Santa Barbara Coastal Long Term Ecological Research (SBC LTER)

Information Management Overview and Plan

Li Kui
Margaret O'Brien
2021

EXECUTIVE SUMMARY

The primary responsibility of the Santa Barbara Coastal LTER Information Management System (IMS) is to handle data and information produced by our NSF grant. The primary goal of our documentation is to inform interested audiences and to assure continuity. This document (the "IM Plan") serves as the overview and timeline, and its intended audience is a reviewer, either external or the SBC IMS advisory committee. It contains brief descriptions of all IMS components. Other types of documentation include component-specific descriptions and schematics, and guides for IMS staff, assistants and scientists. Some documentation is available only to SBC members. SBC's IMS is closely integrated with the UCSB Marine Science Institute (MSI), and the information manager (author) also collaborates closely with IMS personnel from Moorea Coral Reef LTER (MCR).

Major changes to this document

This document was initiated in this format in 2009, and the text is updated as major directions or tasks changed. Appendices may be updated more frequently. A summary of major changes in 2021 below.

Section 2.2 (Integration of data processing and data publication) and 3.2 (Metadata) - incorporates the use of SBC Metabase, formalized in 2014.

Section 3.4 Removed data package levels. These were used in an older scripted method of generating EML, and supplanted by Metabase.

Section 3.11, Multiple descriptions of past projects moved to History.

`

1. INTRODUCTION

1.1. Mission Statement

The primary objectives of the Santa Barbara Coastal LTER Information Management System (IMS) are to 1) preserve high-quality and well-documented data collections that are both secure and easily accessible, 2) serve the SBC and broader community with data products and tools, and 3) participate in relevant committees and activities at both the site and LTER Network level.

The SBC IMS currently meet or exceed all standards listed in the LTER Network's Review Criteria for LTER Information Management (https://lternet.edu/wp-content/uploads/2012/01/LTER_IM_Review_Criteria_V1.1.pdf). We anticipate that IM standards of the LTER Network will continue to evolve with emerging technologies and information needs, and will maintain our leadership role in this area to ensure that the SBC IMS is well positioned to meet the future expectations for LTER IM.

1.2. IMS Requirements

The SBC IMS is primarily expected to oversee the housing and security of all SBC data, allow sharing among SBC participants, and to design, produce and catalog data products for broader dissemination. Additionally, the IMS handles contributions from the SBC to the LTER Network catalogs, some material for reports to the funding agency (NSF), and collects or develops useful data related tools to serve the SBC participants and external audiences. Nearly all SBC information is accessible via the internet, and so design and maintenance of website(s) also falls under the purview of the IMS. The SBC IMS is not responsible for hardware or system administration, as this is provided by MSI, and includes infrastructure for file, application, database and web servers, with appropriate support software, described in this document where appropriate.

1.3. IMS Documentation

SBC's IMS is documented at several levels. Some documentation is publicly available on the website (https://sbclter.msi.ucsb.edu/).

(1) A general Information Management (IM) Plan (this document). Its intended audience is a new employee, a reviewer, either external or the SBC IMS advisory committee. An IM Plan is required for all LTER sites. This document is occasionally available publicly, or by request.

(2) IM Guides, whose intended audience is the IM staff and assistants, and which assures continuity. These are updated as needed. Video recordings are available for documenting the procedures on various IM-related tasks.

(3) Schematics, descriptions, revision notes, etc., of individual system components as reference material for current and future IM staff. Component documentation is stored with the project's repository, and so included with 'check-out', or via GitHub.

1.4. Personnel

SBC's IMS is closely integrated with the UCSB Marine Science Institute (MSI.ucsb.edu) and the Moorea Coral Reef LTER (MCR.lternet.edu). SBC has a dedicated information manager located at MSI (Li Kui) with contributions from the former information manager (Margaert O'Brien),the project coordinator (Jenny Dugan), MCR information manager (M. Gastil-Buhl, until mid 2021), and MSI IT personnel (Brian Emery). We also collaborate with several other LTER sites on ad hoc projects, and the Ecoinformatics program at the National Center for Ecological Analysis and Synthesis (https://www.nceas.ucsb.edu/), also located at UCSB. Major data contributors designate research staff members to interact with the SBC information manager, and about 80% of researchers' laboratories are located at UCSB, chiefly at MSI (https://msi.ucsb.edu/), the Earth Research Institute (https://www.eri.ucsb.edu/), and Bren School for Environmental Science and Management (https://bren.ucsb.edu/). SBC also employs occasional assistants or undergraduate students for directed tasks as funding permits.

1.5. Policies

For data sharing and publication, SBC's policy is aligned with the LTER General Use Agreement and the Network's "Type I-II" designations, and the policy is included with data packages. Type I data are generally posted publicly within 1 or 2 years of collection, although some ongoing electronic data are available sooner and data requiring complex chemical analyses or data processing procedures may be delayed. All Type I data published since 2016 in SBC is under the CC-BY license. For "Type II" data, our policy is that data will be described in the public catalog, but the tables require authentication before delivery. A policy of requiring authentication may reduce availability of basic package information in federated catalogs (such as DataONE), and so currently, no SBC packages fall into this category. In addition, SBC employs a "Type 0 (zero)" designation for data we have acquired from outside parties and for which Network policies do not apply (e.g., KML files of fire perimeters in the watersheds)). These data may be republished per guidelines from the original producer. SBC also has posted a website Privacy Policy which is aligned with those of the University of California (SBC's local internet provider and host institution) and the University of New Mexico and LTER Network (owners of the DNS registration). All policies are available on the website.

The SBC IMS's primary responsibility is to handle data produced by our NSF grants. As with all LTER projects, SBC leverages and/or collaborates with other research conducted in the Santa Barbara area, and in some cases, these associated projects also leverage the SBC data systems (primarily the file server). The LTER Network is considering a policy under which collaborative projects that make use of an LTER site's IMS resources agree to publish their data with LTER data. In 2011, NSF proposals were required to include a data management plan, and SBC began assisting collaborators with this process. If these collaborators decide to make use of the SBC IMS, their data would be covered by this policy. A list of collaborative projects and their relationship with the SBC IMS is in Appendix IV.

1.6. IT Systems

SBC's holdings are stored in a networked directory system (LDAP) maintained by MSI, and a user account is required to view any data file. Write-access is limited to those responsible for data collection and maintenance. The directories for incoming or “working” data are maintained separately from those for "final" data products that are intended to be shared between disciplines or to be published. With this system, data are available to all SBC members immediately, as well as for processing or publication. SBC account holders are encouraged to use their MSI home directories for work-in-progress to take advantage of regular backups, but this is not required. Because the SBC data processing & publication workflow developed since 2018 has significantly streamlined the data publication process efficiently and accurately, we have limited the number of new users (especially students) and encourage them to access all data from the data catalog on the website.

Our IT components include issue tracking, file system backup, and supported hardware for data processing, all supplied by the Marine Science Institute (MSI). See Appendix I for information about backups and the software stack.

1.7. Training

The information manager (Kui) provides training lectures/workshops to the PIs, graduate students, and lab & field technicians on various subjects:

- SBC data catalog and methodology on data download (several programming languages)
- Data resources in the region and the tools to use (e.g. sea surface temperature, high frequency radar)
- SBC Server structure and access
- Data publication procedure and template (A YouTube video has been recorded and linked to the website)
- Data error checking procedures (A series recordings can be accessed internally)

1.8. Definitions

Data package: data entity (or entities) and metadata. Data entities are most often tables.
Data package update: The addition of new data to an ongoing data package.
Data package maintenance: Enhanced metadata or data to improve presentation or usability, or to keep the package current with standards or best practices.
Data package collection: a group of data packages that is from the same subproject with the same spatial/temporal coverages and sampling frequency.

2. DATA

2.1 Data types

Data come from diverse scientific endeavors in a variety of habitats including terrestrial/riparian, streams, beach, reef and ocean. Examples of measurements are in Table 1, and a complete inventory of current public data can be viewed on Environmental Data Initiative (EDI): https://portal.edirepository.org/nis/identifierbrowse?scope=knb-lter-sbc). As of this writing, 224 packages are publicly available and described in Ecological Metadata Language (EML), the metadata exchange format used by the LTER Network. Public data holdings comprise a variety of data formats, such as data tables (csv or txt), KML files, and images. A data package's status (ongoing or completed time series) may change over time, primarily due to the shifts in research focus. Any completed time series data package that was once on-going typically has had its last package update only to the EML metadata to specify the package end date and whether a subset of data would be continued in any other data package. All data packages are maintained, i.e., kept up to date with current standards and practices.

Table 1 Data types and measurements managed by SBC's information management system

Discipline	Representative measurements
Hydrology and meteorology	Stream discharge, precipitation (terminated in 2019 due to shifts in research focus)
Oceanography	Moored and profiled hydrography (CTD), currents (ADCP), pH, Oxygen, optics, swell
Biogeochemistry	Major nutrients, cations, particulate carbon and nitrogen, and pigments
Populations and community structure	Algae and animal survey data in fixed transects and experimental plots
Ecosystem processes	Rates of elemental flux, primary production (various methods), stable isotopes
Genomics	Organizational Taxonomic Units (OTU), microsatellite markers
Remote sensing	Kelp canopy biomass from Landsat
Model	Regional Oceanic Modeling System (ROMS)

2.2. Integration of data processing with data publication

SBC has deliberately not chosen a system where datasets are "submitted" by scientists and published by the IM staff. Instead, data is co-managed by the information manager and the data owners (i.e., investigators and their research staff), and wherever possible we integrate data publication with data processing. Integration is essential for datasets that are designated as "ongoing" because :

1. scientific personnel are the source for knowledge about data collection (instruments or observational data) and changes to sampling protocols;
2. IM has an understanding of data output formats and metadata principles and repository

requirements to ensure usability by the public or synthesis researchers
3. IM staff is knowledgeable about different data errors that might be encountered in various data types, which optimizes the data quality.

SBC IM has developed workflows and scripts for most of the ongoing datasets, from data collection to final data publication. The lab technicians or field surveyors are trained to run the data quality checking scripts and fix the errors as data is first entered. This management style is complicated, and requires coordination among diverse scientific domains, measurement types, laboratories, and software. The IMS has developed several data processing patterns and accommodates software choices (e.g. SAS, Matlab, R, and MS-Excel), depending on the scientific fields (ecologists tend to use R while oceanographers prefer Matlab). Scientific personnel are trained as necessary in informatics concepts (e.g., data table design, SI units), programming practices, and use of the shared file server and data staging series (from "working" to "final"). Structured metadata in the SBC-Metabase (see more details in a later section) and XML export is handled by the information manager and trained assistants. Coordination and training are the responsibility of the information manager. Figure 1 shows the general pattern of data collection to data package distribution, and the IMS is involved as soon as instrument data have been downloaded or observational data have been entered and uploaded to the server.

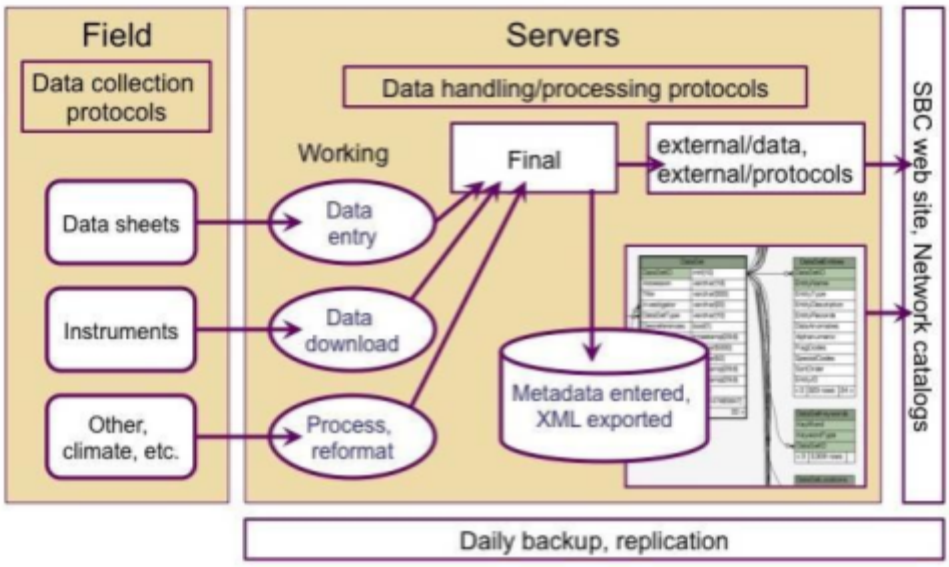

Figure SEQ Figure * ARABIC 1. SBC LTER IMS. Integration between data related activities and file system and eventual data distribution in Network and catalogs.

Two data pathways

There are currently two major pathways for data publication/update: (1) "series" datasets, where data are processed (QA/QC) and go directly into the publication pipeline.This pathway is solely handled by IM staff and periodically reviewed by scientists; (2) "unique" datasets, where scientific staff members produce the data entities and enter metadata, which are then reviewed by the information manager at the last step of data curation and publication.

The “series” pathway, in which IM manages the data processing and publication, is designed for most ongoing datasets where the data format rarely changes and the metadata are stable. Approximately 12% of SBC packages (by number) are ongoing series datasets. Data is entered or downloaded as appropriate in the laboratory by the lab assistants, and then processed by the information manager (see IM projects section for details on what IM team has accomplished on data processing). The first version of the data package, including data table and format are co-designed by the corresponding PIs and the information manager. The integration of the data processing and publication ensures high data quality, and improves the timeliness on core dataset updates. The primary data package collections using this pathway include: kelp forest annual monitoring, long-term kelp removal, annual lobster survey, bottom temperature, light intensity, as well as beach wrack, bird, and invertebrate surveys. R scripts are used at the QA/QC step, and SAS is used for generating some of the derived datasets. A similar sequence is applied to data from moored oceanographic instruments, in which Matlab is used for data checking and processing. Standardizing our processes has the advantage of furnishing a product that can also be used as input for value-added products, e.g., the contributions to the Environmental Data Initiative (EDI) “EcocomDP” project. The ongoing datasets are planned to be updated once a year with more frequent updates as requested by the researchers or data become available. The IMS team and survey team are working together to enhance dataset checking codes to ensure high data quality.

The second pathway (“unique”) is the primary pathway (by data package number) and it is appropriate for data packages that are designed for one-time publication or have limited update frequency. Typically, data are collected by individual scientists. The majority of SBC’s packages are maintained using this pathway, except a few “ongoing time-series” that have historically been managed by individual labs (e.g. Kelp canopy area and biomass from Landsat). This pathway is applied to most datasets because it allows flexibility on data processing scripts, data table format, and project design. In this pathway, laboratory personnel propose and assemble the metadata in a template (https://sbclter.msi.ucsb.edu/data/help/ under dataset publication section) , and the information manager reviews the proposed content, and adds it to SBC-Metabase. Prior to data publication, scientists have been encouraged to communicate with IM in different stages of the project:

1. At the proposal stage, the IM helps to draft or review the IM component of the new project, especially the ones that will adopt the SBC IM system.
2. During the initial experimental setup, the IM locates similar experiments and protocols used in the SBC and discusses the possibility of incorporating or combining the survey efforts and provides insight on optimal frequency of data collection to allow compatibility with existing instruments.
3. In the data QA/QC stage, IM shares guidance on data quality checks, and provide checks on consistency on column format(e.g., datetime), NA values, and range of values
4. In the data publication stage, IM provides information on data formatting and metadata documentation.

All data packages (the first versions of the “series” and all “unique”) are presented to the scientists or their staff using a “staging portal” in EDI that displays metadata using the same views as are used by the production catalog. This enables scientists to better understand metadata components and approve the design before the package becomes final.

Most of SBC published data tables are in ASCII standard. As more studies have been carried out in recent years, many non-tabular data are added into our publication pipeline, such as still images, videos, spatial data, codes and models. As of 2021, an LTER working group has drafted the best practices for these types of non-tabular data. SBC has tested the best practices when publishing data with still images and spatial data (in GeoTIFF format). More non-tabular data will be published following the best practices once it is finalized at the network level.

3. STATUS AND MANAGEMENT OF IMS COMPONENTS

SBC has developed the IMS centered around SBC-Metabase, which is an SQL database used to store information for data packages, data catalog, research activities, bibliography, personnel, the website and various other uses. Figure 2 shows the current (as of 2021) status of the Metabase with its major components. Metabase was first introduced into SBC from another LTER site (Georgia Coastal Ecosystems) and it was a joint effort between SBC and MCR (see the history of Metabase in a later section). Since 2018, the Metabase has been further integrated and improved to 1) refine the essential SQL tables designated for the data package productions (EML content), and 2) add/modify various schemas to accommodate website content that require constant updates and tracking. In addition, R scripts play crucial roles in extracting the information from the Metabase and converting to various tables or formats for downstream uses. The R scripts files reside in the Github repository as version control and SBC-Metabase follows the backup procedure handled by MSI.

3.1. Website

Essentially all SBC information is available via the Internet. The public website (http://sbc.lternet.edu) is organized around broad subject areas (Table 2) and complies with LTER Network standards for content, menus and links (LTER, 2009). The SBC website is a hybrid of static HTML pages and browser-side scripting. The site was redesigned in 2019 to modernize content, upgrade scripts, and make better use of central resources like EDI and metabase. The primary web address is an MSI-hosted URL (sbclter.msi.ucsb.edu), and that URL is registered with the Network Office to receive redirects from sbc.lternet.edu. The website framework is Bootstrap 4; custom Javascript was written by undergraduates from the UCSB Computer Science Department, under the direction of the site's data management team. Exports from the SBC LTER database were written by the data manager.

The modular nature of our website is consistent with general recommendations for web design, and also has advantages for SBC's co-management style; page content can be easily edited (or new pages created) by someone with minimal training in web authorship. To facilitate multiple authors, all content is maintained in a version control system (Github, https://github.com/sbclter/sbclter-website-2019/). Responsibility for website coordination and integrity lies with the information manager. SBC also maintains two secure areas for SBC members after login: a) an internal website (https://sbc.lternet.edu) for non-public HTML-based material and b) HTTP-accessible data directories

(https://sbc.lternet.edu/data). Details of the websites' design and implementation and documentation for individual applications are available in the GitHub repository . Planned updates for the website are listed in the timeline (Appendix II).

Table 2. Major sections, content and implementation strategies for the SBC website (http://sbc.lternet.edu)

Section Title	Content	Implementation
Research -Sampling Sites	Descriptions of core research activities Descriptions and maps of sampling sites	Static HTML, with content supplied by Investigators. Sampling site map uses Google Maps API, with content exported from Metabase
About SBC -People	contact info, Personnel diretory, individual profile pages.	Personnel pages are displayed with Javascript, using content exported from Metabase.
Publications	Bibliography, including presentations	Javascript display with content pulled from CrossRef, using DOI exported from Metabase
Data -Information management	SBC's data catalog, plus links to other catalogs of interest. Information plan, guide, system and project documentation	Data catalog front end uses custom categories to gather data packages into collections. Individual package displays pull metadata (EML) from EDI. Dataset search uses code supplied by BLE LTER. All code is javascript.
Education	Descriptions of K12 and higher education programs	Static content maintained by the project coordinator
Community	News and events, media links and other announcements	Static content maintained by the information manager
Internal Areas	Access to the file server, various help pages and site map	HTTPS, strong encryption. Certificate provided by MSI.

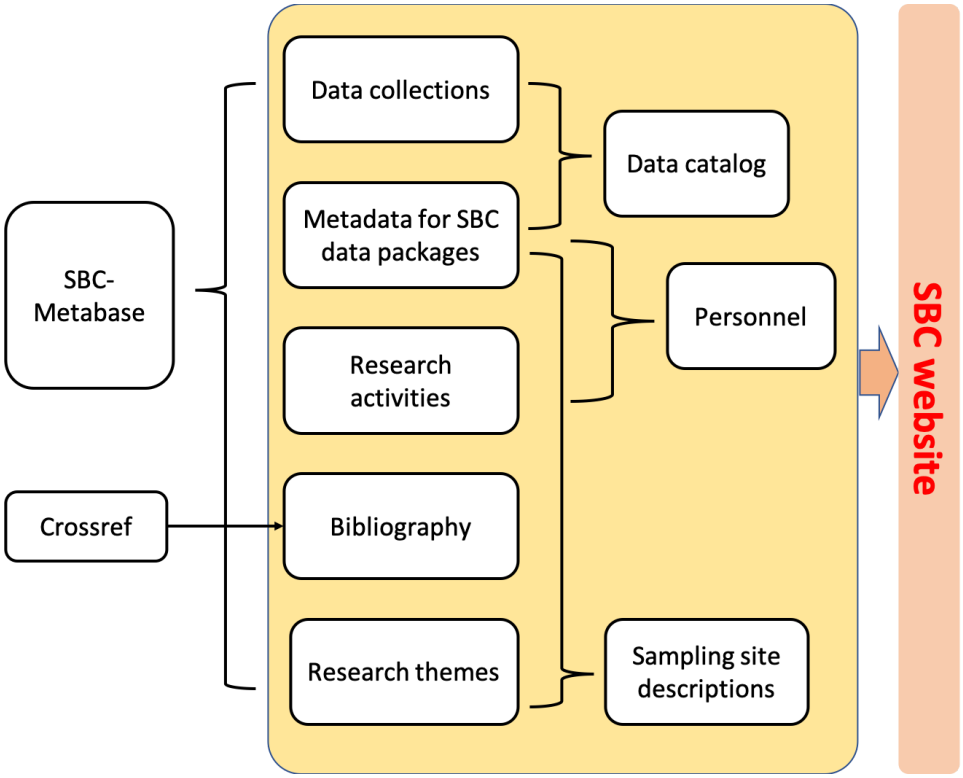

Figure 2, metabase structure and linkage with the website

3.2. Metadata and Metabase

Metabase was adopted by SBC in 2012 as a joint effort between SBC and MCR ("Metabase2", see below). In 2017, Metabase2 was streamlined to a smaller set of tables and schemas that better represented SBC's uses. All metadata for the published datasets is stored and managed in the lter_metabase schema. The metadata contents specifically for EML data package publication are held in a separate schema (mb2eml_r), which is an abstraction layer for our EML generation workflow. R scripts read the abstraction layer and compile content into EML. Some EML-required elements are stored in individual data package directories on the fileserver, together with the data tables (entities), e.g., abstract and method text in MSWord documents. Documentation of metadata SQL tables and the EML generation workflow can be found in the Databits article (Kui and O'Brien, 2018)

3.3. Data Packages

Data and metadata together comprise a "data package." Metadata documentation is published in the XML specification EML (exported from SBC-Metabase) and the corresponding data entities together are published in EDI. EDI is housing almost all SBC data packages, with few packages published in Dryad (suggested by journal publisher) or NCBI (designated genetic related repository). The data packages published outside of EDI should have a companion data package published in EDI. The companion data package includes programming scripts, ancillary environmental data, derived data products, or a link to the external repository.

In most cases, ASCII tables are used for data delivery and archive, as these have proven to be the most flexible and efficient for long-term use of heterogeneous data. Other formats we regularly handle are KML (e.g., for reference datasets like perimeters of recent fires), and native binary formats such as GIS and GeoTIFF for spatial analysis (e.g., Kelp canopy chlorophyll to carbon ratio), and NetCDF for gridded kelp biomass from Landsat.

Our goal as a data provider is to publish data packages in a state appropriate for their intended use, and we recognize that the quality, completeness and complexity of both data and metadata vary. It has been our experience that high-quality data and metadata are required if data are to be confidently used in sophisticated applications. Simply posting data tables is not enough; data providers must consider additional features. SBC has developed definitions for data packages that describe their metadata and data content and quality (Table 3). Substantially more effort is required to produce the highest level ("Integration"), from which the data can be accessed by an application using only the EML metadata (e.g., a relational database).

While adopting Metabase, SBC developed robust, structured and well-controlled vocabularies for our data packages, which are focused on research themes, taxonomy, measurements (units), and sampling locations. The controlled vocabularies are used throughout the EML such as, keywords, observations, and measurements. Where possible, vocabularies are related to existing standardized vocabularies (e.g. Global Change Master Directory, NBII Biocomplexity, Knowledge Network for Biocomplexity, and the LTER Controlled Vocabulary) so that their use as EML keywords can include the vocabulary name. In 2019, we began adding some semantic annotations to datasets, and will coordinate their use with other LTER sites (see below, "Standardized Measurement Descriptions" section).

3.4. Data Package Management

Data packages are managed using "inventory" types in the Metabase. In 2019, an additional schema was added to Metabase to categorize the data packages into collections, for the SBC website's data catalog. Pre-defined groups are discussed and reviewed between the information manager and PIs, with cross-reference tables to assign data packages to collections (a many:many relationship). The integration of the website data catalog with Metabase simplifies data package management as data publication and the website are updated in one place.

The data collections on the website are accessible via web forms keyed to local habitats, measurement types and LTER core research topics. The catalog also contains a search function based on EDI's PASTA API, using code from another LTER site, the Beaufort Lagoon Ecosystem.

3.5 Site Descriptions

Site descriptions appear in both datasets and on the website (in a Google Maps Application).
The geographic information for all deployed instruments (CTD, current, pH, Temperature, Oxygen) s is compiled for periodic review and reference by scientific staff, or to support sampling efforts (e.g boat scheduling).
Similar for dataset collections, SBC sampling sites are grouped into different research themes (land,

beach, ocean, and reef) and displayed using Google maps under the "Research" section. In Metabase, a cross-reference table links the research themes and ongoing time-series datasets;R scripts reads the geospatial coverages (lat and lon coordinates) of the assigned datasets and renders them as YAML files for Google Maps.

3.6. People

SBC personnel information is stored in the SBC-Metabase through cross-reference tables between two schemas. The "Research" schema documents the most up-to-date participants and their current research domains, whereas the "lter_metabase" schema contains personnel's affiliations. Personnel information in SBC-Metabase is integrated into several aspects of site operations: email lists managed by MSI, content for the SBC website, and data package publication. Additionally, personnel information in the LTER network database is updated manually by the information manager annually by comparing the network lists and the Metabase. The network office is considering a web service to allow synchronization from site personnel databases which would alleviate this manual work.

3.7. Bibliography

A research group's bibliography is an important publicly available resource. SBC adds 25-75 new citations annually which also are managed in Metabase. We designed a "bibliography" schema with several purposed in mind: 1) record publications in the SBC; 2) provide annual reports for the funding agencies (e.g., NSF's "report.gov"); and 3) provide a publication list to the LTER Network (e.g. 40th year review). In 2005, SBC moved its bibliography from a static text list to descriptions written in EML and housed in Metacat (O'Brien, 2006). In the process we contributed significantly to the development of both EML and Metacat. While we have found the EML specification to be well-suited to a bibliography, we converted to a simpler system of citations in 2019.

In 2019, we developed a new system for the bibliography by using bibtext fields as table columns that allowed citation formatting downstream. All citations are managed in SBC-Metabase, under the "biblio" schema. Typical workflow to update the bibliography on the website includes: 1. The project coordinator Jenny Dugan collects the publication from all SBC personnel and the publication list is sent to the information manager for review. The reviewing process includes removing the duplicates and converting documents to an archivable PDF format for NSF. 2. IM compiles the DOI and ISBN list, and generates/exports corresponding citations using the Crossref API and R scripts. 3. After validating the output with the original publication list, the records are imported into Metabase. 4. The content is converted to BibTex (for the network office) or a YAML file (for website display) using R scripts. These improvements enabled reporting to NSF to be managed by the IMS, instead of ad hoc by the project coordinator. Citation storage in Metabase is minimal (DOI, ISBN) annual contributions to the network office publication database are streamlined, and citation formats for the web catalog are flexible. Metabase also allows cross-linkages between the personnel and publications, and future plans include cross-links between datasets and publications.

The general criteria for determining the SBC publications have been discussed among the PIs, project coordinator, information manager, and the network office. The criteria are expected to change as the in-depth discussion occurred among LTER IMs.

- Supported by LTER funding, including salary, GSR, support of a faculty member's students, use of LTER boats, staff, supplies, apparatus or equipment, support for travel, conference/workshop fees, publication fees
- Uses LTER datasets in analyses or synthesis
- Uses data or analyses resulting from LTER activities or working groups

3.8. Quality control and Protocols

As data uses one of two pathways for processing and publication, the data quality control also has two pathways. Primarily, quality control is based in the researchers' laboratories. For the ongoing time series datasets, quality control is done by the information manager staff after the standards and checks are established. All data packages include text methods, and the long-term datasets usually have additional protocol documents in PDF format. Data collection and processing protocols are easily accessible on the file server to SBC personnel, and a metadata system for a protocol bibliography available in Metabase. To date, we have outlined current practices and recommendations for managing protocol documents in data packages (https://sbclter.msi.ucsb.edu/data/management/research_protocols/). The information manager works closely with analysis personnel to document quality control in metadata as appropriate for individual data packages. All the quality control at the data entity level has been documented in the programming scripts (R and Matlab) with video recording as a demonstration (available through internal access). Quality control of SBC's community survey data was first documented (O'Brien and Harrer, 2008).

3.9. Standardized Measurement Descriptions

One strategy for improving access to data for synthesis at a Network level, is for LTER sites to have described their measurements in such a way that these can be compared using automated tools, and/or registered their measurements with network data synthesis research projects. The most straightforward strategy is to first standardize measurements at the site level with complete descriptions (including methods, precision and error). SBC is working toward reduced heterogeneity within its own data packages by gradually consolidating measurement descriptions in Metabase, (typically as datasets are updated). However, some real (although subtle) differences remain. An alternative is to link the measurements using semantic annotation that describes the measurement in a way that is widely accepted by scientists from the same research area. As of 2021, we have five datasets that have the semantic annotations describing standard measurement using the "Ecosystem Ontology". We plan to have more measurements and datasets linked to the standard description after coordinating with other LTER sites. We will work with the LTER IM community as this practice becomes more widely adopted. SBC is uniquely positioned to assume a leadership role due to our involvement with ontology development with the Extensible Observational Ontology (OBOE) in the related projects.

4. SBC Informatics projects

4.1. OBOE Ontology

Our work with the LTER Controlled Vocabulary is also related to O'Brien's efforts with ontology development with the Extensible Observational Ontology (OBOE) in the Semtools project (DBI-0743429, Leinfelder et al. 2011). This work also has the capacity to inform similar ontology development at the Network level, for example, in data discovery or the description of standardized measurements, and will also facilitate interoperability with systems beyond the LTER Network, such as the Biological and Chemical Oceanography Data Management Office (BCO-DMO), and the Consortium of Universities for the Advancement of Hydrologic Science, Inc. Hydrologic Information System (CUAHSI HIS). SBC plans to examine the usability of the OBOE ontology (and others) for standardizing SBC measurements during the second half of SBC IV.

4.2 Non-tabular data

In 2019, LTER information managers had discussed the lack of best practices for various types of non-tabular data (e.g. satellite data, model output, image analysis, and genetic data). The "non-tabular data" working group was formed and led by EDI. SBC have provided useful case studies to include in the best practice. In particular, Li Kui has been involved in the genetic data and still image subgroups and helped to draft the best practices. This working group aims for providing guidelines and practices for standardizing and publishing the non-tabular data to the LTER IM community and a broader scientific community. The work is ongoing, and the final products will be a website compiled various types of non-tabular data best practice and the steps to archive them into the EDI repository.

SBC has already published some non-tabular data packages (drone images, model output, still image analysis, and genetic data). We plan to publish more non-tabular data once the best practices are finalized at the network level.

4.3 LTER Core-Metabase and MetaEgress

Metabase was adopted by SBC in 2012 as a joint effort between SBC and MCR ("Metabase2"). In 2017, Metabase2 was streamlined to a smaller set of tables and schemas that better represented SBC's uses. During that conversion, export scripts were created in R; that workflow was introduced to the LTER network in an EDI webinar: https://www.youtube.com/watch?v=k4TnnCQVod0. The system has since been expanded and enhanced through collaborations among five LTER Information Managers as "LTER-Core-Metabase" (https://github.com/lter/LTER-core-metabase). This group effort aims to provide a sophisticated data management tool to the LTER community. The R scripts have been significantly enhanced to accommodate differing needs from different LTER sites ("MetaEgress", https://github.com/BLE-LTER/MetaEgress). These tools will continue to evolve as more LTER sites participate. The system was introduced to the public at the Earth Science Information Partners annual meeting in July 2019, and the concept and framework of LTER-Core-Metabase described in the LTER Databits (Gastil-Buhl et al, 2019).

4.4 Excel-to-EML

SBC data manager, Li Kui, used the SBC-Metabase as a model to build a simplified data archive system called "Excel-to-EML" (https://github.com/lkuiucsb/Excel-to-EML). Instead of having to manage Postgres server and learn to code in SQL language, Excel workbook is a more user-friendly program for most research groups. This tool is intended for small research groups with fewer datasets required for publication. The "Excel-to-EML" allows research groups to structurally store metadata and efficiently streamline the data archive process. As of 2021, this tool has been adapted by several university research groups as well as regional marine biological organizations (e.g. Partnership for Interdisciplinary Studies of Coastal Oceans).

4.5 Data tools

To assist users in understanding data quickly and easily, information managers have developed/collected data related tools. Data visualization is crucial for data assessment. In 2019, a hackathon organized by EDI invited Li Kui to contribute to a data visualization tool called "datapie" https://github.com/IMCR-Hackathon/datapie. This is a R shiny app that allows users to plot exploratory figures as a first glance for any target datasets, either a csv on local computer or a dataset DOI in DataOne repository. For the ongoing dataset, a R script is used to generate time series figures for internal usage: https://sbclter.msi.ucsb.edu/internal/research/Collaborative_Research/Figures_long_term_studies/ (required login). These figures are sent out for external users as requested. SBC plans to display the figures on the website after we finalize figure format and priority.

Accessing some large-scale exogenous datasets have been challenges for some researchers, for example, sea surface temperature that consists of a large amount of daily temperature files, or the ocean current monitored by radar in the Santa Barbara Channel that is in NetCDF format. The information manager, Li Kui, provided scripts in R and Matlab to export the data in a common csv format: https://sbclter.msi.ucsb.edu/publications/software/. These scripts help to speed up the research process and widening the scope of the data analysis (e.g., more variables can be used in the statistical calculations).

SBC also provides data for applications with graphical functions such as ClimDB. ClimDB's replacement, the HyMet project, is under construction. SBC plans to contribute hydrology and climate related datasets to that project, and also to the CUAHSI HIS database: https://data.cuahsi.org/

4.6 Data processing by the IM team

Since 2019, the data processing for the ongoing time-series datasets has been transitioned to the IM team. The primary accomplishments included: a) log books documenting deployment/retrieval of instruments and survey records; b) automated initial data QC by cross-checking log book information and the sensor outputs; c) R or Matlab scripts for detailed QA/QC and data formatting for each of the ongoing datasets, e.g. organism size, survey date, instrument deployment time, species code; and d) created scripts to push data into the publication pipeline.

The data processing is an ongoing project for the IM team and is subjected to constant updates to the programming scripts, primarily due to changes on sample method and instrument brand or version. The IM team works closely with the field crew to ensure the changes get incorporated into the data processing.

5. IMS history

5.1 Contributors to the IMS

See previous IM Plans for more details of contributors activities.

2019 - current: information manager: Li Kui
    2019: website coding: Wei Tung Chen, Nick Duncan

2004-2018: information manager: Margaret O'Brien
    2012: student assistant: Hannah Ake
    2011: assistant information manager: Alex Guerra
    2007-2020: close collaborator: M. Gastil Buhl (MCR LTER information manager)
    2006: EDQ coding: Chad Burt

2003-2004: interim information manager: Chris Jones (PISCO project)

2000-2003: information manager: Wei Yee Luan

5.2 Metabase

For the first 10 years, metadata at SBC was managed ad hoc, and EML generally created manually. In 2013, we finished our migration to a relational database for metadata, with scripted output. This more centralized system meant that metadata could be more easily standardized and accessible for multiple uses. Working with the Moorea Coral Reef (MCR) LTER, we adopted the database model that drives Georgia Coastal Ecosystems (GCE) LTER, "Metabase", and which had also been recently adopted by Coweeta (CWT) LTER. Initial adoption was to house descriptions of research themes with export as LTER-project XML (a subset of EML schema), which quickly demonstrated Metabase's usability and provided an entry point for work with Metabase design. In 2013, we finished code modules for export of EML datasets. These represented a significant improvement over the scripts originally written for project-export, and adhere to more professional programming practices. As this is a collaborative effort, all scripts and database work was coordinated between the two projects. Scripts were designed to work as web services, allowing them to facilitate network databases in the future. Smaller changes to Metabase and export scripts continue and are described above ("Metadata" Section).

5.3 Research Activities DB

High-level SBC research themes were described and added to Metabase in 2011, and previously, this information had been used to create a catalog of research activities (similar to datasets). The system is described in the DataBits article (O'Brien 2011).

5.4 PASTA criteria development

In 2017, the LTER Network replaced the metadata cataloging system ("Metacat") with its own PASTA software. PASTA includes a data quality engine with a list of criteria for package acceptance. This activity was led by SBC information manager O'Brien as working group chair (Servilla et al, 2013; O'Brien, et al, 2015). As of 2015, there were 32 checks to be passed by every data package (with approximately 50 more checks logged or requested).

5.5 Data Query for EML packages

From 2009 IM Plan: Because tables associated with time series can become large and cumbersome, SBC developed a generic tool (the EML Data Query tool, EDQ) for loading data packages into a relational database so that data can be queried with web forms (Figure 4, O'Brien and Burt 2007, (Leinfelder et al 2010). The application is not customized to any single dataset type. It reads EML metadata, uploads the described data table to a relational database and creates a map interface and form that then generates SQL queries based on user input. The application takes advantage of established community standards and accommodates a variety of data tables. This approach allows data owners to control the format of the tables they publish, while accommodating a repository of highly varied scientific data, and still allows the complete table to be archived in ASCII format. Another alternative would be to create custom interfaces and data models for each data type; however, that added complexity would increase maintenance costs and further strain resources.

The EDQ was written in 2006 for EML 2.0.1 and used a prototype of a Java library written by the LTER and the NCEAS Ecoinformatics programming groups. This code library (Data Manager Library, or DML) is now being significantly revised as part of the LTER NIS PASTA framework. A newer version of EML (2.1) has significant advantages over 2.0.1, and all SBC datasets were upgraded to meet its requirements. Consequently, the EDQ must be redesigned or replaced. We had tentatively planned this activity pending certain other activities not under our control, mainly the NIS production release and a revised DML. However, the subsetting of large datasets is a community need, and so no single research group (such as SBC) should develop a custom, local solution. Since SBC has substantial experience in this area, it's more likely that we will be involved in an advisory capacity within a much larger group such as the LTER Network or even DataONE, both of whom have expressed an interest in their systems meeting this need.

A need to subset several large SBC datasets remains. In recent years, as an additional quality control measure, data tables have been ingested into a relational database system mainly to check data typing and to compute data ranges and explore basic features (e.g., descriptive statistics). An added benefit of this practice is that once ingested, data could also be subset manually using SQL queries (by O'Brien).

This capability has not been widely advertised, and such a service would be custom and ad hoc. However, it remains as an alternative if other avenues are not fruitful.

5.6 DataONE Semantics Working Group

From 2015 IM Plan: O'Brien was invited to join the DataONE working group "Semantics and Integration" in 2013. DataONE is a data federation project funded by NSF-BIO (http://dataone.org). LTER is a "primary DataONE node", and consequently, all LTER data appear there. O'Brien's involvement is concerned mainly with describing the context and use cases for scientific data. In 2014, she formally joined the DataONE project to develop an improved search mechanism using LTER primary production data. O'Brien is also involved with the GeoLink project, funded by EarthCube. This effort compares schemas and models common information to enable cross-schema searches of data from disparate contributors such as LTER and the database maintained by the Biological and Chemical Oceanography Data Management Office (BCO-DMO.org).

5.7 LTER Network

From 2015 IM Plan: SBC information manager O'Brien co-chaired LTER Information Managers' Committee (IMC, http://im.lternet.edu, http://intranet.lternet.edu/committees/information-management) from 2009-2013, and has served on numerous IMC working groups. She is a member of two NIS Tiger Teams ("Data Manager" and "Metadata Quality") and the LTER Network Synthesis Data Committee (http://intranet.lternet.edu/committees/synthesis-data).

SBC's contributions to the Network are chiefly concerned with the quality and usability of EML metadata. O'Brien chairs the IMC working group, "EML Congruence Checker." This leadership stems from SBC's early adoption of EML for our own catalog, and our experience with the EML Data Manager Library (section 3.11, above). Additionally, our use of EML in the SBC LTER bibliography was reviewed by the EML development community and contributed significantly to the enhancements to the EML schema version 2.1. O'Brien served as the EML 2.1 release coordinator (O'Brien and Jones, 2008). Our work with LTER-project XML, based on EML and co-led by O'Brien and C. Gries (North Temperate Lakes LTER), is also likely to contribute to EML advancement. O'Brien previously chaired the "EML Best Practices" working group (2009), and also serves on two other IMC working groups, "UnitsDB" and "Controlled Vocabulary". All three are concerned with the standardization of EML dataset content. SBC plans to use the web services of the latter two systems as they become available (and are maintained), most likely after 2015 (Appendix III).

Literature Cited

Kui, L. and O'Brien M. (2018). Postgres, EML and R in a data management workflow. https://lternet.edu/wp-content/uploads/2018/03/2018DatabitsSpringIssue-web.pdf. Long-term Ecological Research Databits. Page 28-31

M. Gastil-Buhl, Margaret O'Brien, Tim Whiteaker, Li Kui. (2019). LTER Core Metabase. Long-term Ecological Research Databits 2019 summer. https://lternet.edu/wp-content/uploads/2019/06/DataBits_Summer2019.pdf.

Leinfelder, B., J. Tao, D. Costa, M. B. Jones, M. Servilla, M. O'Brien, C. Burt. 2010. A metadata-driven approach to loading and querying heterogeneous scientific data. Ecological Informatics, 5: 3-8.

LTER 2009. Review Criteria for LTER Information Management Systems.

LTER 2009a. Guidelines for LTER Website Design and Content.

LTER 2010. Strategic and Implementation Plan. http://intranet.lternet.edu/documents/lter-strategic-and-implementation-plan

O'Brien, M., M. Servilla and D. Costa. 2016. Assuring the quality of data packages contributed to the LTER Network Information System. Ecological Informatics. DOI: 10.1016/j.ecoinf.2016.08.001

O'Brien M. 2010a. Using the OBOE Ontology to Describe Dataset Attributes. LTER Databits, Fall 2010. http://databits.lternet.edu/fall-2010/using-oboe-ontology-describe-dataset-attributes

O'Brien M. and S. Harrer. 2008. Processing and quality control of kelp forest community survey data. Proceedings of the Environmental Information Management Conference, Albuquerque, September 2008.

SBC LTER. 2009. Attributes and Units in LTER Data Packages, v0.9. http://sbc.lternet.edu/external/InformationManagement/documents/SBC/Attributes_Units_LTER_data_packages.pdf

Servilla, M., M. O'Brien and D. Costa. 2013. Assuring the quality of data packages contributed to the LTER Network Information System. American Geophysical Union Fall 2013 Meeting, Poster IN53C-1576.

Literature no longer cited

Included here for history of SBC IMS development:

Leinfelder, B., S. Bowers, M. O'Brien, M. B. Jones, M. Schildhauer. 2011. Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data. Proceedings of the Environmental Information Management Conference. Santa Barbara, September 2011.

O'Brien, M. C. 2006. Using EML and Metacat for a site bibliography. at: 2006 LTER All Scientists Meeting, Estes Park, Colorado.

O'Brien M. and C. Burt, 2007. A Query Interface for EML dataTables. LTER Databits, Spring 2007. http://databits.lternet.edu/spring-2007/query-interface-eml-datatables.

O'Brien M. and M. Jones. 2008. EML 2.1.0 to be Released Soon. LTER Databits, Spring 2008. http://databits.lternet.edu/spring-2008/eml-210-be-released-soon

O'Brien M. 2011. The Santa Barbara Coastal (SBC) LTER's implementation of projectDB using Metabase. LTER Databits, Fall 2011. http://databits.lternet.edu/fall-2011/santa-barbara-coastal-sbc-lters-implementation-projectdb-using-metabase

O'Brien M. 2010. Using EML in Your Local Data Catalog. LTER Databits, Spring 2010. http://databits.lternet.edu/spring-2010/using-eml-your-local-data-catalog

O'Brien M. 2012. EML and Google Maps. LTER Databits, Fall 2012. http://databits.lternet.edu/fall-2012/eml-and-google-maps

SBC LTER. 2008. Guide to Creating SBC Datasets with Morpho, v0.9. http://sbc.lternet.edu/external/InformationManagement/documents/SBC/Morpho_userGuideSBC .pdf