available at www.sciencedirect.com

www.elsevier.com/locate/ecolinf

Meta-information concepts for ecological data management

William K. Michener

University of New Mexico, Department of Biology, Albuquerque, NM 87131-0001, United States

ARTICLE INFO

Article history:
Received 15 August 2005
Accepted 15 August 2005

ABSTRACT

Ecological databases continue to grow in volume, breadth and complexity. Higher level descriptions of data (i.e., metadata) and information derived from subsequent data processing and analyses (i.e., "meta-information" in the broadest sense) are essential for understanding and using the increasingly complex and voluminous data and information. The concepts of meta-information, in general, and metadata, in particular, have evolved in concert with the increasing needs for functionality by the community. From a scientific perspective, metadata may be characterized as having developed from initially supporting data discovery; to facilitating acquisition, comprehension and utilization of data by humans; and, most recently, to beginning to enable automated data discovery, ingestion, processing and analysis via metadata-enabled scientific workflow systems. The continued conceptual and operational developments in metadata required to support comprehensive automated scientific workflow systems portend many challenges and opportunities. For example, there are significant opportunities for collaboration among ecologists and computer scientists in developing domain-specific controlled vocabularies and ontologies that provide the basis for semantic mediation—the "glue" technologies that enable automated data discovery, ingestion, processing and analysis. Similarly, there are opportunities for computer scientists and engineers to develop new mechanisms that support automated metadata encoding—such as providing the information that would be necessary to understand the end-to-end flow of sensor data from in situ data collection, streaming through quality assurance filtering, aggregation, transformation and additional processing, analysis, and publication of digital products. As the technologies mature, we still have many sociological barriers to overcome including the needs for increased attention to software usability testing and engineering to enhance user-friendliness of metadata management software, new capital investments in ecological data archives, and increasing the metadata management benefit–cost ratio for the average scientist via incentives and enabling tools.

1. Introduction

Ecology is defined as the study of organisms in relation to their environment, including the many possible interactions among organisms (e.g., predation, parasitism, and predation). Ecological data encompass the biological, chemical, physical and social sciences and many of their sub-disciplines. Because of the complex interactions among organisms and between organisms and their environment, ecology has evolved as a science and increasingly addresses questions at broader spatial and temporal scales, and at multiple scales of biological organization—i.e., the full range of "biocomplexity" (Michener et al., 2001). Expansion of the depth, breadth and complexity of ecology has been accompanied by significant changes in the types, magnitude and complexity of data that are acquired and analyzed by ecologists.

E-mail address: wmichener@LTERnet.edu.

1574-9541/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.ecoinf.2005.08.004

The term "meta" is commonly used to denote a higher level description. Thus, metadata refers to a higher level description of data and meta-analysis indicates a higher lever description or synthetic analysis of multiple analyses. Over the past two decades, there has been an increased awareness of the importance of such meta-information about ecological data, information and analyses. Much of the recent focus has been on developing metadata standards and associated software to facilitate management and understanding of the large, diverse and complex data collected as part of the ecological scientific enterprise. In this paper, I define metadata and summarize how the meta-information concept has evolved operationally over the past two decades. Finally, I present some of the challenges and opportunities associated with further enriching ecological meta-information and provide a vision for future meta-information.

2. Metadata—a definition

Metadata may be defined as "information about data"—i.e., the information required to understand data, including data set contents, context, quality, structure, and accessibility (Michener et al., 1997). In short, metadata describe the "*who*, *what*, *when*, *where*, and *how*" about every aspect of the data.

Metadata benefit science in many ways (Michener, 2000; Scurlock et al., 2002a). First, data longevity is increased. Comprehensive metadata counteract the natural tendency for data to degrade in information content through time (i.e. information entropy *sensu* Michener et al., 1997; Fig. 1). This is particularly important for long-term studies where the database outlives the original investigator or where data are collected by scientists from many disciplines over a broad area, requiring considerable data integration and synthesis. Second, data reuse by the originator and data sharing with others are facilitated. Scientists often find that a data set they previously collected for a specific purpose can be reused to answer new questions. Sufficient documentation of sampling and analytical procedures, data quality, and data set structure are necessary so that the data can be correctly interpreted or reinterpreted. Third, well-documented data may be used to expand the scale of ecological inquiry. Examples include data sets from short-term studies that evolve into long-term data sets (Magnuson, 1990) and data sets that are used to address unanticipated questions.

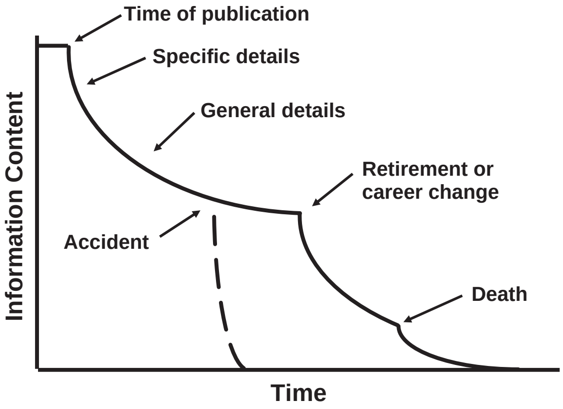

Fig. 1 – Illustration of the natural degradation in information content associated with data and metadata—information entropy (from Michener et al., 1987, by permission of the Ecological Society of America).

3. Evolution of the metadata concept

From an operational perspective, the concept of "metadata" has evolved to include support for the increased functionality needed by the scientific community. There are at least three levels of increasing metadata functionality that may be easily categorized from a scientific perspective:

(1) support data discovery;
(2) facilitate acquisition, comprehension and utilization of data by humans; and
(3) enable automated data discovery, ingestion, processing and analysis.

Data discovery is the most basic level of functionality that metadata can support. For most studies, a scientist will first be interested in ascertaining whether pertinent data already exist. Prior to the 1990s, such data discovery was frequently accomplished via word-of-mouth (e.g., presentations at meetings, information exchange with professional colleagues) or the "Methods and materials" section of publications. As research organizations evolved and the ecological sciences community expanded, it sometimes became necessary to "catalog" the various databases that were being collected and maintained within organizations and networks of organizations. For instance, a data set catalog was created and published in 1990 for the United States Long-Term Ecological Research (LTER) Program to provide basic metadata (e.g., title, data originator, abstract, spatial and temporal context, keywords, contact information and usage constraints) for the bourgeoning number of LTER data sets (Michener et al., 1990). This publication provided an initial mechanism for supporting discovery of LTER data and served as the basis for future automated approaches.

Advances in database and web-based technologies enabled traditional published data catalogs to be replaced by electronic and searchable data catalogs and data directories (e.g., Kanciruk et al., 1999). Presently, many such databases that support data discovery are available through the web and provide numerous searching options (e.g., keyword, data, and location). Examples from the United States include NASA's Global Change Master Directory and the USGS National Biological Information Infrastructure. These databases provide either a controlled vocabulary or a thesaurus that facilitates the standardization of keywords and subsequent data discovery. Many scientists now routinely use commercial search engines for their searches, including discovery of ecological data, although data directories and data catalogs provide important functionality and many offer value-added options that are not available through commercial search engines.

Manual data acquisition, comprehension and utilization require much more metadata than is needed to support data discovery alone. In addition to basic data set descriptors that may be required for a data directory or data catalog (e.g., data set title, associated scientists, abstract, and keywords) a scientist needs access to all relevant metadata that relate to:

(1) the research context (e.g., hypotheses, site characteristics, experimental design, and research methods);
(2) the status of the data set and information related to data accessibility;
(3) the physical structure of the data; and
(4) facilitating comprehension and utilization, including relevant data citations (Michener et al., 1997).

Content standards for ecological metadata have been relatively slow to emerge and be broadly adopted, although this trend is changing with the increased recognition of their value. Possibly the earliest listing of metadata descriptors for ecological data was published in 1987 and included 30 unique parameters that encompassed facets related to the research context, accessibility, and physical structure of data (Michener et al., 1987). This was followed by two related activities associated with scientists in the United States LTER Network: (1) the aforementioned LTER core data set catalog (Michener et al., 1990) and (2) an effort to establish LTER documentation standards (Kirchner et al., 1995).

The next comprehensive set of metadata descriptors for ecology was published a decade later and included 64 unique metadata descriptors (Michener et al., 1997). These metadata descriptors represented a revised listing that was derived from a seminal research effort (Michener et al., 1995) supported by the Ecological Society of America's Committee on the Future of Long-term Ecological Data (FLED, chaired by Dr. Katherine Gross)—the goal of which was to better preserve valuable ecological data well beyond the lifetime of the data originator. The ecological metadata descriptors have served as the standard for data papers submitted to *Ecological Archives* (the Ecological Society of America's electronic journal that publishes data and other materials that supplement the print journals).

It is noteworthy that the activities of the FLED committee and associated publications coincided with similar efforts by the United States Federal Geographic Data Committee to establish metadata standards for geospatial data (Federal Geographic Data Committee, 1994, 1998)—efforts that preceded the development of international metadata standards for geospatial data (Technical Committee ISO/TC 211, 2003). In a related effort, the USGS NBII added numerous metadata descriptors to enhance the value of the geospatial metadata content standards for the biological and ecological sciences (FGDC Biological Data Working Group and USGS Biological Resources Division, 1999; Frondorf et al., 1999).

Automated data discovery, ingestion, processing and analysis require comprehensive and structured metadata. Ecological Metadata Language (EML) is a comprehensive metadata standard that is particularly applicable for a broad range of ecological data and is sufficiently structured to support many of the automated functions listed above (Jones et al., 2001, Fegraus et al., 2005). EML is organized into a suite of modules (e.g., "dataset", "access", "physical", "party coverage", "project", "methods", "physical", "attribute", "datatable"). The modules significantly enhanced the granularity and expanded the number of ecological metadata descriptors previously identified (Michener et al., 1997), as well as those descriptors included in the USGS Biological Data Profile (FGDC Biological Data Working Group and USGS Biological Resources Division, 1999) and other relevant standards. The modules include descriptors that support:

(1) data discovery (e.g., theme and spatial, temporary and taxonomic domains of the data set and its accessibility),
(2) interpretation and appropriate use (e.g., research objectives, experimental design, sampling procedures, site selection, parameter descriptions and units, and data processing [semantics]), and
(3) automated use (e.g., structural attributes of the data [schema] and format of the data [syntax]).

EML is implemented in XML (eXtensible Markup Language), which defines the structure of the text file that contains the metadata and is machine-parsable. EML is, therefore, well suited for research applications ranging from simple data discovery to advanced data processing and can be modified to meet domain-specific needs. A variety of software tools have been developed to support EML-compliant metadata entry and management (i.e., Morpho and Metacat; see Fegraus et al., 2005). More information on EML and associated metadata tools can be found at www.ecoinformatics.org, a community-based resource for ecoinformatics information and software.

EML provides a framework for scientists to encapsulate rich semantic descriptions of data in their metadata, including, for example, units of measurement, sampling area, and precision. Software can then be encoded to ingest semantically enriched data and to automatically perform various transformational and analytical steps such as converting soil temperatures in Fahrenheit to Celsius or calculating density of organisms per square meter based on the number of organisms in a sampling quadrat and the size of the quadrat. Such capabilities represent a subset of the functionality offered by comprehensive metadata-enabled scientific workflow systems. For example, the Kepler workflow system is being developed to support automated data ingestion, data transformation, and analyses, as well as capture and retain the details of the end-to-end procedure, thereby enabling reproducibility and reusability (Michener et al., 2005; Pennington and Michener, 2005). A principal objective of the Kepler workflow system is to create a smart analytical environment whereby most routine data processing steps including data discovery and ingestion, data transformation steps, quality assurance and quality control, as well as many analyses can be largely automated, thereby freeing the scientist to perform other activities.

4. Challenges and opportunities

The automation of data discovery, ingestion, processing and analysis afforded by enriched and structured metadata, coupled with scientific workflow systems will transform

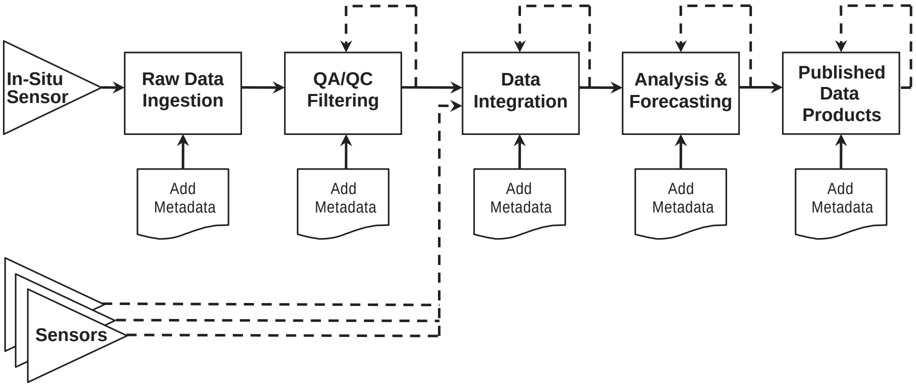

Fig. 2 – Example of end-to-end flow of in situ environmental sensor data from ingestion through quality assurance/quality control (QA/QC) filtering, integration with additional data streams, analysis and modeling, and publication of data products. Metadata may be added at every step of the process and each major step from QA/QC through publication of data products may be performed multiple times.

ecology. One can easily envision a future whereby data, metadata and scientific workflows are integrated into information products that are transferred in their entirety to interested parties and, even, routinely associated with peer-reviewed publications to facilitate verification, enable scientific use by others, and serve education needs. Such a future will afford many opportunities, but will also entail costs.

In the short-term, there is much enabling research that is needed to support the continued development of comprehensive workflow solutions. First, there are significant opportunities for ecologists to work collaboratively with computer scientists in developing the controlled vocabularies and ontologies that provide the basis for semantic mediation—the technologies that underlie smart data processing and scientific workflow systems. Second, there are many opportunities for computer scientists and engineers to develop new concepts and mechanisms that support automated metadata capture. Fig. 2 illustrates an end-to-end multi-step scientific process whereby sensor data are ingested, filtered, transformed, analyzed and converted into publishable information products, each step of the process being accompanied by the automated capture and encoding of relevant metadata. Many of the steps are iterative and highlight the fact that metadata acquisition can be a dynamic, long-term process, especially when the data are required for long-term ecological hypothesis testing, forecasting and prediction.

It has been previously argued that a useful goal for metadata is for it to enable comprehension and use of data by those other than the data collector for a minimum of 20 years after the data were archived, based solely on the information provided in the metadata (NRC, 1991). Such a goal may now seem too conservative as long-term research programs are unraveling multi-decadal phenomena (Magnuson, 1990). Regardless, most scientists recognize that incomplete and inadequate metadata are a significant technical barrier to data integration and analysis efforts (Hale et al., 2003; for best practices, see Michener, 2000 and Cook et al., 2000). Such recognition is often taken to mean that “more metadata is always better.” On one hand, it is true that functionality improves with increasing quantity, quality and structure of metadata. Conversely, though, more, higher quality and better structured metadata are costly, especially in relation to personnel time. For example, academia does not adequately reward the time and energy required to comprehensively document a database or information product, preferring instead to assign a much higher value to funded grants and peer-reviewed publications. In addition, there are very few public archives where scientists can deposit data and metadata (Olson and McCord, 2000; Scurlock et al., 2002b) and the software tools that support metadata acquisition and management have been developed by sophisticated computer programmers for a technically adept user group and are, arguably, not especially user-friendly. Clearly, every challenge can also represent an opportunity and these last two challenges argue for more attention by funding agencies and the scientific community to support ecological data archives and to support the software usability engineering and testing that are normally required for good commercial-off-the-shelf software.

There are many scientific benefits to be gained from continued evolution and enrichment of metadata and the meta-information concept. We must, however, work to provide more compelling incentives like automated scientific workflow systems that benefit scientific productivity, instill a deeper appreciation for the value of meta-information via education, and ease the burden associated with providing data and metadata. In doing so, it may be useful to recognize that frequently “more and better” can be the enemy of the “good enough”—meaning that mandates for more comprehensive metadata may disenfranchise scientists unless such mandates are accompanied by real and tangible benefits (Porter and Callahan, 1994; Callahan et al., 1996) and unless significant attention is given to easing the burden of providing metadata (e.g., enhanced software usability, increased automation of metadata encoding).

Acknowledgments

Much of the material presented here is based on support from the National Science Foundation (NSF) under grants DEB-0236154, ITR-0225665, DBI-0129792, and ATM-0332827.

Additional support came from DARPA (N00014-03-1-0900) and the Andrew Mellon Foundation.

REFERENCES

Callahan, S., Johnson, D., Shelley, P., 1996. Dataset publishing—a means to motivate metadata entry. Proceedings of the first IEEE metadata conference. IEEE Computer Society, Silver Spring, MD, USA. http://www.computer.org/conferen/proceed/meta96.

Cook, R.B., Olson, R.J., Kanciruk, P., Hook, L.A., 2000. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82, 138–141.

Federal Geographic Data Committee, 1994. Content standards for digital geospatial metadata (June 8). Federal Geographic Data Committee, Washington, D.C., USA.

Federal Geographic Data Committee, 1998. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee, Washington, D.C., USA.

Federal Geographic Data Committee Biological Data Working Group and USGS Biological Resources Division, 1999. Content Standard for Digital Geospatial Metadata - Biological Data Profile, FGDC-STD-001.1-1999, Federal Geographic Data Commitee. Washington, D.C.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86 (3), 158–168.

Frondorf, A., Jones, M.B., Stitt, S., 1999. Linking the FGDC geospatial metadata content standard to the biological/ecological sciences. Proceedings of the Third IEEE Computer Society Metadata Conference, Bethesda, MD, http://computer.org/proceedings/meta/1999/.

Hale, S.S., Miglarese, A.H., Bradley, M.P., Belton, T.J., Cooper, L.D., Frame, M.T., Friel, C.A., Harwell, L.M., King, R.E., Michener, W.K., Nicolson, D.T., Peterjohn, B.G., 2003. Managing troubled data: coastal data partnerships smooth data integration. Environmental Monitoring and Assessment 81, 133–148.

Jones, M.B., Berkley, C., Bojilova, J., Schildhauer, M., 2001. Managing scientific metadata. IEEE Internet Computing 5 (5), 59–68.

Kanciruk, P., Gentry, M., Rhyne, T., 1999. MERCURY—a web-based metadata search and data retrieval system. EOGEO99: Earth Observation (EO) and Geo-Spatial (GEO) web and internet workshop '99. Committee on Earth Observation Satellites (http://webtech.ceos.org/).

Kirchner, T., Chinn, H., Henshaw, D., Porter, J., 1995. Documentation standards for data exchange. In: Ingersoll, R., Brunt, J. (Eds.), Proceedings of the 1994 LTER Data Management Workshop. Long-Term Ecological Research Network Office, University of Washington, Seattle, Washington, USA, pp. 5–8.

Magnuson, J.J., 1990. Long-term ecological research and the invisible present. BioScience 40, 495–501.

Michener, W.K., 2000. Metadata. In: Michener, W.K., Brunt, J.W. (Eds.), Ecological Data: Design, Management and Processing. Blackwell Science, Oxford, pp. 92–116.

Michener, W.K., Feller, R.J., Edwards, D.G., 1987. Development, management, and analysis of a long-term ecological research information base: example for marine macrobenthos. In: Boyle, T.P. (Ed.), New Approaches to Monitoring Aquatic Ecosystems. ASTM STP, vol. 940. American Society for Testing and Materials, Philadelphia, pp. 173–188.

Michener, W.K., Miller, A.B, Nottrott, R., 1990. Long Term Ecological Research Core Data Set Catalog. Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC.

Michener, W.K., Brunt, J.W., Helly, J., Kirchner, T.B., Stafford, S., 1995. Demystifying metadata. In: Gross, K.L., Pake, C.E., Allen, E., Bledsoe, C., Colwell, R., Dayton, P., Dethier, M., Helly, J., Holt, R., Morin, N., Michener, W., Pickett, S.T.A., Stafford, S. (Eds.), Final Report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED): Volume I. Text of the Report, Ecological Society of America, Washington, DC, pp. 40–62.

Michener, W.K., Brunt, J.W., Helly, J., Kirchner, T.B., Stafford, S.G., 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7, 330–342.

Michener, W.K., Baerwald, T.J., Firth, P., Palmer, M.A., Rosenberger, J.L., Sandlin, E.A., Zimmerman, H., 2001. Defining and unraveling biocomplexity. BioScience 51, 1018–1023.

Michener, W., Beach, J., Bowers, S., Downey, L., Jones, M., Ludascher, B., Pennington, D., Rajasekar, A., Romanello, S., Schildhauer, M., Vieglais, D., Zhang, J., 2005. Data integration and workflow solutions for ecology. Data Integration in the Life Sciences (DILS), San Diego, July 2005. Lecture Notes in Computer Science, vol. 3615. Springer, pp. 321–324.

National Research Council, 1991. Solving the Global Change Puzzle: A U.S. Strategy for Managing Data and Information. National Academy Press, Washington, DC.

Olson, R.J., McCord, R.A., 2000. Archiving ecological data and information. In: Michener, W.K., Brunt, J.W. (Eds.), Ecological Data: Design, Management and Processing. Blackwell, Oxford, UK, pp. 117–141.

Pennington, D.D., Michener, W.K. 2005. The EcoGrid and the Kepler Workflow System: a new platform for conducting ecological analyses 86 (3), 169-176.

Porter, J.H., Callahan, J.T., 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. In: Michener, W.K., Brunt, J.W., Stafford, S.G. (Eds.), Environmental information management and analysis: ecosystem to global scales. Taylor & Francis, London, England, pp. 193–202.

Scurlock, J.M.O., Kanciruk, P., McCord, R.A., Michener, W.K., Olson, R.J., 2002a. Metadata. In: Munn, E. (Ed.), Encyclopedia of Global Environmental Change: Volume 2. The Earth system: biological and ecological dimensions of global environmental change, Ecosystems Section (Editors: H. Mooney and J. Canadell). John Wiley, Chichester, pp. 409–411.

Scurlock, J.M.O., Olson, R.J., McCord, R.A., Michener, W.K., 2002b. Data banks: archiving ecological data and information. In: Munn, E. (Ed.), Encyclopedia of Global Environmental Change. John Wiley &and Sons, New York, pp. 248–259.

Technical Committee ISO/TC 211, 2003. Geographic Information – Metadata – ISO 19115:2003.–http://www.iso.org.