Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research

Peter H. MCCARTNEY
Center for Environmental Studies
Arizona State University
Tempe, AZ 85282, USA

And

Matthew B. JONES
National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara
Santa Barbara, CA, USA

1. ABSTRACT

Metadata provide information on the structure and meaning of data. It is one of the most basic components for building a scalable, networked infrastructure for accessing ecological data. Several partnering groups from ecology have collaborated to define a standardized format for metadata that is machine-parseable and extensible. This has enabled new projects focusing on the development of tools for managing metadata archives and for accessing and processing the datasets they describe. Ecological Metadata Language and its associated tools will have a significant impact on the integration and synthesis of ecological data at a global level.

2. THE ROLE OF METADATA IN ECOLOGICAL INFORMATICS

The goals of ecological informatics are to ensure the long-term availability of ecological data and to enhance the usability of those data in the pursuit of knowledge about our environment. The use of digital media to capture, store, and process increasingly larger volumes of data has contributed significantly to these goals, but this has in turn created new challenges for indexing, navigating and documenting this sudden wealth of information[1].

A critical tool for meeting this challenge is *metadata.* Metadata is the documentation that transforms data from a stream of numbers and characters into information. All of us who work with data have relied upon metadata such as column labels, data type declarations, etc., even if we didn't recognize those things by that name. Metadata provides information at many levels to support many phases of our interaction with ecological data. Information such as catalog identifiers, title, originator, etc. provides the base citation information for *identifying* a dataset. Search engines rely on keywords and coverage descriptors for spatial, temporal, or thematic domains to assist with the *discovery* of datasets. Information on the research context that produced the data assists in the *evaluation* of the dataset. Connection details, filenames, and access control information enable *acquisition* of a dataset. Finally, detailed descriptions of entities, attributes, and data quality enhance the *usability* of the dataset for analysis.

Members of the ecological research community have been compiling metadata as part of the data archive process for over a decade. Notable examples include the Long -Term Ecological Research network [2] and the Oak Ridge National Labs[3]. In 1997, following Michener's paper on ecologically relevant metadata [4], researchers at the National Center for Ecological Analysis and Synthesis (NCEAS) began implementing the first version of Ecological Metadata Language (EML), which was revised several times and culminated in EML version 1.4.1 [5]. As experience with this initial version of EML grew, it became apparent that it needed a revision to increase its usability and flexibility for the ecological community. The Knowledge Network for Biocomplexity (KNB) project thus began an effort to revise the EML specification to produce a second version that was

even more broadly useful. Simultaneously, in summer 1999, the LTER information management committee evaluated the status of metadata within the LTER network in light of a series of long term goals for the future of informatics in ecology (see Brunt et al this volume). The committee found that (1) there was a need for standardization in both content and presentation format, and (2) that metadata needed to be presented in a machine-parseable form to support advanced development of automated data search and processing tools [6]. As a result, a metadata committee was formed to work with the two independently funded projects (Knowledge Network for Biocomplexity [7] and Arizona State University's Networking our Research Legacy project [8]) that had begun the process of revising Ecological Metadata Language (EML).

3. ECOLOGICAL METADATA LANGUAGE

Development of EML has followed several guiding principles. (1) It should be encoded in a machine-parseable format, with strong industry support and independence from particular platforms or software. (2) Extensive prior work in metadata standards both within and outside ecology should be used as a basis to enhance compatibility and reduce redundancy. (3) The standard should serve to integrate, rather than dictate, individual site solutions for creating, storing and managing metadata.

eXtensible Markup Language (XML) was selected for the encoding format XML is an SGML-based text syntax (UNICODE) for marking up data and documents. It bears similarities with HTML, but is designed for tagging the content of a document with a means for validating that content against a formal schema. Tools for parsing XML documents are available for all modern development languages and XML documents are easily transformed into other formats for display through the related eXtensible Stylesheet Language (XSL) specification. The XML Schema specification is itself an XML file and provides a powerful medium for designing and sharing content models through the use of commercial design tools or custom XSL style sheets.

A significant amount of prior research was reviewed in designing EML. Within the ecological community, the seminal paper by Michener et al. [4] had established guidelines for metadata content that was reflected in the text and HTML formats designed by various individual LTER sites. NCEAS encoded the content model developed by Michener et al. in XML in what was first released as the EML 1.0 specification. Outside ecology, extensive work on geospatial metadata standards by the Federal Geographic Data Commission (FGDC)[9] and the International Standards Organization (ISO)[10] resulted in comprehensive content models released as text and Universal Modeling language (UML) specifications respectively. The National Biological Information Infrastructure (NBII) extended the FGDC standard to accommodate biological datasets [11]. The resulting NBII standard adopted substantial portions of the original EML version 1.0 specification. Other standards such as the Dublin Core Element Set [12] for internet resources, the Global Change Master Directory DIF standard, and the Mercury metadata standard used by Oak Ridge National Labs [13] were also reviewed.

Considerable diversity existed across the 24 LTER network sites in terms of the content and format of metadata, and in the manner in which metadata catalogs were integrated into other aspects of site management. The goal in creating EML was to define a common standard and format that could be generated easily from existing metadata without burdening sites with significant alteration of their existing system.

EML 2.0 Design

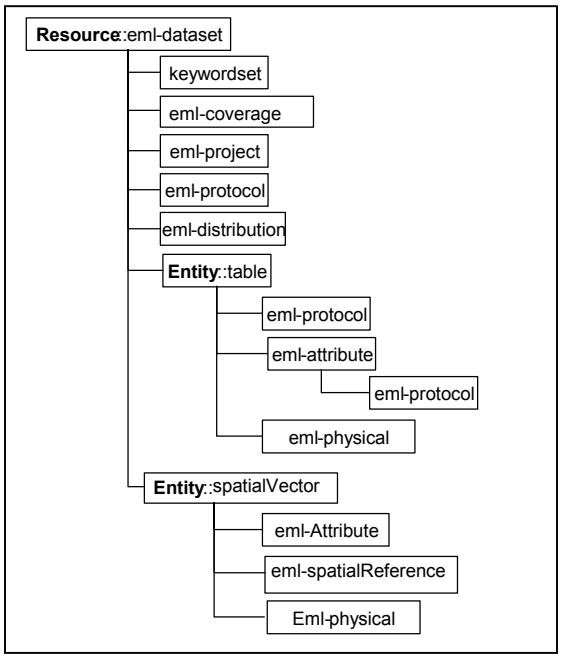
The resulting draft specification for EML 2.0 is a complete revision of the original EML 1.0. Detailed information on its development and downloads of draft specifications are available online[14]. EML 2.0 has several significant design features. (1) It is modular, with separate schemas defining sets of descriptors that relate to a specific category of information. (2) It uses XML Schema complex types to enable an object-oriented approach in which abstract classes are defined and then extended to create specific variants. Using this approach, EML defines several information resource types including "dataset" and "literature" (and potentially many more) that each inherit a common set of elements that correspond to the basic identification and discovery elements found in most metadata standards. (3) It is extensible by linking multiple modules within a package. Features derived from XML and form Resource Description Framework (RDF) allow subject-object relations to be defined between metadata document without modification of existing module schemas. (4) EML modules are organized to separate the description of the *logical* content of an information resource from that of its *physical*

instance. This feature automatically abstracts the details of physical formatting from users, allowing them to focus on the information itself and simplifies the maintenance of metadata, as disk formats or storage locations change through time.

Overview of content models for EML dataset
The super class for all EML documents is Resource. This set of elements defines those identifier and discovery elements that are common to any information resource and is based closely on the Dublin Core metadata standard. Resource is never directly instantiated – it is extended by several schemas including eml-dataset, eml-literature, and eml-software. Still others may be defined such as eml-model, or eml-collection.

Eml-dataset introduces several elements for describing a dataset and serves as the association point for a series of modules used in defining certain types of data or properties of data. Eml-project provides information on the research context that produced the data. A dataset is associated with one or more entities, each of which is described with a module that is extended from a basic Entity class. These may include tables, gislayers, images, grids, views, or stored procedures. Depending on the type of entity, other modules (such as attributes, constraints, spatial reference, spatial organization, data-quality) may be associated. The information provided in the entity and associated modules focuses on the logical information of the data. A related module, eml-physical, provides the descriptions of the actual digital instance of that entity (such as file information, connection information, column parsing instructions, etc). Changes in the format or location of a file can be made without altering the logical description represented in the entity section.

Other modules (such as protocol or responsible party) may be associated with several different modules whenever a particular class of information is appropriate. These modules define a consistent structure for specific kinds of information that could potentially apply in many different contexts within a metadata document. An EML metadata package would consist of one document such as eml-dataset or eml-literature, plus any other associated modules, and optionally the data objects as well. Resolving the linkages between modules specified by the triple statements would yield a nested tree of documents that can be easily traversed to locate any given element of information (Figure 1).



4. APPLICATIONS

The most significant feature of any metadata standard is the advances it enables for data discovery, access, and analysis. The projects responsible for developing EML have been working simultaneously on several software products that will facilitate access and use of environmental data.

Metadata creation
One of the most limiting obstacles to building networked data archives is getting past the learning curve and time burden of filling in metadata descriptions. EML is a fairly complex set of elements numbering in the 100's, many of which are not applicable for any given dataset. Two similar products are being developed to provide a simpler interface that would encourage scientists to prepare metadata without needing to either learn an entire management system or hire a data manager. *Morpho* is a Java-based metadata management tool developed by the KNB project. Building on an earlier XML editor developed by NCEAS[15], Morpho combines a user friendly forms environment for editing EML documents with a management client for submitting, maintaining and searching metadata packages on a networked storage system. Extensive configuration enables Morpho to accommodate changes or extensions to the EML schemas without requiring modification of the program code. Morpho includes a reverse engineering module for interactively walking users through the documentation of ASCII data files by parsing the file and allowing the user to view,

change, or provide more detail on the results. ASU is developing a related project called Xylographa which will be a web-based application consisting of three main components 1) a collection of reverse engineering modules running as either web services or Java applets (currently a relational database module is completed), 2) an import utility for parsing and importing other metadata formats into EML via XSLT style sheets or Java servlets, and 3) an interview wizard that walks a user through the metadata entry process in a step by step manner that provides navigation guides and access to contextual help. A metaphor for the design of Xylographa is modern tax software, such as *Taxcut* and *Turbotax* that use an interview mode and automated retrieval of information from related documents.

Metadata management

XML's flexibility and extensibility pose new challenges for metadata storage and query. Standardization of format allows search expressions to be constructed using a universal syntax, sparing users to log on to specific archive web pages or learn a specific catalog's query syntax.

The KNB project has developed a native XML database storage system, *Metacat*, that decomposes the XML document object model into a linked list of node descriptions that are stored as individual records in a relational database [16, 17]. The *Metacat* servlet receives queries expressed in Xpath syntax and transforms them into SQL statements to locate matching nodes. The relational pointers are then resolved back up to the document's root so that Metacat can return the entire document. Metacat forms the basis for what KNB hopes will be a national network of replicated metadata servers that can be queried by Morpho or other client applications using an XML messaging format to encode Xpath search expressions and subsequent XML responses. For sites or individual researchers without an existing metadata management system, the Metacat/Morpho combination is an ideal solution as it is a complete system that is already configured to participate in a national network (Figure 2).

Xanthoria is a project developed by ASU to provide a threaded client-server search system for querying multiple, heterogeneous XML data catalogs. The goal in developing Xanthoria was to build an easily configured solution to provide an EML interface to existing SQL-based metadata catalogs. Xanthoria works very similar to Z39.50, a text search system used in many library networks,

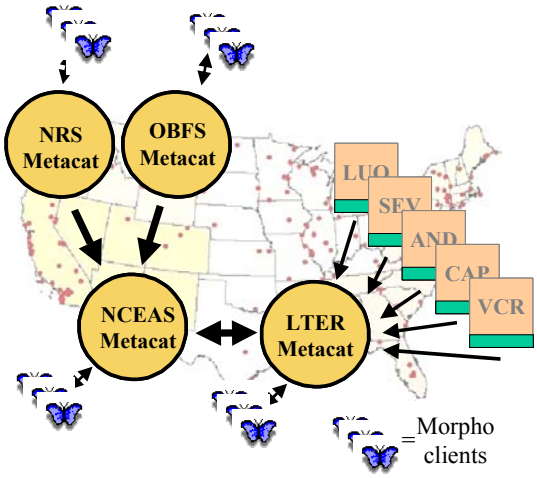


Figure 2. National network of Metacat servers.

but is based on XML and Xpath. Xanthoria services connect to several types of storage systems including SQL databases, XML file folders, Xindice XML database, and Metacat. In the case of the SQL connector, a user-configurable Java bean performs the SQL to XML translation. In all connectors, differences in content schema are handled by user-supplied XSLT stylsheets that translate the native schema to EML. Each connector runs as a web service, listening for Simple Object Access Protocol (SOAP) requests from a client application (Figure 3). The structure of these requests is an extension of the XML messaging format used by Morpho to communicate with Metacat. The query application communicates with the targets, and collates and paginates the results for the user. The query form for generating the search expression is generated from the schema itself and can thus accommodate search on any XML target for which an XML Schema file has been provided. It uses an external configuration file identifying available targets and the schemas

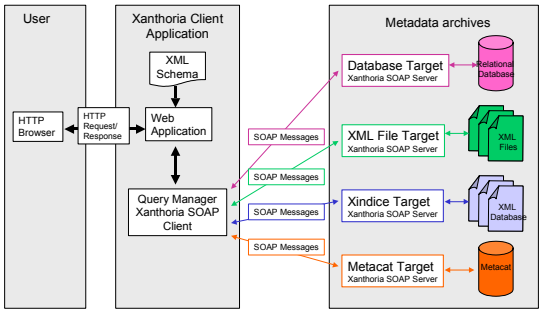


Figure 3. Xanthoria Web Service for XML Metadata.

supported. This configuration file also provides a map to the schema hierarchy so that the client is aware that a given target can support queries based on its own schema as well as on a more generic schema from which it was extended. For example, the client will send queries based on EML-Resource to both dataset and literature targets, but queries based on EML-Dataset will be sent only to targets for that schema.

Processing and Analysis

Impacts that more directly affect researchers are to be made by other current projects that seek to make use of machine-parsable metadata to enable more automated processing and analysis of data. At ASU, a collection of web-based data access and processing tools are being created to provide users with a richer array of exploratory and download tools than is currently available in most data clearinghouse systems. Basic GIS operations such as resample, reproject, or clip can be requested before data are downloaded. Tabular data will be queryable via a JDBC connection using a wizard that helps users construct SQL queries. Exploratory Data Analysis (EDA) functions (such as charts, plots, and cross-tabluations) will be available as an enhanced data browsing package.

ASU is also working on applications that build upon this basic data access infrastructure to target specific user groups. One example is a biodiversity applicaton for the Ecology Explorers educational program at CAP LTER. In this application, a series of XML configuration files will be used to provide a map between the EML descriptions of several datasets and some fundamental parameters that K-12 users will explore through this guided application. Users will be able to analyze parameters such as species richness using a choice of input data such as birds, arthropods, vegetation without needing to understand how to extract that particular query from the different source databases.

The KNB project is working on a project called *Monarch* that provides an exploratory data analysis and modeling environment for data described by EML metadata. Monarch uses an XML configuration system to describe analyses and models that are implemented in commonly used analytic tools (such as SAS, Matlab, etc). Monarch uses the information from an EML dataset description to generate the appropriate command scripts for a particular analysis and then executes the analysis using a plug-in architecture for the target execution environment. As a result, any data that is accessible and described in an EML format can be automatically analyzed over the web using these powerful statistical packages, which dramatically speeds up the process of understanding and interpreting data in synthetic and collaborative analyses. Monarch is expected to be a very useful technology for network developers to provide distributed access to common data processing and analysis functions.

5. FUTURE DIRECTIONS

The applications described above illustrate some of the initial efforts to draw upon the power of standardized, machine-readable metadata. Within them, several common themes point to future directions for informatics development.

One apparent goal is the development of integrated, networked applications that provide users with access to the full range of analytic functions without the need to install or learn specialized statistical or GIS software. This not only benefits many researchers, it has a profound effect on our ability to make ecological data available to a broader community including educational users, policy makers, and the general public. Standardized metadata, combined with online network access to data, will enable many applications to be constructed for the same data sources, each targeting a specific kind of audience.

Another clear trend is the abstraction from the physical details of data organization, encouraging the user to express their analytic requests in a syntax that is much closer to the logical content of the data. File formats and storage solutions are constantly evolving. One of the functions of metadata should be to provide the linkage between physical storage and information in a manner that frees the user from tracking changes or details within the physical component. Future research aims at higher levels of abstraction still. While EML provides a consistent *syntax* for addressing datasets, it does little at the present to overcome the *semantic* differences between datasets. New goals for metadata research will turn to ontology-based solutions for linking the EML descriptions of data to inquiry-based concepts that come closer still to the parameters by which we define ecological knowledge.

6. CONCLUSIONS

Standardized metadata is a significant step forward in ecological informatics. It provides the means for cataloging the growing base of data archives and

for addressing these data through a common syntax. This in turn is leading to the development of much more versatile applications that enable users to contribute to, navigate, and make use of, networked archives of ecological data.

REFERENCES

[1] Nature. 1999. It's sink or swim as a tidal wave of data approaches. Nature 399:517-520.

[2] Long Term Ecological Research. http://www.lternet.edu.

[3] Olson, R., and R. A. McCord. 1998. Data archival. Pp. 53-58 in W. K. Michener, J. H. Porter, and S. G. Stafford, eds., Data and information management in the ecological sciences: A resource guide. Long-Term Ecological Research Network Office, Albuquerque, NM.

[4] Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7(1):330-342.

[5] Nottrott, R., M.B. Jones, and M. Schildhauer, 1999. Using XML-structured metadata to automate quality assurance processing for ecological data, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD, April 6-7, 1999.

[6] McCartney 2000. Report of the Long-Term Ecological Research Metadata Committee Meeting, February 2000, NET Office, Albuquerque. [http://caplter.asu.edu/data/metadata/report2k.doc]

[7] KNB, 1999. The Knowledge Network for Biocomplexity. [http://knb.ecoinformatics.org/]

[8] NRL, 2000. Networking our Research Legacy. http://ces.asu.edu/bdi

[9] FGDC 1998. Content Standard for Digital Geospatial Metadata. Federal Geographic Data Committee.

[10] International Standards Organization, 1999. CD 19115, Geographic information – Metadata. Norwegian Technology Standards Institution, Oslo, Norway.

[11] Frondorf, A., M.B. Jones, and S. Stitt, 1999. Linking the FGDC geospatial metadata content standard to the biological/ecological sciences, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD. April 6-7, 1999

[12] Dublin Core Metadata Initiative. [http://dublincore.org/].

[13] Olson, R., L. Voorhees, J. Field, and M. Gentry. 1996. Packaging and distributing ecological data from multisite studies. Pp. 93-102 in Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November 1996, Lake Buena Vista, FL. Environmental Research Institute of Michigan, Ann Arbor.

[14] Ecoinformatics.org [http://www.ecoinformatics.org]

[15] Nottrott, R., M.B. Jones, and M. Schildhauer, 1999. Using XML-structured metadata to automate quality assurance processing for ecological data, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD, April 6-7, 1999.

[16] Berkley, C., M.B. Jones, J. Bojilova, and D. Higgins, 2001. Metacat: a Schema-Independent XML Database System, 13th International Conference on Scientific and Statistical Database Management, IEEE Computer Society.

[17] Jones, M.B., C. Berkley, J. Bojilova, and M. Schildhauer, 2001. Managing Scientific Metadata, IEEE Internet Computing 5(5): 59-68.