DATABITS

The Newsletter of Ecological Information Management
Summer 2019

A full slate for the Summer 2019 Databits! Site Bytes makes a triumphant return after a long absence. There is also plenty of information on new initiatives, including upgrades to PASTA, Best Practices for using Zotero to manage bibliographic data, development of a core metabase for metadata management, a chance to catch up on a former colleague, some reports and article reviews and finally, an extended primer on semantic annotation. Enjoy! – John Porter & Sven Bohm, editors

TABLE OF CONTENTS

SITE BYTES

KELLOGG BIOLOGICAL STATION LTER

This last year KBS welcomed Hsun-Yi Hsieh to the data management team. She will be helping with LTER and GLBRC data management and has already helped to clear the backlog of un-uploaded data. This year we plan to better incorporate links to the EDI catalog into our metadata system. – Sven Bohm

FLORIDA COASTAL EVERGLADES LTER

The big news at FCE LTER is that we are updating our website from a mostly hand-coded version to a hybrid site using a content management system (CMS) for our static content and the Foundation CSS Framework for our dynamic content. Moving to a CMS will allow

members of the FCE team, besides the IM and project manager, to manage their own section of the website. Florida International University has invested in the Cascade CMS and moving FCE static web content to that platform was straightforward. Cascade, unfortunately, does not readily support FCE's custom query interfaces to data, bibliography, and personnel pages. Program Manager Mike Rugge has implemented the Foundation Framework for the dynamic part of the FCE website. He has created a template that mimics the appearance of the Cascade site, into which FCE's Perl scripts will be embedded. The new FCE website will have several features that the current one does not. It will be responsive, integrate social media, and have a site-wide search capacity. – Kristin Vanderbilt

Central Arizona–Phoenix LTER

As is the case with many, if not most, LTER sites, the CAP LTER employs two separate but parallel systems for publishing site research data. CAP maintains a database of research data sets, and associated data and metadata files that are accessible through a data catalog on the CAP website. At the same time, these same data and metadata files are uploaded to the Environmental Data Initiative (EDI) data repository. CAP's current infrastructure requires this parallel approach to (1) meet the NSF requirement that data are deposited in a national, public data repository (the EDI in this case), and (2) showcase CAP data products, and provide convenient access to CAP data to project investigators, decision makers, and the regional public through the data catalog on the CAP LTER website. These two approaches meet the project's goals but the redundancy is inefficient.

To streamline CAP's data-publishing efforts, web programmers at the Julie Ann Wrigley Global Institute of Sustainability (GIOS), CAP's home department at Arizona State University, are developing a new data portal for the CAP website that will feature an inventory of and links to CAP data sets based on a custom implementation of the PASTA+ architecture. This approach will eliminate the need for maintaining a separate infrastructure (database, file server, XSLT functionality, etc.) and workflow to feature CAP data products on its website. Additionally, leveraging the PASTA+ architecture provides much improved search and viewing functionality than CAP is able to achieve through its current implementation. Leveraging the PASTA+ architecture will also enable GIOS to expand, in a very visible way, the reach of its data-management services beyond CAP. GIOS has almost forty faculty appointments in the School of Sustainability, many of whom conduct primary research in an ecological or environmental domain, yet only a small number of whom are affiliated with CAP. Most of these faculty have data-publishing needs, yet lack the resources available to CAP investigators. Once constructed, the same code base used to build the new CAP data catalog on the CAP website can be adapted easily to build a catalog of GIOS data sets on the GIOS website. Non-LTER data sets will be uploaded to EDI with an EDI scope then harvested using a keyword unique to these data sets. Much of the effort involves rebranding the PASTA+ architecture with a ASU theme, a non-trivial task, and it is a testament to the sound design of PASTA+ that this is possible. CAP and GIOS hope to debut the new data catalogs in spring 2019. – Stevan Earl

Niwot Ridge LTER

Niwot is in the process of overhauling it's website, moving from a locally-hosted CMS (Expression Engine) to Squarespace. In August, we updated our local data catalog to used EDI as the back-end rather than our local database (with credit going to code provided by the Tim Whiteaker at BLE: check out his code https://github.com/BLE-LTER/PASTA-JavaScript-Search-Client) if you had not yet had a chance!). This workflow allows us to improve efficiencies in metadata preparation and submission visavis maintaining a local system and submitting to EDI separately. It takes advantage of the speed and enhanced search capabilities of LTER-wide resources provided by EDI (specifically, PASTA's solr search client); ensures all datasets are downloaded with a citable DOI improve usage tracking; and enforces versioning to encourage reproducible workflows. The Squarespace website (currently in development) will continue to use this method for data access. Separately submitting datasets to EDI and updating metadata in our local database had led to some divergence in recent years, so in the meantime we are working hard to get everything on EDI up to date with complete EML before fully deprecating our current system. – Sarah Elmendorf

Beaufort Lagoon Ecosystems LTER

The BLE team completed its inaugural field season on the Beaufort Sea coast last August. Since then, we have been analyzing data and preparing for the winter field season, which involves training to prepare for cold climate operations and a bevy of logistics to sort

out. We're excited to see what these lagoons reveal about seasonal variability once we've completed the winter and spring sampling.

On the information management front, we welcome An Nguyen who will serve as Assistant Information Manager. Four days after she joined our team, we received our first BLE dataset to archive, so the timing couldn't be better. In the coming year we plan to employ the LTER Core Metabase as our internal data catalog and contribute insights and improvements back to its GitHub repository. We also plan to develop Zotero best practices for tagging publications based on how strongly they are related to BLE and improve the Zotero JavaScript Search Client on GitHub accordingly. We'll deploy a data catalog interface and an improved Zotero search interface to our website later this year. We've been very happy with our static HTML website thus far, and love talking about it and data catalogs and Zotero and things IM, so feel free to reach out to Tim and An! – Tim Whiteaker

Konza Prairie LTER

The major accomplishment of KNZ IM team during last year included redesign and improvements to the information management infrastructure, to support data and information services for the KNZ LTER research.

We manage the KNZ website (http://lter.konza.ksu.edu) in the Drupal Environmental Information System, which was launched in March 2017. The website now provides access to 129 projects with total of 415 data files (250 are GIS shapefiles and KMZ file). Our structured metadata allows linking datasets, publications and personnel and better search capabilities. The EML-export system automatically generates PASTA-compliant attribute level EML files with all KNZ datasets. All of our online data with metadata are uploaded regularly to the repository of the Environmental Data Initiative (EDI). Using Google analytics, the website receives a high number of visits (~ 900 visits and 2800 pages views per month for the past year), with approximately 71% as returning visitors.

During the past year, we updated all KNZ LTER-supported/related publications, thesis and dissertations (total of 1807 publications online), searchable by key word, author, year, and publication type. We continue to make improvements to our stream data workflow infrastructure, upgrading scripts, documentation, archive, and data integrity checks. We edited and updated our metadata and procedural protocols to ensure any changes in technique or structure of our datasets are accurately reported.

We have redesigned our outreach webpage to increase the visibility of KNZ impacts on the education, outreach, and training activities, to increase the relevance of our long-term research to broader society (http://lter.konza.ksu.edu/konza-lter-outreach). In addition, our website now includes our diversity, inclusivity, and non-harassment statements.

We have recently upgraded the core LTER file/web server, added a new windows server (2016 Hyper-V cluster) using LTER site supplement funding. These upgrades conclude a replacement of the 2016 domain controller, switches, storage, management machine, 15 virtual machines, and all hardware.

We continue to provide service and expertise within the broader LTER IM Committee (IMC). KNZ IM (Yang Xia) participated in all LTER IMC activities including the IM annual meeting, IM water cooler, EDI hosted workshops, and maintained the ClimDB/HydroDB as a volunteer administrator.

We will continue adding new KNZ projects and data sets within our local IM program as well as at the LTER network level, continue website development, maintenance of high data quality, aggregation, and synthesis to meet the requirements for the LTER NIS – Yang Xia

Sevilleta LTER

There are several projects underway to overhaul and modernize the Sevilleta LTER information management system. We are in the early stages of transitioning our data from on-premise, aging servers into the cloud. Plans are underway to develop a GitHub repository for SEV-related scripts, including QA/QC code for our datasets. Along those lines, I have been working to streamline the meteorological data processing pipeline that consists of a combination of Perl, SAS and SQL scripts into a simplified pipeline written only in R. Planning for a new, static website is also in the works. The field crew are ramping up for their spring field season, and work is underway to build and implement a new mean-variance experiment that will investigate potential changes in dryland biomes under increasingly dry and variable climate conditions. – Kristofer Hall

Andrews Forest LTER

The Andrews Forest LTER is in its 5th year of its 7th NSF LTER grant. We are gearing up to submit our 8th proposal in March of 2020. A significant amount of our IM efforts have been working with our data workflows to improve the processing efficiency and quality of our core collections into our local information system. We continue to test and validate our processes. In particular we have new systems written in Python for calculation of stream discharge (hydrology) and tree biomass (vegetation). The process of capturing and processing the metadata and more than 40 million streaming hydroclimatological measurements per year has been further refined to the point of being nearly manageable!

Communication with several PIs to improve study data quality and documentation has resulted in new databases being uploaded into PASTA. We held two effective metadata-related events recently: the first was a 'metadata party', which was advertised as such as an incentive for site researchers to provide details, methods, and data abstracts for their studies; the second was primarily directed to graduate students affiliated with the Andrews LTER as a means to familiarize them with the general process of preparing and submitting data. In both sessions Information Managers provided step-by-step guidance, hands-on practice opportunities, and basic tools to help organize and archive data.

Don Henshaw, USFS Information Manager with the Andrews Forest will retire after more than 40 years on the job. The government shutdown has delayed the retirement date until later in the year. Cross-training has been initiated with other LTER staff in preparation.

In addition to representing IM on the Andrews Forest LTER Executive team, Suzanne also serves as co-chair on IMEXEC. Both Don and Suzanne continue to be involved in Network level activities.

Andrews Forest completes high-speed wireless pilot project. The Andrews Forest was awarded a small grant from the Oregon State University Research Office to perform a high-speed wireless networking pilot project. The installation began in October 2018 and was completed before the first high elevation snow fall. The radio link uses low-cost and low-power microwave radios in the 5.8Ghz unlicensed frequency band along with a precision alignment across three towers to traverse eighty-miles of complex mountain terrain. Through this effort we have connected the Andrews Forest headquarters to the College of Forestry at Oregon State University. Total capacity has increased from 1.2 mbps to 163 mbps download and 76 mbps upload. The new wireless network increases the total download capacity at the Andrews Forest by over 10,000% compared to its previous capacity. – Suzanne Remillard, Don Henshaw and Adam Kennedy

Bonanza Creek LTER

Things continue to evolve at the Bonanza Creek LTER as we work to provide quality data management support for our research team and others interested in our data. While our basic infrastructure has remained relatively static over the past year or so, we have tried to make improvements wherever possible to improve operational efficiencies and to better leverage available resources. We maintain a comprehensive MySQL database for all aspects of site and information management which is now administered through our host institutions' computing services unit. This database provides detailed content for our website and data inventory. As needs arise and best practices evolve, we have made necessary modifications to the database structure to best meet our operational goals.

The current iteration or our website, which is built in PHP but also uses Perch software to include certain managed content sections, is operated in collaboration with other projects in our institution to better leverage support and infrastructure. Additionally, our research program has taken on a broader landscape perspective and our network of climate stations has expanded to include even more remote locations. This means we must make some upgrades to our streaming climate communications infrastructure. We are still currently using the Vista Data Vision Software system to provide a graphical user interface to all of the streaming climate data collected across our network.—Jason Downing

Northern Gulf of Alaska LTER

As a new site, this has been a busy year for the Northern Gulf of Alaska (NGA) site. The primary scientific focus of this first year has been planning and executing our three seasonal cruises along sampling lines in the Gulf of Alaska. However, we have also completed the design and fabrication of two new moorings that will extend our monitoring and process work year-round. These moorings, the Gulf of Alaska

Ecosystem Observation (GEO) Moorings, will be placed about halfway out along the Seward Line. Each of the two moorings serves a different purpose. GEO1 will return data in real-time via Iridium satellite communications from a surface buoy to a shore side receiver hosted by Pacific Gyre, Inc. Oceanographic data such as temperature, salinity, currents, CDOM, oxygen, and nitrogen will be collected and transmitted. And GEO1 will do all this while surviving ocean swells up to 30 meters! GEO2's internally recording instruments include all the usual parameters, as well as a pCO2 sensor, Acoustic Zooplankton Fish Profile, sediment trap, particle imaging system, and a passive acoustic recorder for listening for whales.

In the information and data management world, we've been just as busy. Data from the first year is being processed, so we've started a redesign of the data portions of our otherwise completed website. This started with an all-hands review of other LTER websites by the PIs to get design tips. Data from our site is stored in Axiom Data Science's Research Workspace, which is connected to the Research Workspace DataONE member node. This allows us to explore using DataONE's API to display legacy data from previous projects alongside the newest additions automatically. More on this as it is developed. There are other features of the Research Workspace (RW) that will enable integration into the wider LTER environment. For instance, RW integrates Jupyter notebooks (Python or R) which we will use as our primary method of generating EML for datasets. In the future, we will investigate expanding RW's form-based ISO 19115 metadata editor to allow the creation and export of EML records. PI's can also use notebooks to standardize analysis and visualizations of the disparate data types from our cruises. Notebooks from the Gulf Watch Alaska program exemplify standardized data plotting across a large, heterogeneous project. We hope to learn from that work, further generalizing the best practices and applicability of those types of scripts.

Finally, we have some highlights from our education and outreach component. We were fortunate to have a K-12 educator aboard each 2018 cruise. They participated in sampling activities and created logs and videos of their experience as educational aids for their students. The educator on our Spring cruise brought a team of filmmakers and developed a series of 40 short videos that introduced members of the NGA science team and described our work. These videos make up part of a series of learning modules called "Expedition Gulf of Alaska: an Online STEAM Experience", which we posted as a playlist on our YouTube channel. -- Chris Turner

Plum Island Ecosystems LTER

Much of the past year and preceding year, since migrating our web site from a Drupal 6 version to a DEIMS Drupal 7 version, has been focused on making high quality data and metadata available and searchable on both the PIE web site, http://pie-lter.ecosystems.mbl.edu/data, and the LTER/EDI Data Portal.

We use a variety of data and metadata quality checks prior to submission to the DEIMS PIE Data Catalog and LTER/EDI Data Portal. PIE uses an Excel spreadsheet template for researcher's initial metadata and data entry. A macro designed for the Excel template is used to generate a csv file of the data and to do some quality checks on the data and metadata before manual entry into the DEIMS PIE Data Catalog. EML metadata is generated from the PIE DEIMS site and submitted to the LTER/EDI Data Portal for a variety of congruency checks and evaluation before uploading the data csv file to the portal. If there are no errors in the initial evaluation step the EML and data are allowed to be uploaded to the Data Portal for a final evaluation and data upload check.

The process and numerous iterations involved with quality and congruence checks has been onerous at times but it has improved the quality of data made publically available. A continual challenge is to keep up with new congruency checks and subsequent required recoding of scripts for EML generation in addition to tracking various versions of metadata files with varying congruency checks depending upon the legacy of recent data submission. Whew...! Such is the life of an information manager!

We are looking forward to EML 2.2 with new Project Funding fields and also exploring how semantic ontologies can be integrated with our data to hopefully provide more intuitive approaches to search and discovery of appropriate data. – Hap Garritt

Northeast U.S. Shelf LTER

This is the first Site Byte from the Northeast U.S. Shelf (NES) LTER. We appreciate all the help received from other IM's in developing the plan for our Information Management System in our proposal. The first year

was very challenging as we had 3 broadscale, 5 transect, and 14 day cruises with incoming data, while needing to establish local repositories for these data in real time. We spent much of the first year initiating workflows to automate the parsing, cleaning, and extraction of metadata from ship-provided data. A highlight from our first year was launching our project website (https://nes-lter.whoi.edu/). We capped off the first year with our first contribution to EDI, an 11-yr dataset from the Martha's Vineyard Coastal Observatory (MVCO) (knb-lter-nes.1.2).

We kicked off the second year with a successful hands-on session with PIs and students during our annual project meeting. This served as the first "release" of our nes-lter-ims Python library publicly available on GitHub, developed with the Applications Group of WHOI's Information Services department. We demo'd Jupyter Notebooks to access and perform some analysis of cleaned ship-provided and PI-provided data. In particular we compared in situ fluorescence profile data with chlorophyll a from water samples. PIs were interested in these sorts of comparisons to improve the sampling program for future cruises. We are continuing to develop workflows for handling PI-provided data from post-cruise sample processing.

Our latest highlight is that we will soon release a web-based API to the PIs for a subset of cleaned data in a local repository. The API is intended to address the challenge of providing data to our PIs within one month after a cruise and prior to availability from other repositories.

The API provides the ability to import data directly into code and is language agnostic to accommodate the PIs' use of a variety of programming languages (Matlab, Python, R).

This summer we look forward to hosting an EDI Data Fellow, and we plan to engage IM's from other pelagic LTERs in telecons for this Fellow's project with data from plankton imaging systems. We would like to compare EDI's ecocomDP tables with tables formatted for the Ocean Biogeographic Information System (OBIS). In September 2019 we'll kick-start the third year by presenting a poster at the OceanObs'19 Conference related to a manuscript that we co-authored: "ILTER - the International Long-Term Ecological Research network as a platform for global coastal and ocean observation." We look forward to contributing more datasets to EDI knowing the importance of our datasets for the Global Ocean Observing System Essential Ocean Variables (GOOS EOVs). -- Stace Beaulieu

Santa Barbara Coastal LTER

In 2018, SBC LTER had completed our data manager transition, from Margaret O'Brien to Li Kui, some tasks are still fulfilled by O'Brien (e.g., website). During this transition, components of our dataset production workflow (metadata database and R scripts) were streamlined and modernized. Li Kui gave two online webinars about SBC LTER Information Management System (IMS) during summer 2018 in the Environmental Data Initiative (EDI) webinar and training series.

One of our signature data collections, the ongoing kelp forest community survey, has been heavily used by synthesis working groups. Hence, we have spent considerable effort on data cleaning and update. We validated our taxonomic information through World Register of Marine Species (WORMS) and tagged each SBC taxon (usually species) with the WORMS ID, making our datasets available for broader research groups. We also developed a workflow for observational data entry, such as adding more stringent controls to reduce the data entry errors and running R scripts for error detections. These improvements enhance usability and reduce likelihood of misinterpretation.

SBC LTER's ocean and watershed sensor data have been processed in the corresponding research labs. In 2019, we plan to migrate all data cleaning and update tasks over to the SBC LTER data manager, in an effort to develop a centralized data management system. – Li Kui

Moorea Coral Reef LTER

MCR concentrated on updating our corpus of timeseries as well as archiving data for journal articles. Both these activities are in preparation for our upcoming mid-term review next June. With each timseries update, the metadata has also been upgraded to incorporate features such as checksums, ORCiD information, and alignment of datetime formats with the ISO-8601 standard. Next year's round of timeseries updates hopefully will include semantic annotation using the soon to be released EML 2.2. But at MCR we have not yet set a procedure for when to revisit datasets which are not timeseries and would not otherwise receive revisions. With a more automated

EML generation process such upgrades could be done more frequently.

Dataset curation seems like doing the dishes. The state of being done passes so ephemerally. No sooner is a dataset revision published than a data user discovers some issue. The EML Congruence Checker (ECC) continues to save time during the data publishing process but it is limited in what it can check. Some things surface only with actual use of the data in analysis by others. In the course of cross-domain synthesis using MCR data, we discovered one community survey dataset's sampling design was misinterpreted; we decided to revise all our methods descriptions of our core time series to clarify sampling transect geometry to avoid future data misuse.

To scientists the significant metric of progress, or worth of our corpus of data archives, is how much those data are used leading to publications where authors outside our site have acknowledged use of our data, A simple search on scholar.google.com for the scope of package IDs, knb-lter-mcr, shows a rise in citation but leveling off in the last three years. This may be due to a move toward citing datasets by their DOI rather than their packageId. Not as simple an inquiry but I expect one of my fellow IMs will share a script for that within a year.

The day to day mechanics of publishing datasets does have its tedious aspects once all the tools are built and the learning phase is past. So this past year I have taken inspiration from the discoveries our investigators have published. Russ Schmitt wrote "In the last decade, a predator outbreak and cyclone devastated coral across the seascape, yet recovery was more rapid than ever before observed anywhere in the world. MCR research has unlocked the secrets of this unprecedented resilience. Critical insights gained into the governing processes, connectivities and feedbacks provide the basis for general management strategies that can help restore and strengthen resilience of coral reef communities today and into the future." I like to think that, in a small way, my efforts with their data helped make this happen. -- M. Gastil-Buhl

Virginia Coast Reserve LTER

Recent VCR activities have focused on consolidating and streamlining some of our workflows and systems. One of the challenges of a long-term research project that really is long-term is that you accumulate legacy systems and software. A top-of-the-line piece of software in one year may no longer be anywhere near top-of-the-line a decade or more later. Even worse, that older software may no longer be supported or fail to function in newer operating systems.

For that reason we went through our main servers and upgraded them to Ubuntu 18.04 LTS, which has a free support lifetime until April 2023, and a paid lifetime support through April 2028. We then ported all our critical applications to the new OS, with an eye towards identifying those that could not be updated. Fortunately, all our critical applications were able to be moved forward. However, for new programming we are focusing primarily on Python and R, with some minor work with PHP (for web forms etc.). We had also moved from Drupal as our content-management system to WordPress. WordPress is generally less powerful, but also easier to maintain and upgrade. We also try to use as few add-on modules as we can get by with, again to minimize potential obsolescence. Our system is highly modularized with workflows implemented in Linux shell scripts, so that each component in a workflow can be refactored or improved without requiring rewriting of the entire workflow. – John Porter

LTER CORE METABASE

M. Gastil-Buhl, Margaret O'Brien, Tim Whiteaker, Li Kui

Introduction

What is the motivation for using a relational database for metadata? It is an investment of time to install, load and maintain metadata in a relational database, but one that can pay off tremendously in efficiency, consistency and the ability to migrate and reuse content. A relational database (rDB) can be accessed by code that generates dataset EML as well as code for other purposes such as back-end scripted website pages or annual reports. Such reuse means less maintenance of duplicate content for those purposes. But those benefits aside, the ability to control vocabulary and migrate content, plus central editing that cascades to all uses is enough to consider using a metadata database. This is an example of adding energy to a system to control - or even lower - entropy. For a research project the size and complexity of an LTER site, the demands on the information management system (IMS) plus the scope of its datasets warrant that effort. However, if a group has a small number of datasets (e.g., ten or twenty) and little

expectation of growth, the use of a relational database is probably not justified.

NSF leadership has encouraged LTER sites to leverage existing tools within the network. There exist several data models currently in use in the LTER network to store dataset and other metadata in a relational database. Metabase is just one of those; others include DEIMS, and the custom solutions at various sites. Metabase originated at GCE LTER, and predates EML; the model is quite general. The entire GCE LTER software suite has a long track record of continual upgrades to meet increasing expectations and it had already been installed for use by a second LTER (Coweeta, CWT), which makes its components ideal candidates for reuse. The LTER-core-metabase model is based on one of those components; however, users should note that LTER Core Metabase will not work with GCE Toolbox "off-the-shelf".

The LTER-core-metabase model represents a collaborative cross-site design carried out by the Moorea Coral Reef (MCR) and Santa Barbara Coastal (SBC) LTERs, over several years at the Marine Science Institute (MSI), UCSB (Gastil 2013, O'Brien and Gastil-Buhl 2013, Kui 2018), and further vetted by the Santa Barbara Channel Marine Biodiversity Observation Network (https://sbc.marinebon.org). During all adaptations, we have invested extra time to carefully design and describe enhancements that could be merged back into a shared model or other instances, (e.g., at GCE or CWT), while keeping the data model itself loosely coupled to the related procedural code. Much of the original MCR and SBC work is included in this project.

The project at GitHub, https://github.com/lter/LTER-core-metabase, represents a subset of essential tables from the original metabase rDB, plus schemas for inventory control contributed from MCR and SBC and an abstraction layer (rDB views) for export to EML records using the EML R package available from rOpenSci (Boettiger, n.d.). The name "LTER-core-metabase" was chosen because this project started in and is maintained by the "LTER" network; it is anticipated that this database will serve as the basis for extensions and customizations (hence "core"), and "Metabase", to recognize its origins at GCE LTER.

Status and Content

A Git-clone or download includes a set of five SQL scripts which create the three schemas (DDL), set permissions, and then insert controlled and sample content. Installation assumes the user has postgreSQL installed and a mechanism to run SQL commands (e.g., GUI or command line), and is also conversant in database use, e.g., basic rDB concepts and their implementation in posgreSQL.

- 3 schemas adapted and integrated from working databases at MSI, UCSB
    - Schema lter_metabase - 23 tables based on GCE Metabase2
    - Schema pkg_mgmt - for data package inventory tracking
    - Schema eml_views_r - the abstraction layer, for export to EML records using R code
- Controlled content
    - LTER Units dictionary (V1)
    - LTER Controlled vocabulary (V1), for keywords and named thesaurus
    - Descriptions of commonly used file types (e.g., CSV, TSV, zip, KML)
    - Parent tables to drive EML features
- Required elements, e.g., measurementScale and their domains
- Optional content, e.g., typing for keyword, number and storage elements

Often the easiest way to understand a database is by examining its content and usage, and so we included a script to load example content from SBC LTER. The last script includes all metadata for four datasets, both single and multiple-entity, plus additional site-specific keywords, thesauri and sampling locations. The sample datasets were given IDs that are easy to filter so they would not collide with user-content. However, users could choose to examine or load only parts of the examples when needed (of course, being cognizant of parent-child constraints), or load them into a second, sandbox database.

Migration-based Approach

Code revision control for declarative code cannot be handled the same as for procedural code. Repairs and upgrades are incremental and applied to an existing installation with the data in place. In rare cases it is necessary to perform a complete database archive, followed by drop, re-install and re-load data (as you would after a catastrophic failure or a postgres server migration), but that is a tedious process best avoided under normal circumstances. We are using a code repository (Git) designed for procedural code, and adapting it for database code.

Because you cannot simply install a new version of a database schema after data are inserted, the general approach to development and deployment works with "migrations" or "transitions" (Nitsche, 2018). Instead of a code-repository saving a certain "state" of your

database, it saves all the steps needed to get there, and the database itself serves as the reference-state. In practice then, you script every change in SQL, and store these scripts in a manner that they can be applied to the target database, leading to the same behavioral and functional state. Thus, what is available today in the Git repository is the Version 1 starting database. Our migration code will consist of ALTER statements.

This was the process used by MCR and SBC in 2013 during their original porting and adaptation of GCE Metabase. It is consistent with the need to customize locally (as is common with LTER sites), and with components that can be added in a modular fashion (e.g., to reintroduce a "project" schema, so that "funding" can be structured). Over time of course, incremental migrations mean that new local installations will take longer, because an increasing number of scripts must be applied to bring the database to the desired state. We anticipate that eventually the community will wish to create a Version 2 starting database. At that point, a process similar to what we describe here would take place. A Version 2 process would also plan for content migration from Version 1 to 2.

Adoption

Versions of the SBC/MCR-adapted model and related code were shared as early as 2015. Those experiences and the likelihood of uncontrolled spin-offs, plus the establishment of several new LTER sites with an interest in adopting existing tools for their own centralized metadata prompted the establishment of this project. For example, as information manager of a new site, Beaufort Lagoon Ecosystem LTER (BLE), Tim Whiteaker was anxious to establish their data catalog. Whiteaker was attracted to the simplicity of the database now used by SBC compared to other database solutions (Kui 2018), its implementation as free and open source software, and the related R scripts for exporting EML. But perhaps the most appealing aspect was the potential for collaboration. Rather than a custom, independent solution, BLE preferred to invest in a project that several sites could rally around to solve a common IMS need, with sites working together to keep the database and related tools current, and to plan for changes such as for the imminent release of EML 2.2. As BLE adopts LTER-core-metabase, it will also act as a test bed for its implementation. Having just received the first BLE dataset for archiving, the timing couldn't be better!

The BLE information management team is new to PostgreSQL, and so the exercise of instantiating the database schema and loading data will be a test of not only the database design, but also the workflows and documentation related to its use, taken from the perspective of the novice PostgreSQL user. BLE has forked the LTER-core-metabase repository in GitHub and has successfully loaded sample data included in the repository. As BLE makes progress in loading its own data and generating EML, the team will suggest enhancements to the LTER-core-metabase project either through the issue tracker or changes within its own fork.

Next Steps

To date, work has been mainly on the postgreSQL model to harden ad hoc SQL and content for more general use. This phase is nearly complete, with a few issues raised by BLE still to be addressed. We envision several channels for future work. For efficient usage, additional documentation will also be necessary, for aspects of table population order, or for creating ALTER statements for your local instance.

A. LTER Working group: We recommend that LTER establish an IMC working group to be responsible for the code in GitHub https://github.com/lter/LTER-core-metabase. This group could take on planning (such as for projects listed here), or for additions to the core model, e.g., to take advantage of EML 2.2. It should also establish guidelines and vetting for suggested changes (e.g., via ALTER statements), and develop a decision tree for when these should be incorporated into the master branch and when they should remain as local forks.

B. Export EML metadata records: Currently, LTER Core Metabase has one schema for EML export using R code. That code (Kui 2018) is modular and well organized, however, it needs significant hardening for uses beyond the original narrowly-prescribed R-studio workflow under MS-Windows. An older system using Perl for EML export is still in use at MCR and provided the template to code the views used by R (Gastil 2013). The choice of procedural language is immaterial, and users may adapt any of these examples ad hoc. Work to generalize site-based procedural code is currently not planned.

C. Common EML profile: Work on a common pattern for use of EML elements as exhibited by EDI and LTER datasets - an "EML profile" - has already begun. Concurrently, EDI is considering steps toward using alternate sources of metadata for input to its EML Assembly Line (Smith n.d.) and aligning input with the EML profile. Another schema can easily be added to LTER Core Metabase to hold views matching

this profile (when it becomes available), so that it can be used with the EDI EML Assembly Line.

D. EML-to-metabase: Tools for populating tables from existing EML would be essential to an IMS at an existing LTER site considering migration to LTER Core Metabase. SBC and MCR have been through this process, and created export scripts and loading patterns that are partly reusable (Martin, pers. comm). The tasks of generalizing these will be much faster to work through after EML docs are examined and/or their structures converge.

REFERENCES

Boettiger, C. n.d., EML - Create and Manipulate Data using the Ecological Metadata Language. https://github.com/ropensci/EML (accessed 2019-03-15)

Gastil-Buhl. M. 2013. Data Package Inventory Tracking: Slicing and Dicing with SQL. Spring 2013 LTER Databits https://lternet.edu/wp-content/uploads/2018/01/2013-spring-lter-databits.pdf (accessed 2019-02-25)

Kui L. 2018. Postgres, EML and R in a data management workflow. Spring 2018 LTER Databits https://lternet.edu/wp-content/uploads/2018/03/2018DatabitsSpringIssue-web.pdf (accessed 2019-02-25)

Nitsche, S. 2018. "One does not simply update a database" – migration based database development. https://dev.to/pesse/one-does-not-simply-update-a-database--migration-based-database-development-527d (accessed 2019-02-25)

O'Brien, M and M. Gastil-Buhl 2013. Metabase Adoption by SBC and MCR. Spring 2013 LTER Databits https://lternet.edu/wp-content/uploads/2018/01/2013-spring-lter-databits.pdf (accessed 2019-02-25)

Smith, C. n.d. EML Assembly Line: A workflow and set of functions to make EML using R. https://github.com/EDIorg/EMLassemblyline (accessed 2019-03-15)

BEST PRACTICES FOR ZOTERO

Tim Whiteaker

Zotero is a free and open source reference manager with functionality similar to EndNote, Mendeley, and CiteULike. It features a desktop application, plug-ins for Web browsers and word processors, and an API to facilitate searching your publications within your Zotero library synced to the cloud. The LTER Network Communications Office (NCO) and a few LTER sites use Zotero, and a working group formed in 2018 to establish Zotero best practices in the context of an LTER site or similar project. During the February-2019 LTER IMC Virtual Water Cooler I gave an update from the working group, and as a result of much feedback from the webinar we added several items to the best practices document which you can find on EDI's External Data Management Resources page. In this DataBits article I summarize the main topics in the current best practices and highlight recent additions.

In the **Getting Started with Zotero** section, you'll find tips on keeping your usage of the 300 MB of free cloud storage in check. We suggest naming LTER Zotero groups LTER-ABC, where ABC is the three-letter acronym for the LTER site. We recommend tagging items that are suitable for sharing with the LTER Network bibliography with LTER-ABC, and items oriented toward information management with LTER-IMC.

There is a new section on **Sharing with the LTER Network Bibliography** inspired by recent efforts to create a network-wide bibliography in preparation for the 40-yr LTER review. While the NCO is currently accepting BibTeX files describing publications from individual sites for inclusion in the bibliography, Zotero users can skip the BibTeX export+cleanup step and update the LTER Network Zotero group directly. Don't forget to include the LTER-ABC tag and a DOI when available! This section also describes how Zotero makes adding the required LTER-ABC tag simple.

In the section on **Using BibTeX To Streamline Reporting to Research.gov**, you'll find instructions on how to describe your items in Zotero so that information in exported BibTeX files is properly translated by Research.gov when uploading a list of conference presentations, theses, dissertations, book sections, or books. New to this section is a Bash command crafted by Mary Martin of Hubbard Brook

LTER which removes extraneous (at least in the eyes of the Research.gov parser) curly braces within a BibTeX entry. The braces are carried through to the final presentation of the item by Research.gov which can look ugly, and so until their parser learns to remove the braces, it's up to you, or Mary's one-liner, to clean them up.

Another addition is a subsection on handling **Unpublished Journal Articles in Research.gov**. Published peer reviewed items are handled separately in annual reports to the National Science Foundation, but for unpublished items, Research.gov asks for a bibliographical entry for each item. Zotero can export bibliographical entries in many different styles, making it a simple matter of copying and pasting the entire entry for an item into Research.gov.

Thanks to free cloud storage and the Zotero API, you can include an interactive bibliography on your website using one of the tools listed in the **Supporting an Online Searchable Bibliography** section. Solutions, or at least some handy tools, exist for WordPress, static websites, and Python sites. You can view bibliographies on the LTER websites for VCR, NGA, and BLE to see some of these solutions in action.

You'll find **other goodies in the appendices** such as a workflow going from EndNote to ZotPress via Zotero and some example Python code.

The Zotero Best Practices document is still a work in progress. For example, we need help in determining how to describe an undergraduate thesis since that is not an item type recognized by Research.gov. We would also appreciate ideas on a systematic way of classifying items based on whether they were written by your project personnel about your project, written outside of your project personnel but uses data from your project, and so on. Want to share your ideas or chat about Zotero? Come join the conversation on the Zotero channel within the LTER Slack workspace. Happy Zotero-ing!

ILTER INFORMATION MANAGEMENT WORKSHOP

John Porter

A half-day Information Management Workshop entitled "*Facing the Future: Developing Science-based Information Management with New Technology*" was organized by Yiching Lin. It included presentations on the past and future of ILTER Information Management by Chau-Chin Lin and John Porter, and some examples of new and challenging data types such as voluminous audio streams resulting from soundscape research and images and their resulting processed data products from Unmanned Aerial Vehicles (UAVs) by Yu-Huang Wang and Sheng-Shan Lu.

The workshop concluded with a panel on: Developing new ILTER Programs on Science-based Information Management.

A panel consisting of Chau-Chin Lin (Taiwan), Hideaki Shibata (Japan) and Wim Hugo (South Africa) and moderated by John Porter (USA) focused on three questions:

- What have been the biggest advances in scientific information management which enable new scientific discovery?
- What areas remain challenges and what might be done to overcome them? and
- What would be the ideal program for increasing ILTER scientific discovery?

The discussion was wide ranging from the growing availability and use of cyberinfrastructure, to the need for improved tools for creating metadata, to the desirability of providing more advanced semantic linkages between datasets and standards-based data collection. Successes were noted in the areas of improved software for managing data and metadata, and increasing availability of data. Nonetheless the panel noted the challenges of properly documenting and sharing ILTER data. The need for tools to better support different languages, and provide better semantic linkages to the same underlying concepts across languages were identified as important. There is also the cultural challenge of convincing researchers that sharing their data is the best way to advance both science as a whole, and their own individual careers through increased citations and increased collaborations. There were significant audience contributions regarding the desirability of simplified interfaces for creating the metadata needed to publish data, and discussion of alternative interfaces and approaches to metadata generation. A theme throughout the discussion was the need for data scientists and information managers to collaborate directly with ILTER researchers.

WHERE ARE THEY NOW?

Don Henshaw

Karen Baker

LTER Site(s): Palmer Antarctica LTER (1990-2011), California Current Ecosystem LTER (2004-2011)

LTER key roles: Data Management Task Force (now IMExec), 1994-1997, 1998-2001; Governance Committee (Terms of Reference document); Participatory Design work; Baker et.al. 2000

Current status: Independent scholar and contractor, Chicago, Illinois

(The following is compiled by Don Henshaw but is based on direct quotes and words from an interview with Karen with some reorganization and paraphrasing to provide necessary context)

What accomplishments, challenges, and experiences (local site or network level) do you remember encountering while an LTER IM?

Karen Baker was working as an oceanographer and computer programmer when she became the first Information Manager at the Palmer (PAL) LTER site in 1990 and later at the California Current (CCE) Ecosystem site in 2004. She witnessed dramatic changes in LTER that enhanced scientific research and its Information Management (IM) community. Now after many years working with social scientists and armed with a PhD in Information Sciences, Karen is able to look back on her experience with the LTER.

The LTER site-network configuration enabled the growth and dynamics of 'infrastructuring'. Local sites developed 'collective data practices' led by an 'embedded' data manager at each site while a broader, multi-project data environment existed with the all-site IM Committee (IMC) that fostered active engagement of participants and learning through joint design activities (Baker et.al. 2000). Data sharing was a milestone that LTER achieved early on with changes in data practices. A second milestone followed increasing responsibilities for the data management (DM) role when the IMC governance developed from informal practices to the Terms of Reference (Bylaws) that provided structure and allowed more formal representation of the IMC within the LTER science committees.

Piloting of data sharing by LTER preceded agency mandates for DM plans in 2013. In the 1990s, each site designed and managed data and systems using different approaches depending upon local circumstances. These activities meant the community was poised some years later to contribute to development of the LTER Network Information System (NIS). "I thought we were fortunate that the design of the early NIS came after local site data management developed - some say site differences resulted in inefficiency but I saw how site-based activities contributed to our effectiveness in terms of providing workforce experience with 'collective data management' across a diversity of settings. Developing 'modules' for the very early NIS educated us all." IMs across the sites developed an understanding of both local and network needs, aligning site-based data systems with remote systems. Karen was amazed to be involved with the "balancing, mediating, negotiating, facilitating and tailoring of work-arounds" related to data activities associated with the "unfolding and growth of DM over time".

Specifically, Karen's team at PAL LTER developed the initial site catalog (SiteDB) with site profiles. Mason Kortz and James Connors were key architects of the PAL and CCE local information system DataZoo (Baker et al., 2011) as well as in the continuing design of the LTER Unit registry (Karasti et al., 2010). Lynn Yarmey, now part of Research Data Alliance staff, worked on metadata conventions and standards, and with Karen described a 'web of repositories' (Baker and Yarmey, 2009).

The early site and IMC work was informative, but vocabularies and ontologies to describe 'data work' were lacking. Karen's early efforts in inviting social scientists to partner with PAL, CCE, and LTER addressed the "articulation of data and design issues

that inform our understanding of the invisible labors of local data specialists". Karen's close associate Helena Karasti, a frequent attendee of IMC events circa 2000, who studied LTER data management, described the IMC as a Community of Practice and introduced notions of participatory design and co-design (Karasti and Baker 2004, 2008). Another close colleague, Florence Millerand, studied the development and enactment of EML, writing that she heard two stories: the developers' story of success given development of a standard, and the information managers' story of success-to-come given the many years of site-specific work required to implement the standard (Millerand and Bowker, 2009). Florence also wrote about when local site troubles are recognized across many sites, they can become recognized as community issues to be discussed and acted upon collectively (Millerand et al, 2013). Helena and Florence both brought insights and descriptive vocabularies from other fields while other collaborators wrote about the LTER site-network configuration as an approach that learned from earlier efforts (Aronova et al, 2010).

A collaboration with Geoffrey Bowker initiated Karen's focus on the sociotechnical dimensions of data management and information infrastructures (Bowker et al, 2010). Today she is exploring data infrastructures which include facilities, services, and dynamic interactions among all of the elements that support data work and data flow. The term 'infrastructuring' is used in social sciences to emphasize the continuing process involved in making and maintaining infrastructure.

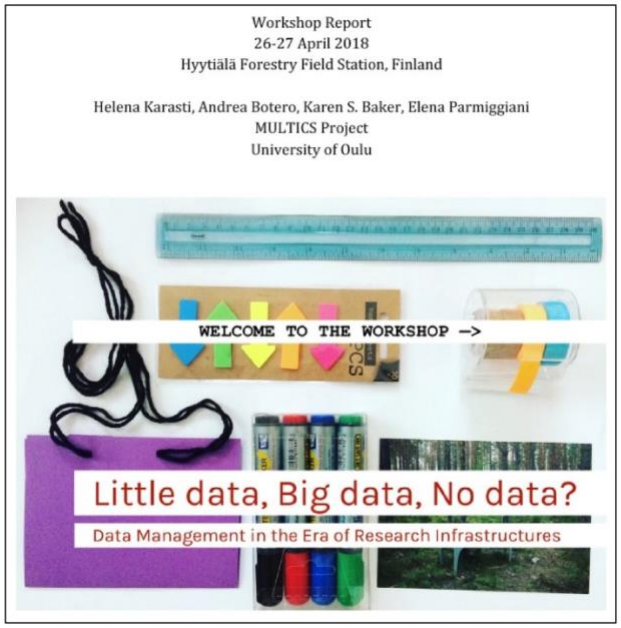
-What are you working on now or have done since your time as an IM? Have your IM skills been applied?

Karen's IM skills led to a fellowship offer for a PhD in the School of Information Sciences program at University of Illinois Urbana-Champaign (see photo insert). The offer prompted her retirement from University of California San Diego (UCSD) after more than 35 years, and provided a path to broaden her skills and experience base. This also allowed Karen to share LTER experiences within the academic realm. It was "an aspect of being part of an academic environment relating to DM that I had not imagined previously".

Karen's time at UIUC allowed her to investigate data management in other venues and to delve into the concept of 'data work'. A number of long-term sites involved in earth and environmental sciences were studied at different phases of development including Emiquon Preserve on land owned by the Nature Conservancy which drew upon the Illinois Large Rivers LTER site experience (1981-1986), the National Center for Atmospheric Research (NCAR) which provided Karen a one-year fellowship with NCAR in Boulder, Colorado, and Yellowstone National Park in partnership with biogeoscience researchers and park resource managers. In addition, at the Shortgrass Steppe (SGS) LTER Karen partnered with Nicole Kaplan, the site's information management team leader, during the SGS three-year decommissioning, an important period of disturbance and transformation as their information system was shut down, their web site disassembled, and their data migrated (Kaplan et al, 2014). At each of these sites with embedded data management, local data efforts contributed to the production of data and knowledge.

After graduation Karen opted for the freedom of choosing her work, travel, and pace. She now works virtually as an independent scholar and contractor with a variety of colleagues. As an information scientist, she continues her research with collaborators, including Helena studying infrastructure-making in multiple arenas, with Florence on ecology and data management, and with a distributed stakeholder team focusing on institutional change. The LTER data infrastructure that started as a bottom-up, multi-site approach, provides an interesting contrast with happenings in Europe where research infrastructures

are developing top-down with national, regional, and international roadmaps. A data infrastructure workshop for a multi-stakeholder consortium in Finland was recently conducted entitled "*Little Data, Big Data, No Data? Data Management in the Era of Research Infrastructures.*" (Karasti et al., 2018; see photo insert). A vocabulary is being developed to address data activities including the notion of 'data care' where "…data care in scientific workplaces refers to more than the virtue of hard work, available expertise, or smart choices in data arrangements" (Baker and Karasti, 2018). Data care is a proactive approach to data work that creates a forum for refining understandings, conducting comparative cross-checking, discerning differences in data vocabularies, and reflecting on the ramifications of data decisions. The concept of data care foregrounds the political and ethical dimensions as well as the feelings, insights, and intuitions at the heart of science and its work with data.



-What are your personal goals for the future?

Karen's training in science, then as a programmer, then as site IM team leader, and finally as information scientist partnering with social scientists, makes being a research scientist and an independent scholar a great way to continue learning and contributing. Karen plans to continue engaging with earth and environmental scientists as well as social scientists, including ethnographers, designers, artists and others. She feels fortunate to be able to expand her experience base and to balance her personal life with her work life. Her collaborations today are virtual so location is no longer critical. She moved to Chicago to be closer to family including a first grandchild.

Karen feels her career has allowed her to continue working and learning together with talented colleagues and interesting people. "With this idea of continuing to learn and to adapt in the digital era, which became evident to me while working with the LTER community, my current approach illustrates that there are opportunities to contribute to science and society in many different ways."

Citations:

Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present. *Historical studies in the natural sciences*, *40*(2), 183-224.

Baker, K. S., & Karasti, H. (2018). *Data care and its politics: Designing for a neglected thing.* Proceedings of the 15th Participatory Design Conference (PDC'18): Full Papers - Volume 1, 20-24 August 2018, Hasselt and Genk, Belgium.

Baker, K. S., & Yarmey, L. (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation*, *4*(2), 12-27.

Baker, K. S., Benson, B. J., Henshaw, D. L., Blodgett, D., Porter, J. H., & Stafford, S. G. (2000). Evolution of a multisite network information system: the LTER information management paradigm. *BioScience*, *50*(11), 963-978.

Baker, K. S., Kortz, M., & Conners, J. (2011). *DataZoo: an Oceanographic Information System Supporting Scientific Research.* Retrieved from http://escholarship.org/uc/item/139019q8

Bowker, G. C., Baker, K. S., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 97-117). Dordrecht, Netherlands: Springer.

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. http://hdl.handle.net/10217/87239

Karasti, H., & Baker, K. S. (2004). *Infrastructuring for the long-term: Ecological information management.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS), IEEE.

Karasti, H., & Baker, K. S. (2008). Digital data practices and the long term ecological research program growing global. *International Journal of Digital Curation, 3*(2), 42-58.

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term matters in collaborative development. *Computer Supported Cooperative Work (CSCW), 19*, 377-415.

Karasti, H., Botero, A., Baker, K. S., & Parmiggiani, E. (2018). *Little Data, Big Data, No Data? Data Management in the Era of Research Infrastructures. Workshop Report, May 2018. Workshop 26-27 April 2018 at Hyytiälä Forestry Field Station, Finland*. Retrieved from University of Oulu, Finland. ISBN 978-952-62-2006-2. Available at: http://hdl.handle.net/2142/100870 and http://urn.fi/urn:isbn:9789526220062.

Millerand, F., & Bowker, G. C. (2009). Metadata standards: Trajectories and enactment in the life of an ontology. In M. Lampland & S. L. Star (Eds.), *Standards and their stories. How quantifying, classifying, and formalizing practices shape everyday life* (pp. 149-165): Cornell University Press.

Millerand, F., Ribes, D., Baker, K., & Bowker, G. C. (2013). Making an issue out of a standard: Storytelling practices in a scientific community. *Science, Technology & Human Values, 38*(1), 7-43.

The "Where are they now?" series is a means of remembering and reacquainting ourselves with past LTER Information Managers. This series is intended to highlight former LTER Information Managers (IMs) by exploring what they recall of LTER meetings, events or other memorable moments during their years of involvement, and update their activities in the following and current years. The intent is to provide an opportunity for any current Information Manager or DataBits reader to contribute profiles of former IMs that they may know and wish to highlight in future DataBits editions. Susan Stafford was interviewed for the first in this series (Spring 2014 DataBits) and Nicole Kaplan provided a guest article (Spring 2018). Here, Karen Baker provides some insights into her career as the PAL and CCE LTER Information Manager and with subsequent involvement in socio-technical activities.

10 DAYS IN THE LIFE OF THE CODE GENERATION WEB SERVICE

John Porter

We don't keep comprehensive records on the use of code generation services on the EDI and LTER portals. But I did take a recent look to see how they were being used. During the 9-day period (May 27-June 4) there were 2,166 uses of the code generation services. Not surprisingly, R code constituted the largest single percentage with 210 base-R programs and 535 tidy-R programs (34% overall). However, use of the SPSS (540 programs, 25%) and SAS (505 programs, 23%) code generation services was also substantial. The Matlab service generated 207 programs (10%) and the newly added Python service accounted for 154 programs (7%).

GOOD READS: "TEN SIMPLE RULES FOR DIGITAL DATA STORAGE"

John Porter

It is always interesting to read an article drawn from another research community that examines best practices for managing data. This 2016 article, drawn from the collections community has some recommendations that may seem familiar to the ecological data community. Some, such as: "Keep raw data raw," "Store data in open formats," "Data should be uniquely identifiable," "Have a systematic backup scheme" and "Link Relevant Metadata" will seem familiar because they are (hopefully) part of daily practice. However, they also have some recommendations that are more challenging, such as: "Anticipate how your data will be used" and "Know your use case" where the wide array of data types and data uses by ecologists make it difficult to achieve exact solutions. Nonetheless, they ask some useful questions regarding looking for community standard formats and software tools in guiding how data should be stored and how changes to the data will be tracked.

There are other recommendations that are a bit less relevant to the ecological community such as "Adopt the proper privacy protocols" since most of ecological data have few privacy issues associated with them. One recommendation that might be controversial is that "Data should be structured for analysis,"

endorsing the use of normalization (specifically Codd's 3rd Normal Form), because for many ecological users doing relational joins are not a normal part of their analytical workflow. Thus although LTER sites might manage data in relational tables, data are typically published as flat files.

They finish up with the general dictum that "The location and method of data storage depend on how much data you have." Fortunately most ecological data does not demand petabyte databases (yet!).

In addition to the recommendations, the article also contains some good tables describing various file formats, programming tools and algorithms and persistent identifiers. Overall it is a useful paper to read, especially for ecological information managers starting new projects or encountering new kinds of data.

Hart, E.M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Poisot, T., Woo, K.H., Zimmerman, N.B. and Hollister, J.W., 2016. Ten simple rules for digital data storage. PLOS Computational Biology https://doi.org/10.1371/journal.pcbi.1005097

SEMANTIC ANNOTATIONS IN EML 2.2

Margaret O'Brien (EDI, SBC LTER), Steven Chong (NCEAS), Mark Schildhauer (NCEAS)

INTRODUCTION

A semantic annotation is the attachment of semantic metadata to a resource. It provides precise definitions of concepts and clarifies the relationships between concepts in a machine-readable way. The process of creating semantic annotations may seem tedious, but the payoff is enhanced data retrieval and discovery. Semantic annotations will make it easier for others to find and reuse your data.

For example, if a dataset is annotated as being about "carbon dioxide flux" and another annotated with "CO2 flux" the information system should recognize that the datasets are about equivalent concepts. In another example, if you perform a search for datasets about "litter" (as in "plant litter"), the system will be able to disambiguate the term from the many meanings of "litter" (as in garbage, the grouping of animals born at the same time, etc.). Yet another example is if you search for datasets about "carbon flux", then datasets about "carbon dioxide flux" can also be returned because "carbon dioxide flux" is considered a type of "carbon flux".

EML 2.2 will have the capacity to add semantic statements as annotations to datasets (EML Development Committee, 2019). Here, we describe those features briefly and give examples of two common annotations - at the dataset-level and attribute-level. The EML development committee is developing a Semantic Primer with greater details about creating annotations in EML. It will be released along with other EML 2.2 documentation.

SEMANTIC TRIPLES

Semantic annotations enable the creation of "triples", which are 3-part statements composed of a **subject**, a **predicate** (that can be object properties or datatype properties), and an **object**.

```
[subject] [predicate] [object]
```

These components are analogous to the parts of a sentence; the **subject** and **object** can be thought of as nouns in the sentence and the **predicate** is akin to a verb or relationship that connects the **subject** and **object**. The semantic triple expresses the statement about the associated resource. After processing the EML into a semantic web format, such as RDF/XML, the semantic statement becomes interpretable by machines.

RDF is *not* designed to be displayed to people. It is designed so that components are accessible through the Web, for applications to look up precise definitions and relationships between these resources and other concepts. To simplify their use, the three components of a semantic triple should be HTTP URIs (uniform resource identifiers), which are

- globally unique and persistent, and
- resolvable or dereferenceable

In addition to unique and non-ephemeral (URLs are considered ephemeral) the definition referenced by the URI should not change substantially, so that the resources referencing them may rely on their annotations remaining consistent and truthful.

The simplest triple statement is a sequence of (subject, predicate, object) terms, separated by whitespace and terminated by '.' (Prud'hommeaux & Carothers, 2014). Below is the semantic statement for the relationship between Spiderman and the Green Goblin, with fictional URIs:

```
<http://example.org/#spiderman> <http://example.org/#enemyOf> <http://example.org/#green-goblin> .
```

Below is a true triple created for a Jornada LTER dataset. The URIs resolve, to say that a dataset (subject) "is about" (predicate/property) a "desert area" (object/value). The subject dereferences to a dataset in PASTA, knb-lter-jrn.210327001.1, the predicate to a relationship ontology which defines the concept "is about", and the object to a concept in the Environment Ontology, which contains a complex definition and cross references for “desert area”.

```
<https://doi.org/10.6073/pasta/06db7b16fe62bcce4c43fd9ddbe43575> <http://purl.obolibrary.org/obo/IAO_0000136> <http://purl.obolibrary.org/obo/ENVO_00000097> .
```

The “is about” relationship means that more precise searches can be constructed, e.g. a computer can return this dataset alongside other datasets that are "about" a precisely defined area called a "desert" -- not just related to deserts in some unknown way, which is all that is possible with keywords.

Semantic Annotations in EML 2.2.0

In **EML 2.2.0** there are 5 places where annotation elements can appear in an EML document:

- **top-level resource** -- an `annotation` element is a child of `dataset`, `literature`, `software`, `protocol`
- **entity-level** -- an `annotation` element is a child of a dataset's entity (e.g., `dataTable`)
- **attribute** -- an `annotation` element is a child of a dataset entity's `attribute` element
- **eml/annotations** -- a container for a group of `annotation` elements, using references
- **eml/additionalMetadata** -- `annotation` elements that reference a main-body element by its id

Annotation element structure

All annotation nodes are defined as an XML type, so they have the same structure anywhere they appear in the EML record. Here is the basic structure. Specific examples are below. The annotation node holds the `predicate` and `object`; the `subject` of the semantic statement is the parent element of the annotation.

```
<annotation>
    <propertyURI label="property label here">property URI here</propertyURI>
    <valueURI label="value label here">value URI here</valueURI>
</annotation>
```

The EML annotation is used to create semantic triples. The table below shows how the triple components `subject`, `predicate`, and `object` map to EML annotations using the JRN statement above.

Triple component	EML location	Note	Example
`subject`	Parent element of the annotation (element just above it in the XML "tree")	To be a subject, an element must have an `id` attribute	https://doi.org/10.6073/pasta/06db7b16fe62bcce4c43fd9ddbe43575
`predicate`	`//annotation/propertyURI`	the "verb" in a statement	http://purl.obolibrary.org/obo/RO_0001025
`object`	`//annotation/valueURI`	"object" of the "verb"	http://purl.obolibrary.org/obo/ENVO_00000097

When are IDs required in the EML? Annotations at the dataset, entity or attribute level presume that the parent element is the *subject*; hence, if an element has an annotation child, an id is required, so that the element can become the subject of the triple. Annotations using `eml/annotations` or `eml/additionalMetadata` will have subjects defined with a `references` attribute or `describes` element, so as for other internal EML references, an `id` is required. The EML-2.2 parser checks for an `id` attribute on the parent element if an annotation is present.

Labels: It is recommended that the label field of the annotation is populated by the value from the label field (`rdfs:label`) or preferred label field (`skos:prefLabel`) from the referenced vocabulary.

EXAMPLE 1 - RESOURCE LEVEL (TOP-LEVEL) ANNOTATIONS: DATASET

The top-level resources in EML are `dataset`, `literature`, `protocol`, and `software`, and any of them can be annotated. This example is for a dataset. **A top-level annotation applies to the entire resource (dataset)**. The `annotation` element is the last element of the resource group (i.e., it appears right after `coverage`).

- The *subject* of the semantic statement is the parent element of the annotation, the dataset. It must have an `id="` attribute.
- Each annotation consists of a `propertyURI` element and `valueURI` element, which define an *object property* and the *object* (value), respectively.
- `propertyURI` and `valueURI` elements
  - must have a `label` attribute that is suitable for application interfaces.
  - should have URIs that point to terms in controlled vocabularies
- Labels should be populated from the label field (`rdfs:label`) or preferred labels field (`skos:prefLabel`) in the referenced vocabulary.

In the following dataset annotation,

- the *subject* of the semantic statement is the parent element `dataset id="dataset-01"` (which in the resulting triple would use the resolvable HTTP URI for the global ID for the dataset, that includes its DOI)
- the *object property* is "http://purl.obolibrary.org/obo/IAO_0000136"
- the *object* (value) is "http://purl.obolibrary.org/obo/ENVO_01000177"

Taken together, the semantic statement can be translated to "the dataset is about a grassland biome"

Example 1: Top-level resource annotation (dataset)

```
<dataset id="dataset-01">
    <title>Data from Cedar Creek LTER on productivity and species richness for use in a workshop titled
        "An Analysis of the Relationship between Productivity and Diversity using Experimental Results from
        the Long-Term Ecological Research Network" held at NCEAS in September 1996.</title>
    <creator id="clarence.lehman">
        <individualName>
            <salutation>Mr.</salutation>
            <givenName>Clarence</givenName>
            <surName>Lehman</surName>
        </individualName>
    </creator>
    ...
    <coverage>
        ...
    </coverage>
    <annotation>
        <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_0000136</propertyURI>
        <valueURI label="grassland biome">http://purl.obolibrary.org/obo/ENVO_01000177</valueURI>
    </annotation>
    ...
</dataset>
```

EXAMPLE 2 - ATTRIBUTE-LEVEL ANNOTATIONS: ATTRIBUTE

An attribute annotation applies to the measurement or observation in the data entity attribute, such as a column in a spreadsheet or table. It associates precise measurement semantics such as the feature or "thing" being measured, and the measurement standard or property for interpreting values for the attribute. The simplest annotation is a single reference to a complex measurement described in an ontology (as in this example).

An attribute annotation is an `annotation` element contained by an `attribute` element.

- The *subject* of the semantic statement is the parent element of the annotation, the `<attribute>`. The XML node must have an `id="`.
- Each annotation consists of a `propertyURI` element and `valueURI` element, which define an *object property* and the *object* (value), respectively.
- `propertyURI` and `valueURI` elements
    - must have a `label` attribute that is suitable for application interfaces.
    - should have URIs that point to terms in controlled vocabularies
- Labels should be populated from the label field (`rdfs:label`) or preferred labels field (`skos:prefLabel`) in the referenced vocabulary.

In the following dataset annotation,

- the *subject* of the semantic statement is the parent element `attribute id="att.4"` (which in the resulting triple would be a fragment of the URI)
- the *object property* is "http://ecoinformatics.org/oboe/oboe.1.2/oboe-core.owl#containsMeasurementsOfType"
- the *object* (value) is "http://purl.dataone.org/odo/ECSO_00001197" which resolves to the "Plant Cover Percentage" term in the ECSO Ontology (https://github.com/DataONEorg/sem-prov-ontologies/tree/master/observation).

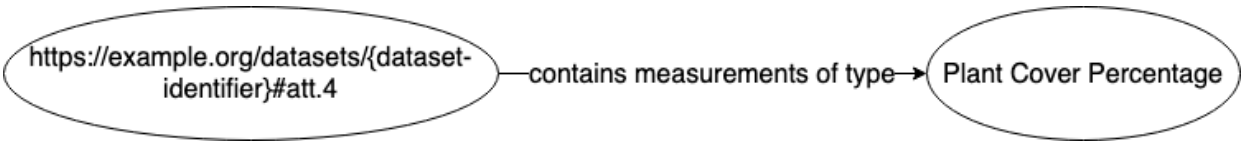
Taken together, the semantic statement indicates that "the dataset-attribute with the id 'att.4' contains measurements of type plant cover percentage".

- Example 2: attribute annotation

```
<attribute id="att.4">
    <attributeName>pctcov</attributeName>
    <attributeLabel>percent cover</attributeLabel>
    <attributeDefinition>The percent ground cover on the field</attributeDefinition>
    <annotation>
        <propertyURI label="contains measurements of type">http://ecoinformatics.org/oboe/oboe.1.2/oboe-core.owl#containsMeasurementsOfType</propertyURI>
        <valueURI label="Plant Cover Percentage">http://purl.dataone.org/odo/ECSO_00001197</valueURI>
    </annotation>
</attribute>
```

RDF GRAPHS

Below is an example of how an annotation is converted to a graph of the RDF triple. The parts of a triple (subject, predicate, and object) become nodes and links in a graph.



CONCLUSION

For complete documentation, see the EML Semantics Primer, which can be found with the soon-to-be-released EML 2.2 documentation. It contains several more examples, and other RDF material for reference. It is important to keep in mind that **Semantic statements are not simply a set of loosely structured keywords; they must be logically consistent**. Inconsistent annotations could have dreadful consequences. So the primer will also have examples of how things can go wrong.

The professional scope of our occupation will continue to grow, adding both challenges and opportunities. Communities like LTER and EDI will need to make decisions about what ontologies to adopt and priorities for applying them to their datasets, based on many aspects including (but not limited to) dataset importance, vocabularies' domain coverage, content, complexity, longevity and maintenance plans, plus their technical structure. This should be a joint venture, between data managers (with scientific input), repositories, and the designers or maintainers of ontologies and other vocabularies.

As data managers, we will need to be able to understand the concepts in the ontologies chosen by our communities, in addition to the concepts in the datasets we manage, and to create logical annotations between the two. Concurrently, repository and other search-system managers will need mechanisms to interpret the implied subjects and create the RDF triples from EML annotations when needed. Repositories will need to provide mechanisms to navigate the structure of the ontologies chosen by their communities, and provide technical guidance to their communities as they choose ontologies and other vocabularies.

References

Prud'hommeaux E. and G. Carothers. 2014. W3C Recommendation: RDF 1.1 Turtle. http://www.w3.org/TR/2014/REC-turtle-20140225 (accessed 2019-03-15)

EML Development Committee. 2019 https://github.com/NCEAS/eml (accessed 2019-03-15)

A BRIEF TOUR OF THE EDI DASHBOARD

Mark Servilla, Duane Costa, and James Brunt, Environmental Data Initiative, University of New Mexico

INTRODUCTION

As a user of the EDI Data Repository have you ever tried to upload a data package or access data through either the LTER or the EDI Data Portal and received an error message indicating that PASTA+ or another subsystem is not responding? Have you ever uploaded a data package and wanted to see if it is still being processed by PASTA+ or determine when it completed processing and was registered by PASTA+ as a published and archived data package, or for that matter, if it was successfully synchronized and indexed by DataONE? Were you ever interested in simply seeing how many new or updated data packages had been added to the EDI Data Repository over the last 24 hours, week, or month? If you responded "yes" to any of these questions, there is some help out there to answer them in the form of the "EDI Dashboard" website at https://dashboard.edirepository.org/dashboard. The EDI Dashboard was initially created as an internal tool for us to monitor the state-of-health of PASTA+ and related systems that keep the EDI Data Repository running smoothly. It has since blossomed into an EDI "Swiss Army knife" for reporting on and managing information about EDI, PASTA+, and the data packages that are under our care. The following article is a brief tour of what the EDI Dashboard provides to us, as EDI administrators, and you, as users of the EDI Data Repository.

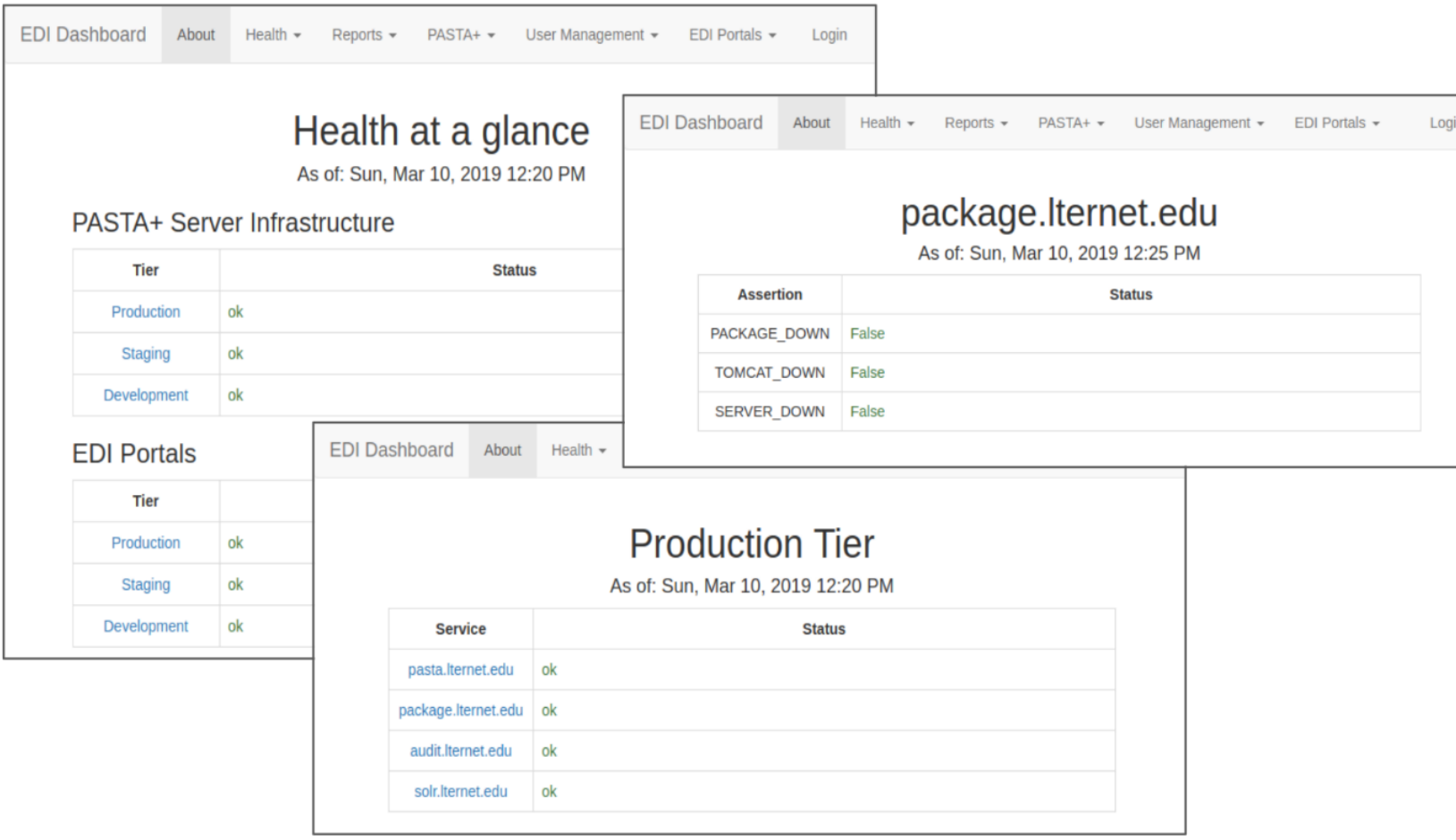
The EDI Dashboard is currently partitioned into four major sections that are accessible from the website banner: Health, Reports, PASTA, and User Management, along with the typical website "About" page and a section of convenience links to the different EDI Data Portal environments (i.e., production, staging, and development). You may also notice a user login link on the right-side of the banner section. For the most part, the EDI Dashboard is accessible to the public. There are some actions within sections, however, that do require an administrative login for privacy reasons (e.g., under User Management or Reports), but we'll still describe them just so you are aware of their purpose. Before we jump off into this tour, we do want to emphasize that this site is provided with no expectations and it is continually changing, mostly with new tools and features. Oh, and if you are the inquisitive sort, you may find an Easter egg or two sprinkled about.

HEALTH AT A GLANCE

The first major section of the EDI Dashboard is Health. In fact, you'll notice that the default display you see when you reach the EDI Dashboard home page is the "Health at a glance" page, the grand view of all critical systems under EDI management. It is divided into six subsections that cover the different areas of EDI cyber-infrastructure: PASTA+ Server Infrastructure, EDI Portals, LTER Portals, EDI GMN, LTER GMN, and Related Services. Each subsection is defined by a hierarchy that is classified by one or more of the deployment environments we manage or by a particular service. At each level, the dashboard will display the state of that environment or service by indicating whether it is "ok" (in green) or "down" (in red). Drilling down further into one of the environment links (e.g., the Related Services subsection) shows you individual services, also with "ok" or "down" status indicators.

These status indicators will let you know immediately if there is a problem with a particular service, but the real details are found yet one layer down when you select the individual service link. At this lowest layer, the EDI Dashboard breaks down the "state-of-health" into component level states that comprise the duty of the service. For example, PASTA+'s Data Package Service is composed of three hierarchical components: from the highest level component to lowest are the Ubuntu Operating System, Apache Tomcat, and the Data Package Java application (which operates under Apache Tomcat). Each of these components must be functioning correctly for the overall "state-of-health" to be "ok". At this point, the "state-of-health" check begins with an evaluation of the highest level component first, and only then proceeds to the next lower component if the higher level is healthy (i.e., it doesn't make good sense to check if Apache Tomcat is running when the Operating System is not responding). You'll notice that the "state-of-health" indicators in this view have changed from "ok" and "down" to assertions, such as "SERVER_DOWN" or "TOMCAT_DOWN", followed by either "True" or "False". For system administrators, these assertions are more

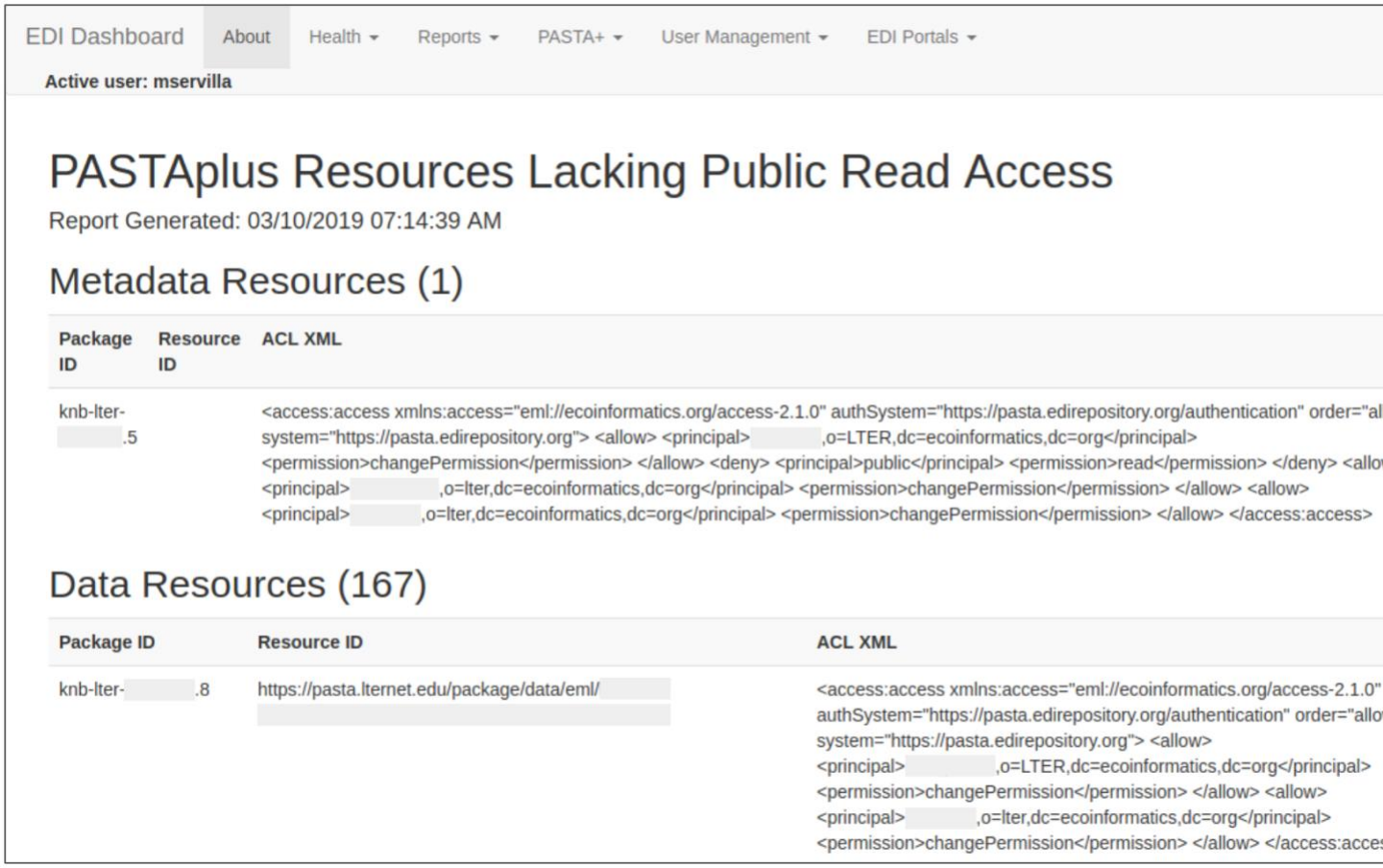
meaningful because they indicate a specific state condition that tells us where to begin looking for a problem if one exists on a particular PASTA+ or related service.

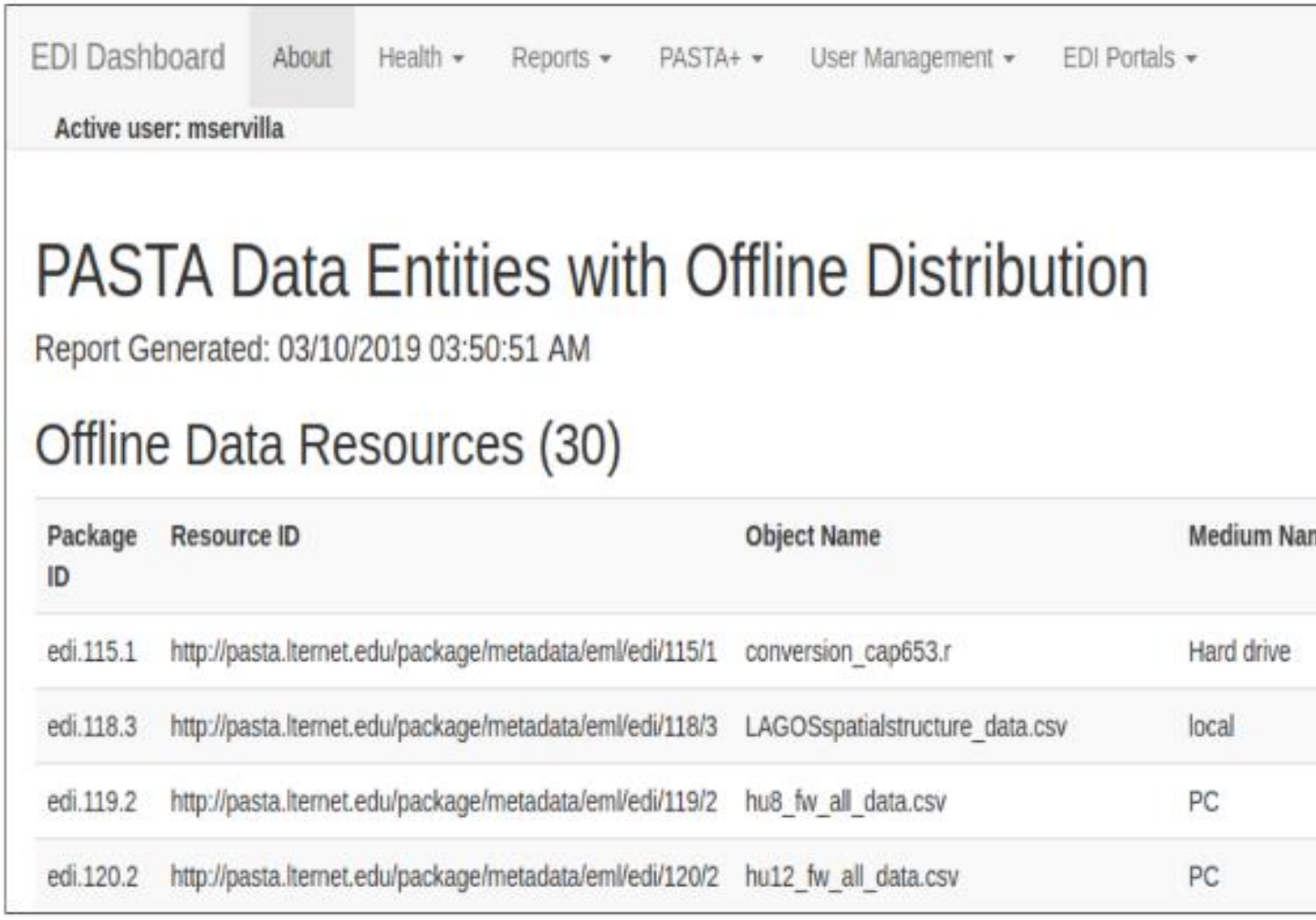


The "state-of-health" process checks on all of our critical infrastructure once every 5 minutes. If you watch any of the "state-of-health" web pages, they too update every 5 minutes so that you may see, at a glance, the health of our systems. The sub-system performing the 5 minute health check also sends an email to us whenever there is a change in status to any server we monitor. This capability complements our use of Nagios, which will eventually be phased out. Like Nagios, the "state-of-health" service and the EDI Dashboard web application run on a server in Amazon's EC2 cloud so that they may continue to function if our local infrastructure or network become compromised.

Reports

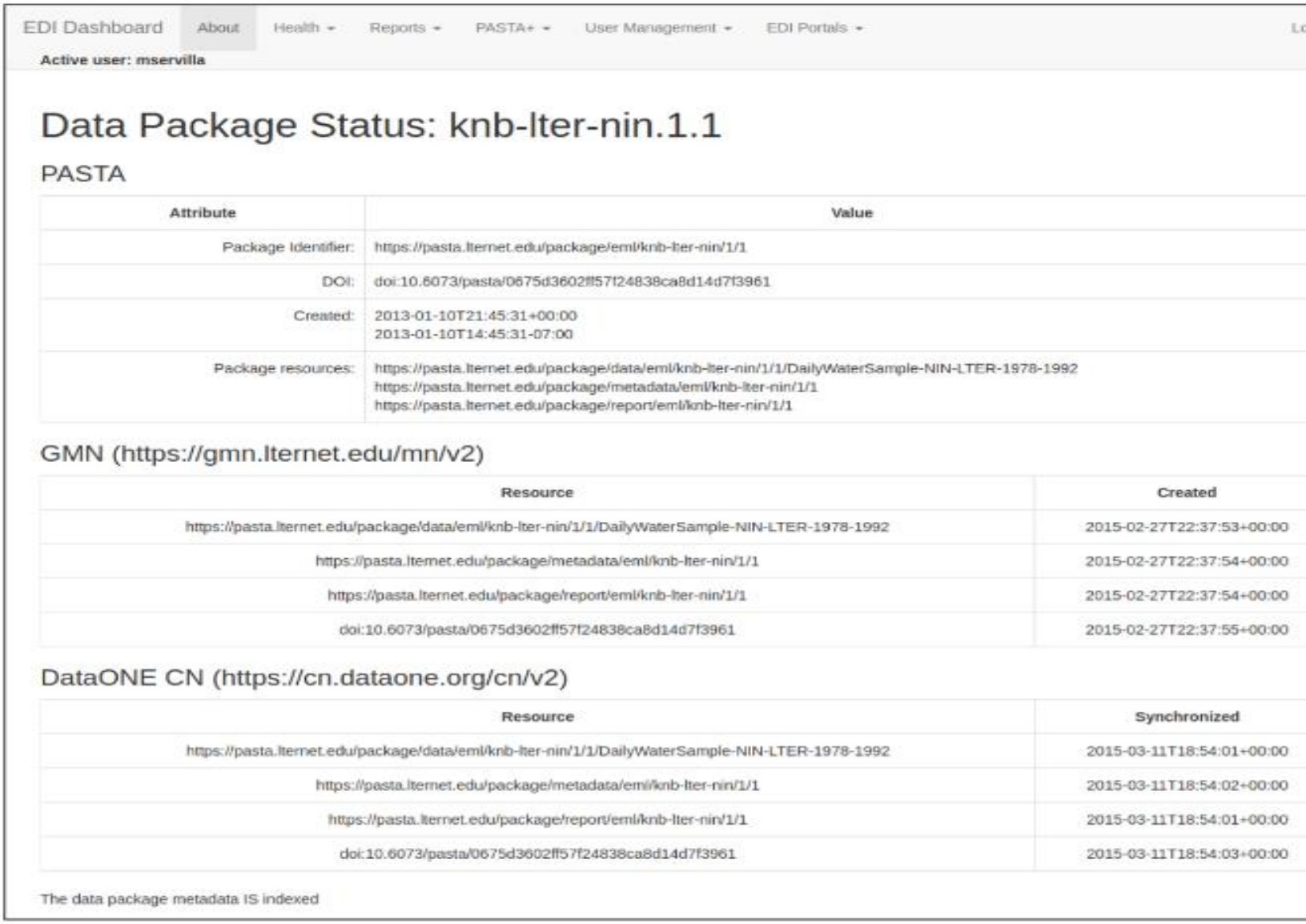
The Reports section of the EDI Dashboard is somewhat of a "catch-all" section for displaying various information about the EDI Data Repository, PASTA+, and data packages. The first two reports are accessible only by EDI administrators since they may expose what some may consider to be sensitive information. The "No Public Access" report lists data packages in the EDI Data Repository that contain access control elements that do not allow public access to one or more data entities or the entire data package itself. It is important for us to keep track of data packages being submitted without public access since we strongly believe that all data should be open and accessible unless circumstances require privacy.


EDI Dashboard
About
Health
Reports
PASTA+
User Management
EDI Portals
Active user: mservilla
PASTAplus Resources Lacking Public Read Access
Report Generated: 03/10/2019 07:14:39 AM
Metadata Resources (1)
Package ID
Resource ID
ACL XML
knb-lter- .5
<access:access xmlns:access="eml://ecoinformatics.org/access-2.1.0" authSystem="https://pasta.edirepository.org/authentication" order="al
system="https://pasta.edirepository.org"> <allow> <principal> ,o=LTER,dc=ecoinformatics,dc=org</principal>
<permission>changePermission</permission> </allow> <deny> <principal>public</principal> <permission>read</permission> </deny> <allo
<principal> ,o=lter,dc=ecoinformatics,dc=org</principal> <permission>changePermission</permission> </allow> <allow>
<principal> ,o=lter,dc=ecoinformatics,dc=org</principal> <permission>changePermission</permission> </allow> </access:access>
Data Resources (167)
Package ID
Resource ID
ACL XML
knb-lter- .8
https://pasta.lternet.edu/package/data/eml/
<access:access xmlns:access="eml://ecoinformatics.org/access-2.1.0"
authSystem="https://pasta.edirepository.org/authentication" order="allo
system="https://pasta.edirepository.org"> <allow>
<principal> ,o=LTER,dc=ecoinformatics,dc=org</principal>
<permission>changePermission</permission> </allow> <allow>
<principal> ,o=lter,dc=ecoinformatics,dc=org</principal>
<permission>changePermission</permission> </allow> </access:acces



Similarly, the "Offline Data" report shows us the data packages that are using the "offline" attribute in the EML metadata. Offline data may be used in some cases where the data are too large for online access or the data is so very sensitive that it must be protected at an offsite location. Both reports are refreshed on weekly basis. Thankfully, the number of records in either the "No Public Access" and the "Offline Data" reports is fairly low.

The next report, which is open for all to access, is the "Package Tracker". This report takes a PASTA+ package identifier in the form of "scope.identifier.revision" as input and returns state information about that data package,



EDI Dashboard | About | Health | Reports | PASTA+ | User Management | EDI Portals

Active user: mservilla

Data Package Status: knb-lter-nin.1.1

PASTA

Attribute	Value
Package Identifier:	https://pasta.lternet.edu/package/eml/knb-lter-nin/1/1
DOI:	doi:10.6073/pasta/0675d3602ff57f24838ca8d14d7f3961
Created:	2013-01-10T21:45:31+00:00 2013-01-10T14:45:31-07:00
Package resources:	https://pasta.lternet.edu/package/data/eml/knb-lter-nin/1/1/DailyWaterSample-NIN-LTER-1978-1992 https://pasta.lternet.edu/package/metadata/eml/knb-lter-nin/1/1 https://pasta.lternet.edu/package/report/eml/knb-lter-nin/1/1

GMN (https://gmn.lternet.edu/mn/v2)

Resource	Created
https://pasta.lternet.edu/package/data/eml/knb-lter-nin/1/1/DailyWaterSample-NIN-LTER-1978-1992	2015-02-27T22:37:53+00:00
https://pasta.lternet.edu/package/metadata/eml/knb-lter-nin/1/1	2015-02-27T22:37:54+00:00
https://pasta.lternet.edu/package/report/eml/knb-lter-nin/1/1	2015-02-27T22:37:54+00:00
doi:10.6073/pasta/0675d3602ff57f24838ca8d14d7f3961	2015-02-27T22:37:55+00:00

DataONE CN (https://cn.dataone.org/cn/v2)

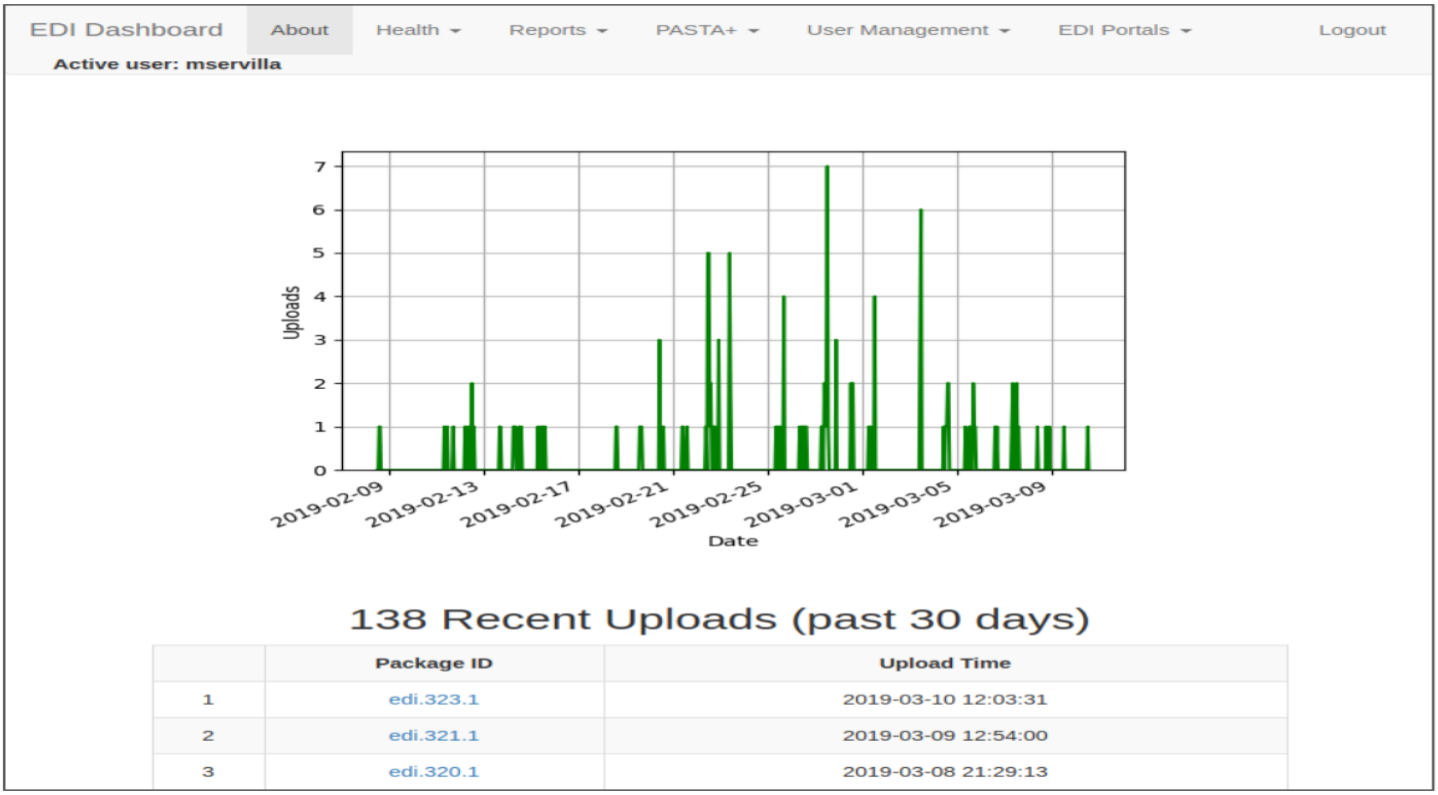
Resource	Synchronized
https://pasta.lternet.edu/package/data/eml/knb-lter-nin/1/1/DailyWaterSample-NIN-LTER-1978-1992	2015-03-11T18:54:01+00:00
https://pasta.lternet.edu/package/metadata/eml/knb-lter-nin/1/1	2015-03-11T18:54:02+00:00
https://pasta.lternet.edu/package/report/eml/knb-lter-nin/1/1	2015-03-11T18:54:01+00:00
doi:10.6073/pasta/0675d3602ff57f24838ca8d14d7f3961	2015-03-11T18:54:03+00:00

The data package metadata IS indexed

including when it was uploaded and registered in PASTA+, if and when it was uploaded to the DataONE Generic Member Node (either the LTER or EDI GMN, respectively), if and when it was synchronized to the DataONE Coordinating Node, and if it had been indexed by DataONE's Solr search engine. This report may be helpful to those users who would like to know more information about their data packages beyond an acknowledgement that it has been published into the EDI Data Repository. This particular report is both new and evolving, so the information displayed today may be very different from the information displayed tomorrow—new information may include the date and time when the data package was copied to Amazon's Glacier storage and the date and time of the last checksum verification of any disk stored resource (metadata, data, and report) of the data package. Stay tuned for updates.

The last set of reports that you may view in this section are the "Recent Uploads" reports. These reports are divided into queries for the past 24 hours, week, and month, and display a time-series plot of the upload frequency for the time

period specified, as well as a list of the recently uploaded data packages and the date and time they were uploaded. We find this report most helpful to quickly see how active the EDI Data Repository has been in the recent past.

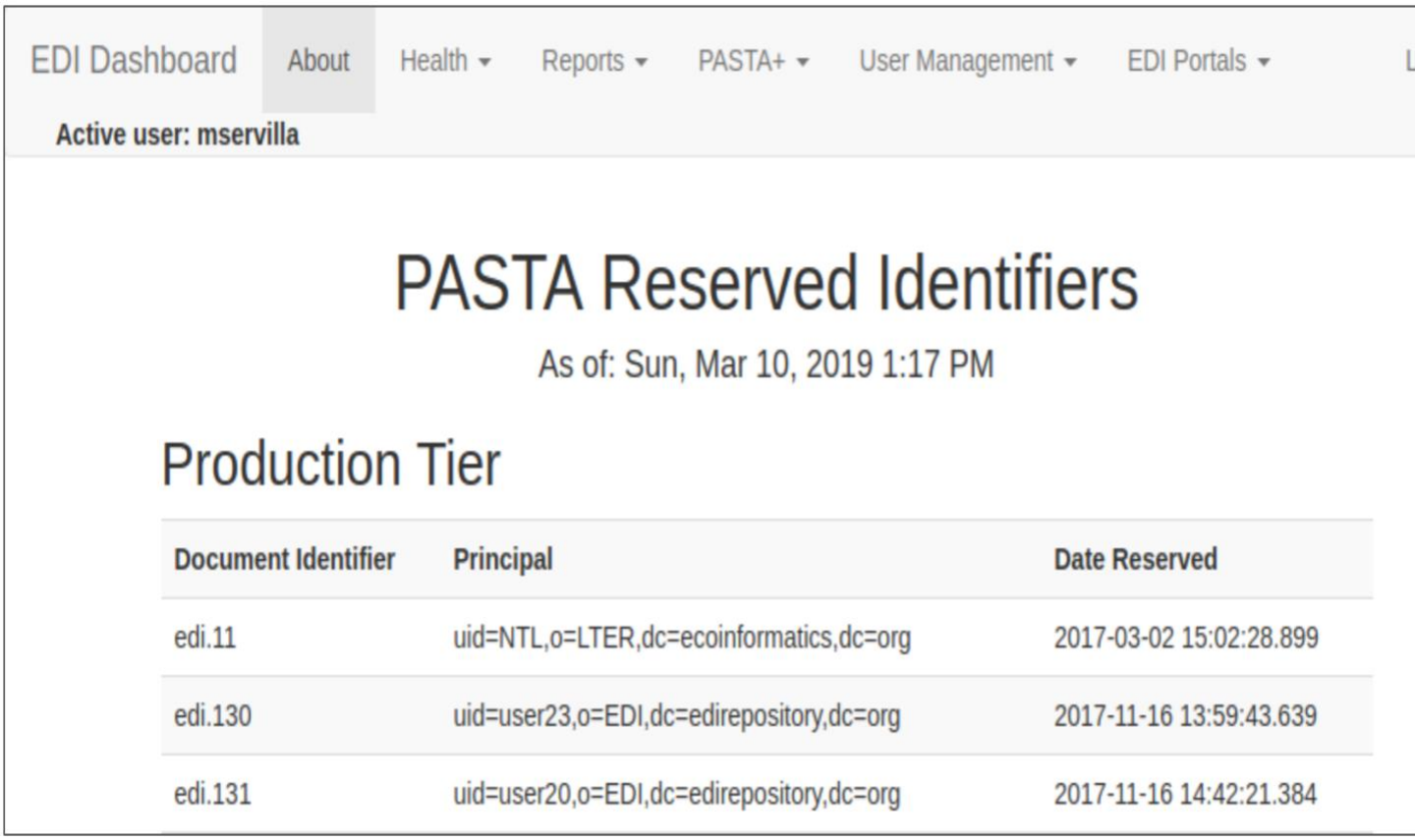


PASTA+

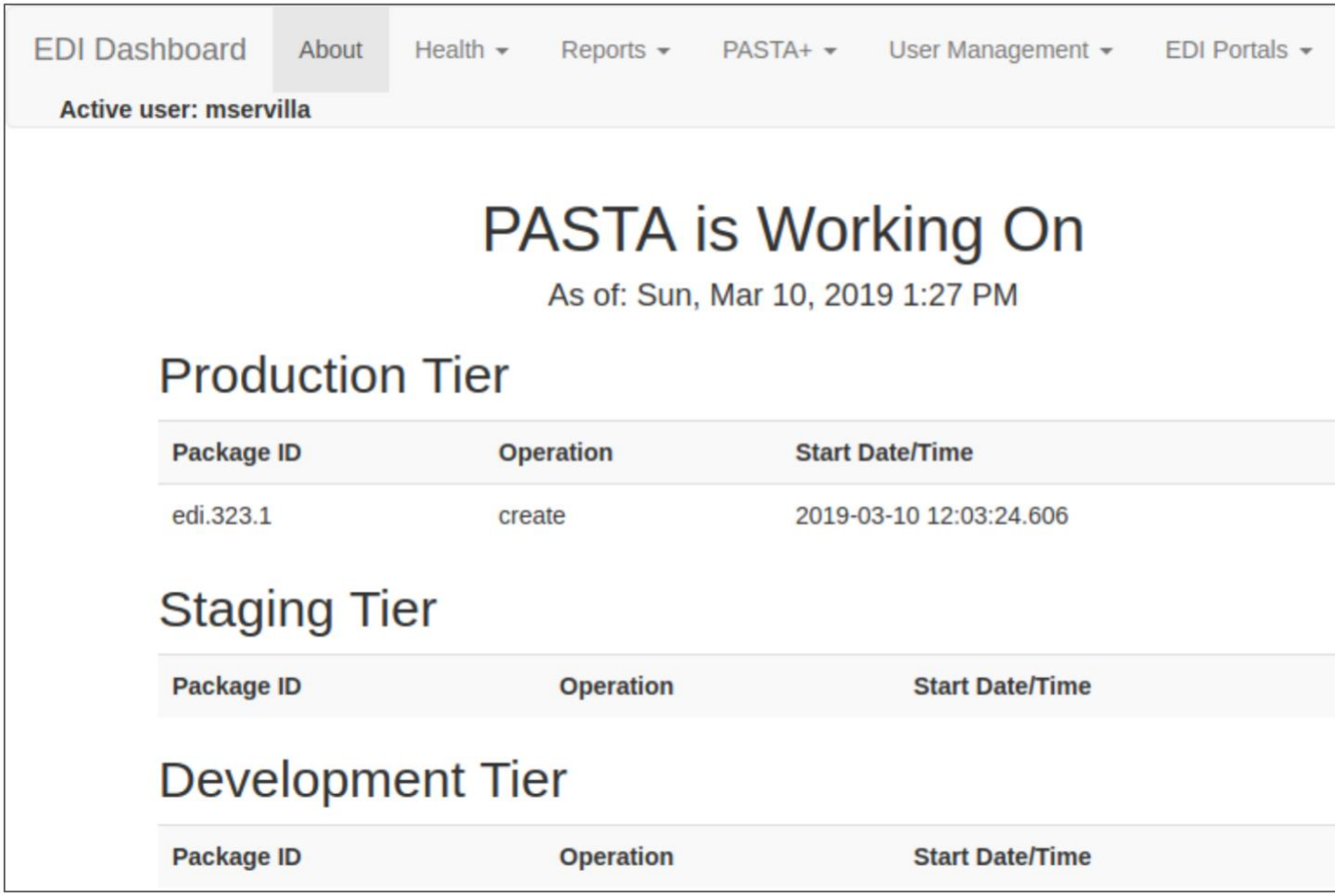
The third section of the EDI Dashboard contains two convenience functions: the first displays a list of data package identifiers (applicable only to the "edi" scope at this time) that have been reserved by an individual and the second shows any data package that is actively being processed by PASTA+ as a result of either an evaluation or upload.

Because the "edi" scope is shared across so many individuals and organizations, we found it helpful that users could set aside and reserve package identifiers for which they could use with a future data package. Unfortunately, some users would reserve a set of package identifiers, but immediately forget what identifiers they had reserved. The "Reservations" function displays a list of all reserved package identifiers that are not associated with a data package in the EDI Data Repository. The list shows the scope and identifier value of the package identifier, the full principal identity of who made the reservation, and the date and time of when the reservation occurred. This list is divided into

sections designated for the production, staging, and development environments that we support.

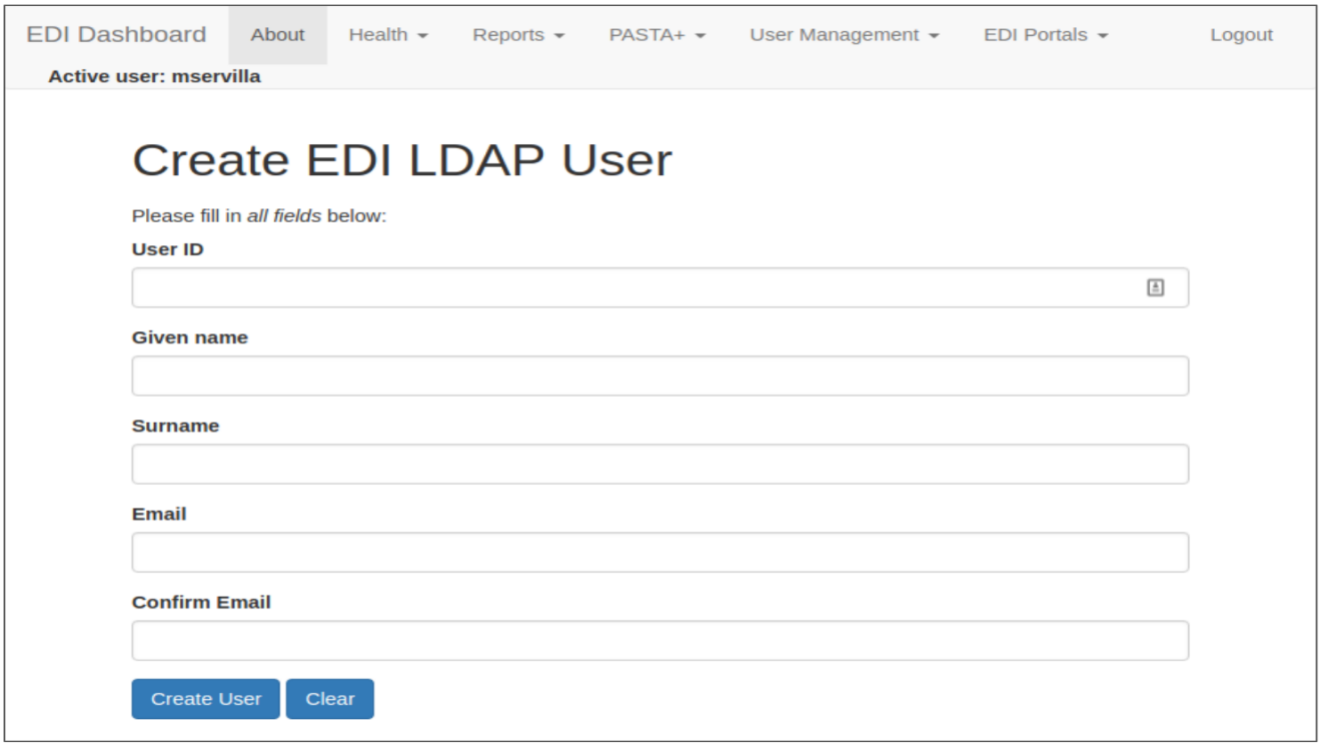


The second function is the "Working On" table, which displays data packages that are actively being processed by PASTA+ as either an evaluation or an upload (labeled as a "create" operation in the table) and the date and time processing began. This table is also divided into sections for production, staging, and development environments. Large data tables can take extra time during processing to ensure its congruence quality. For anyone who has just started an evaluation or upload process through either the LTER or EDI Data Portal or PASTA+'s REST API, the "Working On" table is invaluable to see if your data package is still in the processing state. As EDI administrators, we often consult this table before we begin our Wednesday evening system patching or if we need to deploy software to fix a critical bug. You may find that your data package lingers in the "Working On" table if it requires extra time during the quality checking phase of processing, especially if it contains many or large data tables.



User Management

The last major section of the EDI Dashboard is "User Management". At present, functions under "User Management" pertain only to users registered in the EDI LDAP user directory. Since EDI has broadened its scope to include communities outside of the LTER Network, and because we do not explicitly manage the LTER LDAP, we had to deploy an LDAP system that allowed us to register non-LTER users. To simplify the management of these users, we have developed a user management functions that gives EDI administrators the ability to create and delete users, but also allows individual users the ability to modify their account information. The first three functions under "User Management" are restricted to EDI administrators: "Create User", "Delete User", and "List Users". User account information is limited to login identifier, given name, surname, and email. Once a user account is initially created with the "Create User" function it is seeded with a random password, and a one-time password reset request is sent to the user's email address. As EDI administrators, we do not manage the user's password. The "Delete User" function does what is says, it deletes a user's account permanently, and the "List Users" function simply provides a list of user login identifiers in the form LDAP distinguished names.



Users, on the other hand, can modify their account information using the functions "Update User", "Change Password", and "Reset Password". The "Update User" function allows an authenticated user to change only their given name, surname, or email address. The "Change Password" allows a user to change a current password to a new password. Both the "Update User" and "Change Password" functions, as expected, require a current password to be successfully processed. The "Reset Password", like the "Create User", sends a one-time password reset request to the email address currently registered with the user's account.

Summary

In summary, the EDI Dashboard provides a collage made up of vignettes into EDI cyberinfrastructure that is helpful to both EDI users and administrators. The website has evolved (and continues to evolve) to accommodate new tools and services necessary to perform our jobs. We do want to set an expectation that this website should be viewed as an ongoing development that may change without notice. With that in mind, we are also eager for new ideas to incorporate into the EDI Dashboard that will help improve the overall curation of environmental and ecological data in the EDI Data Repository. Just drop us a line.