Santa Barbara Coastal Long Term Ecological Research (SBC LTER)

Information Management Overview and Plan

Margaret O'Brien
January 2013

1. INTRODUCTION

1.1. Mission Statement

The primary objective of the Santa Barbara Coastal LTER Information Management System (IMS) is to facilitate diverse research and outreach goals by focusing on ease of access, data organization and integrity, and long-term preservation. Its general, our IMS is:

1. Cross-platform and largely based on Internet standards rather than on vendor-specific technology,
2. Integrated with laboratory science to keep participants up to date on changes and advancements, and
3. Modular, allowing us to incorporate the skills of a broad community, enhance components individually, and to integrate SBC activities with those of the LTER network and our other collaborators.

The SBC IMS currently meet or exceed all standards listed in the LTER Network's Review Criteria for LTER Information Management. 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 Documentation

SBC's IMS is documented at several levels. Some documentation is publicly available on the website (http://sbc.lternet.edu).

(1) A general Information Management (IM) Plan (this document). The intended audience is a reviewer, either external or the SBC IMS advisory committee. An IM Plan is required for all LTER sites. This document also contains three appendices that may be updated more frequently than the IM Plan itself:

I. Description of the IMS hardware and software stack.

II. An inventory of data packages organized by management needs and classified according to both local and Network categories.

III. Timeline for improvements to the IMS components

(2) An IM Guide (Wiki), whose intended audience is the IM staff and assistants, and which assures continuity. The IM Guide is updated frequently, as needed.

(3) Schematics, descriptions, etc., of individual system components as reference material for current and future IM staff. Component documentation is stored in dedicated directories in each project's repository.

1.3. Personnel

SBC's IMS is closely integrated with with UCSB Marine Science Institute (MSI) and the Moorea Coral Reef LTER (MCR.lternet.edu). SBC has a dedicated information manager located at MSI (Margaret O'Brien) with contributions from the project coordinator (Jenny Dugan), MCR information manager (M. Gastil-Buhl), and MSI IT personnel (Jim Woods, 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 (NCEAS.ucsb.edu/ecoinfo), also located at MSI. 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 MSI, the Earth Research Institute (ERI.ucsb.edu), and Bren School for Environmental Science and Management (Bren.ucsb.edu). SBC also employs occasional assistants or undergraduate students for directed tasks as funding permits. See the IM Guide (available after log in) for a history of contributors to the SBC IMS.

Communication between SBC scientists and Information Management is fostered by the SBC IMS Advisory Committee (IMSAC), which includes SBC co-PI John Melack (chair), the information manager and two rotating Investigators from different research fields (currently D. Reed and D. Siegel). The IMSAC establishes priorities for SBC's IMS activities and increases the scope of researcher involvement

in the SBC IMS.

1.4. 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. Type I data are generally posted publicly within 1 or 2 years of collection, although some ongoing electronic data are available much sooner and data requiring complex chemical analyses or data processing procedures may be delayed. For "Type II" data, our policy is that data will be described in the public catalog, but distribution information for the tables set to "offline", and interested parties instructed to contact the researcher. 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. USGS stream flow). These data may be republished, but if not, the browser redirected to the original resource. 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 our own data. 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 system (primarily the file server). The LTER Network is considering a policy under which collaborative projects that make use of an LTER site's IM 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.

1.5. IT Systems

SBC's holdings are stored in a networked directory system, and a user account is all that is required to view any data file. Write-access is limited to those responsible for data collection and maintenance. The directories for incoming 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. Our common data areas have been stable for several years so returning users will remain familiar with the structure, and the directory structure is published internally. SBC account holders are encouraged to use their home directories for work-in-progress to take advantage of regular backups. See Appendix I for information about backups and the software stack.

1.5. Definitions

Data package: data entity (or entities) and metadata.
Data package update: The addition of new data to an ongoing data package.
Data package maintenance: Enhanced metadata or data to keep the package current with standards or best practices.

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 and anticipated data products (as of Jan 2012) is in Appendix II. Currently, 148 packages are

publicly available and described in Ecological Metadata Language (EML), the metadata exchange format used by the LTER Network. Public data holdings comprise 263 data entities (e.g., data tables, KML files, images). Approximately half the inventory is ongoing time-series, in which the package is regularly reviewed and data added (i.e., updated, see definitions). 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
Oceanography	Moored and profiled hydrography (CTD), currents (ADCP), 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, AVIRIS (anticipated)

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 are 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" since scientific personnel are the source for knowledge about changes to sampling protocols and data output. This management style is complicated, and requires coordination among diverse scientific domains, measurement types and laboratories. The IMS accommodates several data processing patterns and software choices (e.g. SAS, Matlab, and MS-Excel). Scientific personnel are trained as necessary in the use of editors (e.g. Morpho, an EML editor), formats (XML), informatics concepts (e.g., data table design, SI units), programming practices, and use of the shared file server. Coordination and training is the responsibility of the information manager. There are currently two major pathways for data publication: (1) data are published as the last step of data processing and (2) scientific staff members describe the data using templates, and final products are built by the information manager using scripts and centralized metadata. Data packages using these pathways are designated respectively as "template" and

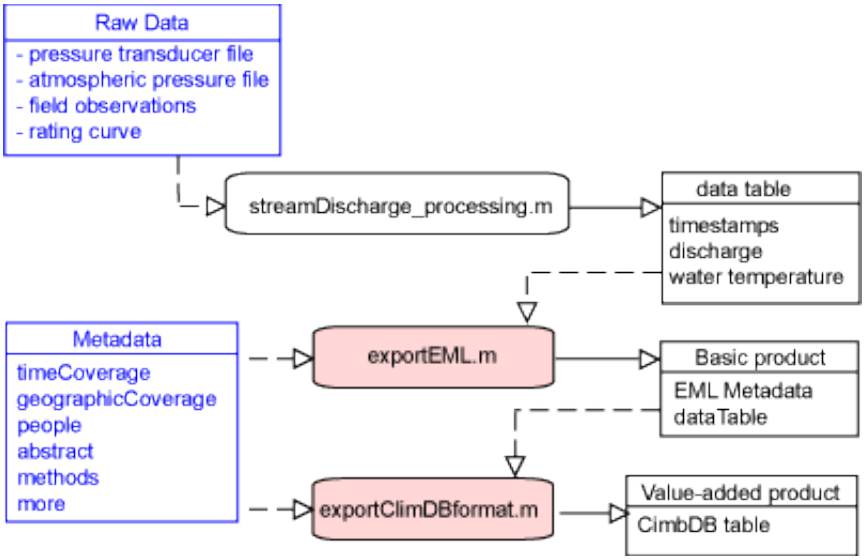
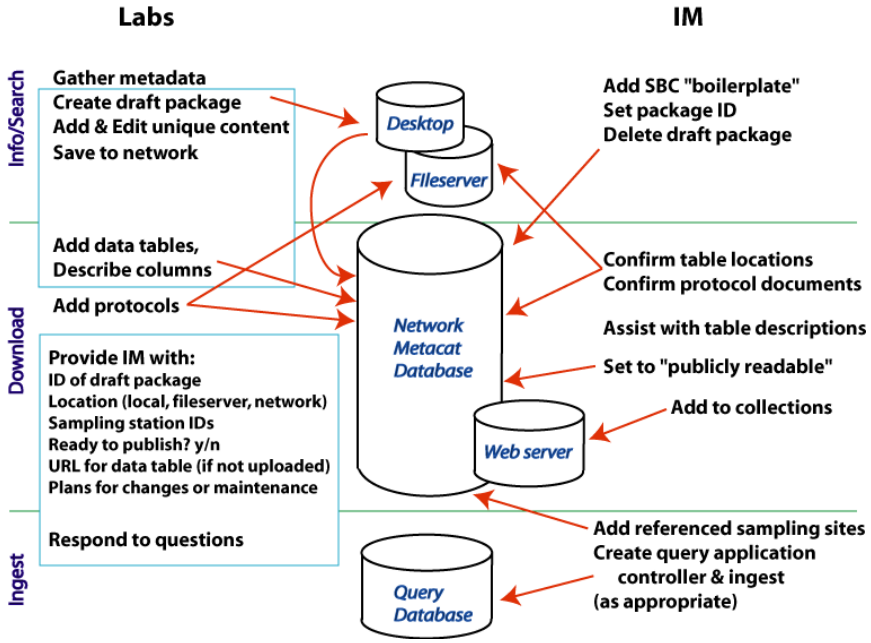


Figure 1. Data Processing flow for SBC's stream discharge data. Components contributed by the IM team are in pink boxes. Data inputs are in blue, and data products are on the right.

"unique" in the inventory (Appendix II).

The "template" pathway, in which data are published as the last step of processing is appropriate for data in which the format rarely changes and the processing language has good support for the XML document object model (required if scripts must access an XML template). An example of this pathway is our process for publishing stream discharge data using Matlab (Figure 1), and a similar sequence is used for sensor data from moored oceanographic instruments. This pathway is appropriate for data that can be considered part of a series, in which multiple tables have the same format and use the same metadata template. As of 2012, approximately two-thirds of SBC packages (by number, in six series) are updated or maintained in this manner. Data are processed in the laboratory as appropriate, and a table output format is co-designed by the data owners and the information manager. The information manager creates an EML metadata template, which is filled in using semi-customized export scripts written in Matlab, by either by laboratory personnel or by the information manager. Using the same language as is used for data processing means that data publication can be more easily integrated with processing, and may include a step to upload to the data catalog. Standardizing the process has the advantage of furnishing a product that can also be used as input for value-added products, e.g., a contribution to the network database, "ClimDB", in its required format (http://www.fsl.orst.edu/climhy).

The second pathway ("unique") is appropriate for data packages in which table descriptions are unique (e.g., experiments) or more likely to change from year to year, e.g example, community survey data (Figure 2). As of early 2012, SBC has 52 packages maintained using this pathway, and of those, 36 are "ongoing". This pathway is also necessary when the processing language does not have adequate support for the XML-DOM, which is the case for data processed by SBC ecologists using SAS. In this pathway, laboratory personnel gather the unique content into metadata documents (e.g., title, abstract, column descriptions), and the information manager adds standardized material and other components from the metadata system. Scientific personnel occasionally use the EML editor "Morpho", and guides have been written to assist with editing of SBC datasets, and to train them in appropriate attribute and unit models (SBC LTER, 2008, 2009a). We expect that processing software will progress to better XML-DOM support, and we plan to migrate some manually built data packages to the automated pathway.



0Figure 2. Scheme for coordined creation of data packages. Laboratory personnel gather metadata and (e.g., with Morpho) and IM personnel add standardized material and manage contributions to the data catalog. Arrows indicate areas of the fileserver or catalog affected.

3. STATUS AND MANAGEMENT OF IMS COMPONENTS

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, scripted pages and web applications. The unique content for static pages is maintained separately from the standardized content, menus and styling, and page assembly is scripted. As much as possible, material for links and menus is parameterized in configuration files. 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 very little training in web authorship and static sections can be easily replaced by scripts or web applications as these are developed. To facilitate multiple authors, all content is maintained in a version control system (Subversion). 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 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 separate documents. Planned tasks associated with the website are listed in Appendix III's timeline.

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

Section title	Content	Implementation
Research	Descriptions of core research activities	Backend storage for research themes is Metabase with export in projectDB. Browsed using Perl CGI script. (O'Brien, 2011).
Site	Descriptions and maps of sampling sites	Sampling sites are presented using Google maps. Static content is deprecated.
People	Personnel directory and individual profile pages	PERL CGI application using LDAP database as the backend.
Publications	Bibliography, including presentations	Storage is EML in Metacat with Perl CGI to display and filter by author.
Data	SBC's data catalog, plus links to other catalogs of interest	The data catalog can be browsed using a Perl CGI script. Backend storage and search uses Metacat (O'Brien, 2010)
Education and Outreach	Descriptions of K12 and higher education programs	Static content maintained by the project coordinator
News	Recent events of general interest, archived past events	Static content on the main page is updated quarterly by the project coordinator. News archive is a PERL CGI application with storage in text and RSS feed available.
Information management	Information plan, guide, system and project documentation	Static content maintained by the information manager
Internal area	Access to the fileserver, various help pages and site map	HTTPS, strong encryption. Certificate provided by LTER network.

3.2. Metadata

Several metadata components (e.g., personnel, methods, sampling sites) are necessary to describe data or link between areas of a website. Metadata at SBC has historically been managed ad hoc, but we are developing a more centralized system, with metadata in a relational database that is accessible for many uses. Working with the Moorea Coral Reef (MCR) LTER, we adopted the database model that drives Georgia Coastal Ecosystems (GCE) LTER, "Metabase2", and which has also been recently adopted by

Coweeta (CWT) LTER. In 2011, Metabase2 was ported from MS-SQL to an open source database (PostgreSQL) as "SBC-Metabase". Our initial use was to house descriptions of research themes and export these in the LTER-project XML schema (a subset of EML schema), which quickly demonstrated Metabase's usability and provide an entry point for work with Metabase design. Scripts for project-export were designed with reuse for EML datasets in mind, and their completion is a major focus of work planned for in 2012. As this is a collaborative effort (with MCR) all scripts and database work is coordinated between the two projects. Also in 2012, we will begin the work on the web-services switchboard, which will allow us to incorporate new material in dataset as additional components of the Metabase system are integrated. Tasks for integrating Metabase with the SBC IMS are listed in the timeline (Appendix III). Because the NIS (Network Information System) "PASTA" framework (http://nis.lternet.edu) will require data packages with consistent, high-quality metadata, many of these activities are timed with that development in mind. Activities will include Metabase population from existing EML datasets and re-export as regularized EML 2.1.

3.3 Data Packages

Data and metadata together comprise a "data package." Metadata documentation is published in the XML specification EML, and ASCII tables are used for data delivery and archive, as these have proven to be the most flexible and efficient for heterogeneous data. All EML metadata are housed in the NSF-funded XML database "Metacat" maintained by the LTER Network Office (LNO.lternet.edu). 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). Approximately 70% of SBC's data holdings are at this level. As data package maintenance shifts from the current semi-manual creation model to more efficient creation via scripts during 2012, information manager time will become available to address other issues required to bring the remaining 30% of data packages up to "Integration" level.

Data packages may be classified or sorted according to many other criteria, in addition to those above for metadata content and usability. We refer to these groupings as "inventory types", and several are incorporated in to our data package management software (Section 3.3.1). The inventory in Appendix II is grouped according to some of these categories.

Table 3. Levels of data package complexity with potential uses. Asterisks indicate that related issues and needs are being considered at a network level. The conceptual "LTER metadata completeness levels" given in the last column are obsolescent, but are included for comparison.

Level	EML Metadata content	Data content	Uses
Information	Boilerplate: project description, publisher, contacts, access and use statements Unique: title, abstract, personnel*, publication date, Temporal, geographic and taxonomic coverage, keywords*	Data are optional	Information only, e.g., LTER Type II data LTER "Identification" or "Discovery" (see text) Searches by time, location, taxonomy Dataset citation
Download	**Information** + Data description, including column names, definitions & units*, physical description*, download URL	Data are required, but inconsistencies are allowed	Data are available, but the user may need help with interpretation. LTER "Evaluation", "Access", "Integration" (with caveats)
Integration	**Download** + Metadata congruent with data*	Data are required, and must be clean and congruent with metadata	1. Integration, workflows, further automated processing 2. Query applications 3. Contribute to Network databases LTER "Integration"

We are also planning for the incorporation of more advanced features that will further enable streamlined discovery and use, such as semantic annotation, unique and regularized identifiers, detailed methods and sampling design in metadata markup, and quality-controlled data values (Table 4). Such features do not necessarily require metadata or data of a certain complexity (as in Table 3), and potentially could be added to any data product. However, incorporating these features will require the development of more sophisticated tools, or are most effective to implement after significant standardization has occurred at a network level. Generally, it is expected that these features will not appear in SBC data products until after the challenges of basic data ingestion/integration have been met. SBC participates in community and network projects that affect or promote these enhancements.

Table 4 Enhanced features of data and metadata. These metadata features can be incorporated at various levels, and data/metadata requirements vary. Most aspects of enhanced features involve inter-site collaboration.

Level	EML Metadata content	Data content	Benefits
Linked	Protocols Detailed project* descriptions	any level, integration level is preferred	Links on website between data packages and other website areas, (research, publications)
Quality controlled	QC metadata* methods	Integration level, plus data values subjected to quality control measures*	Confident integration
Annotated	Semantic annotation*	Data optional, integration level is preferred	Confident and accurate discovery, automated integration

3.3.1 Data Package Management

In collaboration with MCR, software tools for managing our sites' (SBC's and MCR's) data package inventory were begun in 2012, and added to Metabase. As with other Metabase work, SBC and MCR design and implement together, and then populate their individual instances. Tools are outlined in Table 4, and are being designed with the following use cases in mind: a) to share package-status information with an IM assistant tasked with updates, b) to sort/filter packages by inventory type, e.g., to generate an appendix for this document, and c) to display lists of packages on the web by status or type, e.g., packages in 'draft'. A static inventory showing some features of data packages is available in Appendix II.

Table 5 Data inventory management tools in use at SBC.

Name	Description	Implementation	Status
package management database tables	SQL tables for tracking status of data packages and their inventory types	postgreSQL,	Tables fully populated for ongoing time-series packages
Web views	HTML views of data packages	PHP	In development, e.g., dataset_status_tracking.php
SBC activity log	Links between sampling sites and research activities since the project's inception.	postgreSQL table with cross-references between research activities and sites	Populated time-series activities from watershed group, and reef group for 2000-2013, based on data holdings. Info to be checked and/or confirmed by laboratories.

3.4 Site Descriptions

All SBC time-series sampling sites have been described in a standard XML format that is compatible with EML and KML. This geographic coverage information is added to data packages using semi-automated scripts. Locations are displayed on the SBC website in two ways. First, each EML data package's HTML display shows a map as it is rendered by the data catalog (O'Brien, 2012). Secondly, our XML-formatted site descriptions are used to populate a Google map in the "Site" area of the SBC website, to provide

spatial context for research (Table 2). Geographic information will also be made available as KML, which is appropriate for many external applications including mapping tools being developed by the LTER Network (Appendix III). We continue to review policies and practices for managing SBC sampling sites (https://sbc.lternet.edu/info_management/research_sites.html).

3.5 Personnel

The current backend storage for our personnel database also houses the system for fileserver access (LDAP), and is maintained by MSI personnel and co-managed by MSI, MCR LTER and SBC. Personnel information from LDAP was ported to Metabase during 2011 as part of the adoption of Metabase for research projects. It is planned that personnel information for data packages will be exported from Metabase (not LDAP), and that by 2015 SBC's LDAP will be used only for fileserver access and that the web directory will be served by Metabase content (Appendix III). Until then, personnel information must be maintained in both LDAP and Metabase, and also in the Network personnelDB (which is currently undergoing redesign). Ideally, site personnel information will be synchronized with the Network database via web services. The SBC/MCR web service client switchboard planned for development 2013 will enable these activities during the second half of SBC-III (Appendix III).

3.6 Research Activities Catalog

High-level SBC research themes were described and added to Metabase in 2011. These are exported for web display in the LTER-projectDB XML specification and are linked to the personnel directory. The LTER-project schema is based on EML, and is in use at several LTER sites. To facilitate linkages between SBC research themes and relevant data, the project catalog uses the same menus and keywords as those used for datasets. We plan to expand our use of projectDB during 2015 to include details of specific research projects and sampling activities, which will enable website cross-links between research, data, sampling sites and/or publications (Appendix III).

Knowledge of ongoing research activities is also applicable to data package updates and management, e.g., so that IM personnel can anticipate new data. In 2012, initial content was added to Metabase for ongoing research activities and sampling sites, and these two tables cross-referenced in the package management schema (Metabase, section 3.3.1, above). Ongoing time-series are planned to be the next group of SBC research activities to be fully populated and displayed with projectDB.

3.7. Bibliography

A research group's bibliography is an important publicly available resource. SBC adds 50-75 new citations annually which also must be managed in the databases of funding agencies (e.g., NSF's "Fastlane") and at the LTER Network. 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, the current capability of Metacat is not ideal for their search and display. Consequently, we are examining other options. Currently, there are three avenues to evaluate: (1) develop of an application using a native XML database with XQuery, (2) port the XML-structured bibliography to the SBC-Metabase, or (3) port the XML-structured bibliography to the database used by the Network which is slated for redesign. We are considering the advantages of each option to SBC, and details and/or proof-of-concept projects are available on the internal website. The timing of improvements depends on the evaluation of options and the availability of the improved Network bibliographic database, and so SBC will not embark on these upgrades work until after 2014.

3.8. Quality control and Protocols

Quality control is generally based in the researchers' laboratories. All data packages include methods, usually in the form of protocol documents (PDF). Data collection and processing protocols are easily accessible on the fileserver 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://sbc.lternet.edu/info_management/research_protocols.html). The information manager works closely with analysis personnel to document quality control in metadata as appropriate for individual data packages. Quality control of SBC's community survey data has been documented (O'Brien and Harrer, 2008). Documenting quality control at the data package level is an enhanced feature of data packages (Table 4), and planned for the second half of SBC III (Appendix III).

3.9. Standardized Measurement Descriptions

In order for successful synthesis to occur at a Network level, LTER sites will need 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 method is to first standardize measurements at the site level with complete descriptions (including methods, precision and error). Because this is an enhanced feature of data (Table 4), we anticipate this activity occurring during the second half of SBC-III. We will work with the LTER IM community as we standardize our measurements in preparation of this becoming a LTER Network wide requirement (O'Brien, 2010a). SBC is uniquely positioned to take on this work due to our involvement with ontology development with the Extensible Observational Ontology (OBOE) in the Semtools project (DBI-0743429, O'Brien co-PI).

3.10 Data Catalog

SBC's data catalog has been based on EML metadata since 2003. The catalog was decoupled from a local, legacy Metacat system in 2010 and now draws SBC EML directly from the LTER Network for display on the SBC website (http://sbc.lternet.edu/data). Our holdings are categorized into "collections" accessible via web forms keyed to local habitats, measurement types and LTER core research topics. The Metacat catalog itself can also be searched by keyword and dataset owner (investigator). Because our EML is drawn from the Network Metacat, local searches return the same results as searches at the network. As part of the migration in 2010, EML-dataset XSL templates were redesigned in a "tabbed" format to highlight major sections such as geographic area and methods. The SBC templates are used by two other LTER sites (Moorea Coral Reef, Virginia Coast Reserve), the Network Controlled Vocabulary website (http://vocab.lternet.edu), and are being considered for adoption in the redesign of the Network-wide data catalog. Major changes to the catalog are not anticipated at this time, although some maintenance will be required as EML content is enhanced and associated XSL templates mature. In 2014, we will consider replacing our current search implementation (to the Network Metacat) with searches to our local Metabase database.

3.11 EML Data Query application (EDQ)

Because tables associated with time series can become large and cumbersome, we have developed a generic tool (the EML Data Query tool, EDQ) for loading Integration-level datasets into a relational database so that data can be queried with web forms (Figure 4, O'Brien and Burt 2007, (Leinfelder et al 2010). This application is not customized to any one dataset type. It reads EML metadata, uploads the described data table to a relational database and creates a map interface and form which then generate 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 uses a prototype of a Java library written by the LTER and the NCEAS Ecoinformatics programming group. 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 so all SBC datasets should be upgraded to meet its requirements. Consequently, the EDQ should be redesigned. We are planning this activity for 2014, after: a) the NIS production release (and a revised DML is available), b) all SBC datasets have been upgraded to EML 2.1 (Appendix III), and c) appropriate MSI server and software system upgrades (Appendix I).

4. SBC Informatics projects

4.1 LTER Network

SBC information manager O'Brien currently co-chairs LTER Information Managers' Committee (IMC, http://im.lternet.edu, http://intranet.lternet.edu/committees/information-management), 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 datasets. O'Brien chairs two IMC working groups, "EML Best Practices", and "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 now available in the EML schema version 2.1, and 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 also serves on two other IMC working groups, "UnitsDB" and "Controlled Vocabulary". Both are concerned with the standardization of EML dataset content. SBC plans to use the web services of these two systems after its development of the web service switchboard (Appendix III).

4.2. 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 useability of the OBOE ontology for standardizing SBC measurements during the second half of SBC III (Appendix III).

Literature

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.

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 Web Site Design and Content.

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

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

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

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.

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. C., C. Gries, W. Sheldon, J. Walsh, J. Porter, S. Bohm, J. Brunt, S. Remillard, K. Ramsey, J. Downing, K. Vanderbilt, R. Aguilar and M. Servilla. 2009. A collaborative database for LTER projects at sites. Poster at: 2009 LTER All Scientists Meeting, Estes Park, CO.

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

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

Appendix I. SBC Hardware infrastructure

Computing systems at MSI

Data management for the project has the advantage of utilizing the computing capabilities of the Marine Science Institute (MSI). MSI has a 1000Mb/s connection to the UCSB campus backbone, which provides shared access to a 622Mb/s CALREN-2 connection, which in turn provides access to the Internet. MSI supports the research servers. The main data server providing network file sharing (Samba and NFS) is a running RedHat Enterprise Linux 5 (64-bit). The data server also runs SVN for revision control systems, SAS, Matlab, GSLIB and PERL for scientific applications. Currently we have 11.5 TB of storage (expandable) available on that system. The second server is running RedHat Enterprise Linux 5 (64-bit), which runs the Apache web server, and the Tomcat java servlet engine. A third server running RedHat Enterprise Linux 4 (64-bit) is the primary database server, running PostgreSQL, MySQL and the personnel database (LDAP). The Server room is connected to E-Power, and redundant power is provided by an APC 6000 UPS battery backup. Distributed server backups (via Amanda) are coordinated with MSI.

File system backup

Backups are carried out by the MSI computing staff. Full backups (level 0) are performed monthly, with incremental (level 5) and progressive incremental (level 9) backups weekly and daily, respectively. Five months of disk-to-disk backups are stored on the server, with storage space allocated to the /backup partition as necessary. Off-site backups are copies to external hard drives, stored in another office on campus. The entire public data catalog is also archived on DVD.

Appendix II. SBC Data Inventory

January 2013

The inventory of data packages is presented in two parts: Appendix II.1, Cataloged data package, and Appendix II.2, Data packages currently in draft, undergoing re-design, or for which data are not yet released but expected. The inventory is online as of early 2013; please note date on each page. Information for both types is likely to change.

Appendix II.1. Inventory of publicly available SBC data packages

Data packages are classified by several criteria, including SBC-III proposal type, use/intent, management scheme, and network type. This inventory is organized by management needs. Packages are listed according to temporal types: 'ongoing time-series', 'short-term data', 'terminated time-series', and 'non temporal' (generally, supplementary material). 'Terminated time-series' may be reinstated.

Appendix II.1: Cataloged data packages (requires SBC LTER login):
https://sbc.lternet.edu/internal/research/Metadata/EML_data_packages/00_pkg_management/html/cataloged.html

Appendix II.2. Inventory of draft, anticipated or Network Type II data packages

Data packages whose metadata are not yet pubic, under-described packages, or packages that do not yet have publically available data (ie, will be moved from Network Type II to Network Type I) are listed here. Type II packages are likley to have had metadata already released (cataloged), and data manually distributed with permission from the PI. Data are available for to those with an SBC login. ALL INFORMATION HERE IS PROVISIONAL.

Appendix II.2: Draft and anticipated data packages (requires SBC LTER login):
https://sbc.lternet.edu/internal/research/Metadata/EML_data_packages/00_pkg_management/html/draft_anticipated.html

Appendix III. Time line of SBC IMS Improvements

(Last update: January 2013)

Table 1. IMS improvements as proposed in SBC-III Proposal Supplement 2-4, reproduced in table form.

year	*Proposed plans*	*Deliverable*	*Progress*
2011-2012	All SBC time-series sampling sites have been described in standard format and applications are in development to map these on the website to provide spatial context for research and data. Geographic information will be available as KML, which is appropriate for many external applications (e.g. mapping tools being developed by the LTER Network).	Mapped sampling sites	maps from both /sites/sampling, and in EML dataset templates
2012	Re-build all EML datasets from SBC-Metabase. We began building datasets with Metabase in 2011. We plan to update our time-series datasets in 2012 from Metabase, and upgrade all static datasets to EML 2.1	Higher quality metadata	Populating metabase. Awaiting server/software upgrades (Jan 2013)
2012	The development version of the NIS Provenance-Aware Synthesis Tracking Architecture (PASTA) is scheduled to be available in mid-2012. We will submit all of our datasets to the NIS	SBC data tested against PASTA system	Ongoing time series are tested against staging system as part of update; all have passed.
2014	The production PASTA is scheduled to be released in 2014, and We will submit all of our datasets at that time (dependent on NIS software meeting its own milestones)	Data in PASTA	Production schedule accelerated to Jan 2013.
2015	Expand the content of SBC's projectDB to include details of specific research projects and sampling activities	Website cross-links between research, data, sampling sites and/or publications	see table 2
2016	In order for successful synthesis to occur at a Network level, sites will need 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 method is to first standardize measurements at the site level with complete descriptions (including methods, precision and error). We will work with the LNO and the LTER IM community as we standardize our measurements in anticipation of this becoming a LTER Network wide requirement.	Standardized SBC attributes	

Table 2. Following are the specific activities required to produce IM System deliverables. Proposed deliverables from Supplement 2 are in boldface. For the most part, the timeline focuses on the period before the mid-term review in 2015. A reference to description of affected component(s) in the IM Plan is included. If no name is given, then activity is carried out by site information manager and/or assistant (as appropriate).

These activities are not included, but are ongoing in all years:
a) Design and cataloging of new data packages, updates of current time-series.
b) Server and software upgrades and migration
c) Participation in Network-level projects (some are noted)

year	*Deliverable*	*Affected component*	*Activity*	*Progress as of Jan 2013*
2012	**Mapped sampling sites**	Website, site descriptions (3.1, 3.4)	Finish web scripts to map SBC time-series sampling sites on the website to provide spatial context for research and data. Make geographic information available as KML	Done
2012	Metabase populated from existing datasets	Metadata (3.2)	From various sources (existing EML, LDAP, manual insertion) populate the Metabase tables necessary to export EML datasets	In progress
2012	Consistent look-up tables	Metadata, data packages (3.2, 3.3)	Regularize values of keys by converging on consistent content in lookup tables. Via queries to Metabase, analyze components for inconsistencies. Includes: personnel, data package titles, abstracts and keywords. entity-level names, physical descriptions, abstracts, delivery URLs (e.g., Network Data Access System, DAS)	Progressing with database population
2012	Efficient export of EML from Metabase (**Higher quality metadata)**	Data packages (3.3)	Expand scripts initiated in 2011 for EML export from Metabase to build complete datasets. All datasets will be exported as EML 2.1.	Delayed till 2013, pending server upgrade and supplement hire.
2012	**SBC data tested against PASTA system**	Network Information System (NIS)	Submit SBC datasets to the PASTA-prototype system scheduled to be available in 2012	Ongoing; all datasets are tested during draft or update.
2012	GCE Matlab toolbox workshop participation	Planning, data packages (3.3)	GCE LTER has developed Matlab tools that streamline some aspects of dataset description and manipulation, and which work with Metabase. A workshop is planned to introduce the toolbox to other LTER sites.	Done, November 2012 (O'Brien, Gotschalk)
2013	Website upgrades	Website (3.1)	Redesign some static sections. Integrate new menus. Incremental changes to dataset and project views. Update research-theme content for SBCIII	Some menu components available (M. Thompson)
2013	GCE Matlab toolbox evaluated for SBC use	Planning, data packages (3.3)	Evaluate usability of GCE Matlab toolbox for SBC, with respect to a) our implementation of Metabase and b) SBC's existing data package publication workflows in Matlab.	In progress, by Chris Gostchalk (supp. funds) and O'Brien

year	*Deliverable*	*Affected component*	*Activity*	*Progress as of Jan 2013*
2013	Outline possible pathways for EDQ upgrade	Planning, website, data package management, EDQ (3.1, 3.3, 3.11)	The EML dataset query application (EDQ) was written in 2006 for EML 2.0.1 using a prototype of the Data Manager Library (DML, Java code). It requires upgrade and/or redesign. Both EML and the DML have been significantly revised (e.g., EML 2.1, and the DML as a PASTA component). The upgrade scenarios, or plans for replacing this functionality will be outlined.	In progress, as of mid-2012
2013	SBC/MCR web service client switchboard	Data packages (3.3)	SBC has supplement funds to refine software for export of EML datasets from Metabase. Work includes a "switchboard" to use other available EML content via web services. This activity will require collaboration with MCR, and hiring of a part-time or short-term enterprise level programmer. As of 2012, web services are available centrally from the Network unitsDB and vocabDB (keywords).	Individual identified, to be hired with 2011 supplement funds
2013	Semi-automated data update workflows for ongoing datasets	Data packages (3.3)	Refine existing scripts (Perl or Matlab, TBA) or adopt new tools (e.g., GCE toolbox) for semi-automated updates of ongoing datasets.	Individual identified, to be hired with 2011 supplement funds
2014	Upgraded EDQ	Website, EDQ (3.1, 3.11)	The EDQ will be upgraded or its functionality replaced	
2014	All SBC datasets are EML 2.1 or better	Data packages (3.3)	SBC datasets that have remained at EML 2.0.1 to be compatible with the EDQ must be upgraded because PASTA will not accept EML 2.0.1.	
2014	**SBC Data in PASTA**	NIS	Submit all SBC datasets to the production PASTA system scheduled for release in 2014	** planned for 2013, per new NIS timeline
2014	Lower technical bar to maintaining project information	Metadata (3.2, 3.6)	ProjectDB was populated in 2011 with research project themes by the lead IM. We would like non-IM personnel to assume this task, and so web forms for input to Metabase are needed.	
2014	SBC project themes on the web are linked to activities	Website, research projects (3.1, 3.6)	Expand use of projectDB schema and associated scripts/templates for displaying scientific activities in Metabase on the website.	2012: began populating ongoing time-series activities in metabase
2014	Metabase evaluated for SBC bibliography	Bibliography (3.7)	SBC's bibliography is currently stored in extensive EML, with additional metadata. Possibly, management could be more efficient with different storage. Examine Metabase for this use, particularly with regard to required reporting and linkages on the SBC website.	

year	*Deliverable*	*Affected component*	*Activity*	*Progress as of Jan 2013*
2015	Cross-linked Metabase tables	Metadata (3.2)	Complete any additional cross-links between research, data, sampling sites and publications in Metabase. Cross-links will have been initiated as part of earlier work with data set production.	
2015	**SBC website cross-links between research, data, sampling sites and/or publications**	Website (3.1, 3.2, 3.3, 3.5, 3.6, 3.10)	Expand and harden website scripts, and standardize configuration to streamline display and maintenance of cross-linked information in Metabase. Specific activities TBD.	
2015	Mid-term review	All	Work-to-date will be highlighted for the mid-term review	
2015	Exploit centrally-stored metadata	Data packages (3.3, 3.5)	It is expected that by 2015, additional production NIS modules will supply content for EML datasets via web services (personnelDB, biblioDB). Also, all central database are also planned to synchronize with site systems via web services.	
2015	Evaluate SBC measurement ontology	Planning, standardized measurement descriptions (3.9)	Plan for use of existing knowledge models for measurement standardization needs (e.g., hierarchical vocabularies, ontology). This work follows from earlier work with SBC datasets and the OBOE ontology (2009-20xx).	
2016	**Standardized SBC attributes**	Data packages, NIS (3.9)	Standardize SBC measurements with complete descriptions, including analysis methods, precision and error.	Examine data associated with ocean acidification collaborations
2017	SBC-IV proposal IM section	All		

Appendix IV. Research collaborators and their relationship with the SBC Information Management System

(Last update: January 2013)

Table 1. The following projects collaborate with SBC LTER, and their relationships between these projects and the SBC Information Management System (IMS) vary. There are several patterns for the extent of the involvement of the SBC IMS; these examples are not exhaustive, and some projects fall into more than one:
a) The SBC IMS has no involvement with the project's data management.
b) Data are housed on SBC server, and arrangements are informal. No plans exist for publication by SBC. This is likely to be true for projects that predate NSF's required data management plan for all proposals (2011).
c) Data are housed on SBC server, and SBC plans to publish data; priority determined ad hoc.
d) Project has a data management plan that leverages SBC IMS and will publish data through our catalog.

Years active	*Project*	*Funding*	*PIs*	*Data description*	*Extent of management by SBC IMS*
19__ - 20__	CODAR	varies	Washburn	Surface currents from radar	None. CODAR maintains it's own data catalog
2005 - 2007	CEQI (Coastal Environmental Quality Initiative)	UCOP (UC Marine Council)	?Reed, ?Gaylord, Stewart?	Moored instruments co-located at SBC reefs	On file server. SBC will publish data. also see: http://escholarship.org/uc/search?entity=ucmarine_ceqi
19__ -	PISCO (http://piscoweb.org)	Moore, Packard	Gaines, Warner	Sampling area overlap SBC near Pt. Conception	None. PISCO maintains it's own data catalog.
1994-	Plumes and Blooms (P&B) http://pnb.eri.ucsb.edu	NASA	Siegel	Semi-monthly profiles of CTD and optics across the Santa Barbara Channel	P&B maintains its own data catalog. Data has been offered, and SBC IMS is considering best path for inclusion
20__	CALobster		Lenihan	Fisheries surveys	
	CDIP (SIO)	MMS		Modeled ocean swell and circulation in the Southern Cal bight	Model results for SBC reefs are retrieved nightly and republished by SBC IMS
	Channel Islands National Park	NPS		reef community surveys	None
2012	Channel Islands National Park	NPS	Kapsenberg	nearshore pH	Data will be managed and published by SBC
	Channel Islands National Marine Sanctuary				None
	Santa Barbara Land Trust (NGO)		Melack	Water chemistry data in Arroyo Hondo	Water chemistry analyzed and published with other SBC data
	Santa Barbara Channelkeepers (NGO)		Melack	Ventura River water chemistry	Water chemistry analyzed and published with other SBC data

Years active	*Project*	*Funding*	*PIs*	*Data description*	*Extent of management by SBC IMS*
	City of Santa Barbara	City of SB	Melack	Water chemistry in Mission and Arroyo Burro creek catchments	Water chemistry analyzed and published with other SBC data
	County of Santa Barbara	County of SB	Melack	Water chemistry	Water chemistry analyzed and published with other SBC data
	Friends of the Santa Clara River (NGO)				
	USGS			Stream discharge	SBC publishes processed data from USGS gauges in our area.
2010	Coastal-trapped waves	NSF	Fewings	Physical oceanography	Housed on SBC/MSI server; data may be removed. No plans for SBC to publish data
2010	Regional upwelling relaxation	NSF	Washburn, Fewings	Physical oceanography	Housed on SBC/MSI server; data may be removed. No plans for SBC to publish data
2010 –	Reef foodweb	NSF	Page, Miller	C&N Isotope content of producers and consumers	SBC will publish data
2005 – 2007	Kelp Biomechanics	NSF	Gaylord		On file server.
2012 –	Ambient pH	State of California	Passow	Bi-weekly pH of seawater for instrument calibration	SBC will publish data
2010 –	AUVs		Siegel	CTD deployments on autonomous vehicles	SBC will publish data
	Surf grass		Reed, Blanchette		On file server.
2004	Kelp Dispersal		Reed		On file server.
	Regional reef community		Rassweiler		On file server.
2010 –	Diatom exudates and high CO2	NSF OCE	Passow, Brzezinski, Carlson		None. Project uses BCO-DMO
2012 –	Kelp genetics	NSF	Alberto, Reed		SBC will publish data
2012 -	OMEGAS	NSF	Hofmann, Blanchette, Washburn	pH	SBC does not manage data. Collaborating on processing and publication methods. Project to use BCO-DMO.