Ecological Informatics 5 (2010) 3–8

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

A metadata-driven approach to loading and querying heterogeneous scientific data

Ben Leinfelder [a,*], Jing Tao [a], Duane Costa [b], Matthew B. Jones [a], Mark Servilla [b], Margaret O'Brien [c], Chad Burt [c]

[a] National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, United States
[b] Long Term Ecological Research Network, University of New Mexico, United States
[c] Santa Barbara Coastal LTER, University of California Santa Barbara, United States

ARTICLE INFO

Article history:
Received 9 February 2009
Received in revised form 29 August 2009
Accepted 31 August 2009

Keywords:
Heterogeneous data
Metadata
Data query
Data integration
Synthesis
Informatics

ABSTRACT

The Ecological Metadata Language is an effective specification for describing data for long-term storage and interpretation. When used in conjunction with a metadata repository such as Metacat, and a metadata editing tool such as Morpho, the Ecological Metadata Language allows a large community of researchers to access and to share their data. Although the Ecological Metadata Language/Morpho/Metacat toolkit provides a rich data documentation mechanism, current methods for retrieving metadata-described data can be laborious and time consuming. Moreover, the structural and semantic heterogeneity of ecological data sets makes the development of custom solutions for integrating and querying these data prohibitively costly for large-scale synthesis. The Data Manager Library leverages the Ecological Metadata Language to provide automated data processing features that allow efficient data access, querying, and manipulation without custom development. The library can be used for many data management tasks and was designed to be immediately useful as well as extensible and easy to incorporate within existing applications. In this paper we describe the motivation for developing the Data Manager Library, provide an overview of its implementation, illustrate ideas for potential use by describing several planned and existing deployments, and describe future work to extend the library.

1. Introduction

Synthetic studies that span disciplines and data types are increasingly important in ecology and environmental science (Green et al., 2005). Synthesis across long temporal periods (e.g., Jackson et al., 2001), across disciplines like ecology and economics (e.g., Costanza et al., 1997), and across populations spanning large spatial areas (e.g., Worm et al., 2006) all require data that have been collected by many investigators spanning decades of research. The ability to standardize and simplify the process of data loading, querying, and integration for these data sets that are collected at different times and places is critical to efficient synthesis. Nevertheless, today most synthetic analyses are conducted using a laborious and manual process of data collation and integration. Metadata-driven data repositories such as the Knowledge Network for Biocomplexity (KNB, http://knb.ecoinformatics.org) and the National Biological Information Infrastructure (NBII) Metadata Clearinghouse (http://mercury.ornl.gov/nbii/) are collating collections of data to simplify the process of discovering and obtaining data, but combining these data for integrated analysis is still mostly a manual process.

Even a cursory examination of the ecological data housed in the KNB and NBII data repositories reveals the wide variety of data structures used by researchers to capture their observations (Andelman et al., 2004; Parr and Cummings, 2005; Jones et al., 2006). The strength of the metadata-driven model employed by the KNB is that it easily supports data storage without prescribing a particular serialization mechanism or imposing structural constraints on data files. Ecological Metadata Language (EML) allows data owners to preserve their original data format by *describing* it rather than *conforming* to a standardized schema (Fegraus et al., 2005). This is accomplished using highly structured EML metadata that describes both the physical format and the logical schema for each data set, along with semantic information needed to properly interpret the data. This feature differs from many data warehouses such as VegBank (http://vegbank.org) that use a single fixed schema for data and therefore only accommodate data that match the schema; any other data must be transformed before they can be loaded into the warehouse, thereby discarding some information in the transformation process.

Although the KNB is able to provide an effective storage solution for heterogeneous data, accessing that data has heretofore been an *ad hoc* process of downloading the original data files and processing them manually using data loading and query processes customized to each data format as described in the metadata. Because managing these datasets is so taxing on human and technological resources, many scientists resort to creating customized data integration solutions that

[*] Corresponding author.
E-mail address: leinfelder@nceas.ucsb.edu (B. Leinfelder).

1574-9541/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2009.08.006

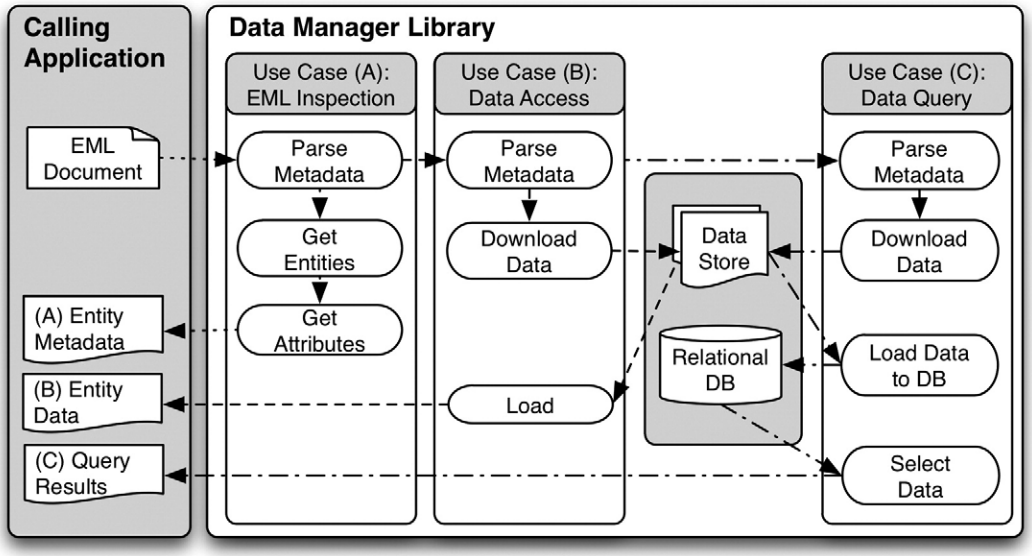


Fig. 1. Three major use cases supported by the Data Manager Library are highlighted. A host application that incorporates the library first makes a request to parse an EML document (A, B, C). The application can then use the API to request information from the metadata (A), download the data to the host data store (B, C), and create backing tables in an associated relational database (C). After these tables are created, the library can be used to load data into the database (C) and process database queries on behalf of the application (C).

frequently employ unscripted data management and irreproducible analytical processes. However, the information needed to automate this process is available in EML metadata. In this paper, we describe how the EML Data Manager Library utilizes this structural metadata to improve the efficiency and accuracy of data integration in the service of synthetic science.

2. Data Manager Library

The EML Data Manager Library provides a common software library for parsing EML metadata and using the structural information about the data to create relational database tables, load associated data and metadata into these database tables, and facilitate query and selection operations on the data (Fig. 1). The library also allows metadata elements describing the dataset to be included as data attributes even if they were not originally stored as such in the data. Thus, without incurring the costs of custom database development, researchers can leverage relational database query features on scientific data that may (and usually does) have an entirely different native format (e.g., a text file).

Two EML modules ('eml-dataset' and 'eml-dataTable') provide explicit descriptors that allow humans to document datasets and software to accurately process the data structures. These modules are typically used as part of EML but can also be included in custom metadata schema. The library exposes an Application Programming Interface (API) for utilizing the query capabilities of the Data Manager Library (Table 1), allowing it to easily be incorporated in multiple software systems without duplicating development efforts or requiring any single architecture.

3. Parsing and loading

The Data Manager Library relies on high-quality metadata about the physical format and logical schema of data in order to support data query features. Metadata authors should provide complete and accurate documentation using the eml-dataset module to maximize the utility of the query capabilities. EML documents can be accessed in a variety of ways by the Data Manager Library parser. Commonly they are served from a remote storage system such as KNB's Metacat that handles the metadata and the data storage as well as providing versioning capabilities (Jones et al., 2001). A Datapackage represents a parsed version of EML's dataset element, which is a collection of data entities including tabular and other types of data structures. These entities are further described by metadata about the attributes that they contain. Thus, each entity described in EML maps to a table in a shared relational database used by the Data Manager Library and the attributes of each entity correspond to the columns of those tables. The data are retrieved and the table[s] populated using details about

Table 1
Principal operations available in the Data Manager Library programmatic interface.

API operation	Function description
DataManager.parseMetadata()::DataPackage	Process an EML document, extracting key information about the data tables and their attributes, making this information available for further operations
DataPackage.getEntityList()::Entity[]	Get the list of Entities described in the EML document
Entity.getAttributes()::Attribute[]	Get the list of Attributes for an Entity
DataManager.downloadData()::boolean	Download the data associated with an Entity into the calling application's data cache
DataManager.loadDataToDB()::boolean	Create backing tables and load data for an entity into a backing relational database
DataManager.selectData()::ResultSet	Select data using a SQL query from the relational database tables for one or more Entities
DataManager.setDatabaseSize()::void	Set the total allowable size of data tables in the backing database
DataManager.setTableExpirationPolicy()::void	Change the policy determining when the backing tables for an Entity can be released from the backing database

the physical format of the data that are specified in the eml-dataTable module where field and record delimiters can be identified and other complex data structures defined.

The library supports many protocols for downloading data. These include File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), and Storage Resource Broker (SRB) for distributed storage systems, the local file system, and web services that implement the KNB EarthGrid API.

3.1. Queries for integrating data

Using the Data Manager Library, a client application can assemble a query against the tables and metadata in an EML data package. Client applications and analysis systems like Kepler (Altintas et al., 2004) can seamlessly access all or part of the dataset and merge data with metadata. The Data Manager Library's selectData() operation is used for executing a query on the tables in a single data package or across data packages. The query specification contains the select, project, and constraint clauses associated with a SQL query, allowing the calling application to request particular rows and columns from the associated data tables to be returned. The information about the entity and attribute structures comes solely from the EML, therefore the Data Manager Library can be used to build dynamic data query applications without hard coding queries against any single data schema. Query construction does require a thorough human understanding of the datapackage and constituent data tables — the Data Manager Library can facilitate queries, but cannot intuit them (see the Discussion section).

In addition to basic single-table queries, multiple tables from within a single EML document or from multiple EML documents can be integrated using join and union operations provided in the Data Manager Library query specification (Fig. 2). The union capabilities can be used to concatenate data across many tables that share a similar schema. For example, time-series data are frequently stored in the KNB in many files, each with the same structure but representing different yearly time periods. The union operation could be used to combine multiple years of data into a single table before applying any other selection constraints on the data. Similarly, if a data package contains multiple tables that are related to one another, then the tables can be merged by specifying the join condition in the query request. For example, in a data package that contains separate tables for site-level data and organism-level data, one could join the tables in order to relate site properties to organism properties (e.g., rainfall to height).

3.2. Promoting metadata to data

Within any given data package, contextual information that is constant across all of the records is often recorded in the metadata rather than repeating the information unnecessarily in all of the data records. For example, if all of the measurements occurred at one site or on one day, this information will likely be recorded in EML's temporal coverage and geographic coverage fields rather than in the data tables themselves. When taken individually certain attributes are redundant: a table known to contain temperature readings for September 2008 need not have a "month" or "year" column because it is unnecessary to repeat the same value for every reading. However, when one wants to combine this data table with others that were collected at different times or in different contexts, one needs to add the contextual information to the data tables in order to track the differences among records. Many researchers would manually cut, paste, and repeat the information into the data tables because the metadata and data are represented in completely different systems and are hard to combine. The Data Manager Library streamlines this process by allowing queries that promote information from metadata records into their associated data tables, thereby allowing the tables to be more efficiently unioned and joined into integrated data products. Fig. 2 shows how data from one package (Date and Height from package foo.1.1) can be easily combined with data and metadata from another package (Date and Height from package bar.1.1). To continue our previous example, by issuing a query to the Data Manager Library, researchers can "promote" the temporal coverage metadata and include it as an additional column in query results.

Development of the library is increasingly focused on streamlining access to shared data and eliminating the need for manual *ad hoc* processes. The Data Manager Library provides an alternative to the heretofore-accepted mode of creating custom individual solutions for querying and synthesizing data on the desktop and discourages unscripted data management techniques that quickly lead to irreproducible analyses and impose undue burdens on both human and technological resources.

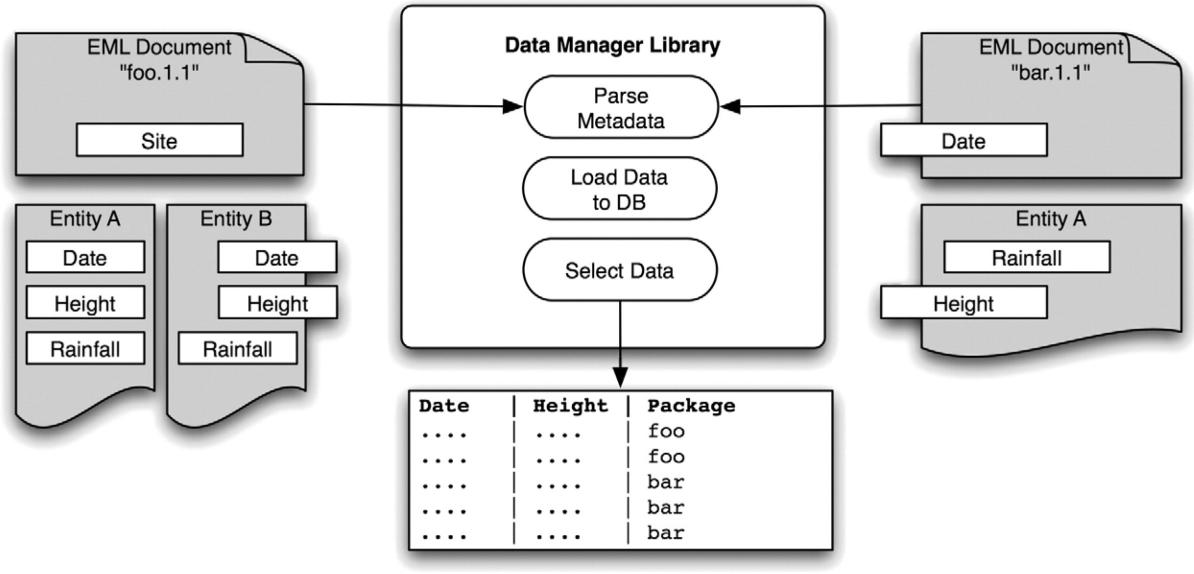


Fig. 2. A query across multiple EML Datapackages (foo.1.1 and bar.1.1): the pertinent associated tables (Entity B and Entity A, respectively) are joined. Date and Height attributes from foo's Entity B are included with Date metadata and Height data from bar's Entity A in the query results (foo's Entity A is ignored in this particular query). The Data Manager is metadata-driven and accommodates arbitrary table schemas such that entities can be structurally dissimilar and metadata can be incorporated in the synthetic data product.

4. Metacat integration

4.1. Dataquery specification

In addition to the Java-based API, the Data Manager Library supports an XML-based mechanism for performing query operations. A document conforming to the dataquery specification can express all the retrieval features of the Java-based API including: complex queries joining multiple data packages, unions across many tables, selection refinement with subquery clauses, and metadata promotion to data. Using XML to serialize the query operation allows client applications to be implemented in non-Java-centric software environments and encourages a broader adoption of the Data Manager Library's query facilities. The schema for the XML query syntax is available with the Data Manager distribution (https://code.ecoinformatics.org/code/eml/).

4.2. Metacat

The Data Manager Library has been integrated into the Metacat server (http://knb.ecoinformatics.org/software/metacat/) so that locating and querying heterogeneous data from the KNB network is provided as a standard service and requires no direct client-side use of the Data Manager Library. Metacat 1.9 contains the first prototype for handling XML-based "dataquery" requests. When clients submit a data query, the Metacat servlet acts as a proxy to the Data Manager Library by parsing the XML request and translating it into an object-oriented query using the Data Manager API. As with other mechanisms for submitting requests, the pertinent metadata and data are retrieved and loaded in the Data Manager Library's database, and the query results are returned as a tabular data file (Fig. 3). This mechanism allows simple web applications that can formulate XML requests to utilize Metacat and the Data Manager Library to handle complex data requests and have the integrated data result be returned directly to the browser application (as used in the FIRST project described below).

5. Case studies

5.1. FIRST

A large component of the Faculty Institutes for Reforming Science Teaching (FIRST) project involves the extension of existing EML, Data Manager Library, Metacat, and Morpho (Higgins et al., 2002) technologies to enable data analysis in the domain of science education. FIRST researchers expect to create the tools needed for both capturing and querying educational assessment data (e.g. tracking how students do on exams from year to year) and to provide these tools to professors, instructors and researchers.

By incorporating the eml-dataset module in the design of the FIRST educational assessment metadata schema, the features included in the Data Manager Library can be exploited to quickly provide query capabilities without the substantial overhead of creating a custom relational database solution. Student response data from FIRST are contained in many small and relatively simple tables and then registered in the Metacat system. The prototype system uses the Data Manager Library to perform queries that integrate metadata information (e.g., Course name, Question text and classification) with the student response data that is represented as data tables. Fig. 4 shows the FIRST web interface that is used to select metadata fields to be merged with student response data files. The ability to combine metadata with data when retrieving synthetic datasets that span multiple courses or institutions – a requirement for tracking and comparing student performance metrics – is of particular interest for FIRST researchers.

5.2. PASTA

The Long Term Ecological Research (LTER) Network Information System is now developing both short- and long-term strategies to enable a wide range of synthesis research within the LTER Network and the broader scientific community. One such strategy is the Provenance Aware Synthesis Tracking Architecture (PASTA) framework (Servilla et al., 2006). This modular framework is designed to support automated extraction and loading of site-based data into a permanent and persistent archive, which can be used as a data resource for synthesis research.

A foundation module of the PASTA framework, called the "Parser/Loader", uses the Data Manager Library. There are over 6000 EML documents available in community Metacat servers that have been contributed by LTER as of early 2008. Most of these EML documents describe tabular data that are easily accessible by functions of the Data Manager Library. It is the goal of the Parser/Loader module to automatically update PASTA's archive content when new versions of data are available at each participating LTER site. Components of the PASTA framework are now in use by the EcoTrends Web Portal (http://www.EcoTrends.info), a collaborative project (Peters and Laney, 2006; Laney and Peters, 2006) between the LTER Network and other local, state, and federal agencies and institutions to promote the use of long-term ecosystem data for synthesis research.

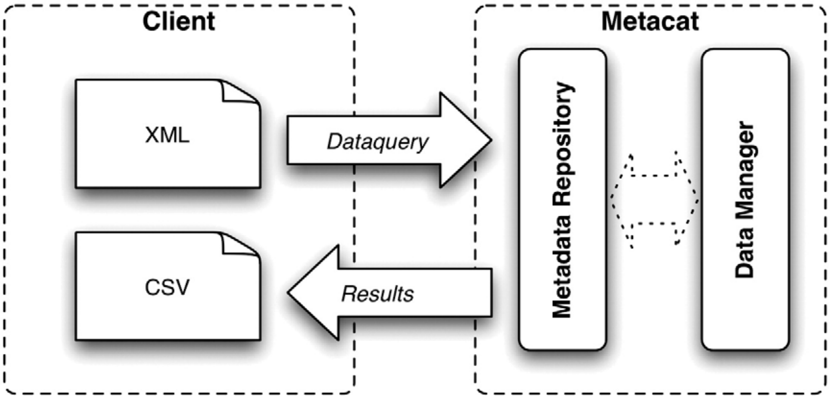
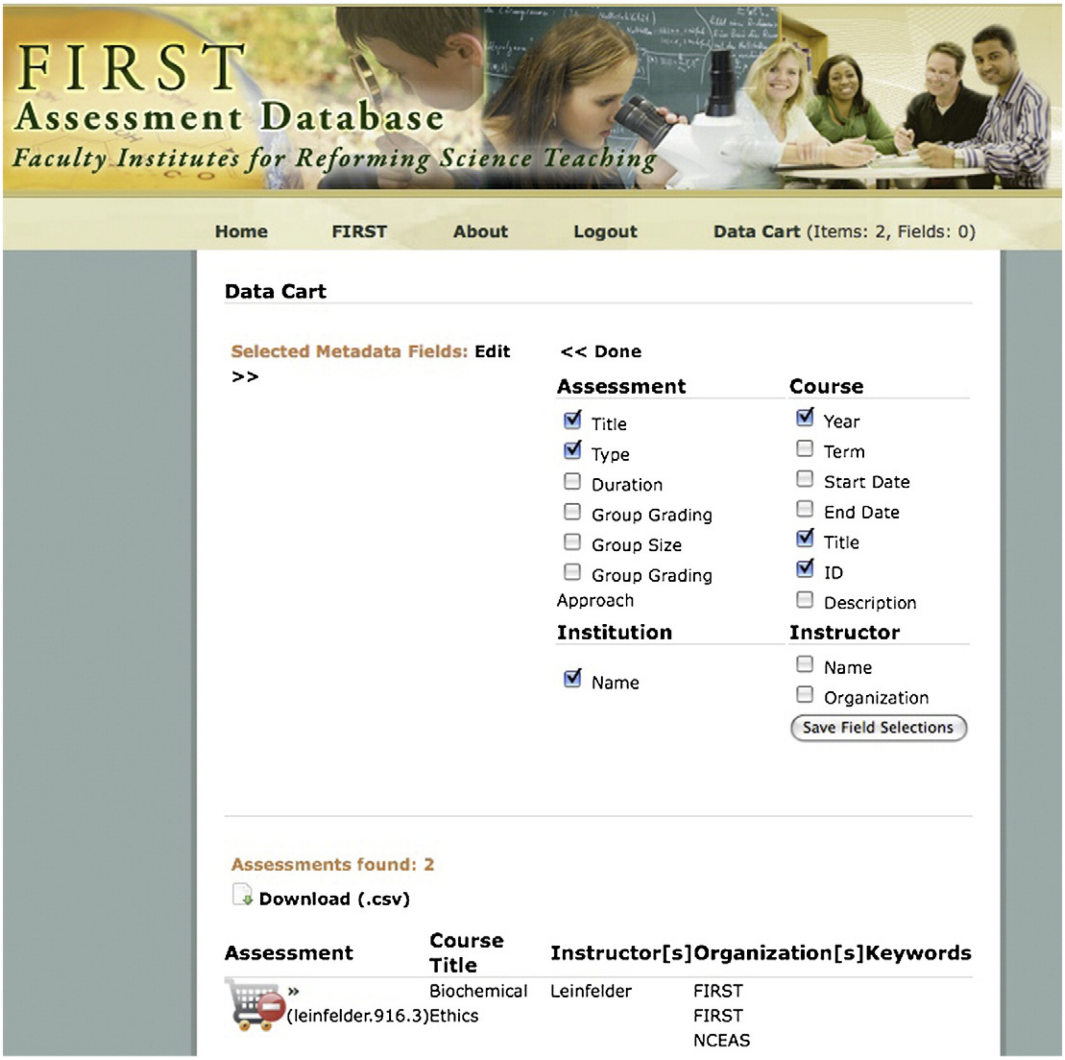


Fig. 3. Data Manager/Metacat integration. A lightweight client constructs a document conforming to the dataquery specification and submits it to the Metacat servlet. The request is passed to the embedded Data Manager where the XML is parsed and used to generate a tabular result set using data and metadata retrieved from Metacat's Metadata Repository. The synthetic data set is returned to the client.



FIRST
Assessment Database
Faculty Institutes for Reforming Science Teaching

Home FIRST About Logout Data Cart (Items: 2, Fields: 0)

Data Cart

Selected Metadata Fields: Edit >>

<< Done

Assessment
- ☑ Title
- ☑ Type
- □ Duration
- □ Group Grading
- □ Group Size
- □ Group Grading Approach

Course
- ☑ Year
- □ Term
- □ Start Date
- □ End Date
- ☑ Title
- ☑ ID
- □ Description

Institution
- ☑ Name

Instructor
- □ Name
- □ Organization

Save Field Selections

Assessments found: 2

Download (.csv)

Assessment	Course Title	Instructor[s]	Organization[s]	Keywords
» (leinfelder.916.3)	Biochemical Ethics	Leinfelder	FIRST FIRST NCEAS	

Fig. 4. FIRST Screen Shot. The web interface for employs a "Data Cart" model for discovering and downloading student response data and assessment (exam) metadata as a single integrated data product. Users select pertinent datasets for their research objectives and optionally include context information about the assessment tools used.

5.3. SBC-LTER

At the Santa Barbara Coastal LTER, the Data Manager Library is employed as part of a web application (http://sbc.lternet.edu/data/) written in Ruby that allows users to query EML-described data stored in Metacat. By integrating the library within the application's architecture, the SBC-LTER has eschewed custom relational database development and embraced a flexible solution for providing access to LTER data.

Spatially and temporally specific SBC-LTER datasets with high observation frequency spurred development of the query interface so that users – be they scientists, students or other informed parties – could limit the number of records retrieved based on date, time and location parameters (O'Brien and Burt, 2007). Queries may further restrict results to specific data columns before the output is finally delivered in a zipped CSV file format. While the SBC-LTER tool does not currently exploit the metadata promotion feature, it could be potentially useful for creating synthetic datasets that span multiple observation sites or that include metadata-dependent attributes.

6. Discussion

The Data Manager Library is an effective tool for efficiently accessing relational data from data repositories like the KNB that loosely couple metadata and data. It can be used to increase efficiency of loading, accessing, and querying heterogeneous data, and is especially useful in synthetic analysis activities that require integration of data sets from a broad range of providers.

However, users of the Data Manager Library will find that it operates optimally when processing data packages with extremely high-quality metadata. Any errors in the metadata regarding the physical and logical structure of the data can prevent data from being loaded and queried. Without complete data descriptions, it becomes difficult or impossible to intuit reasonable table schemas in which to house the data. Moreover, the actual data must be relatively "clean" in that data types must match between metadata and data. Thus, reliance on human entered metadata frequently compromises the utility of the tool.

As the Data Manager Library becomes more ubiquitous among analysts, we hope to see a pattern of metadata and data quality improvement develop throughout the KNB. Early adopters might be frustrated by the scarcity of well-described data in the KNB and could react by seeking alternate, roll-your-own solutions for acquiring, merging, and querying data. However, if the community can increasingly build a reliance on and a demand for Data Manager Library features that require accurate and complete metadata, then quality metadata will follow.

Another major issue is that the Data Manager Library only deals with the mechanical issues associated with integrating heterogeneous data. The library does not assist with determining whether two or more data tables can be appropriately integrated for analysis because the EML record does not expose sufficient semantic information about the contents of the data. As a consequence, to effectively retrieve meaningful results from Data Manager queries, a researcher must have intimate knowledge of the data semantics in order to build appropriate queries for the library to execute. Semantic compatibility is especially important when joining or unioning tables, and is currently only available via human interpretation of natural language descriptions of the data. Several efforts in the broader community are producing semantic annotation approaches that would allow machine-based reasoning about data compatibility (e.g., Madin et al., 2007; Madin et al., 2008). In future work, we expect to be able to leverage these advances in semantic representation so that the Data Manager Library can streamline the selection and integration of semantically compatible data packages.

While these semantic systems evolve, the current EML Data Manager Library can be used to efficiently integrate and query heterogeneous data and could substantially reduce the labor needed to undertake synthetic analysis. The library rewards researchers who invest in metadata entry and data annotation by making high-quality, meaningful data sets available for analysis with minimal overhead.

Acknowledgements

This material is based upon work supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #DEB-0553768), the University of California, Santa Barbara, and the State of California. Additional support came from the National Science Foundation under award 0618501 for FIRST, by the Andrew W. Mellon Foundation, and by the University of New Mexico Long Term Ecological Research Network Office.

References

Altintas, I., Berkley, C., Jaeger, E., Jones, M.B., Ludäscher, B., Mock, S., 2004. Kepler: an Extensible System for Design and Execution of Scientific Workflows. Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on, pp. 423–424.

Andelman, S.J., Bowles, C.M., Willig, M.R., Waide, R.B., 2004. Understanding environmental complexity through a distributed knowledge network. BioScience 54, 240–246.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., van den Belt, M., 1997. Value of the world's ecosystem services and natural capital. Nature 387, 253–260.

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. Bull. Ecol. Soc. Amer. 86, 158–168.

Green, J.L., Hastings, A., Arzberger, P., Ayala, F.J., Cottingham, K.L., et al., 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. BioScience 55 (6), 501–510.

Higgins, D., Berkley, C., Jones, M.B., 2002. Managing Heterogeneous Ecological Data Using Morpho. In: Kennedy, J. (Ed.), Proceedings of the 14th International Conference on Scientific and Statistical Database Management, July 24–26, 2002. ISBN: 0-7695-1632-7. ISSN 1099-3371.

Jackson, J.B.C., Kirby, M.X., Berger, W.H., Bjorndal, K.A., Botsford, L.W., Bourque, B.J., Bradbury, R.H., Cooke, R., Erlandson, J., Estes, J., Hughes, T., Kidwell, S., Lange, C., Lenihan, H., Pandolfi, J., Peterson, C., Steneck, R., Tegner, M., Warner, R., 2001. Historical overfishing and the recent collapse of coastal ecosystems. Science 293 (5530), 629–638 July 27.

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

Jones, M.B., Schildhauer, M., Reichman, O.J., Bowers, S., 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37, 519–544.

Laney, C.M., Peters, D.P.C., 2006. EcoTrends in Long-Term Ecological Data: a collaborative synthesis project, introduction and update. LTER DataBits, Spring 2006. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/.

Madin, J.S., Bowers, S., Schildhauer, M., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23 (3), 159–168. doi:10.1016/j.tree.2007.11.007.

Madin, J.S., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observational data. Ecological Informatics 2 (3), 279–296. doi:10.1016/j.ecoinf.2007.05.004.

O'Brien, M., Burt, C., 2007. A Query Interface for EML dataTables. LTER DataBits, Spring 2007. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/.

Parr, C.S., Cummings, M.P., 2005. Data sharing in ecology and evolution. Trends in Ecology and Evolution 20, 362–363.

Peters, D.P.C., Laney, C.M., 2006. EcoTrends in long-term ecological research project. Jornada Trails 10 (1) http://jornada-www.nmsu.edu/site/pubs/newsletr/jornv10i1.pdf.

Servilla, M., Brunt, J., San Gil, I., Costa, D., 2006. PASTA: a Network-level Architecture Design for Generating Synthetic Data Products in the LTER Network. LTER DataBits, Fall 2006. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/.

Worm, B., Barbier, E., Beaumont, N., Duffy, J., Folke, C., Halpern, B., Jackson, J., Lotze, H., Micheli, F., Palumbi, S., Sala, E., Selkoe, K., Stachowicz, J., Watson, R., 2006. Impacts of biodiversity loss on ocean ecosystem services. Science 314, 787–790.