OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2008.0015

Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes

Inigo San Gil,[1] Wade Sheldon,[2] Tom Schmidt,[3] Mark Servilla,[1] Raul Aguilar,[4] Corinna Gries,[4] Tanya Gray,[5] Dawn Field,[5] James Cole,[3] Jerry Yun Pan,[6] Giri Palanisamy,[6] Donald Henshaw,[7] Margaret O'Brien,[8] Linda Kinkel,[9] Katherine McMahon,[10] Renzo Kottmann,[11] Linda Amaral-Zettler,[12] John Hobbie,[13] Philip Goldstein,[14] Robert P. Guralnick,[14] James Brunt,[1] and William K. Michener[1]

Abstract

The Genomic Standards Consortium (GSC) invited a representative of the Long-Term Ecological Research (LTER) to its fifth workshop to present the Ecological Metadata Language (EML) metadata standard and its relationship to the Minimum Information about a Genome/Metagenome Sequence (MIGS/MIMS) and its implementation, the Genomic Contextual Data Markup Language (GCDML). The LTER is one of the top National Science Foundation (NSF) programs in biology since 1980, representing diverse ecosystems and creating long-term, interdisciplinary research, synthesis of information, and theory. The adoption of EML as the LTER network standard has been key to building network synthesis architectures based on high-quality standardized metadata. EML is the NSF-recognized metadata standard for LTER, and EML is a criteria used to review the LTER program progress. At the workshop, a potential crosswalk between the GCDML and EML was explored. Also, collaboration between the LTER and GSC developers was proposed to join efforts toward a common metadata cataloging designer's tool. The community adoption success of a metadata standard depends, among other factors, on the tools and trainings developed to use the standard. LTER's experience in embracing EML may help GSC to achieve similar success. A possible collaboration between LTER and GSC to provide training opportunities for GCDML and the associated tools is being explored. Finally, LTER is investigating EML enhancements to better accommodate genomics data, possibly integrating the GCDML schema into EML. All these action items have been accepted by the LTER contingent, and further collaboration between the GSC and LTER is expected.

Background

The LTER and the Genomics Standards Consortium (GSC) initiated joint discussions in November 2007 during a Long-Term Ecological Research Network (LTER) meeting held at Michigan State University (MSU). James Cole, head of the Ribosomal Database Project, gave an overview of the difficulties associated with annotating 16S sequences,

[1]Department of Biology, LTER Network Office, University of New Mexico, Albuquerque, New Mexico.
[2]Department of Marine Sciences, University of Georgia, Athens, Georgia.
[3]Department of Microbiology & Molecular Genetics, Michigan State University, East Lansing, Michigan.
[4]Global Institute of Sustainability, Arizona State University, Tempe, Arizona.
[5]NERC Centre for Ecology and Hydrology, Oxford, OX1 3SR, United Kingdom.
[6]Oak Ridge Natl. Lab, Oak Ridge, Tennessee.
[7]USDA Forest Service, Pacific NW Research Station, Corvallis, Oregon.
[8]Marine Science Institute, University of California at Santa Barbara, Santa Barbara, California.
[9]Department of Plant Pathology, University of Minnesota, Saint Paul, Minnesota.
[10]Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, Wisconsin.
[11]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Bremen, Germany.
[12]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts.
[13]The Ecosystems Center, Marine Biological Laboratory, Woods Hole, Massachusetts.
[14]Department of Ecology and Evolutionary Biology and University of Colorado Museum of Natural History, University of Colorado, Boulder, Colorado.

genomes, and metagenomes with contextual information and outlined current GSC activities in the area. Significant discussion followed, and given that the GSC already had an interest in developing linkages to researchers generating genomes and metagenomes in the LTER, the LTER was invited to send a representation to the fifth GSC workshop. In particular, the GSC was keen to understand how its Genomic Contextual Data Markup Language (Kottman et al., 2008) (GCDML) might relate to the LTER's Ecological Metadata Language (EML) (Jones et al., 2001). The GSC also strives to learn from the experience of other communities and was therefore keenly interested in how the best practices of the LTERs were applied in building EML and how it achieved NSF adoption of EML. Finally, the GSC was interested, through these linkages, to increase participation of LTER members in GSC goals. This report details these discussions, and elaborates on subsequent interactions that are helping to define how and why these two communities should work together further in the future.

LTER input into the fifth GSC workshop

At the workshop, Inigo San Gil presented a talk that covered essential background to the LTER Network, EML, and shared experience on developing this XML standard, gaining adoption, and general issues of data integration.

The LTER

The LTER was formed in 1980 with support from the NSF. The LTER today has over 2000 researchers associated with 26 sites representing diverse ecosystems and research emphases in continental North America, islands in the Caribbean, the Pacific, and Antarctica—including deserts, estuaries, lakes, oceans, coral reefs, prairies, forests, alpine, and Arctic tundra, urban areas, and production agriculture. The LTER mission is to provide the scientific community, policy makers, and society with the knowledge and predictive understanding necessary to conserve, protect, and manage the nation's ecosystems, their biodiversity, and the services they provide. When the first six LTER sites in the LTER Network were funded in 1980, the idea of studying specific ecosystems at temporal and spatial scales was revolutionary. Currently, all 26 LTER sites participate in this endeavor, and there are over 30 other countries that formed LTER networks with presence in all continents. Many other countries are in the process of forming LTER networks, most of them following similar procedures and frequently adopting the same metadata standards as in the US LTER. To learn more about the status of the non-US LTERs, one can google the Internet, that is, international LTER networks.

One of the challenges that the LTER Network has faced since its inceptions has been data synthesis. LTER started generating and offering to the public metadata and data very early on (Michener et al., 1997); however, different LTER sites adopted different protocols to divulge their metadata. Researchers focusing on cross-site comparisons faced formidable challenges to pursue synthesis projects. To address some of these challenges, LTER adopted a network-wide metadata standard in 2001 and gradually has standardized nearly all existing metadata records (San Gil and Baker, 2007). The LTER Network metadata standard, EML, is recognized by NSF (LTER Information Managers Executive Committee, 2005), and EML rules compliance is used by the NSF review teams to evaluate each LTER site's performance. Particularly, the official NSF LTER site review criteria document states that: "(1) Metadata shall be of sufficient quality and completeness to ensure long-term (>20 years) usability of data. (2) Metadata shall be EML-compliant at level 2 (a document shall be discoverable; metadata content is well beyond the minimums dictated by the EML schema rules). Metadata should be EML-compliant at level 5 (level 5 is also known as the integration level, when the metadata enables machine readability for associated data). LTER Site EML shall comply with LTER best practices." Note that positive site reviews are critical for continuing funding of the on-going long-term programs at each LTER site.

The EML

EML (Jones et al., 2001) is a comprehensive standard that has been adopted by a sector of the larger international ecological research community. EML has had two major revisions, and the development committee is currently working on a new release. Even though the EML standard enables data integration at the machine level (with little or no human intervention), it is thought to lack adequate information holders for more specialized genomic datasets.

EML is implemented as a series of XML schema document types that can by used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset. EML has four general descriptors at the top of the hierarchy. One can choose to describe a dataset, a protocol, a citation, or software. The EML-dataset module contains general information that describes dataset resources. It is intended to provide overview information about the dataset: broad information such as the title, abstract, keywords, contact information, maintenance history, and distribution of the data themselves. The EML-dataset module also imports many other modules that are used to describe the dataset in fine detail. Specifically, it uses the EML-methods module to describe methodology used in collecting or processing the dataset, the EML-project module to describe the overarching research context and experimental design, the EML-entity module to provide detailed information about the logical structure of the dataset. A dataset often is composed of a series of data entities (tables, shape files, photos, and the like) that are linked together by particular integrity constraints. The EML-literature module contains information that describes literature resources. It is intended to provide overview information about the literature citation, including title, abstract, keywords, and contacts. Citation types include: article, book, chapter, edited book, manuscript, report, thesis, conference proceedings, personal communication, map, generic, audio visual, and presentation. The "generic" citation type would be used when one of the other types will not work. The EML-software module contains general information that describes software resources, while the EML-protocol module is used to define abstract and prescriptive procedures for the dataset.

Other EML modules that are not at the top of the EML hierarchy, but can be used in several contextual places in a dataset are the EML-coverage, EML-data table, EML-spatial raster or vector, EML-physical, and EML-attribute modules.

There are more EML modules, but for the sake of briefness, we shall focus only on these four. The EML-coverage module contains fields for describing the coverage of a resource in terms of time, space, and taxonomy. These coverages (temporal, spatial, and taxonomic) represent the extent of applicability of the resource in those domains. The geographic coverage is expressed via a set of bounding coordinates that define the North, South, East, and West points in a rectangular area, optionally including a bounding altitude, or using a G-Ring polygon definition, where an irregularly shaped area may be defined using a ordered list of latitude/longitude coordinates. The temporal coverage section allows for the definition of either a single date/time, or a range of dates/times expressed according to the ISO 8601 Date and Time Specification. The taxonomic coverage section allows for detailed description of the taxonomic extent of the dataset or resource.

The EML-dataTable module is used to describe each tabular set of information in a dataset. A series of comma or tab-separated text files may be considered a dataset, and each file would subsequently be considered a data table entity within the dataset. Data tables may be ASCII text files, relational database tables, spreadsheets, or other type of tabular data with a fixed logical structure. The EML-dataTable module allows for the description of each attribute (column/field/variable) within the data table through the use of the EML-attribute module. Likewise, there are fields used to describe the physical distribution of the data table, its overall coverage, the methodology used in creating the data, and other logical structure information such as its orientation, case sensitivity, etc. The EML-attribute module describes all attributes (variables) in a data entity: dataTable, spatialRaster, spatialVector, and other EML-entity modules. The description includes the name and definition of each attribute, its domain, definitions of coded values, and other pertinent information, including child modules to describe units, precision, quality controlled procedures, missing codes and the like. An attribute-list encompasses a list of attributes that go together in some logical way, while the attribute structure within the attribute list is used to define a single attribute.

The EML-physical module describes the external and internal physical characteristics of a data object as well as the information required for its distribution, that is, filename, delimiters, encoding methods, header/footer lines, authentication of a file, etc. Distribution information describes how to retrieve the data object. The retrieval information can be either online (e.g., a URL) or offline (e.g., a data object residing on an archival tape) or inline (attached to the metadata document). The EML-spatialRaster and the EML-spatialVector modules allow for the description of entities composed of data values that are usually georeferenced to a portion of the Earth's surface. Specific attributes of a spatial raster and vector can be documented here including the spatial organization of the raster cells, the cell data values, vectors, and the relationships among them.

Specific input into MIGS/MIMS and GCDML

Inigo San Gil contributed directly to discussion at the workshop on finalizing the "Minimum Information about a Genome/Metagenome Sequence" (MIGS/MIMS) specification (Field et al., 2008). He also provided considered suggestions about the nascent GCDML project on behalf of the EML project. Specifically, he expressed concerns about committing to a finite set of data parameters to describe a habitat. A workaround is to leave open the number of parameters, allowing the user to enter the parameter name and definition. He also suggested that GCDML currently lacks a place holder for additional metadata. That is, sometimes the researcher may want to document some details that cannot be captured elsewhere in the specification. A possible alternate solution is to include a place holder for additional information or additional metadata. Other suggestions for aiding the adoption of MIGS/MIMS and GCDML were drafting "best practices" documents, as well as proposing sponsored trainings, and finally, providing assistance in developing of customized GCDML compliant metadata ingestion tools through cooperation.

He also pointed out that here is a potentially synergistic overlap between LTER and GSC activities with respect to their choice of software development projects to enable the ingestion and export of data and metadata. The GSC and LTER communities have independently chosen the same technologies for the task to build metadata capture tools. These XML based tools are centred on Orbeon and XForms (Ajax, or the engines behind Web 2.0). The GSC's GenCat (Field et al., 2006) software takes an XML schema (e.g., originally the MIGS.xsd schema and now GCDML) and autogenerates an online data capture and repository system on the fly. GenCat currently includes Input, Edit, Browse, and Search functions, and offers access to a range of the repository contents through REST-style Web services, as well as consuming external SOAP-based Web services. Similarly, the LTER new *metadata editor project* (Aguilar et al., 2008) provides a full-fledged Web-based editor and entry form tool for metadata. The required future features of this editor will include: compliance with multiple metadata standards [such as EML and Federal Geographic Standard Committee (FGDC) and its Biological Data Profile or (BDP)], autosave feature, content validation, examples, metadata best-practices guidance, authentication system, integration of thesauri and taxonomic Web services, and Google maps mashups. Given the overlap and the use of the same technologies, the LTER and GSC have agreed on exploring the possibility of joining forces and develop either only one tool, or leverage common modules and expertise the projects. LTER also develops crosswalks (a mapping correspondence) between its and metadata standards used by other scientific communities. LTER is exploring a possible crosswalk between the GCS and LTER metadata standards, and some possible options were highlighted at the workshop. Because at this point both implementations are not tuned sufficiently, it seems wiser to wait until both standards evolve to favor a more seamless integration than in its current state.

Roadmap and Achievements

The future: actions and recommendations

The LTER Network endorses the activities of the GSC and will make the following specific contributions to this community. It will help

- support the outreach efforts of the GSC by advocating MIGS/MIMS adoption within the LTERs

- provide GSC guidance on the development and implementation of the GCDML
- recommend enhancements to EML to support genomic data,
- explore LTER collaboration with the GSC on training and;
- continue efforts to build on the similarities between the GSC's GenCat and the LTER's Metadata Editor projects.

The LTER will also increase efforts to reach out to researchers generating genomes and metagenomes. A fundamental ongoing activity initiated to support these goals is an in-depth evaluation of how many LTER research projects involve sequencing genomes. The LTER microbial ecology program that began in 1999 serves as an excellent case study (Microbial LTER Project, 1999). LTER principal investigators associated with the microbial ecology program are being asked to provide some guidance and their own research plans as examples of LTER interest in genomic research. Once we receive feedback from all the LTER microbial projects as well as from others, we will propose to the LTER Information Managers Executive Committee (IMExec) and LTER Network Information System Advisory Committee (NISAC) an extended work plan that goes beyond the collaboration items laid out in this report. NISAC guides the work performed by the Network Information System personnel, a substantial part of the LTER network wide informatics resources.

The LTER also propose draft designs for the extension of EML. The ecological community with the GSC would greatly benefit from using parts of the EML schema to document their reports, as there is substantial support and knowledge in the community. However, EML lacks adequate information placeholders to properly describe genomic and metagenomic data. There are several possibilities for enhancing the EML schema to accommodate this data and benefit from GSC work in this area.

Three key options are being considered:

1. An additional module (incorporation of GCDML) could be proposed for genomic data, parallel to EML's four current modules. This top-hierarchy branch addition is particularly interesting from the point of view of technical schema merging aspects, as well as related tools to edit, ingest or use metadata in an analytical framework.
2. A second possibility would be to expand the existing EML dataset module to accommodate a new entity type for genome sequences, which for portability, would follow the GCDML core schema.
3. A third option could be implemented without changing the current EML schema at all, but instead, would make use of its built-in extensibility and internal references. A GCDML data description can be included as "additional metadata" and its content referenced by an "otherEntity." A similar method is currently used in EML to refer to measurement units described using the Scientific, Technical, and Medical Markup Language (STMML) (Murray-Rust and Rzepa, 2002) schema.

The last option would result in fully compliant EML that contained descriptions of genomic data; however, any machine interpretation would require that standardized practices first be established. We argue that a full integration of the GCDML schema into EML (option 1) is the best long term solution.

The future: integration of ecological and genomic/metagenomic data with other data types

Data integration must extend beyond considerations involving EML and GCDML. There is a clear need in the LTER community for sharing microbial diversity data, whether sequence, metagenome or genome-based. There are many LTER participants doing broad-scale microbial studies, discussed more below, that include a wide range of research outputs in a wide variety of measurement and file formats. Imagine now comparing across these multiple studies; doing so without standardization would prove an exceedingly difficult task. As opposed to forcing researchers into making *a posteriori* comparisons, why not solve the problem from the ground up? The following case studies provide a view on the needs and complexity of LTER-associated microbial diversity projects.

The first case study is Linda Kinkel's (2008) broad-scale microbial studies, which include sequence data, details of organism isolation procedures, HPLC, mass spec profiles, nutrient utilization data, and other phenotypic data. Linda's group is considering a number of data integration tools that would enable them to integrate diverse types of data (such as images, matrices of quantitative and qualitative information, sequences, etc.). The North Temperate Lakes (NTL) LTER site provides another case study that identifies the needs to link environmental datasets and genomics databases (Jacob et al., 2007). The NTL-Microbial Organisms (MO) Environmental Sequence Database effectively link microbial community composition and environmental data from north temperate lake ecosystems. NTL-MO dataset includes community fingerprints, 16S rDNA sequences, phylogenetic assignments, and environmental data collected using high-throughput techniques. The data are stored in a relational database that is fully integrated with the North Temperate Lakes LTER datasets, linking molecular data with ecological data. The NTL microbial laboratory need to develop their own local system to manage the geospatial nature of their microbial genomics research places the lack of an adequate metadata standard in the spotlight. EML, as the LTER Network standard would have to be enhanced to address the work conducted at NTL-MO's microbial research. We feel confident that with an integration of a mature GCDML, the data and projects conducted at NTL-MO will be adequately documented. NTL-MO's McMahon, with the support of her coworkers, has pledged help in developing adequate metadata standards as outlined in this report.

Another case study example is the Alpine Microbial Observatory (AMO), which is associated with the Niwot Ridge LTER. Research at the alpine microbial observatory focuses on studying the diversity and function of soil micro-organisms across extreme environmental gradients in alpine ecosystems. In order to address these questions, AMO researchers take samples from different locations and collect soil biogeochemistry and sequence data, all of which is stored in a locally developed database. The AMO database carries the x, y, z, and t measurements that are core to MIGS/MIMS (Field et al., 2008), and therefore also GCDML

(Kottmann et al., 2008). The sequence and environmental data structures in the MIGS/MIMS specification, and therefore the GCDML schema, have their parallels in AMO database tables—the attributes of georeference, sampling event, sequence, and environmental data can be mapped between AMO and the metadata standards discussed in this paper. Among the elements that AMO's workflow has required are basic technical attributes of data such as descriptors, quantities, units of measure, and error estimates. In addition, methodology, administration, and attribution information is also present in the AMO workflow and research priorities. The AMO database will soon be able to import, export, and query MIGS/MIMS compliant data in GCDML.

Perhaps the best representative case study that reflects the value of the standards and tools sponsored by the GSC community is Linda Amaral's Microbial Inventory Research Across Diverse Aquatic (MIRADA) project (Amaral, 2008). This project proposes to establish a Microbial Biodiversity Survey and Inventory across all the major aquatic (marine and freshwater) LTER sites. MIRADA's proposed Biodiversity Survey and Inventory takes advantage of the aquatic sampling locations that are part of the established LTER network of sites and builds on existing infrastructure for coordination at the Marine Biological Laboratory (MBL) in Woods Hole, Massachussets, under the project International Census of Marine Microbes (ICoMM). MIRADA will adopt ICoMM's massively parallel, 454-based rDNA tag sequencing strategy that allows extensive sampling of both common and rare members of microbial populations, and provides a common metric for integrating studies of microbial diversity across aquatic LTER sites. This strategy, based on sequencing of hypervariable regions of the small subunit ribosomal RNA gene, has the ability to recover large sample sizes (100 to 1000 times the amount of information recovered from a typical clone library survey approach) of all components of the microbial community—*Bacteria*, *Archaea*, and *Eukarya*. This will not only enable cross-site comparisons, but also provide valuable baseline data for integrating population structures with ecosystem change, and understanding microbially mediated trophic dynamics and biogeochemical processes—areas of study already underway at many of the LTERs. Amaral's specific objectives are to:

1. Document and describe both microbial (*Bacteria*, *Archaea*, and *Eukarya*) baseline diversity and relative abundance data for microbial operational taxonomic units (OTUs) as defined by SSU rDNA hypervariable tags at aquatic LTER sites.
2. Determine which microbial OTUs are common to both freshwater and marine LTERs.
3. Determine whether diverse aquatic LTER sites possess "signature" assemblages characterized by space, time, and environmental parameters.
4. Discover novel tag sequences that likely represent novel micro-organisms that LTER researchers and students can further characterize and study.

Some of the projected MIRADA LTERs program measurable impact includes the publication of primary data by the participating LTER partners, and release of data in a variety of formats for the wider community.

The MIRADA project involves designing and developing a custom metadata and data repository for this cross-synthesis project. The MIRADA participants will investigate a way to bridge the content details for the MICROBIS database and the MIGS/MIMS, GCDML, and/or EML reports.

Wade Sheldon developed a comprehensive relational database for managing environmental sequence data and metadata at the Sapelo Island Microbial Observatory (SIMO) in 2001 (Sheldon and Moran, 2001; Sheldon et al., 2002). The database schema was designed to reflect the natural hierarchy that exists among samples and information and materials derived from them and to maintain the complete research context for data stored in the system, including environmental context of samples, field and laboratory methodology, and analytical post-processing. Interactive web applications support querying the database by both environmental and taxonomic characteristics, and retrieving results in standard bioinformatics file formats for analysis in other systems. The SIMO database is also coupled to automated bioinformatics pipelines for classification of 16S rRNA sequences and submission of annotated sequence data to the NCBI GenBank database, including all environmental and research context modifiers supported by NCBI. Sequence records and metadata can be retrieved using either the SIMO ID or GenBank Accession, and GenBank links are displayed on sequence detail pages to support bi-directional queries between both systems. A public version of the SIMO database has also been developed to provide access to the SIMO 16S rRNA classification pipeline and GenBank submission tool, including basic metadata forms for entering environmental, geographic and research context descriptors to accompany the sequence data set.

The success of the SIMO environmental sequence database has inspired similar efforts by other NSF-funded Microbial Observatories, including the North Temperate Lakes MO, Red Layer MO, and Alpine MO. The SIMO database design was also influential in the development of the International Census of Marine Microbes MICROBIS database (Neal et al., 2006). The SIMO database could potentially serve as a prototype for web-based systems that provide useful analytical services as part of the data submission process, and support dynamic generation of GCDML-compliant metadata as a value-added product.

Finally, note that there are many more case studies at LTER sites that we would have included in this short report, and such supplemental information can be found at the GSC WiKi and links therein.

Final message and specific actions

LTER is committed to a number of action items that will bridge the technical gaps it currently faces in capturing (meta) genomic data. By collaborating with GSC, LTER will be able to leverage the efforts in designing a comprehensive metadata standard for genomic and metagenomic data that is currently needed by the ecological genomics community. In addition, LTER brings experience to the table on successfully championing a community standard. We see LTER as a stakeholder in this wide community effort initiated by the GSC, and we feel the importance of being involved in a process that will directly affect how the LTER community plans and conducts site and network-oriented research.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

Aguilar, R., Gries, C., and San Gil, I. (2008). LTER Metadata Editor Project. (http://intranet.lternet.edu/im/project/MetadataEditor).

Amaral, L. (2008). (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0717390).

Field, D., Morrison, N., et al. (2006). Meeting Report: eGenomics: Cataloguing our Complete Genome Collection II. OMICS **10,** 100–104.

Field, D., et al. (2008). Towards a richer description of our complete collection of genomes and metagenomes: the "Minimal Information about a Genome Sequence." Nat Biotechnol **26,** 541–547.

Jacob, C., Brent, A.D., Benson, B.J., Newton, R.J., and McMahon, K.D. (2007). Proc. 9th World Multiconference on Systematics, Cybernetics and Informatics.

Jones, M., et al. (2001). EML—http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html

Kinkel, L. (2008). Spatial Variation, Diversity and Genetic Composition of Microbes in Prairie Soils (http://www.cedarcreek.umn.edu/microbo/index.htm).

Kottmann, R., Gray, T., et al. (2008). A standard MIGS/MIMS compliant XML Schema: Towards the development of the Genomic Contextual Data Markup Language (GCDML). OMICS (this issue).

LTER Information Managers Executive Committee (2005). http://intranet.lternet.edu/im/im_requirements/im_review_criteria

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G. (1997). Nongeospatial metadata for the ecological sciences. Ecol Appl **7,** 330–342.

Microbial LTER Project. (1999). (http://www.lternet.edu/microbial_ecology/).

Murray-Rust, P., and Rzepa, H.S. (2002). STMML. A markup language for scientific, technical, and medical publishing. Data Sci J **1,** 128–192.

San Gil, I., and Baker, K. (2007). Databits. http://intranet.lnternet.edu/archives/documents/Newsletters/DataBits/07fall/#fa2

Sheldon, W.M., Moran, M.A., and Hollibaugh, J.T. (2002). Efforts to link ecological metadata with bacterial gene sequences at the Sapelo Island Microbial Observatory. Pages 402–407 in: Proceedings of the 6th World Multiconference on Systemics, Cybernetics, and Informatics. Information Systems Development II. International Institute of Informatics and Systemics, Orlando, Florida.

Sheldon, W.M., and Moran, M.A. (2001). Sapelo Island Microbial Observatory Environmental Sequence Database (http://simo.marsci.uga.edu/public_db/).

Neal, P., Patterson, D., and Bordenstein, S. (2006). MICROBIS: The ICoMM Marine Microbes Database (http://icomm.mbl.edu/microbis/).

Address reprint requests to:
Inigo San Gil
LTER Network Office
University of New Mexico
MSC03 2020
Albuquerque, NM 87131

E-mail: isangil@lternet.edu