A. Scott Reed, Timothy R. Seastedt, Robert B. Waide, Diana H. Wall; *Long-term Agricultural Research: A Research, Education, and Extension Imperative*, BioScience, Volume 58, Issue 7, 1 July 2008, Pages 640–645, <https://doi.org/10.1641/B580711>

Waide, R. B., Brunt, J. W., & Servilla, M. S. (2017). Demystifying the Landscape of Ecological Data Repositories in the United States. *BioScience*, 67(12), 1044-1051.

## Featured Articles – USA Authors

---

### A history of LTER Information Management Committee (IMC) Meetings: venues and participation

**Don Henshaw**

**Andrews LTER (AND), Oregon State University**

Data management was a primary emphasis at the outset of the Long-Term Ecological Research (LTER) program, which was initially funded by the National Science Foundation in 1980, and each site incorporated a data manager to properly address the expected long-term data collections. The Long-Term Ecological Research (LTER) Information Management Committee, originally just a group of site data management representatives, started meeting annually in 1982. The first “Data Management Workshop” was held in Urbana-Champaign, Illinois, in November 1982, and included representatives from each of the eleven sites.

The intent of this article is not to address the work or leadership of the committee or the themes of the meetings, but to review the history of meeting locations and site participation. It can be reported that this inaugural meeting featured topics such as data access, archiving, data set documentation, and cross-site exchange of data, perhaps surprisingly, topics which are all still relevant today.

Interestingly, the final recommendation bullet from the meeting summary stated “that the LTER data managers should meet annually so intersite cooperation and discussions can be continued.” As of the end of 2017 there have been approximately 35 annual meetings of the IMC.

Research for this article was stimulated by the discovery of historical meeting notes from early meetings held in 1982, 1983 and 1988. Meeting notes from 1989 to present have been stored in the document archive at the LTER Network Office, although only sketchy notes exist for 2000 and 2002. No meeting notes have been discovered for years 1985-1987 and the meetings listed for 1985 and 1986 are speculative based on participant memories. There is no evidence for any meeting in 1987. The summary tables in this article were developed by documenting individual participation and site representation mined from meeting participant lists.

## Meeting venues

IMC meeting venues have typically been located in conjunction with other LTER-sponsored meetings, meetings of interest, or at hosting LTER sites (Table 1.). In the earlier years meetings were held immediately before Ecological Society of America (ESA) meetings as several of the Information Manager representatives (IMs) frequently would attend ESA. Association with ESA brought the IMC twice to Canada and once to Hawaii. Fourteen meetings were held in conjunction with ESA from 1989 to 2007. Seven meetings have been held in conjunction with the LTER All-Scientist's Meeting (ASM) beginning with the very first ASM in 1985 at Cedar Creek.

Five meetings have been co-located with LTER-sponsored meetings that explored various advances in information management and provided a forum for discussion:

- Research Data Management in the Ecological Sciences Symposium, Hobcaw Barony, SC, 1984
  - Symposium book (Michener 1986)
- Spatial Data Workshop, Seattle, WA, 1994
- Data and Information Management in the Ecological Sciences (DIMES), Albuquerque, NM, 1997
  - <http://www.ecoinformatics.org/pubs/guide/frame.htm>
- Environmental Information Management conference (EIMC), Albuquerque, NM, 2008
  - <https://eim.ecoinformatics.org/eim2008>
- Environmental Information Management conference (EIMC), Santa Barbara, CA, 2011
  - <https://eim.ecoinformatics.org/eim2011>

**TABLE1. IMC MEETING LOCATIONS AND NUMBER OF PARTICIPANTS**

Year	Meeting location	Associated meeting	Total ptcps	Sites present	Sites not present	Guest (s)	Notes
1982	Urbana-Champaign, IL	IMC only	23	11	0	2	First IMC meeting; ILR host
1983	Corvallis, Oregon; OSU McDonald Forest	IMC only	25	11	0	2	Second IMC meeting, AND host
1984	Hobcaw Barony, Georgetown, SC	Data Mgmt Symposium	?	?	?	?	NIN host; DM Symposium book but no IMC notes
1985	*Cedar Creek LTER	1 <sup>st</sup> ASM	?	?	?	?	CDR host; no notes
1986	*Konza Prairie LTER	IMC only	?	?	?	?	KNZ host; no notes
1987	*Unknown		?	?	?	?	No known meeting
1988	Sacramento, CA	ESA	19	15	0	0	Possibly w/ ESA, Davis CA
1989	Toronto, Canada	ESA	20	17	0	0	
1990	Snowbird, Utah	ESA	24	17	0	1	

1991	San Antonio, Texas	ESA	25	17	1	0	
1992	Honolulu, HI	ESA	23	18	0	0	
1993	Madison, WI	ESA	26	18	1	2	NTL host
1994	Seattle, WA	Workshop	33	19	0	1	Spatial Data Workshop; LNO host
1995	Snowbird, Utah	ESA	28	17	1	0	
1996	Archbold Biological Field Station, Lake Placid, FL	Eco-Informa	25	16	2	2	Eco-Informa Conference, Orlando, FL
1997	Albuquerque, NM	DIMES	27	18	2	1	DIMES Symp.; LNO host
1998	Baltimore, MD	ESA	33	20	1	4	BES host
1999	Spokane, WA	ESA	30	19	2	5	
2000	Snowbird, Utah	ASM/ESA	41	21	2	5	With ASM and ESA
2001	Madison, WI	ESA	42	24	0	6	NTL host
2002	Orlando, FL	SCI2002	34	20	4	3	SCI2002 conference
2003	Seattle, WA	ASM	42	23	1	7	
2004	Portland, OR	ESA	42	23	2	5	
2005	Montreal, Canada	ESA	50	24	2	9	
2006	Estes Park, CO	ASM	59	23	3	23	
2007	San Jose, CA	ESA	43	23	3	7	
2008	Albuquerque, NM	EIMC	42	23	3	1	LNO host; EIMC 2008
2009	Estes Park, CO	ASM	54	24	2	12	
2010	Kellogg Biological Station, MI	IMC only	49	26	0	0	KBS host
2011	Santa Barbara, CA	EIMC	36	25	1	0	SBC host; EIMC 2011
2012	Estes Park, CO	ASM	48	24	2	9	
2013	Fairbanks, AK	IMC only	34	24	2	0	BNZ host
2014	Copper Mountain Resort, Frisco, CO	ESIP	32	24	1	0	
2015	Estes Park, CO	ASM	34	25	0	0	
2016	Santa Barbara, CA	IMC only	31	24	0	0	SBC host; NCO/ NCEAS visit
2017	Bloomington, IN	ESIP	29	21	6	2	

\* No specific evidence for this meeting or its location

Four meetings have been held in conjunction with other organizational meetings:

- Eco-Informa 1996: Global Networks for Environmental Information, Lake Buena Vista, FL, 1996 (<https://link.springer.com/article/10.1007%2FBF02986968> )
- SCI2002 6th World Multi-Conference on Systematics, Cybernetics and Informatics, Orlando, FL, 2002
- Earth Science Information Partners (ESIP), Copper Mountain Resort, Frisco, CO, 2014 (<http://esipfed.org/2014SummerMeeting>)
- Earth Science Information Partners (ESIP), Indiana University, Bloomington, IN, 2017 (<http://www.esipfed.org/meetings/upcoming-meetings/esip-summer-meeting-2017> )

Several meetings have been held in close association with a local LTER hosting site or field stations. These include Urbana-Champlain 1982 (ILR), Corvallis 1983 (AND), Hobcaw 1984 (NIN), Cedar Creek 1985 (CDR), Konza Prairie 1986 (KNZ), Madison 1993 (NTL), Archbold 1996, Baltimore 1998

(BES), Madison 2001 (NTL), Kellogg 2010 (KBS), Santa Barbara 2011 (SBC), Fairbanks 2013 (BNZ), Santa Barbara 2016 (SBC, NCO). Three additional meetings were conducted in close proximity to the LTER Network Office (LNO): Seattle 1994 and Albuquerque 1997 and 2008.

## Participation

The importance of this meeting to the IMC is evidenced by the high level of participation over the past 36 years. Current bylaws state that the site information manager is “strongly encouraged” to attend the annual meeting. Well before any IMC bylaws were developed there was a general understanding that site representation at the annual meeting was mandatory. For the 27 sites with more than 10 years of participation, 17 sites have been present for every meeting (9 sites with perfect attendance) or have missed just one. A representative from the Network Office or Communications office has been present at every meeting since 1988. All 32 sites in the 38-year history of LTER along with the LNO and Network Communications Office (NCO) have participated in IMC meetings. All of these sites and their years of participation are listed (Table 2).

All sites were represented in 11 of the documented 32 annual meetings (Table 1). Twenty-seven of these meetings were only missing two sites or less. All 26 sites were present for the 2010 meeting at KBS, which are the most sites ever represented. The summary tables of participation are developed based on meeting notes, participation lists and in some cases group photographs. No participant lists were discovered for the years 1984-1987 and 2002. Remote participation in 2016-2017 has not been included.

Besides the designated representative for each site, additional site personnel including other data managers, site managers or PIs have typically been included at the meeting. Beginning in 1989 LNO representatives started to attend with an average of more than 6 per year from 2000 to 2010. By the mid-1990's growing interest by other organizations to participate led to the invitation of guests. From 1998 to 2009 an average of more than 7 guests per year were present with interest peaking during ASM meetings. While the meeting agenda has generally included a remote teleconference with NSF, 11 individual NSF officers have attended 16 of the meetings in person.

A summary chart of average meeting attendance by year is as follows:

- 1982-1990: 22 participants including 1 LNO representative and 1 guest
- 1991-1995: 27 participants including 2 LNO representatives and 1 guest
- 1996-2000: 31 participants including 3 LNO representatives and 3 guest
- 2001-2005: 42 participants including 7 LNO representatives and 6 guests
- 2006-2010: 49 participants including 5 LNO representative and 8 guests
- 2011-2017: 35 participants including 4 LNO/NCO representatives and 2 guests

The largest meeting was the ASM in 2006 at Estes Park with 59 participants including 23 guests and an NSF officer. The smallest meetings were in 1988 and 1989 with 19 and 20 participants, respectively. The 2017 meeting included 29 participants and was the smallest in 20 years (1997)

TABLE 2. SITES AND YEARS OF PARTICIPATION IN THE IMC ANNUAL MEETING

Site	Site name	IMC years of participation
AND	Andrews Experimental Forest	1982-2017
ARC	Arctic	1988-2017
BES	Baltimore Ecosystem Study	1997-2017



BLE	Beaufort Lagoon Ecosystems	2017-2017
BNZ	Bonanza Creek	1988-2017
CAP	Central Arizona-Phoenix	1997-2017
CCE	California Current Ecosystem	2004-2017
CDR	Cedar Creek	1982-2017
CWT	Coweeta	1982-2017
FCE	Florida Coastal Everglades	2000-2017
GCE	Georgia Coastal Ecosystems	2000-2017
HBR	Hubbard Brook	1988-2017
HFR	Harvard Forest	1989-2017
ILR	Illinois Rivers	1982-1986
JRN	Jornada Basin	1982-2017
KBS	Kellogg Biological Station	1988-2017
KNZ	Konza Prairie	1982-2017
LUQ	Luquillo	1989-2017
MCM	McMurdo Dry Valleys	1993-2017
MCR	Moorea Coral Reef	2007-2017
NES	Northeastern Shelf	2017-2017
NGA	Northern Gulf of Alaska	2017-2017
NIN	North Inlet	1982-1994
NTL	North Temperate Lakes	1982-2017
NWT	Niwot Ridge	1982-2017
OKE	Okefenokee	1982-1986
PAL	Palmer	1991-2017
PIE	Plum Island Ecosystems	1998-2017
SBC	Santa Barbara Coastal	2001-2017
SGS	Shortgrass Steppe	1982-2013
SEV	Sevilleta	1988-2015
VCR	Virginia Coast Reserve	1988-2017
LNO	LTER Network Office	1988-2016
NCO	Network Communications Office	2015-2017

The annual meeting has been attended by 267 distinct attendees including 101 primary site IMs, 18 LNO representatives, 4 NCO representatives, 11 NSF officers, 64 guests and 70 additional site personnel. The average working span of the primary information manager at one site is 7+ years. 23 primary IMs have or had 10+ years of service in LTER history with 12 current IMs with this level of experience.

### Concluding remarks

While this article has only investigated and summarized meeting locations and site participation, IMC meeting notes over the lifetime of LTER provide a rich history of the people, events, themes and products of this committee. Efforts are currently underway to chronicle IMC activities and impact over this historical period for information management, and these meeting notes will be an invaluable help in any portrayal. Currently, most of the IMC meeting notes can be found online (<https://intranet2.lternet.edu/committees/information-management/>). Other web documents that provide insight into IMC leadership are also available:

[http://im.lternet.edu/news/committees/working\\_groups/governance/history](http://im.lternet.edu/news/committees/working_groups/governance/history) and <https://im.lternet.edu/imexec/leadership>. Specific participation tables used to build the summaries in this article will be available as Google docs on request.

## Acknowledgement

The author wishes to acknowledge: Karen Baker for establishing a timeline of IMC history and discovering the inaugural 1982 meeting notes; James Brunt for discovering the 1988 meeting notes and recovering meeting notes for 2001 to 2004; Susan Stafford and Bill Michener for their collective memories in piecing together the missing years of 1984 to 1987; John Porter for his invaluable meeting notes and all of the LTER IMs that have made these meetings so special for the past 36 years.

## Citation

Michener, W. K. (ed.). 1986. Research Data Management in the Ecological Sciences. Belle W. Baruch Library in Marine Science, No. 16. University of South Carolina Press, Columbia, SC. 426 pp.

# Baltimore Ecosystem Study Metadata Management Facility

**Jonathan M. Walsh (BES)**

## Introduction

With the introduction of rich metadata, it became clear early on that the Baltimore Ecosystem Study (BES) needed an automatic metadata generating system. Work was started in 2003.

This article will describe the system starting with the database of metadata that is at the core, a script written in Active Server Pages (ASP) that handles the generation of the metadata files and related files, and an explanation of how they work together.

A little background: The Long Term Ecological Research Network system Ecological Metadata Language (EML) is a customized XML schema that represents a subset of the FGDC Metadata Standard with some additional features not found in the FGDC standard.

The underlying idea behind my system is that since we must create metadata files that follow a certain format (EML), a script can be created to take the records in the metadata database and write out a file that “wraps” each field in a given record in the EML language. In other words, each field value in the BES metadatabase, for example,

Database Field Name	Value
=====	=====
Dataset_id	BES_0543-1
Description	Data table for BES dataset BES_0543-1
Filename	bird-survey-2001-2015-birds.csv
...	...

(e.g. Dataset\_id, Description, Filename are the metadatabase field names)

will be written out with the proper EML tags surrounding them, for example,

<entityName>BES\_0543-1</entityName>

<entityDescription>Data table for BES dataset BES\_0543</entityDescription>

<objectName>bird-survey-2001-2015-birds.csv</objectName>

To do so, I merely included the necessary “wrappers” (e.g. <entityName></entityName>) in the script and the script simply grabs a value from the database and first writes out the line’s “beginning” or opening tag (e.g. <entityName>), followed by the field value, (e.g. BES\_0543-1) and then the line’s “end” or closing tag (e.g. </entityName>).

For example, in EML, the way to represent the organization name for the Cary Institute in an EML document is as follows:

<organizationName>Cary Institute of Ecosystem Studies</organizationName>

So, the script, as it’s writing out the file, first writes <organizationName>, which is hard coded into the script. Then it gets the organization name value from the current record in the metadata database (Cary Institute of Ecosystem Studies) and writes that, and then it writes the closing tag - </organizationName>.

It steps through all the fields in the database like this, and writes out the values to the metadata file, “wrapping” them all in their EML tags. The metadata file is also given an EML “header”, i.e. the “top” of the file. The header contains the XML schema information and other standard XML top-of-file information.

A marker field in the database tells the script whether or not to write a record. This way, we can recreate a single metadata record as the dataset is updated, or even recreate them all to handle a global change, such as a change in the EML schema declaration.

Some lines in an EML document contain information that is EML-specific only, such as the schema declaration. These lines are easily generated by the script. All the script has to do is send the lines to the file in order, including elements from the metadata database when appropriate.

The script also handles the file naming convention. Each file is named

knb-lter-bes-xxxx.xml where xxx is the BES dataset ID number associated with that dataset.

Note, the EML “wrappers”, or tags, could be stored in another database instead of being hard coded into the script. This is a little neater and I’ve done it in the past on similar projects, and might do it here someday. It’s easier to maintain but harder to code. I made a choice to do it this way because I don’t expect the form of EML to change drastically. So far that’s been true!

### **The Database of metadata (Metadatabase)**

Here is the structure of the database. There are many more fields than I list here but for the sake of brevity I’ve left them out.

#### **Main Table**

1. BES ID
2. Revision - This is what we increment when we re-submit to PASTA
3. Title
4. Filename
5. Path - URL of dataset on beslter domain

6. Attributes linked table
7. Author
8. Data "Category" - which core group is it?

#### Attributes Linked Table

1. Attribute ID
2. Attribute Name
3. Attribute Definition
4. Storage Type
5. Measurement Scale (Standard units when possible, custom units work too.)
6. Missing Value Code
7. Missing Value Code Explanation

#### Attributes Linker

1. Attribute ID
2. Dataset ID

The “linker” table is an easy way to reuse attributes as well as to assign many attributes to a single dataset.

#### Methods

1. Method ID
2. Method Name
3. Method Description
4. Instrumentation

#### Access Permissions

1. Dataset ID
2. Allow/deny
3. Principal
4. Permission

### The Script

The script generates XML ready for upload to PASTA - As described earlier, the script writes EML compliant files and makes them available to upload to the LTER system. It also performs other interesting functions.

1. It generates the data search page on the study's website (beslter.org). For the time being, BES also lists its data collection on the BES website. It is a good way of searching for BES data because you can search the collection with some useful filters and you can sort the results.
2. It generates a “harvest list” – This was for the old Metacat system, where the LTER system would “pull” the files from your server, rather than the current model of “pushing” files to the LTER system. The harvest list was a list whose URL was known to the LTER “harvester”. In it

were the URLs of the metadata files to be harvested that day. This capability is being retained for possible future use.

3. It's designed to connect to any database engine – The database connection is a standalone module in the script. Thus it is very easy to connect to different database platforms, depending on which database engine you have.
4. It's modular so each task is in a discrete module, especially the "file opener" module so the script can open different databases - SQL, MySQL, Oracle, etc. Basically grouped into major EML nodes:
  - a. Dataset
  - b. Methods
  - c. Attributes
  - d. Access
  - e. Authors
  - f. Additional Metadata
  - g. Etc.
5. It can be ported to other scripting languages. The logical flow, if I've planned it correctly, should fairly easily convert to any other language.

## Execution

Run the script and it writes the EML files for the selected records in the database. Records are selected by setting a binary field called Armed\_for\_PASTA to [True]. It also writes a file of the URLs of the EML files it creates so you can paste them into the PASTA uploader interface. It also writes the BES public data page ([https://beslter.org/data\\_browser.asp](https://beslter.org/data_browser.asp)).

FIGURE 1. EXAMPLE ASP CODE TO WRITE BEGINNING OF THE <METHODS> NODE.

```

<?
>> If rs("armed_for_PASTA")=1 Then
>>   objTextFile.WriteLine indent4 & "<methods>" & ""
>>   do while not rsmethodlink.EOF
>>     objTextFile.WriteLine indent5 & "<methodStep>" & ""
>>     objTextFile.WriteLine indent6 & "<description>" & ""
>>     objTextFile.WriteLine indent7 & "<section>" & ""
>>     objTextFile.WriteLine indent8 & "<title>" & rsmethodlink("methodname") & "</title>" & ""
>>     m_methoddescription=Replace(rsmethodlink("methoddescription"), vbcrLf, "</para> <para>" & vbcrLf & indent8)
>>     objTextFile.WriteLine indent8 & "<para>" & m_methoddescription & "</para>" & ""
>>     objTextFile.WriteLine indent7 & "</section>" & ""
>>     objTextFile.WriteLine indent6 & "</description>" & ""
>>     objTextFile.WriteLine indent5 & "</methodStep>" & ""
>>     rsmethodlink.movenext
>>   loop
>>   objTextFile.WriteLine indent4 & "</methods>" & ""
>> End if

```

Here's basically what happens and an idea of the different modules:

1. File Opener
  - a. Connect to Metadatabase
  - b. Filter selected record(s)
  - c. Create and open text files
    - i. Metadata file knb-lter-bes.xxx.xml (This file will go to the LTER system)
    - ii. Harvest List
    - iii. List of URLs for metadata files being created during this run (paste into PASTA file uploader)



- iv. BES website searchable data page ([https://beslter.org/data\\_browser.asp](https://beslter.org/data_browser.asp)) and first of the linked sub-pages for each dataset – (full metadata page with text description, etc. and links to EML file and dataset.)
2. Construct EML file name (“knblter-bes-“ + BES Dataset ID number + “.xml”)
3. Write EML header
4. Get access permissions from the table and write the <access> node in the metadata file
5. Begin writing <dataset> node to file
6. Insert <title> node
7. Get persons/organizations info from the corresponding table in the database and write <creator> node(s)
8. Get publications date from table and write the <pubDate> node
9. Get abstract from table and write <abstract> node
10. Get keywords from table and write <keywordSet> node
11. Repeat as above for the <intellectualRights>, <distribution>, <coverage>, and <contact> nodes.
12. Get methods from the methods table for that dataset and write the <methods> node
13. From main table get the dataset information and write the <dataTable> node, including the attribute information for each data table in the collection. Note that there can be many tables in a single dataset.
14. Close the <dataset> node
15. Write additional metadata into the <additionalMetadata> node (if there is any for this dataset).

Figure 2. SOME EML OUTPUT FOR THE CODE IN FIGURE 1.

```

<dataset>
  <methods>
    <methodStep>
      <description>
        <section>
          <title>Bird Observation</title>
          <para>Point Count Survey Protocol</para>
          <para>(April 2013)</para>
          <para></para>
          <para>
            Adapted from the BES LTER Point Count Protocol (Katti, Shochat, and Warren 2002)
          </para>
          <para></para>
          <para>Objectives:</para>
          <para>
            1. Determine the occurrence and abundance of all bird species observed.
          </para>
          <para>
            2. Investigate the species composition of a community and the relative abundance of different species
          </para>
          <para></para>
          <para>Survey Overview:</para>
          <para></para>
          <para>...</para>
          <para>...</para>
          <para></para>
          <para>Required Equipment:</para>
          <para></para>
          <para>Binoculars, data sheet, clip board, and pencil</para>
          <para></para>
          <para>General Survey Guidelines:</para>
          <para></para>
          <para>...</para>
          <para>...</para>
          <para>
            * Do not conduct point counts in extreme noisy conditions (i.e. heavy construction, noise exceeding the t
          </para>

```

Now just go find the file “pastaurls.txt” and paste those URLs into the LTER PASTA uploader for evaluation and upload. That’s it.

Here’s a link to the script: <https://github.com/jonathanmwash/metadatabase> and here’s a link to the EML file I used for my example:

<http://beslter.org/docdrop/sample-metadata/knb-lter-bes-543.xml>

If you like the idea of this system and would like to try this script, please do feel free and don’t hesitate to get in touch with me about it. Thanks!

## Hello from the BLE LTER information manager

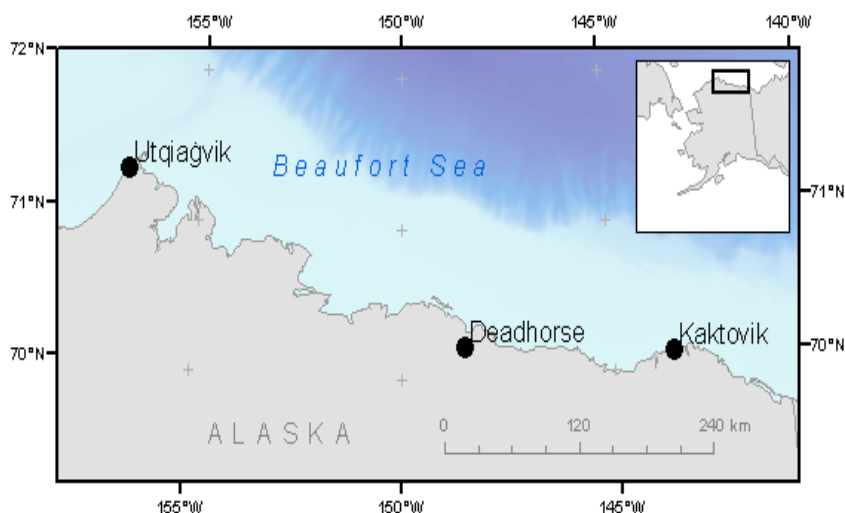
**Tim Whiteaker**

**BLE Information Manager**

Greetings Databits readers! I’m the information manager for the new Beaufort Sea Lagoon Ecosystems (BLE) LTER site, and I wanted to introduce the site and myself to you. We’re still spinning up operations at our site, but with the new Environmental Data Initiative’s Data Center and the Arctic Data Center coming online within the past couple of years and this great community of LTER information managers, I can already tell that this is a great time to be joining the network.

In this brief article, I’ll include some initial thoughts about building an information management system from scratch. At best, it may illuminate possible paths forward or areas where we need more help for new information managers, and at worst, I can come back next year with a more hardened perspective and provide lessons learned.

**FIGURE 1. THREE NODES FOR BLE STUDY ALONG THE ALASKAN COASTLINE. BATHYMETRY DERIVED FROM IBCAO V3 (DOI: 10.1029/2012GL052219), FROM NATUREARTH.**



### About the BLE Site

We’ll be studying six lagoon ecosystems on the Alaskan Arctic coastline along the Beaufort Sea. These lagoons are located near Utqiagvik (formerly Barrow), Deadhorse, and Kaktovik.

Tremendous seasonal variations make them fascinating places to study. In the winter, the lagoon surface is completely covered with ice, which can lead to significantly decreased dissolved oxygen

within lagoon waters. A great influx of freshwater from the spring freshet can drastically drive down salinity and the breakup of sea ice in summer months alters the circulation dynamics between lagoon and sea. High rates of coastal erosion and a pronounced change in climate over the past few decades also contribute to a challenging environment for wildlife.

We'll be collecting data from a wide variety of instruments and seasonal sampling campaigns to help us understand community structure and the resilience of food webs as conditions change across seasons and decades. Sampling at this remote location, especially in winter months, is fraught with difficulty, and so we'll also use hydrology, biogeochemical, and ocean circulation models to supplement sample and sensor results. We'll use these datasets to advance our understanding of how input of materials from land and oceanographic conditions influence coastal food webs. For more about the project, check out our website [1].

### About the Information Manager

FIGURE 2. RESEARCHER JAMES MCCLELLAND SAMPLING WATER DURING ICE BREAK-UP IN KAKTOVIK LAGOON, ALASKA. CREDIT: KENNETH DUNTON, MARINE SCIENCE INSTITUTE, THE UNIVERSITY OF TEXAS AT AUSTIN



I'm a research scientist at the Center for Water and the Environment at The University of Texas at Austin, working on projects ranging from water rights and modeling to paleogeographies during ankylosaurus times. While my degree is in Civil Engineering, I feel like I am as much a developer and GIS user as anything else these days.

I've served as information manager for a handful of other Arctic projects in the Chukchi and Beaufort Seas. My toolchain for those projects included Excel metadata templates, VBA macros for data cleanup and formatting, and tools/controlled vocabularies from the CUAHSI software stack [2] (which is now considered legacy as they have moved to cloud solutions), along with static HTML for each project's website [3]. I work in Python, ArcGIS (including developing add-ins with C#), and Excel, and am willing to discuss any of this work.

### Building a New Information Management System

Datasets from BLE will far exceed anything I've worked with in the past in terms of variety and sheer volume. It's been very interesting learning how other LTER sites handle information management, and to see common practices evolve in a sort of grass roots way with DEIMS and from the top down

with the Environmental Data Initiative (EDI) and the R EML package. Here I share my outsider perspective on possible ways forward for a new site.

As Co-PIs at the various institutions involved with BLE will be handling local data QA/QC and management, I see my responsibilities as follows:

- Getting Metadata from PIs
- Managing Datasets and Metadata Submitted to the information manager
- Generating EML
- Maintaining Our Website

While I've seen some sites that gather metadata via Web forms, I'll likely adopt some form of Excel template like what is used at FCE or PIE, since that seems like a fairly simple solution. However, I hear great things are coming from the Arctic Data Center and EDI in terms of new Web forms for entering metadata and automatically generating EML. I intend to explore the possibility to somehow leverage those forms to gather metadata from my PIs. So, when clicking Submit, instead of the data being submitted to the data center, I would get a notification for reviewing the metadata and data before finally submitting it. I am considering using a spreadsheet or lightweight database for maintaining a local data catalog, but plan to use an API like PASTA or Solr to retrieve records from EDI or the Arctic Data Center's repository to support an online catalog. Then the local (offline) catalog would be maintained for local use only and could be kept simple and tailored to local needs. For EML generation, I see a lot of viable options out there. I favor the R EML package since R is free and EDI seems to be actively supporting the package. However, it may be another two years before my site has data to submit, so I can wait to see what kind of traction the package gets.

### Fun with Our Website

In considering website design, the "Guidelines for LTER Web Site Design and Content" [4] devised by members of our community is very useful. A website template similar to DEIMS to go along with that would be a great addition. The drawbacks of DEIMS are its learning curve and the ordeal of upgrading Drupal versions. Typically each site has its own unique website requirements, institutional hosting support, and local skillsets, but an official cloud platform (hosted at EDI?) for LTER websites where content and styling could be added to an existing template would be a great option. In the absence of such an option, I've created the BLE website using static HTML.

I like **static HTML** for the simplicity of setting up a site without much content, as is the case for our new BLE site. I also appreciate the quick page loads and the fact that I don't have to worry about updating my site to address security vulnerabilities as Drupal and WordPress admins do. There's also the low cost. My institution would charge me thousands of dollars per year to host a custom WordPress site, whereas the BLE static HTML site is currently hosted for free on Netlify. Among its many features, Netlify provides a distributed content delivery network so that if one server hosting my site goes down, another one will pick up the slack. Another benefit of static HTML is that it's easy to share. You can literally download and unzip our source code from GitHub [5], double-click index.html, and a working site should open in your browser. I envision this as a way to help other new LTER sites create their websites in the future. Simply fork my repo (that's Git speak for make your own version controlled copy of my code) and customize as you please. We could even collaborate to fix bugs, improve the styling, or add new functionality.

Speaking of functionality, I've created additional repositories in GitHub to handle some dynamic elements of an LTER website, namely, site search [6], data search [7], and bibliography search [8]. **Site search** uses Lunr for a static search index and is currently active on our website. **Data search**

uses the Arctic Data Center's Solr API so that no local online data catalog is necessary. Similar code could be created for EDI's PASTA interface. **Bibliography search** uses the Zotero API, which is like EDI's PASTA interface except for bibliographical entries. The bibliography itself is hosted in Zotero and can be associated with each LTER site.

Since BLE is just underway, we don't have a bibliography or dataset archive yet, so I haven't had a chance to test the related functionality in the wild. I would love it if some of you wanted to take the code for a spin. If you're interested, let me know and I'd be happy to walk you through any of the GitHub repositories I've mentioned here. By collaborating, we improve these resources for all interested LTER sites to use.

I'm very grateful for the support from other LTER IMs, EDI, and the Arctic Data Center thus far. Once my information management system has fully taken shape, I'll be sure to report back on Databits about the experience.

## References

- [1] <http://ble.lternet.edu/>
- [2] <https://www.cuahsi.org/data-models/legacy-tools>
- [3] <http://arcticstudies.org/>
- [4] [https://im.lternet.edu/im\\_requirements/webdesign\\_guidelines](https://im.lternet.edu/im_requirements/webdesign_guidelines)
- [5] <https://github.com/twhiteaker/LTER-website>
- [6] <https://github.com/twhiteaker/Lunr-Index-and-Search-for-Static-Sites>
- [7] <https://github.com/twhiteaker/Solr-JavaScript-Search-Client>
- [8] <https://github.com/twhiteaker/Zotero-JavaScript-Search-Client>

## Postgres, EML and R in a data management workflow

**Li Kui and Margaret O'Brien**

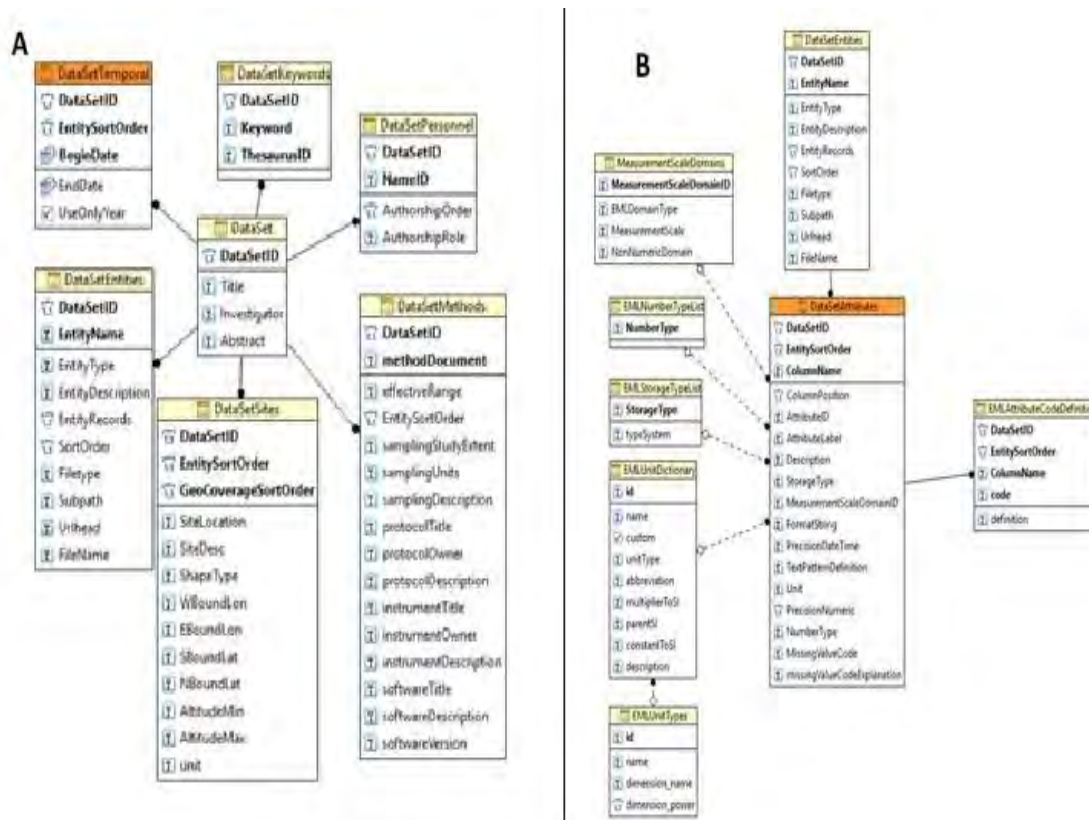
**Marine Science Institute, University of California, Santa Barbara, CA, USA**

The workflow for managing ecological data presented here was designed for the MBON (Marine Biodiversity Observation Network) data management plan but can be adopted to other environmental projects. The workflow consists of three components: 1. Metadata storage in a relational database; 2. EML (Ecological Metadata Language) file generation using R (R programming language); 3. Archiving data packages in EDI (Environmental Data Initiative), an online data repository. Our workflow combines efficient EML record generation using the package developed by the R community with the advantages of centrally-controlled metadata in a relational database.

### Metadata storage and management



**FIGURE 1** TABLES IN TWO OF THE SIX MODULES IN THE **BON** MINI-METABASE: (A) THE DATASET MODULE WITH SEVEN TABLES HOLDS HIGH-LEVEL MATERIAL USED BY EVERY DATASET; (B) THE ATTRIBUTE MODULE, USED FOR CSV OR TXT FORMATTED DATASETS.



The Metabase data model was originally designed by Wade Sheldon for storing metadata from the Georgia Coastal Ecosystem LTER with 86 tables managed in SQL-Server (Sheldon et al, 2012). Two LTER projects (SBC LTER and MCR LTER) ported Metabase to Postgres with minimal changes to the table structure (O'Brien and Gastil, 2013). For use by the MBON, we reduced the SBC/MCR Metabase to 23 tables. In this “mini-metabase”, we kept only the essential tables, that is, those sufficient for dataset metadata storage and routine management.

The mini-metabase is organized in six modules, including one essential dataset module (Fig 1A), four supplemental modules, and one package management module. All seven tables in the dataset module receive new rows when inserting a new dataset. The four supplemental modules consist of an attribute module (Fig 1B), file type module, keyword module, and people module. These modules are updated at a lower frequency as they accommodate broader needs for additional data types or vocabulary controls.

The attribute module (Fig 1B) is used for documenting the detailed attributes for column formatted data (e.g. csv and txt). Most tables in this module are for vocabulary controls. The file type module specifies details for several commonly used file types (e.g. csv, txt, tiff, mat, and R files) in data management. The people module stores all personnel information for various projects. The keyword module includes a list of theme keywords and their associated thesaurus, which functions as a keyword library as well as a controlled vocabulary. Using the metadata stored in the mini-metabase, we can generate EML files using R for future dataset archiving.

Five modules (dataset, attribute, file type, keyword, and people modules) store sufficient metadata to generate EML. The sixth module, package management module, stores material for managing and inventorying datasets, based on work by SBC and MCR LTER (O'Brien and Gastil, 2013). The package management module contains information such as data receipt and archive dates,

management categories (e.g., temporal types such as ongoing time-series data or short-term study data), and data storage locations. This information is not required by EML but used by the project for long-term data management. Typical information we can extract from this module include the latest data update date and version, anticipated datasets and their contact person, or a list of time-series datasets that require frequent updates.

## Generating EML in R

In 2016 R users developed an “EML” package in R for creating EML files. A brief introduction of the structure for EML and a simple application of the “EML” package in R is provided by the [R Project for Statistical Computing](#). However, as the number of datasets increased, it is time-consuming to type in a large amount of metadata. Keeping metadata in a relational database has many benefits for a project the size of the MBON (or LTER), e.g., control of content in parent tables, and because a relational database is often used for other tasks, such as website content or bibliography. It was to our advantage to extract the meta-info directly from Postgres, and this task is relatively simple in R. Four R functions are required to automate the EML generation processes: 1. access the mini-metabase and read information; 2. construct attribute table(s) and create “dataTable” child node in EML; 3. assemble high-level EML components; 4. wrap all of the functions and specify the target dataset ID. All of the R functions and examples of dataset codes can be found on GitHub: [https://github.com/lkuiucsb/EML\\_R/](https://github.com/lkuiucsb/EML_R/).

For security within the Postgres server and to provide an abstraction layer, the R code reads “views” which gather information from all modules rather than querying table directly. The views are tailored to EML elements. For the MBON, there are nine views generated from the mini-metabase: dataset, keyword, method, personnel, attribute, entity, unit, temporal range, and geolocation. Other views could be constructed for other types of exports, e.g., metadata adhering to a different schema, such as ISO-19115-2.

Other than extracting the metadata from mini-metabase, three other pieces are required to run the R code and complete the EML generation (first two are in MS-Word documents): 1. an abstract for the dataset; 2. a method description; 3. XML files with static information such as access, intellectual rights, contact, publisher, and project nodes. Keeping the abstract and method in MS-Word is an excellent way for project scientists to directly contribute the textual components of a dataset.

## Data archive into EDI

After generating the EML file via R, we use the EDI data portal to archive a dataset. Before uploading the data, there is an evaluation process done on both the EML file and the data. The validation system makes sure that the EML meets the EML 2.1.1 standard and that the data structure matches the information in the EML. After successfully uploading the data, a DOI is generated as the permanent identifier.

Datasets created and archived with this method can be found on the Santa Barbara Marine Biodiversity Network website: <http://sbc.marinebon.org/data>, and in datasets recently added to the SBC LTER collection. This Postgres-EML-R workflow is flexible enough to accommodate multiple design patterns, since Postgres views and R modules can be added for additional styles or EML trees.

## References

Sheldon, W.M. Jr., Chamblee, J.F. and Cary, R. 2012. Poster: GCE Data Toolbox and Metabase: A sensor-to-synthesis software pipeline for LTER data management. 2012 LTER All Scientists Meeting, 11-Sep-2012, Estes Park, Colorado.

## The EML Assembly Line: A Metadata Generation Tool for Data Providers in the Ecological Sciences

**Colin A Smith**

**Environmental Data Initiative, University of Wisconsin-Madison, 680 N Park St, Madison, WI 53706**

The Ecological Metadata Language (EML) is a metadata standard developed and maintained by the ecological community for persistence, discoverability, and reuse of ecological data. While it is an effective and widely adopted metadata standard, the effort required to make quality EML is not a trivial task. For instance, it requires extensive involvement from the data provider to accurately and effectively communicate what the data are, how they were created, and where and when they were collected. Additionally, it requires a thorough understanding of the EML schema, EML best practices, and familiarity with software tools for translating information into EML. Traditionally these technical aspects of EML generation have been the responsibilities of professional data managers, but this may be changing.

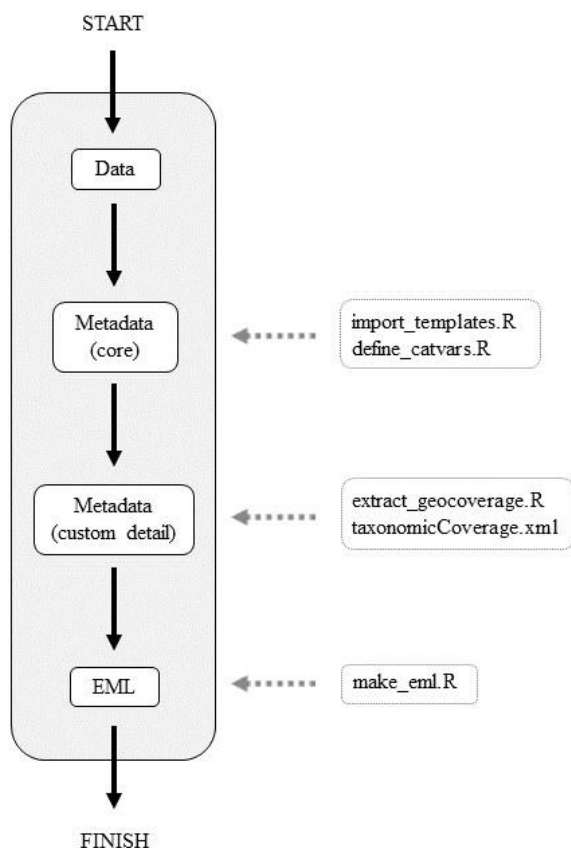
Data publication is increasingly expected by scientific journals, funding sources, and from the cultures of open and reproducible science. This expectation brings to light an issue of scalability, with some potential solutions being: make more data managers, improve efficiencies in data management workflows, or build tools to help data providers take over more of the metadata generation role. This latter option is attractive since data providers have the most comprehensive knowledge of the metadata for their dataset; however, their technical understanding of EML and how to build it may be lacking. Fortunately, there are extant and developing methods to address this problem.

To help data providers generate high-quality EML themselves, the Environmental Data Initiative (EDI) has created the EML assembly line R code library. This metadata generation tool is composed of high-level functions and workflows that are intuitive and easy for the data provider to follow. The user begins the EML assembly line process by consulting a decision tree outlining a set of fixed steps for building metadata given the characteristics of their data. Next, the user executes each of these steps by calling functions and occasionally completing an associated template file. Once all the templates are finished, the user executes a final 'make\_eml' function, which checks user-supplied template files for errors, transcribes the template content into EML, validates the EML against the schema, and finally writes the EML file.

Some noteworthy features of the assembly line include: (1) automated direct extraction of much of the metadata from the data entities, thus requiring less work from the data provider; (2) application of a thick layer of checks throughout the user's workflow to provide meaningful error messages and suggestions on how to fix them; (3) incorporation of EML best practices in the assembly line functions, thus bypassing the need to be fluent in these; (4) user-supplied inputs to the assembly line in the form of data and metadata templates that can be easily edited and rerun as the dataset is revised; (5) the ability for a data provider to select a data repository most fit for publication of their data, because the output EML is not tied to a repository; (6) construction of the assembly line from the R programming language, which is in common usage by data providers in the ecological community; (7) minimal level of R proficiency required by the assembly line, since the user's interaction is restricted to a fixed set of instructions and user inputs are supplied in easy-to-edit template files; (8) development and maintenance of the assembly line as a distributed effort since it is written in the common R language; (9) modularity of all of the supporting lower-level functions of the assembly line, making them available for data managers to use in constructing their own custom EML workflows. In total, these features make the EML assembly line a good metadata generation tool for data providers and one that can be developed and maintained by the ecological community.

The EML assembly line is an open source project hosted on EDI's GitHub (<https://github.com/EDIdorg/EMLassemblyline>). It utilizes the low-level functionality of the EML R library (Boettiger et al. 2017; <https://github.com/ropensci/EML>) and wraps it with functions to abstract required knowledge of EML and the R language generally. We welcome contributions of all forms including new functions, patches, bug reports, and feature requests. On the project's GitHub, you will find the project road map and guidelines for contributions. The EML assembly line is maintained by the EDI core team and has benefited from contributions by members of the LTER, LTREB, and OBFS communities.

Figure 1. An overview of the EML assembly line. Users begin EML assembly by supplying their data.



Next the user calls the `import_templates` function, which looks at the data and autodetects metadata content and writes it to template files for the user to verify and provide additional metadata content. If categorical variables are present in the data, then the user calls the `define_catvars` function which autodetects unique variables and writes them to a template file for the user to define. These completed template files contain the core metadata that is expected of a high-quality EML document. Additional custom metadata may be supplied by running the `extract_geocoverage` function (if the data contain several geographic locations), and by using the `taxonomyCleanr` R package (<https://github.com/EDIdorg/taxonomyCleanr>) to resolve taxonomic data to an authority system and to create the `taxonomicCoverage` node. Once all metadata templates/files have been created, the user runs the `make_eml` function to translate all this metadata into structured EML metadata.

# The Environmental Data Initiative - the first 1.5 Years

## Supporting LTER Information Managers

Kristin Vanderbilt<sup>(1)</sup> and Corinna Gries<sup>(2)</sup>

(1) Florida Coastal Everglades LTER (FCE), (2) North Temperate Lakes LTER (NTL)

Since the Environmental Data Initiative (EDI) project began in August 2016, EDI has engaged in several productive ways with the LTER Information Management Committee (IMC). Major goals of EDI are to maintain the Data Repository holding all LTER datasets, enable IMC members and others to easily publish their high-quality data and metadata to the EDI Data Repository, and to accelerate data synthesis by harmonizing datasets having a common theme. In support of these goals, EDI has upgraded the PASTA software to PASTA+, created new data management software in R, developed a data model for community survey datasets, and provided many training opportunities for LTER IMC members and others.

The EDI Data Repository is built using the PASTA+ infrastructure and can currently be accessed through both the LTER Data Portal and the EDI Data Portal. LTER IMs can submit data packages via either portal. Based on requests from the IMC, new congruence checks for LTER data have been added to the PASTA+ EML Congruence Checker (ECC) (O'Brien et al. 2016) by EDI software engineers Mark Servilla and Duane Costa (Table 1). With feedback from the IMC, a dataset citation format was selected for display on data package summary pages. Another addition to PASTA+ that will be useful to IMs is that provenance relationships between data packages can be specified and publications may be linked to datasets.

**Table 1. The quality checks that EDI has added to PASTA+ Quality Engine after consultation with the IMC.**

Name	Description	Response
<b>PastaDoiAbsent</b>	PASTA will add DOIs to L1 EML. No incoming L0 EML should contain a DOI that resembles a PASTA DOI.	error
<b>ChecksumPresent</b>	Checksums can be used to validate entities. inclusion is good practice.	warn
<b>ChecksumMatch</b>	If a checksum is found and does not match the checksum computed by PASTA, package must be rejected.	error
<b>NumRecordsPresent</b>	Check accompanies the current check to compare the number of records in metadata to number found.	warn

The rOpenSci R package for generating EML (<https://github.com/ropensci/EML>) is used by EDI information manager Colin Smith to develop a workflow for creating EML, the R EML Assembly Line, which makes using R to create EML a more user-friendly process. See an article by Colin in this issue. A training workshop was held in November 2017 at UNM to introduce the software to 17 information managers, including Yang Xia (KNZ), Kris Hall (SEV), and Renee Brown (MCM) from the LTER IMC. LTER Information Management experience was brought to the workshop by instructors Kristin Vanderbilt, Margaret O'Brien and John Porter, plus EDI staff Colin Smith and Susanne



Grossman-Clarke. Other generally useful workflows in R Colin is developing include cleaning taxonomic information and preparing taxonomic coverage data for inclusion in EML.

A secondary, but no less important goal of EDI is to accelerate scientific inquiry. After working with several syntheses working groups at the Network Communications Office (NCO) a process for pre-harmonizing certain LTER long-term datasets was envisioned. Again based on the prevalent interest of these working groups in LTER's many long-term population and community survey datasets members of EDI, LTER Information Managers and synthesis scientists met in 2017 for the "Dataset Design for Community Survey Data" workshop to define how the data could be formatted to make them easier to integrate. This workshop involved LTER IMs Hope Humphries (NWT) and Suzanne Remillard (AND) in the process of mapping data in site specific format to what is now called the EcoComDP data model. More about this project can be found here:

<https://environmentaldatainitiative.org/category/data-package-design/>. This is still in the experimental stage integrating the needs of LTER researchers to maintain datasets in a site-specific format, LTER IM experiences in developing and maintaining cross-site data products, and specific scientific research needs by the synthesis working groups. Furthermore, this approach implements the vision for PASTA+ to support specific data manipulation workflows. In the next phase we will develop a specific data search and access mechanism.

EDI has offered many training opportunities for the IM community for both new and continuing LTER IMs alike. For new IMs, VTCs showcasing three of the existing LTER information management systems were offered by James Connors (PAL, CCE), John Porter (VCR), and Hap Garritt and Jim Laundre (PIE and ARC, respectively). For everyone, there have been VTCs covering a range of topics including "The PASTA + REST API", "Transform and Visualize data in R using the packages tidyr, dplyr, and ggplot2", and "Git and Github". During the ESIP Summer 2017 meeting, Duane Costa partnered with LTER IMs John Porter, Stevan Earl (CAP) and Gastil Gastil-Buhl (MCR) to present how to use PASTA web services to create a local data catalog. Links to all of EDI's training materials can be found on the EDI website (<https://environmentaldatainitiative.org/resources/training-resources/>). Another EDI-organized session at ESIP, attended by several LTER IMs, sought input from the IM community on the development of an Information Management Code Repository. This repository will be launched in 2018 by EDI. A summary of EDI's activities relevant to LTER information managers can be found in Table 2.

At the halfway point of the grant, in January 2018, EDI conducted its in-person Advisory Board meeting, which was led by Aaron Ellison (HFR), chairing the board. Mary Martin (HBR) participated representing the LTER IMs. Other members of the AB are Nathan Booth (USGS), Rebecca Koskela (DataONE), Ian Foster (U of Chicago), and Peter Arzberger (SDSC). The meeting was energizing and very productive with the AB bringing different and important perspectives to the discussion of EDI's various goals and approaches.

**Table 2. Services that EDI provides that support LTER information managers and researchers.**

<b>EDI Services</b>	<b>Support for LTER IMs and Sites</b>	<b>Support for Researchers</b>
Maintains EDI Data Repository	Provides secure location to publish LTER datasets	Makes data open and accessible; meets journal requirements for data publishing.

Provides automated data and metadata quality checking	Helps IMs improve data and metadata congruence and completeness	High quality metadata allow for more confidence in re-using data
Maintains user friendly interface as well as API for upload, search, and download of data	May use API to automate many IM interactions with the portal	Provides valuable information about data in easily accessible format
Provides data security through back up, access control and professional server maintenance	All data are secure	All data are secure
Creates new data management software in R	Offers an open source, user-friendly process for generating EML	Structured metadata supports data extraction directly from the EDI Data Repository
Offers in-person and VTC information management training	Educates new and continuing IMs on a range of topics from git to using PASTA+ web services	Offers researchers opportunity to stay abreast of IM developments
Accelerates Scientific Inquiry through Data Harmonization	Defines a data model to serve as the common structure for a particular research theme.	Makes data more readily integrated for meta-analysis, greatly reducing the time needed to integrate data
Initiates IM Code Repository	Offers a location where IMs and scientists can find and share code for performing IM tasks.	Provides code that may be useful to anyone cleaning and transforming data.

#### References cited:

O'Brien, M. D. Costa, M. Servilla. 2016. Ensuring the quality of data packages in the LTER network data management system. Ecological Informatics 36:237-246. DOI: 10.1016/j.ecoinf.2016.08.001.

#### Resources:

EDI Website: <https://environmentaldatainitiative.org/>

The EDI data portal: <https://portal.edirepository.org>

EDI workflows are developed in GitHub and anyone is invited to contribute: <https://github.com/EDlorg>

Subscribe to the EDI newsletter: <http://environmentaldatainitiative.us14.list-manage.com/subscribe?u=c258a774cbd4d34290410d1ea&id=da346b264c>

The EDI YouTube channel for recorded online trainings: <https://www.youtube.com/channel/UCNZoWPaMG6IkEiH8xRNnrrA>

## Commentaries

# A Message from IMEXEC: where we have been and where we are going

**Stevan Earl<sup>(1)</sup>, Suzanne Remillard<sup>(2)</sup>, Gastil Buhl<sup>(3)</sup>, Wade Sheldon<sup>(4)</sup>, Jason Downing<sup>(5)</sup>**

**(1) Information Manager, Central Arizona–Phoenix LTER, (2) Information Manager, Andrews Forest LTER (AND), (3) Information Manager, Moorea Coral Reef LTER (MCR) , (4) Information Manager, Georgia Coastal Ecosystems LTER (GCE), (5) Information Manager, Bonanza Creek LTER (BNZ)**

We welcome this opportunity to resume publication of DataBits. Our thanks to Eda, the volunteer editors, the LTER Network Communications Office (NCO), and to all the contributors for reinvigorating this venue for sharing ideas, information, news, and other thoughtful contributions. The LTER Information Management Committee (IMC) has published DataBits since 1990, which includes over 40 issues of featured articles, commentaries, news bits, and overall good reads from contributors within and outside of the LTER Network.

The past few years have been a period of change with the IMC and the LTER Network. We have seen the end of the LNO grant and are navigating a shifted landscape with our new partners, the NCO and the Environmental Data Initiative (EDI). These seismic shifts in the LTER landscape are redefining how we operate, but we will continue to foster relationships with our new partners, and advance ideas that have developed from within the IMC, including how we manage the DataBits publication.

Since the last issue of DataBits was published in spring 2014, we have also witnessed some big changes to *sites* within the Network. Three new LTER sites have been added. We welcome these new sites and enthusiastic data managers with open arms. We look forward to the Sevilleta LTER coming back on-line in a revised state, and hope that its reintroduction to the Network will be completed soon. With great sadness, we lament the pending close of the long-running Coweeta LTER.

IMC working groups have been busy creating documents and guidelines designed to aid both new and established sites. Some of the IMC's more recent efforts include: development of IM System Guidelines (1), LTER website recommendations [draft], revised Bylaws (still in draft stage), and contributions to a revised LTER Data Access Policy (2). An IMC working group also produced a training-recommendations document (3) that, in part, cataloged what skills are important for the IM role, and where IMs would most like to receive additional training. This document has helped shape the data management training efforts being developed by the EDI, which feature a series of superb workshops, tutorials, and webinars aimed at supporting scientists with their data needs, and training information managers.

The NCO is also making significant contributions to advance the Network. They have hosted several syntheses working groups, many of which have included some of our LTER IMs. This pairing of scientists and information managers has the potential to advance not only the science that the working groups are considering, but ecoinformatics best practices surrounding those topical areas as well. The NCO has also launched a new website for the public-facing side of the LTER. The new website features a fresh Network logo, and well-designed user interface. The much referenced (by site IMs, anyway) Intranet and document archive remains unchanged for the moment. As a separate effort, an IMC working group is working with the NCO to move and/or recreate the IMC website hosted from the previous LNO (Drupal) framework to the new NCO Wordpress framework.

This is an important time for the IMC: we are navigating profound changes to and within the LTER Network and witnessing a bewildering evolution of the information sciences. As we begin to plan for the 2018 All Scientists Meeting (ASM), we would encourage you to please give some thought to the future of the IMC. Please consider how we should govern ourselves in the new LTER landscape, how we can collaborate most effectively with our new partners, and, critically, how this group can continue to be a leader in the field of ecoinformatics.

We will discuss these and other important issues during the ASM, and we look forward to hearing your ideas. We extend our thanks to those who have contributed to the IMC over its many years and extend our welcome to its new members.

Sincerely,

LTER IMC Executive Committee

### References:

[1] Guidelines for LTER Information Management Systems, Version 2.1, *03 January, 2018*; [http://im.lternet.edu/im\\_requirements/im\\_review\\_criteria](http://im.lternet.edu/im_requirements/im_review_criteria)

[2] Long Term Ecological Research Network Data Access Policy; *Revised by the LTER Science Council, May 19, 2017*; <https://lternet.edu/data-access-policy>

[3] LTER Information Managers Skillset and Training Resources Working Group Recommendations, 5 May 2017; <http://im.lternet.edu/projects/training>

## A survival kit for the shadows from a natural disaster

**Eda C. Meléndez-Colom (LUQ)**

To me, hurricanes are normal events which I have experienced across my life. I was baby during my first hurricane, technically a tropical cyclone, Santa Clara (1956), and thus I remember very little. I will never forget Hugo (1989), a category 3, due to its high impact the island. Other hurricanes have come such as, Georges, (1997, Category 4) and more recently Irma and Maria, a category 3 and 4-5 respectively.

Hugo passed over the Norther East (NE) part of Puerto Rico the same year I started working with the LTER (1989). Hugo had a great impact on “El Yunque”, the Luquillo Mountains, and offered an opportunity for LUQ investigators to collect the first set of post-hurricane data. An entire journal, *Biotropica* (1990), was produced as early as one year after the beginning of the LUQ site.

My memories of Georges are very vague. I remember that it caused floods and landslides across the Island. My memories of Irma and Maria are more vivid. Both hurricanes passed over my house and I remember watching Irma in my terrace with my family.

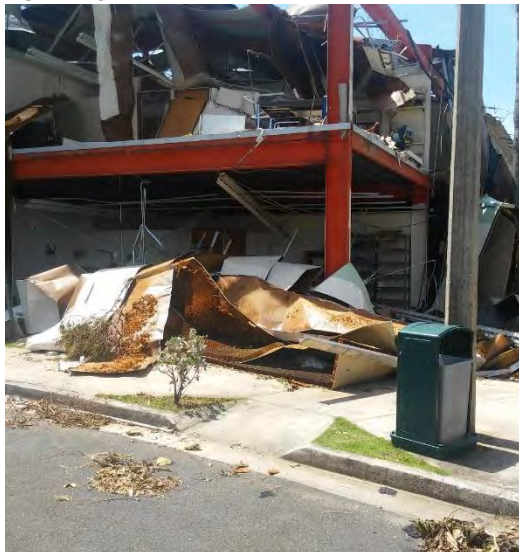
After Irma, I never imagined that another hurricane was going to hit the Island with such a raging force. Maria came 19 days after, and the island, the forest and myself are still recovering from its aftermath!



**PHOTO 1 TAKEN BY LUQ INVESTIGATOR GRIZELLE GONZALEZ-EAST PEAK, A LUQ RESEARCH SITE**



**PHOTO 1 TAKEN BY LUQ IM EDA MELENDEZ-IN THE UNIVERSITY OF PR (UPR)-USED TO BE MY BUILDING**



**PHOTO 1 TAKEN BY LUQ INVESTIGATOR GRIZELLE GONZALEZ-A PALM FOREST?**



**PHOTO 1 TAKEN BY LUQ IM EDA MELENDEZ-IN A RECREATIONAL AREA IN FRONT OF MY HOUSE**



**PHOTO 1 TAKEN BY LUQ INVESTIGATOR GRIZELLE GONZALEZ-LUQ MOUNTAINS**



**PHOTO 1 TAKEN BY LUQ IM EDA MELENDEZ-IN SOMBODY'S FRONT YARD**



Never had I felt our lives in danger while experiencing a hurricane event. My family and I were able to look through the windows and see the “Beast” (some people in Puerto Rico refer to it that way) trying and succeeding to tear apart and destroy some of our already weak homes and electrical infrastructure. The trees were left without leaves or were tipped over, many houses were torn apart and many people were left homeless. We were lucky, our house was left intact and we all still had our jobs. However, we felt the devastation, desolation and a linger sensation that Puerto Rico will never recover from this... I understood then how Europeans and Japanese must have felt after WWII...

But life must continue, and in these situations, there is only a way to survive and keep sane: keep busy! I engaged myself in administrative tasks that I had never done before: an inventory of damages affecting our Department of Environmental Science’s research. This gave me the opportunity to get acquainted with all the professors from my Department, LTER or non-LTER. I even had the opportunity to meet the Dean and other University's staff that I had never met.

I panicked the day after the hurricane. I was able to enter the university only to find that my office was destroyed. Fortunately, I was able to salvage all of the computers of the Administrative and



Information Management staff. I was not worried about loss of data; I had backed up all the data the day before Maria. The server was worrying me: Had the system administrator backed up the server? I had backed up all the LUQ's website database, but he had been out of town for a week and I had no idea if he had backed up the server's file system.

I was thrilled to learn that he had taken the proper precautions and had backup everything. Unfortunately, the computer hosting our website did not boot up again. Fortunately, our system administrator bought a domain (lter.network) and our website and filesystem was up in a Cloud a month after "The Beast" hit us.

So, my professional kit to survive the battering of a merciless hurricane event: Work hard, buy a cloud and have Backups, plenty of them!

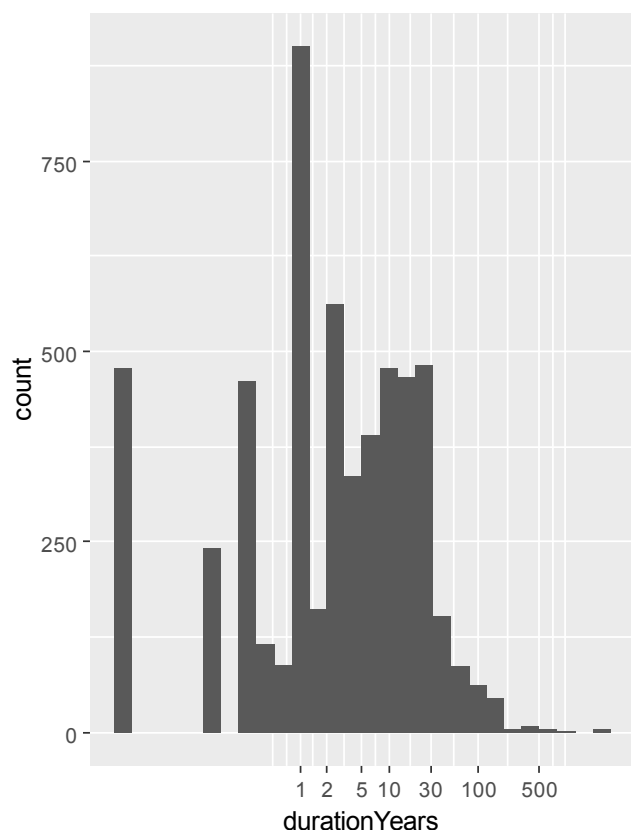
## Duration of LTER Datasets

**John Porter, VCR/LTER**

One of the goals of long-term ecological research is the collection and assembly of long-term data. However, often it is difficult to get a handle on the duration of such datasets for the LTER Network as a whole.

An opportunity to get the "big picture" arose out of a session at the last ESIP meeting that focused on developing web-service-based data catalogs that harvest information directly from the PASTA repository. One of the tools was a web query form that had as one of the output forms a comma-separated-value (.csv) file, and because the output included the starting and ending dates for each dataset, only took a few minutes to generate R code to summarize the duration of 6,106 data packages in the knb-lter-\* scope.

Aggregating across sites there are two major features shown on the histogram of data durations. First



there is a substantial amount of long-term data represented by a bell curve. Overall 26% of LTER data packages were a decade or longer, and an additional 14% fell between 3.5 and 10 years in duration. Second, there are clear spikes at the 1-year, growing season and 0-year (one-time study) time intervals. Data with duration less than 3.5 years in length collectively make up 38% of all data packages.

There are a number of caveats regarding this data. First, a succession of growing-season or single-year data packages that collectively make up a long-term time series would be unrecognized. This analysis focuses only at the data package level, not at the collection-level. Secondly, for ongoing data collections, not all data packages record an "ending" date so a duration cannot be calculated. Here these were lumped with the "one-time" data packages. Finally, the long-term data includes data assemblages

that predate the start of the LTER program – often by hundreds of years.

Despite the caveats, there is a clear way to interpret the histogram. First, there is a large number of datasets in the 10-30 year timeframe that represent long-term data collection efforts at the sites. The peak at around 3 years may represent data from leveraged standard LTER grants whose data is archived by the LTER. The peak at 1 year represents a logical interval for both short-term studies, whereas the peaks above 0, but less than 1, probably constitute studies from individual growing seasons (with growing season length varying by site). The final peak near 0 may represent one-time sampling efforts, such as development of GIS layers, land cover layers, topographic surveys, biodiversity

## Total Eclipse of the Sun at the Andrews Forest

**Adam Kennedy, Mark Schulze (AND)**

The August 21, 2017 solar eclipse was a unique experience at the Andrews Forest (AND). Models estimate that the Primary Meteorological station was positioned with 99.84% obscuration. This “close enough” location provided an opportunity to examine high-resolution (1hz) temperature and shortwave radiation relationships at one of its established long-term climate stations (PRIMET).

To provide a low-cost but high-value dataset, AND added a campaign measurement program to a Campbell Scientific (CSI) CR1000 data logger. The data logger was equipped with an NL115 network module and linked to the AND telemetry server. Solar radiation and temperature data were measured at 1 hz with an Eppley Pyranometer and a RM Young fan aspirated CSI 107-L. The streaming data were retrieved with CSI LoggerNet Admin software and post processed with the GCE Data Toolbox for Matlab (version 3.9.7b) running on Matlab (R2016b).

The start of the partial eclipse began at 0805 PST (Fig. 1. C1). The maximum eclipse occurred over

**FIGURE 1. SOLAR ECLIPSE INFOGRAPHIC OF TEMPERATURE AND SOLAR RADIATION DATA TAKEN DURING AUG 21, 2017 AT THE ANDREWS FOREST PRIMARY MET STATION.**

an hour later at 0918 PST when obscuration at PRIMET reached 99.84% (Fig. 1. MAX). The end of the partial eclipse occurred at 1039 PST (Fig. 1. C4).

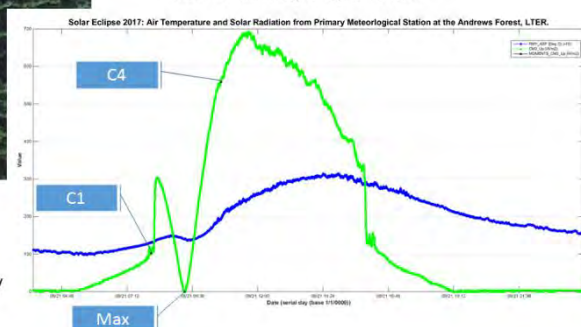
Several interesting features are visible in the resulting dataset. First, between C1 and MAX, we observed a spike in solar radiation rates and a slight decrease in temperature. The former is unrelated to the solar eclipse – rather, it is a function of decreased topographic



**Solar Eclipse 2017**  
Air Temperature and Shortwave Radiation  
Observed at the Andrews Forest, LTER, Primary  
Meteorological Station.

### Solar Eclipse: August 21, 2017

Obscuration: 99.840%  
Start of Partial Eclipse (C1): 08:05:24 [PST]  
Maximum Eclipse (MAX): 09:18:51 [PST]  
End of Partial Eclipse (C4): 10:39:06 [PST]  
Station Lat/Long: 44.21183°, -122.25587  
Station Elevation: 430 meters  
Observation frequency: 1-second



shading as the sun moves above the south bounding ridgeline of the steep, narrow valley in which the climate station is situated. Even though the PRIMET station was just outside the zone of totality, solar radiation approached  $0.0 \text{ W/m}^2$  during MAX.

This was a fun and relatively easy-to-implement data collection campaign because AND has moved to a standardized data stream (both collection and pre/post data processing at most sites) and leverages a long-term climate station that is equipped with a calibrated suite of measuring tools. While we didn't expect to generate ground-breaking new information with this test, this example could serve as a framework for future low-cost, high-value data collection campaigns.

## Visions of LTER IM: A Discussion at the 2017 Meeting

**John Porter, VCR/LTER**

The 2017 LTER Information Managers Meeting included a working group that had as its goal helping to define visions of where we want to go in the future. The group had a broad range of participants, some new to LTER and others with decades of LTER experience. The discussions included some important "lessons learned" and some ideas about areas where innovations and improvements were possible.

In the lessons learned, one new-to-LTER Information Manager with 15 years of non-LTER experience recounted how the Exxon Valdez wreck spurred wide-ranging data collection but had no plans or policies for managing the data accumulated. After 10 years they tried to play catch up, and predictably ran into many barriers with an estimated 1/2 billion dollars' worth of research data lost. However, they have subsequently observed a real change, with clear generational differences, in the research community, leading to an increased recognition in the value of sharing data. Another new IM had experience with the CUASHI system that is semantically-driven using WaterML and noted the pluses and minus with respect to translation of formats, semantic limitations and computational expense. Yet another came from an Ocean Sciences background where there were fewer data portals and web sites. They noted that it was harder to work across, rather than within, disciplines and that IEDA (<https://www.iedadata.org/>) was a common repository for that community, where EDI and LTER were largely unknown.

There was also a discussion of challenges and possible improvements that would help to surmount them. One topic of discussion focused on the desirability of good community standards and development of data models and shared systems to support them. DEIMS was discussed in that context, as was the desirability of improving shared use of units for measurements. In the area of discoverability, there are an increasing number of controlled vocabularies being used or developed in the scientific community, but there are few "crosswalks" between them, hindering semantic mapping of data between systems. That led to a discussion of the need for better descriptions of attributes, perhaps by mapping dataset attributes to archetypal attributes. The European Molecular Biology Laboratory (EMBL) is getting close to doing this and there are also efforts in ontology development in environmental chemistry. The group also discussed how it might be possible to develop attribute descriptions through consolidation of primary-level files coupled to machine learning. However, to achieve machine-level understanding of a dataset we would need to exhaustively document each

attribute. There were many difficult edge cases that would necessitate going from human annotation to machine annotation in order to build up a sufficient knowledge-base to create good algorithms.

Several issues involved the relationship between Information Managers and investigators, such as the desirability of having researchers consult Information Managers (IMs) early in a project, the need for easy-to-use tools for collecting data and the desire to build an ethos that increases end-to-end coordination. Also mentioned were the curious dynamics that sometimes influence IM and researcher interactions. Both need to work together to advance data handling. Often, IMs know more about the range of possible solutions to data handling problems, but researchers may be better at identifying the specific problems themselves. This requires IMs to be proactive and try to anticipate what researchers need to advance management of data.

There was also a discussion of the role of trust in the use of scientific data and ways to indicate to scientists how much they can trust a given dataset. It would be desirable to have some estimate of reliability of a dataset, perhaps from researchers themselves. It was noted that NASA supports ratings for datasets based on rigor of checking and user provided ratings. It was suggested that the qualityControl element in the methods section could be more widely used in EML documents. For example, it is seldom that cross-technician comparisons are noted, although in some cases the identity of the technician collecting the data may have a measurable impact on the data values. There are some groups working on developing data quality assessments, with the goal of generating data quality “badges.” Finally, the reliability of data repositories themselves is also an issue.

The working group concluded by focusing on efforts towards developing standard methods and useful vocabularies. There have been some successful standard methods developed for information management. In some cases, the use of common tools drives the use of common formats for metadata and data. However, the success of a tool (e.g., EML), often depends on how well it is understood by the potential users of the tool. User-oriented tools for EML generation have not been very popular with general users, although that may reflect limitations in the specific tools rather than limitations of structured metadata per se. The group discussed the role education of graduate students on metadata content could play. One opinion was that quality metadata almost always results from a collaboration between IMs and researchers. Researchers need to provide many of the text elements, such as abstracts, whereas IMs deal better with the more technically-oriented parts. Often, iterative development of metadata is needed as documents are passed back and forth between IMs and researchers. Frequently the data itself may be poorly structured, with many inconsistencies that need to be corrected before they can be adequately described in metadata. An increasing use of tools, such as R, that promote consistent data representations, is helping to improve this problem, but either the education of researchers as to what constitutes good archival data, or additional work by IMs, to reformat badly formed data, are still needed. One guideline for telling researchers what to include in their metadata is to ask them “What would you want someone to tell you about the data if you were getting them from someone else?”

There was also a discussion of the existing LTER controlled vocabulary and the need to periodically look for new terms. However, much of the discussion focused on how a richer semantic content to develop data discovery and use could be developed. One option discussed was using links to related publications to try to enrich metadata. Another was to try to engage the community in the effort similar to organizations that use crowdsourcing to annotate images. However, such efforts are challenging and often unsuccessful.

Ultimately, the working group did not come up with specific visions for the future, but it did outline some general issues that we can focus on addressing and laid out some possible approaches and solutions that might form the basis for discussion by future working groups.

## Good Reads

### The FAIR Guiding Principles for scientific data management and stewardship

**Margaret O'Brien (SBC)**

**Marine Science Institute, University of California, Santa Barbara, CA, USA**

Citation: Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data3:160018 doi: 10.1038/sdata.2016.18 (2016)

A diverse set of stakeholders representing academia, industry, funding agencies and scholarly publishers are jointly designing a set of principles referred to as the “FAIR Data Principles”, for Findable, Accessible, Interoperable and Reusable. Their intent is to guide those wishing to maximize the added-value of their data holdings and assist them in evaluating their individual data management choices. One project goal is enhancing automated discovery and use in a data ecosystem that is increasingly decentralized and diverse, which exacerbates discovery and reuse for both humans and computers. This Nature Comment is an introduction to the FAIR project that includes the rationale behind the four principles and some exemplar implementations in the community. It will be of interest to the LTER community, as we have been discussing these issues for many years, and this project provides a framework for those discussions which is also being adopted by the broader scientific community and the repositories which house their data.