Building an Information Management System for Global Data Sharing: A Strategy for the International Long Term Ecological Research (ILTER) Network

Vanderbilt, Kristin L.[1], Blankman, David[2], Guo, Xuebing[3], He, Honglin[3], Li, Jianhui [3], Lin, Chau-Chin[4], Lu, Sheng-Shan[4], Ko, Burke Chih-Jen [4], Ogawa, Akiko[5], Ó Tuama, Éamonn [6], Schentz, Herbert[7], Su, Wen[3], van der Werf, Bert[8]

[1]Sevilleta LTER, Albuquerque, New Mexico USA, [2]Israel LTER, Ben Gurion University, [3]Chinese Ecological Research Network (CERN), Chinese Academy of Sciences, Beijing, [4]Taiwan Ecological Research Network (TERN), Taiwan Forest Research Institute, Taipei, [5]Japan LTER, Kyoto, [6]GBIF Secretariat, Copenhagen, Denmark, [7]Umweltbundesamt GmbH, Vienna, Austria, [8]Alterra, Wageningen, Netherlands

Abstract

The International Long Term Ecological Research (ILTER) Network is a global network of sites arrayed in many ecosystems and countries that aims to address international ecological and socio-economic problems through collaborative research. To facilitate ILTER data discovery, access, and synthesis, a strategy for adopting common information management standards throughout the ILTER has been developed. The proposed strategy suggests the use of Ecological Metadata Language (EML) as a proximate goal for ILTER metadata standardization and ultimately the adoption of ontologies (semantic metadata) to facilitate the integration of ILTER data once standards are agreed upon and tools are developed. This paper presents examples of the information management systems currently in use in the ILTER that are EML-based, ontology-based, or based on a country-specific standard. The advantages of ontology-driven information management systems are discussed, as is the approach ILTER will take to realizing such a system. Also discussed are mechanisms for creating a network-wide information management system that accommodates the different languages used throughout the ILTER.

Keywords: LTER, ontology, metadata, semantic integration, language

Introduction

The International Long Term Ecological Research (ILTER) Network consists of 34 member countries that support long-term data gathering and analysis on a global scale to detect, interpret and understand environmental changes. The strategic plan for the ILTER Network of networks includes these ten-year goals:

1. Foster and promote collaboration and coordination among ecological researchers and research networks at local, regional and global scales
2. Improve comparability of long-term ecological data from sites around the world, and facilitate exchange and preservation of this data
3. Deliver scientific information to scientists, policymakers, and the public and develop best ecosystem management practices to meet the needs of decision-makers at multiple levels

To achieve these goals of collaboration, data compatibility, data exchange, and data preservation will require a significant investment in both ecological informatics research and cyberinfrastructure development. Some ILTER networks have already invested in substantial technology infrastructure that uses different solutions for structuring, storing and analyzing data, and creating and managing metadata.

A viable ILTER information management solution would need to address these different system infrastructures to create an interoperable system of systems. It must also address the challenges of discovering and integrating data that are documented in different languages. To create such a system, ILTER information management and technology specialists from East-Asia Pacific, Europe, and North American regions have recommended that ILTER adopt Ecological Metadata Language (EML) in the short-term as the ILTER metadata standard in order to create a shared metadata catalog and data portal for the network. In parallel, the ILTER should engage in the ontology standardization process that will eventually support the semantic annotation of data.

In this paper, we discuss the rationale for the selection of EML as the ILTER network standard. We then review examples from the East Asia-Pacific ILTER region that illustrate how EML is already being successfully used and also how a country-specific metadata standard can be adapted to generate EML for inclusion in the ILTER metadata catalog. We also discuss why ILTER intends to move toward an ontology-driven information management system that will facilitate semantic data integration, and describe current examples of the implementation of ontologies within the ILTER. We outline the vision for ILTER's future ontology-driven information management system, and conclude with ILTER's strategy for engaging with the international standards development process to ensure interoperability within ILTER and between ILTER and other environmental networks such as GBIF.

Developing an ILTER Data Catalog: EML Implementation in the ILTER

Why Choose EML as the ILTER Metadata Standard?

EML is a standard for documenting ecological data that is implemented as a series of XML modules (EML Specification: http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html). It has already been adopted by several ILTER networks (US LTER, Taiwan Ecological Research Network (TERN), Israel LTER, Mexico LTER, and South African Environmental Observation Network (SAEON), because tools exist to create, manage, and analyze data using EML. The availability of these tools and the considerable experience of ILTER personnel with this standard will make it easier for other ILTER members to adopt if they are initiating a new information management system.

In order to create an ILTER-wide data catalog, all ILTER networks will generate "discovery-level" EML, a core set of elements including title, keywords, abstract, creator, and spatial and temporal domains. Each ILTER member network may choose to manage their metadata entirely as EML, or they may manage the bulk of their metadata in another system from which they generate discovery-level EML. Examples of both approaches are described below. ILTER will have to find solutions for handling metadata written in different languages in order to make all data accessible from a single portal.

Examples of EML Usage and Generation in the ILTER

TERN EML-driven solution to carbon flux data management issues: Currently, no universally accepted method of carbon flux data management has been established that uses a metadata approach for archiving, curating, discovering, retrieving and analyzing data. Instead, each flux research group has formed their own regional network such as CarboEurope, AmeriFlux, and AsiaFlux and each has developed software to address data management issues. Since 2004, Taiwan Ecological Research Network (TERN) has attempted to collect existing EML-based tools to assemble them as a data management system that could be used universally in carbon flux research.

Using this EML-based data management system, a conceptual framework has been developed for flux data management that can be divided into three tiers (Figure 1). The first tier deals with datasets and related information. Data produced by eddy covariance sensors communicating automatically through wired or wireless networks are managed by this tier. In this first tier, all information related to a flux dataset is documented in EML using the Morpho EML editor (Higgins et al. 2002). The second tier relates to information management. Once metadata and data quality have been described and checked, the metadata are stored in the Metacat system (Java servlet, LDAP authentication, and backend schema-independent database). Data are stored using Storage Resource Broker (SRB) (Rajasekar et al. 2003). The third tier consists of web service based scientific workflows that allow easy access to the second tier. The Kepler workflow system (Ludäscher et al. 2006) was adapted for use in this layer.

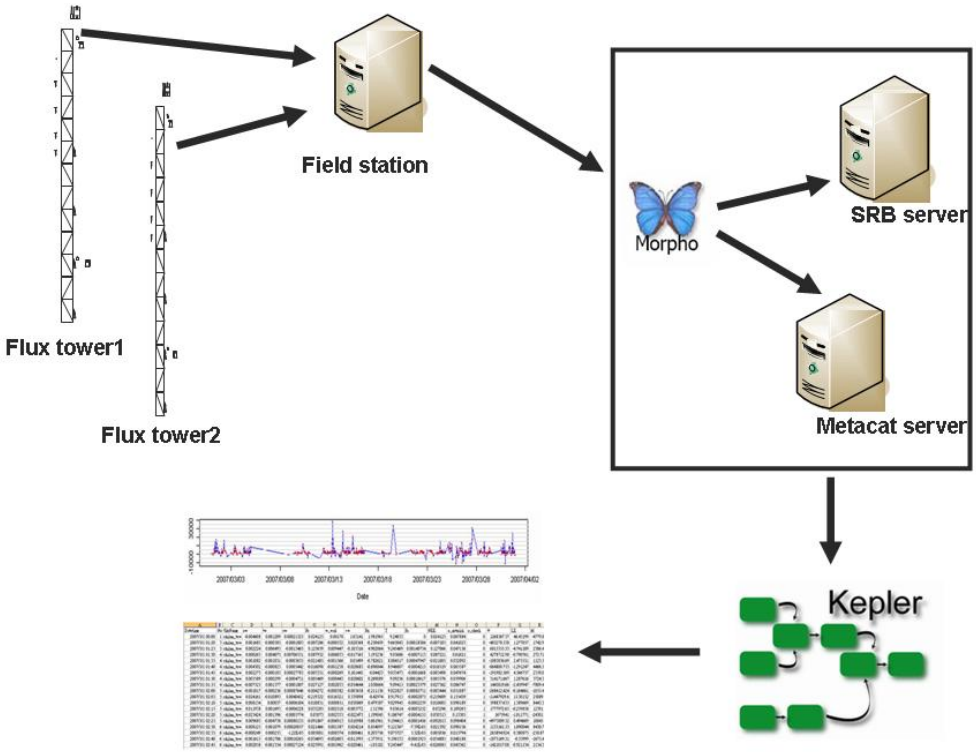

Figure1. Using EML-based tools for carbon flux data management

The use of this EML data management system was applied in Chilan, a TERN site where two flux towers have been set up since 2000. The two towers are equipped with vertical and horizontal wind vectors and the CO_2 mixing ratio at 20 Hz is measured with a sonic anemometer. A desktop computer collects these data. Every 30 minutes, the computer stores the raw data which is downloaded weekly and loaded to a SRB server to be retrieved for analysis. Metadata for these raw data are created and stored in the Metacat. Then, using the Kepler system, five workflows are run that search metadata from the Metacat, download data from the SRB, rotate data coordinates, QA/QC the data, and create Web-Pearman-Leuning (WPL) corrections to standardize the flux data calculation process based on each 30 minutes of data collected.

Output of the final calculation of all flux data are displayed in a text file which reports all the variables and a graphical file which shows the flux trend of a specific period. These secondary data can be saved locally or remotely. The adaptation of the existing tools based on EML from the flux data management experiment has achieved the goal that analyses of sequential ecological data be accompanied by formal process metadata.

Adaptation of CERN metadata standard to generate EML: In China, the Chinese Ecological Research Network (CERN) is the main organization conducting ecological research and data management, analysis, synthesis and sharing. Based on the draft of "The National Metadata Standard for Ecological Data Resources (GB/T 20533-2006)" that considered many metadata standards such as EML and ISO19115 while it was being developed, CERN proposed a metadata standard more relevant to CERN's needs. This standard was adopted in 2006. Although the conceptual framework of CERN's metadata has many elements, CERN trimmed many elements and only reserved those that were crucially necessary for describing ecological data when CERN built its physical metadata database.

CERN's metadata database is composed of seven modules, including (1) dataset identification module, (2) entity identification module, which contains information about each entity (such as a database table or other file) in a dataset, (3) observational plot module, which describes each plot's spatial coverage and geographic background information, as well as management information of the plot, (4) observational method module, (5) data quality evaluation module, (6) project information module, and (7) dataset distribution module. Although modules or elements in CERN's metadata standard and EML are not exactly the same, a valid EML document can be generated from CERN's system. Each EML element logically corresponds to one or more elements in CERN's metadata. CERN's identification module (Figure 2) includes elements that are quite similar to EML: dataset title, identifier, abstract, keywords, creator, date of dataset creation, status of the dataset, language, disk size of the dataset, spatial coverage, and temporal coverage.

The centralized CERN information management system harvests metadata in the CERN format from all CERN sites and stores it in an Oracle RDBMS, and provides a central metadata catalog for all CERN data. Metadata content can be output to XML documents, and CERN can generate EML documents to be included in the ILTER metadata catalog.

♀数据集标识符**dataset identifier**
♀数据集名称 **dataset name**
摘要 dataset abstract
目的 dataset purpose
创建者 creator
其它贡献者 contact person
发布日期 dataset publishing date
状态 status of dataset
语种 language
字符集 charset
存储量 disk size of dataset
记录数 number of records
关键词 keyword set
开始日期 beginning date of the temporal coverage
结束日期 ending date of the temporal coverage
地理边界矩形之西部边界经度 longitude of west boundary of the spatial coverage
地理边界矩形之东部边界经度 longitude of east boundary of the spatial coverage
地理边界矩形之北部边界纬度 latitude of north boundary of the spatial coverage
地理边界矩形之南部边界纬度 latitude of south boundary of the spatial coverage

Figure 2. CERN metadata's identification module contains elements found in EML discovery level metadata.

ILTER Confronts Metadata Language Issues:

EML harvested from different sources may be documented in different languages and character systems, and development of an ILTER-wide data catalog will require that all metadata be represented in one language (assumed to be English for purposes of this paper). Three aspects of the language issue need to be addressed: 1) Internationalization of the metadata exchange standard (format), 2) Localization of the software tools used, and 3) Creation of a multilingual thesaurus.

1) Internationalization of the metadata exchange format: One option for resolving language issues is to generate multiple EML documents, one in English and one in the ILTER Network member's native language. This approach requires the most time and work, but has the advantage of maximizing the semantic integrity of each EML document. A second option would be to include multiple languages in a single EML element. Japan LTER, for instance, would include Japanese and English titles in the <title> element. TERN currently puts both English and Chinese into the <title> (Figure 3) and <abstract> elements of EML.

<title> **Using Genetic Algorithm to Predict Distribution of Taiwan Fir** 運用基因演算法預測台灣冷杉分佈 </title>

Figure 3. Internationalization of the <title> EML element, showing the title in English and Chinese.

A third, but least satisfactory option, would be to include duplicate elements for discovery-level EML, one for English and one for the native language. The drawback to this approach is that the title element will no longer be unique, and could confuse the query engine.

2) Localization of the software tools: Local team members may wish to use their native language to document their data and interact with the metadata editor. People must know the terms that the software uses and make sure that they don't misinterpret their data. The programming design of the toolkit should separate the presentation layer from the logic layer to provide this localization capability. By providing software skins in different languages, people from different nations will be comfortable using the software. For instance, Morpho was developed with an English language user interface, but has been localized to a Chinese version developed by scientists at TERN.

3) Development of a multilingual thesaurus: Each scientific domain should standardize a controlled vocabulary which can then be the basis for a domain thesaurus which maps semantically equivalent terms. Equivalent terms could be translated between languages, as has been done for the General Multilingual Environmental Thesaurus (GEMET) in Europe (http://www.eionet.europa.eu/gemet).

Advanced Data Integration Using Ontology-Driven Systems

Need for Ontology Systems

The metadata catalog described thus far will be effective to broadcast the availability of ILTER data holdings, but it does not fully solve goals (2) and (3) of the ILTER strategic plan. To deliver scientific information to broad science and policy users will require an integrated system with access to data held in ILTER networks. To achieve data compatibility across ILTER systems will require a full understanding of the semantics of data holdings in each ILTER program. Thus, ILTER will have a strong need to participate in large-scale data access and data standardization processes, such as ontology development. An ontology is a description of a set of concepts and the relationships between them that enable data discovery and integration. The ontology system will help ILTER achieve its 10 year goals by supporting global data syntheses through the development of semantic data discovery services, semantic data integration services, and data access services that leverage data semantics.

Although a complete framework for semantic data integration does not yet exist within free software available for ecologists, we present two examples below that show how this concept can be realized and the additional power to find and integrate data that ontologies offer. The first example is the MORIS system developed by LTER Europe and the second is SeMIS, developed at the Chinese Academy of Sciences (CAS). We then describe a vision for the ontology system for the whole ILTER.

Example Ontology System Successes in the ILTER

MORIS: The MORIS system (Schentz and Mirtl 2003) demonstrates the successful use of ontologies within a single framework. Version 1 of MORIS is an information system primarily designed for the Austrian part of the United Nations Economic Commission for Europe (UN –

ECE) project “Integrated Monitoring”, dealing with extremely heterogeneous observations on soil, vegetation, water, and air. In MORIS, metadata are part of the ontology, describing the meaning of observations and measurements in detail with scientific concepts and relations between them. Types of scientific concepts include observation design, parameters, observed entities, methods, treatments, samples, context of observation, people, institutes, and projects. Those ontologies are closely coupled with the measured data values so that scientists accessing the data for synthesis and analysis can access data and metadata through a uniform interface and correctly interpret them.

One of the main differences between the MORIS system and EML is the treatment of methods. In EML, methods are described in natural language. In MORIS each part of a method is described using a controlled vocabulary and a system of relationships. Because methods are defined using an ontology, the MORIS system can determine whether or not two sets of research data can be integrated without requiring the researcher to compare two text documents.

SeMIS: A semantic-based metadata integration system for scientific data: In contrast to the fully-integrated ontology system used by LTER Europe, the Semantic-based Metadata Integration System (SeMIS) demonstrates the successful use of ontologies for integrating data from heterogeneous existing metadata systems. SeMIS, developed by researchers at the Chinese Academy of Sciences (CAS), is a framework that enables the translation of metadata formats that conform to different standards to a global schema so that multiple metadata standards can be accessed, queried and manipulated in an integrated way. Using a domain ontology, metadata can be manipulated in a uniform way based on the common semantics of metadata from different standards regardless of the differences in metadata format and structure.

SeMIS was developed in three steps. First, a global domain ontology was developed by domain experts and computer experts working together. The global ontology has two roles: 1) It provides the user access to the data with a uniform query interface to facilitate the formulation of a query on all the metadata sources, and 2) It serves as the mediation mechanism for accessing the distributed data through any of the metadata sources. Second, metadata elements were mapped to the concepts in the global ontology. Metadata can be originally encoded and expressed in XML format or stored in relational database or some data grid system such as SRB. For the XML format, the path-to-path mapping strategy was used where XPath was mapped to ontology classes and/or property paths. The generated mapping rules were stored in a mapping table. Finally, based on the mapping table built in the previous step, the actions of manipulating the ontology are translated to the actions of manipulating metadata. For example, semantic queries are rewritten into XQuery on each local XML dataset, and then the returned query results are reformatted for end users.

SeMIS is useful for integrated metadata browsing and searching and has been taken into practical use in the Qinghai Lake CERN research site investigation and research database project. Based on an observation ontology, users can easily browse and search animals and plants living in specific environments. Currently, SeMIS mainly integrates XML encoded metadata and researchers are working to make SeMIS support more metadata formats.

A Vision for an ILTER Ontology System

These examples from LTER Europe and CAS illustrate the advances that can be made through ontological modeling of environmental data. However, for all of ILTER to take full

advantage of such a system, ILTER would need to engage in ontology standardization efforts that are occurring within the broader ecological informatics community and build interoperable semantic service implementations across the whole network.

Madin et al. (2008) characterize ontologies as framework ontologies and domain ontologies. Domain ontologies provide the detailed semantic information associated with a particular discipline, such as in sub-disciplines of ecology. For example, Williams et al. (2006) created a domain ontology to describe the specifics of food-web interactions among species. However, if domain ontologies are developed in isolation, they may be difficult or impossible to integrate due to logical inconsistencies in their modeling approaches. Thus, framework ontologies that provide a common modeling perspective and that can be used to integrate extended domain ontologies are critical. One such framework ontology is the Extensible Observation Ontology (OBOE) (Madin et al. 2007). OBOE provides a common modeling framework that can be used to create specialized domain ontologies that address specific aspects of scientific observations, such as what entity was measured in an observation, the characteristic of that entity that was measured, and the context in which the measurement occurred. Another framework ontology is ALTERNet Core (http://www5.umweltbundesamt.at/ALTERNet/index.php?title=Ont:ALTER-Net_Ontology), which models the observation in a similar way. Madin et al. (2008) provide an initial comparison of these approaches, but ecology and the ILTER need to participate in the development of one comprehensive ontology framework for observational science data.

Once an ontology framework and a set of domain ontologies are available, we can associate those ontology terms with data collected in the field by mapping ontology fields onto data measurements, a process termed 'semantic annotation' (Madin et al 2008, Bowers and Ludäscher 2003, Bowers et al. 2004). Such semantic annotations allow software systems to use an ontology for data discovery and integration and then access the associated data via the annotation (e.g., use of semantics in workflow design (Berkley et al. 2005)).

Even with a global ontology framework and broadly accepted domain ontologies, we expect ILTER sites will need to maintain their existing local infrastructures because of the significant investment they represent. Thus, the software architecture for an ILTER ontology system must accommodate those systems by allowing the global ontology to be connected to those local systems. One possible architecture would make use of a mediator whose function is to query local systems based on a mapping between the local (ontology) and the global ontology. In such an integrated system, the mediator process is a query/integration engine that exposes the local data via the concepts in the global ontology. The advantage of this integrated architecture is that the local data-infrastructure need not be changed. Only the mediator between the local infrastructure and ontology needs to be created. We expect that when more and more people use the ontology, the local data infrastructure gradually will adopt and adapt concepts from the global ontology. This will also lead to standardization and unification of concepts in the ILTER community.

An International Community for Developing Ontology Standards

Adoption of core and domain ontology standards is a necessary step towards faciliting data exchange within the ILTER and between the ILTER and other ecological networks, and ILTER must therefore engage with the international community that develops standards. Biodiversity Information Standards (BIS) TDWG (http://www.tdwg.org) is the primary

international organization for developing standards for data access and database interoperability relating to biodiversity data. TDWG has recently developed a new technical architecture and has a defined, formal process for preparing and publishing new standards and for working in alignment with other international standards organisations such as Open Geospatial Consortium (OGC) and International Standards Organisation (ISO). The new technical architecture (http://wiki.tdwg.org/TAG/) is underpinned by three components – the TDWG ontology, a common transfer protocol (TAPIR), and globally unique identifiers for biodiversity objects based on Life Science Identifiers (LSIDs). The move to an ontology was adopted to overcome the limitations of defining standards through XSD schemas which are document-centric, difficult to extend, and difficult to integrate across schemas. By adopting an object based approach, common concepts can be defined and reused across different communities and still expressed in community-specific XSD schemas if required.

TDWG has begun the process of defining core LSID vocabularies by expressing common biodiversity concepts as found in various XSD based schemas (e.g., Darwin Core, Taxonomic Concept Schema, Natural Collections Descriptions) in a formal ontology language (OWL). There is an urgent requirement to provide additional vocabularies for scientific observations – a domain of particular relevance to ILTER. The TDWG standards process, whose primary instrument is the convening of task groups that operate under one of its several Special Interest Groups, can facilitate this, and a task group on Observations is currently being formulated under the existing TDWG Special Interest Group on Observations and Specimen Records. The purpose of the Observations task group is to develop a core semantic model for observational data in the ecological and environmental sciences. ILTER domain scientists and technical personnel will participate in the Observations task group discussion via wiki, email, teleconference and physical meetings. The outcome will be additional LSID vocabularies for observations that build on and extend the set of core TDWG LSID vocabularies, and can be deployed in the ILTER information management system to enable enhanced data discovery, interpretation and integration both within and across disciplines.

Summary:

The goal of the ILTER information management system is to foster broad-scale research synthesis efforts by facilitating the discovery, access and integration of global data resources. In the short-term, ILTER will establish a data catalog based on EML documents contributed by all ILTER member countries. EML is the standard of choice for the ILTER because it is readily accessible, through existing tools and significant ILTER expertise, to emerging ILTER networks as they begin to establish information management systems. To achieve the long-term vision of an ontology-driven ILTER information management system, ILTER scientists will also participate in the development of the semantics necessary for the creation of standard framework and ecological and socio-ecological domain ontologies that will be used to support data integration within the ILTER and between the ILTER and other organizations.

Acknowledgements:
This paper is a product of the "ILTER Information Management Workshop on Ontology/EML Integration" that was held at Lake Taihu Field Station, China, April 7-12, 2008. This workshop was supported by the Chinese Ecological Research Network (CERN). The US National Science Foundation supported travel by US participants to attend this workshop.

References:

Berkley, C., Jones, M.B., Bojilova, J., and Higgins, D., 2001. Metacat: a schema-independent XML database system. Proc. of the 13th Intl. Conf. on Scientific and Statistical Database Management.IEEE Computer Society.

Fegraus, E.H., Andelman, S., Jones, M.B., and Schildhauer, M. 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. July 2005: 158-168.

Higgins, D., Berkley, C., and Jones, M.B., 2002. Managing heterogeneous ecological data using Morpho. Proc. of the 14th Intl. Conf. on Scientific and Statistical Database Management.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, H., Jaeger, E., Jones, M., Lee, E.A., Tao, J., and Zhao, Y., 2006. Scientific workflow management and the Kepler system. Concurrency Comput. Pract. Exp. 18 (10), 1039–1065.

Madin, J.S., Bowers, S., Schildhauer, M.P., and Jones, M.B., 2007. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23:159-168.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2: 279-296.

Rajasekar, A., Wan, M., Moore, R., Schroeder, W., Kremenek, G., Jagatheesan, A., Cowart, C., Zhu, B., Chen, S.Y., and Olschanowsky, R., Storage Resource Broker - Managing Distributed Data in a Grid. Computer Society of India Journal, Special Issue on SAN, Vol. 33, No. 4, pp. 42-54 Oct 2003.)

Schentz, H., and Mirtl, M., 2003. MORIS: a universal information system for environmental monitoring. In: Schimack, G.P. (Ed.), Environment Software Systems, vol. 5. Springer.

Williams, R.J., Martinez, N.D., Golbeck, J., 2006. Ontologies for ecoinformatics. J. Web Semant. 4, 237–242.