Midterm report to LTER

Introduction

The mission of the Environmental Data Initiative (EDI) is to expand and accelerate the curation, archiving, dissemination, and re-usability of environmental data, with emphasis on data from projects funded by the NSF DEB.

EDI's goals are to ensure that environmental data are deposited into a data repository for long-term preservation and data integrity and to ensure that environmental data are documented with rich science metadata to be easily discoverable, seamlessly accessible, and re-usable to advance scientific discovery.

Other goals include building, expanding and strengthening the information management community via training, providing collaboration platforms, identifying and supporting major needs, and bringing IMs together at national meetings.

EDI was funded by NSF in June 2016 combining two proposals (PASTA+, PI Servilla, and Network Information Management Office, NIMO, PI Gries) into one collaborative project. The scope of EDI was expanded to serve the larger ecological community in addition to LTER, specifically projects funded by the NSF DEB programs Long Term Research in Environmental Biology (LTREB), MacroSystems Biology (MSB) and the Organization of Biological Field Stations (OBFS). Having its foundation in LTER data management, EDI is standing on the shoulders of giants as this abbreviated timeline shows

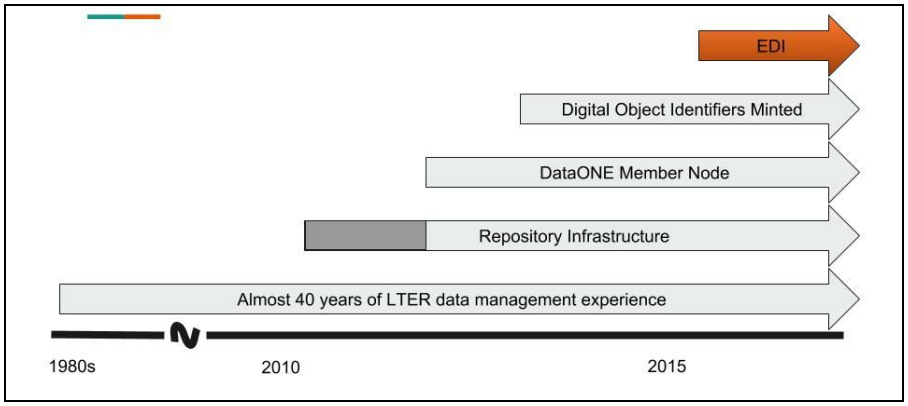

Figure 1: Timeline of LTER and EDI data management and repository developments

The Repository

The EDI (formerly LTER) data repository has been recognized as trustworthy by several journals and organizations, listing it as one option for archiving ecological data. We are also in the early stage of CoreTrustSeal data repository certification; CoreTrustSeal is recognized by most journals, thereby eliminating the need for certification by multiple organizations. Furthermore, we recently applied to the NSF Center for Trustworthy Scientific Computing to conduct an in-depth evaluation of infrastructure cybersecurity with recommendations for improvement, which was last performed on our repository in February 2013.

The basic design of the underlying software has not changed and performs well with very little downtime due to a strict, proactive management and a three tier release schedule, replicating PASTA structure on development, staging, and production platforms. In addition to local and near-site backups, full data replication is now provided through Amazon Web Services' Glacier cloud storage. The original design of a Service Oriented Architecture supporting a REST web-service interface has proven stable and can be successfully used to programmatically access data stored in the repository.

EDI recently contracted with DataCite to be a Digital Object Identifier (DOI) allocating agent, thereby providing and managing DOIs locally; the use of DOIs elevates data packages to citable scientific products that will become interlinked with paper publications, source datasets and eventually code either within the repository or through a third party. Linking data packages to journal articles within the EDI repository is fully functional, but the third party linking of papers and data is still in development by a group called Scholix. Furthermore, both Servilla and Gries have participated in the workflow development involving journal publishers, DataCite, CrossRef, and other repositories through the FAIR (Findable, Accessible, Interoperable, and Reusable) data project at AGU; Servilla is co-chair of the FAIR Workflow recommendations between data

and publishers working group. All of these efforts will make data much better discoverable while also providing appropriate credit to the data creator.

EDI continues to collaborate with LTER information managers to implement quality checks to provide metadata congruent with the data and keep the high standard for the metadata content. Four new quality checks were deployed in May, 2017, with three additional checks planned for deployment in May, 2018 (bringing the total to nearly forty). Currently over 42,700 data packages, occupying 8.5 TB of space, are maintained; 6109 of these are LTER data packages and EDI added ~150 datasets from other communities during the last year.

We are maintaining two data portals, one LTER branded, the other for EDI, both of which are accessing the same underlying data. Either portal may be used for data upload by information managers, as well as search and download. Web statistics for approximately 3 months show that the EDI data portal sees considerably more traffic than the LTER data portal.

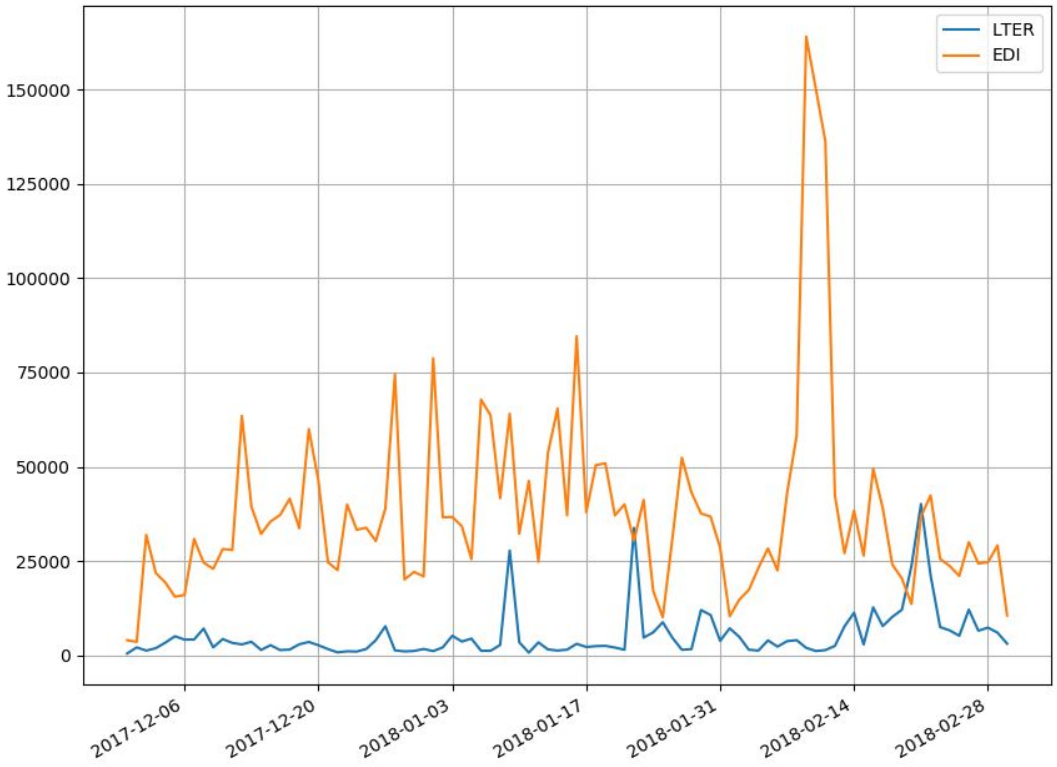

Figure 2: Web traffic analysis through the data portals for approximately 3 months of operation

Technology developments

PASTA software and Data Portal

As mentioned above, the basic design of PASTA software is stable and functioning well and has not changed significantly for EDI; updates were required to make PASTA more generic, including the addition of a new data portal to support EDI (see above). This process involved removing LTER specific branding from the EDI portal and PASTA in support of more neutral consideration of groups other than LTER who may submit data to the repository.

Provenance metadata has now been expanded to include data sources from outside of PASTA, thus providing a more complete view of data provenance described on metadata summary pages located on the EDI data portal. Also new on the EDI data portal is a view displaying journal citations for manuscripts that cite data DOIs from PASTA. Such information is compiled from users and feeds into the the DOI DataCite metadata that will be used by the Scholix project. Tracking data citations in journals is critical from NSFs perspective to better understand the impact of data collection.

We also added a new DataONE member node for EDI, in addition to the existing LTER member node already managed by EDI. Functionality of both member nodes is being improved to better support the needs of our communities.

A new technical effort underway is the development of a system dashboard that provides PASTA administrators the ability to monitor repository activities and system state of health information through a single web application (https://dashboard.edirepository.org/dashboard). This effort is ongoing and will evolve with the needs of the EDI technical team.

All PASTA and related software may be accessed through the PASTA+ GitHub repository at https://github.com/PASTAplus.

Data management support

EDI maintains a GitHub repository that is open to the community: https://github.com/EDIorg and is being used to collaboratively develop the workflow code mentioned in detail below.
In addition to the github, we are proposing to develop a repository for data management specific workflow code that may be either reused, provide examples on how to do certain tasks in code, or develop into publishable code packages. Managing access to this code in a central place will allow IMs to find it, collaborate and learn from each other. We are using software developed by an EarthCube project for this repository OntoSoft (http://www.ontosoft.org/). The opening of this code repository (http://edi.ontosoft.org/) will be celebrated with a 3 day Hack-a-thon at UNM in early June with 10 participants (4 LTER).

Supporting synthesis research

In synthesis research the largest time investment is still in discovering, cleaning and combining primary datasets until all data are completely understood and converted to a similar format. There are two approaches to achieving this data regularity: a) to prescribe the format before data collection starts, or b) to convert primary data into a flexible standard format for reuse. Prescribed formats are impossible to impose on research studies, so we take the second approach: define a flexible intermediate data model, and convert primary data. Because several synthesis working groups supported by the NCO were interested in long-term community observations, we organized a workshop with members of those working groups to developed a standardized data model for such ecological community survey data, called ecocomDP. This is an intermediate between primary observations and the formats needed for each specific synthesis question. This approach allows the data to remain in the format designed by the original researcher, but also become more accessible to multiple synthesis question, which significantly reduces the time investment. The reformatting developed here has no data loss, no aggregations and does not involve any science question driven decisions.

Eight LTER data sets have been transformed into ecocomDP, submitted to the EDI repository and used by the metacommunities working group. Approximately 50 candidate dataset have been identified, and are being prioritized for transformation and upload to the EDI repository, based on needs of NCO and NEON synthesis working groups. The usefulness of this approach is demonstrated in collaboration with the POPLER project. An additional 215 (https://github.com/AldoCompagnoni/popler) datasets are convertible to ecocomDP on demand via a EDI developed conversion script. Furthermore, an EDI/NEON collaboration is underway to convert suitable NEON data products to ecocomDP, with additional plans to explore automated conversion to other common formats for this type of data (e.g., DarwinCore Archive). All resources can be found at https://github.com/EDIorg/ecocomDP

ecocomDP - The data model itself, which is a collection of data tables and accompanying EML metadata.

ecocomDP R code package. A user friendly workflow and set of functions to help data providers reformat data according to the ecocomDP data pattern, create EML metadata, and synthesize ecocomDP data packages stored in the EDI and NEON data repositories.

Taxonomic information is essential to ecological community surveys, and we developed an R workflow to assist with this process using available R packages (r taxize, https://cran.r-project.org/web/packages/taxize/index.html). taxonomyCleanr is a user-friendly R workflow and set of functions to help data providers and managers clean and document taxonomy data (i.e. accessing several authority databases it helps correct misspelled scientific and common names, select accepted names, identify synonyms, and then creates EML metadata). The workflow is used as part of the ecocomDP system, but has broad use, and is available independently (https://github.com/EDIorg/taxonomyCleanr/).

Training, Outreach and Community Building

EDI has published a monthly newsletter since summer 2017 (https://environmentaldatainitiative.org/blog/page/3/) with technical news, a featured dataset (https://environmentaldatainitiative.org/data/edis-featured-data-contributions/) and announcements of engagement opportunities. The newsletter currently has 150 subscribers, 46 identify themselves as LTER and 16 as LTER IMs. The newsletter is usually forwarded to various other mailing lists reaching more than those 150 subscribers.

Our website https://environmentaldatainitiative.org received 5000 visitors and 17,000 views between April 2017 - March 2018. The website provides extensive data management resources, which were migrated from the LTER IM website, edited and updated. A Twitter account was established upen recommendation of the EDI advisory board and has grown to 130 followers. Servilla developed a data upload tweeter which announces every newly uploaded or updated dataset making our data more visible.

EDI has been represented at conferences with presentations (7), posters (2), and booths (2). We have organized special sessions (2) and workshops (8) at ESA, AGU, ESIP, OBFS, and GLEON meetings. In the coming season we will have a ‘data desk’ at ESA in collaboration with ESIP, DataONE, DataCite and iDigBio, where we’ll provide general data consultations, data management plan help, answer questions, and provide a forum to demonstrate data applications. This was started at AGU and we anticipate more visibility at ESA.

EDI provides remote data management consultation to single investigator labs and multi investigator projects and we have responded to inquiries maintaining a database of 340 contacts and a log of ~200 conversations. These conversations are more general inquiries including those by larger research or monitoring groups that are looking to develop their own data management procedures (e.g., Nature Conservancy, California Department of Water Resources, National Park Service). Only some of these conversation lead to inclusion of individuals in the community of data managers which is built more successfully through our training activities.

EDI ran 17 online information/training webinars between June 2017 and April 2018 with an average of 15 participants each and covering data management subjects in general (introduction to EDI’s services, what are metadata, what should go into metadata, introduction to R and RStudio), specific tools develop by EDI (R workflow code for generating EML) and repository topics (uploading data, using the API) https://environmentaldatainitiative.org/events/training-webinars-workshops/). Most of these webinars were recorded and are available on EDI’s YouTube channel https://www.youtube.com/channel/UCNZoWPaMG6lkEiH8xRNnrrA. As mentioned above extensive written data management resources are available on our website and our github:

(https://environmentaldatainitiative.org/resources/five-phases-of-data-publishing/, https://github.com/EDIorg/tutorials, https://github.com/EDIorg/EMLassemblyline/blob/master/documentation/instructions.md)

We conducted one in-person training course for 17 participants (4 LTER IMs) as a three-day session at UNM in Albuquerque. As a result, non-LTER participants have contributed 22 new data packages and 8 projects/sites have adopted EDI's data publication workflow. These adopters remain in close communication with EDI staff as they implement this workflow, which will result in many new data contributions and self-managed maintenance. A second workshop is planned for June 2018.

This summer EDI's first data management fellowship will take place with six fellows who will be working at different field stations. They all will attend our training workshop and then are expected to learn how to assemble, clean and submit data sets from their respective site in close collaboration with site researchers and EDI staff.

EDI in the larger data curation community

EDI staff are actively involved in shaping data curation approaches, best practices, and standards that will help coordinate between repositories, clarify the landscape and overall make data more discoverable and usable. As outlined above, we participated in the AGU FAIR project (making data Findable, Accessible, Interoperable, and Reusable), in which journal publishers, repositories and data managers collaborate to develop guidelines for publishing data.

Prior NSF EAGER funding and current ESIP working groups are discussing measuring Return of Investment for data repositories while EDI is working out its own cost schedules to inform NSF of resources needed to make data archiving mandatory. EDI is a voting member of EarthCube's Council of Data Facilities, staff members participate in EarthCube's Technology Architecture Committee and Research Data Alliance (RDA) working groups to stay informed and help shape the data curation landscape.

A current work in progress is formulating EDI's data policy https://environmentaldatainitiative.org/data/edi-data-policy/ keeping possible legal ramifications in mind and discussing the issues with other repositories.

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.

Several EDI personnel are also members of the EML development committee and are active in developing Version 2.2 release. Specific contributions from EDI are to EML's measurement units and to semantic annotation. The work on measurement units coincides with EDI's work on the LTER unit registry (see below). Since EML 2.2 will be backward compatible, we developed a system of deprecation to shorten the list while still allowing legacy units in EML instances. Semantic annotation will allow elements of EML to be "tagged" to concepts in external dictionaries, and RDF triples constructed for metadata elements. An EML 2.2 release is expected in mid-2018.

Specific services to LTER

SiteDB: It was recognized that the LTER siteDB contained information useful to the development of the new LTER website https://lternet.edu/site/ and the NCO transferred most of the text into the new structure where sites are able to update their information. However, siteDB contained some standardized numeric site descriptors which were archived as a dataset in EDI https://portal.edirepository.org/nis/mapbrowse?packageid=edi.138.3.

Clim/HydroDB is currently maintained by EDI while we research best approaches to upgrade this resource. We initiated a conversation with NCEI to possibly host climate station data from sites and integrate them in their system, while EDI would host annual snapshots for archival purposes. We will propose a working group at the LTER All Scientist Meeting to discuss the future of this LTER resource.

Unit registry: The LTER Information Managers Committee (IMC) created a unit registry for reference and unit reuse. The registry has a relational back end, web services and interface (http://unit.lternet.edu). LTER also created text guidelines for best practices related to measurements and units. As EDI needs measurement description resources for its own EML construction, the LTER unit dictionary is a candidate. We have migrated the LTER text (best practices and other material) from the IMC website to EDI. The Unit database itself has been migrated to modern infrastructure at UNM and the content archived. We are exploring approaches for adapting it for the broader community's needs as "Version 2", mindful of LTER legacy code which depends on Version 1.

Publications

Smith, C. 2018. The EML Assembly Line: A Metadata Generation Tool for Data Providers in the Ecological Sciences LTER Databits Spring 2018
https://lternet.edu/wp-content/uploads/2018/03/2018DatabitsSpringIssue-web.pdf

Vanderbilt, K., C. Gries 2018 The Environmental Data Initiative - the first 1.5 Years Supporting LTER Information Managers, LTER Databits Spring 2018
https://lternet.edu/wp-content/uploads/2018/03/2018DatabitsSpringIssue-web.pdf

Glossary of abbreviations

AGU - American Geophysical Union
DataCite - https://www.datacite.org/
DataONE - Data Observation Network for Earth https://www.dataone.org/
DEB - NSF Division of Environmental Biology
DOI - Digital Object Identifier
EML - Ecological Metadata Language
ESA - Ecological Society of America
ESIP - Earth Science Information Partners http://www.esipfed.org/
GLEON - Global Lake Ecological Observatory Network http://gleon.org/
iDigBio - Integrated Digitized Biocollections https://www.idigbio.org/
LTREB - NSF program ‘Long-term research in environmental biology’
MSB - NSF program ‘Macro Systems Biology’
NSF - National Science Foundation
OBFS - Organization of Biological Field Stations https://www.obfs.org/
R - statistical language
RDA - Research Data Alliance https://www.rd-alliance.org/
UNM - University of New Mexico