The EcoTrends Web Portal: An Architecture for Data Discovery and Exploration

Mark Servilla[1] , Duane Costa[1] , Christine Laney[2] , Inigo San Gil[1] , and James Brunt[1]

[1]LTER Network Office, Department of Biology, MSC03 2020, 1 University of New Mexico, Albuquerque, NM 87131-0001

[2]Jornada Basin LTER, Box 30003, MSC 3JER, NMSU, Las Cruces, NM 88003-0003

Keywords: cyberinfrastructure, synthesis, web portal, long-term ecological research

Abstract

The EcoTrends Project began in 2004 to promote and enable the use of long-term data to better understand processes within the Earth's ecosystems. Collected by local, state and federal agencies and institutions, these data are quality checked and then simplified into a common data format, called "derived data", for the EcoTrends database. A large-format book containing numerous time-series plots, along with vignettes of the derived data, will be published in 2008. Foresight by project coordinators realized the importance of also providing these data via the World Wide Web. A web-based portal would make possible the use of on-line tools that would streamline the discovery and exploration of these data, while at the same time facilitate adding new time-series data to the growing EcoTrends database. The EcoTrends Web Portal is now at its first milestone in production and includes features for data discovery and access by registered users, plotting both derived and smoothed data values, downloading both summary and statistically annotated data in comma delimited and HTML formats, and saving markers to high-value data in an on-line store that simplifies future access. The EcoTrends Web Portal utilizes the Provenance Aware Synthesis Tracking Architecture framework being developed by the LTER Network Office. This framework combines community developed open source tools, such as Metacat and Ecological Metadata Language, into a data warehouse workflow system that, when fully operational, will automate the extraction and uploading of site-based data into a permanent and persistent archive that can be utilized by synthesis projects.

1 Introduction

Research from around the globe is massing large volumes of data that span extended periods of time and come from a variety of different ecosystem settings. These long-term data are the focus of the EcoTrends Project, which began in 2004 as an informal discussion of how to promote such observations for use in broad-scale and significant synthesis projects. To date, 50 research sites, mostly from within the United States, now participate and contribute data to the EcoTrends Project (Peters and Laney 2006, Laney and Peters 2006), including Long-Term Ecological Research Network (LTER) sites, as well as sites supported by the U.S. Department of Agriculture (Agriculture Research Service and Forest Service), the U.S. Geological Survey, the U.S. Department of Energy, and other state institutions.

The contribution of site-collected data to the EcoTrends Project involves considerable management. All data are quality checked for accuracy and completeness, organized into a common data format, called "derived data", and then loaded into the EcoTrends database for use in community research and synthesis projects. The initial set of data are to be published in 2008 as a compendium of plots and vignettes in a large-format book that illustrates significant time-related trends of the derived data.

Project organizers decided that, in addition to the book, all derived data would be made available on the World Wide Web through a web-based portal application called the "EcoTrends Web Portal" (http://www.EcoTrends.info). The EcoTrends Web Portal is a powerful resource for the community, providing a single point of access to an ever-growing set of environmental and ecological data that is organized in a common format, along with a set of tools for simple and quick discovery and visualization of temporal trends inherent within the data. This portal will accompany the publication and later succeed the book as additional data and new sites join the project. We anticipate over 20 thousand data sets will be available through the EcoTrends Web Portal when it is put into general production.

The following paper provides an overview of the EcoTrends Web Portal, including its architectural design and a discussion of salient features, as it meets its first operational milestone in early 2008.

2 Portal Architecture

The EcoTrends Web Portal is designed with two functional goals in mind. First, the portal must manage existing data and support the addition of new data collected from sites that are already participating in the EcoTrends Project, as well as simplify the introduction of new sites and their data. Second, the portal must streamline access to data for users by providing "smart" data discovery and exploratory tools, including functions to quickly plot temporal trends of one or more data sets. The first goal is addressed by using the Provenance Aware Synthesis Tracking Architecture (PASTA) framework (Servilla et al. 2006) as the portal's architectural foundation. The second goal is achieved through extensions of the current LTER Data Catalog (http://metacat.lternet.edu) data discovery interface.

2.1 PASTA

The underlying design of the EcoTrends Web Portal is based on the PASTA framework that is being developed by the LTER Network Office as part of the nascent LTER Network Information System. This modular framework (Figure 1) is a design pattern for automating data collection from spatially separated locations into a centralized and persistent archive, which can be used as a data resource for further synthesis. To date, the framework has only been implemented within a development environment for testing purposes. The framework utilizes community developed tools, such as Metacat (a schema-independent XML database) (Berkley et al. 2001, Jones et al. 2001), Ecological Metadata Language (EML) (Nottrott et al. 1999, McCartney and Jones 2002, Fegraus et al. 2005), and the Data Manager library (Java functions within the EML distribution for extracting and loading tabular data described by EML) (O'Brien and Burt 2007), in addition to the open source PostgreSQL relational database management system and programmable interfaces (e.g., Java Servlets) for discovering and accessing archived or processed data. All software components developed as part of PASTA, including those of the EcoTrends Web Portal, are available as open source under the GNU General Public License version 2 (GPL2) through the LTER Network Concurrent Versions System (http://cvs.lternet.edu).

Data management begins when new "Source" data are added at a site and the corresponding EML document is updated and harvested into a local "Metacat" database. If the EML document identifier is registered in the "Dataset Registry", an update to the EML document will trigger an automated data extraction event by the "Parser-Loader" module by using functions of the Data Manager library. The new data are added to the local "Cache" archive, which is available to the

"Workflow Engine" for further processing. The "Workflow Engine" represents any transformation process that is necessary to generate derived data products from the original source data. Data output from the "Workflow Engine" is stored in the "Derived Data" database, and metadata, as EML, is harvested back into the local "Metacat" database. External applications, such as web-based interfaces, are able to access derived data products by dereferencing links within EML documents discovered through Metacat's client interface. New participating sites only need to commit to the metadata harvest process and have their EML document registered in the "Dataset Registry".

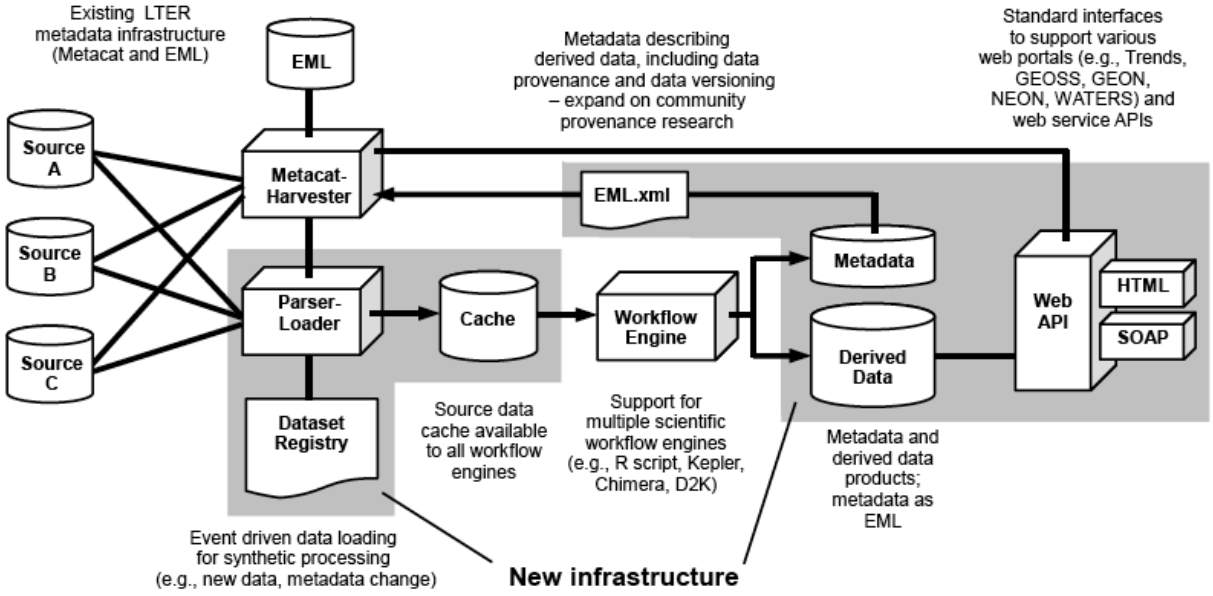

Figure 1 - Major components of the PASTA framework.

Although the PASTA framework is still under development, the EcoTrends Web Portal takes advantage of modules that support derived data products, including Metacat, EML, and a Java Servlet-based web application for data discovery and exploration.

2.2 Data Discovery and Exploration

Data discovery and exploration in the EcoTrends Web Portal starts with the search for derived data products by identifying relevant EML documents contained in the Metacat database. This process is achieved through either a "browse catalog" approach that uses a vocabulary of predefined terms as search criteria or a form-based approach that allows the user to select specific search criteria.

The "browse catalog" is separated into either "topic" or "site" sections, and are presented through separate web pages. The "topic" page is partitioned into biogeochemistry, biotic structure and disturbances, climate and physical variability, and human population and economy categories, each with a unique vocabulary of search terms. The "site" page lists each participating site in alphabetical order and uses the site name as the search term. Documents that are identified by the search are indexed into a list that is associated with each term and are made available through a single link from the corresponding web page.

In contrast to the "browse catalog", the form-based approach supports both a simple "key-word" search page that allows a user to search on phrases containing one or more words and a

complex, multi-field page for fine-tuning search criteria. The multi-field page allows the user to select from a combination of (1) the participating site name, (2) the data variable being measured, (3) the temporal frequency of the derived data, (4) temporal coverage of the derived data, and or (5) the spatial bounding coordinates of the site where the data were originally collected as the criteria to search for EML documents. The site and variable names are displayed in a fixed "dropdown" list that is generated only when new sites or new data variables are added to the EcoTrends Project. Temporal coverage values are manually entered into a form field and are matched against the corresponding categories within the metadata. Spatial searches are matched against bounding coordinates also documented within the metadata. The spatial search tool provides form fields for manually entering bounding latitude and longitude coordinates or a Google map interface (http://code.google.com/apis/maps) that can be interactively adjusted to a visual bounding area.

Data discovery through either approach has the same effect. Search results are displayed as descriptive metadata in a table (Figure 2) for each identified data set, including the name of the participating site and data collection station, the "topic" category, the variable name, the temporal frequency of the derived data, and a set of "tool" icons. The "tool" icons provide the user with a set of functions for saving the selected data set in the local "My Data Store", viewing detailed information about the data set, downloading the data in a "CSV" format, or viewing a time-series plot of the data. In addition, up to four data sets can be plotted together in a single plot (Figure 3). Users have the option to display only the data points, lines between data points, a line that is computed from a moving average of the data points, or any of the three options together. The "My Data Store" is a portal feature that lets users save data sets that are selected from the search results table. The same "tool" icons and plotting capabilities are available to all saved data sets in the "My Data Store". The details of a data set can be viewed in a separate web page, which provides access to data in a "CSV" or "HTML" format and to metadata in the native XML structure of the EML document or in a nicely formatted "HTML" presentation that is generated by the Metacat. Access to all versions of data and metadata and the same plotting capabilities available from the search results table is also found on this "details" page.

3 Discussion and Conclusion

The World Wide Web provides a wonderful abstraction through its server-based applications such that web content can be seamlessly updated without effort for the consumer of that content. The EcoTrends Web Portal is no exception. Its primary goal is to provide access to updated and newly derived data today and into the future. Most of the long-term environmental and ecological data being collected and processed for the EcoTrends Project are part of ongoing research. These projects are continually adding new data to their holdings and provide the EcoTrends Web Portal a rich resource from which derived data products can be generated and new synthesis research may take form. Providing the most up-to-date data to the community is paramount to the success of the EcoTrends Web Portal. At present, site-based data must be manually loaded into the "Cache" database for use in generating derived data products. By using the PASTA framework as its core cyberinfrastructure, the EcoTrends Web Portal is strategically positioned to automate its site-based data loading and generation of derived data products when future standards enable widespread and rich metadata in EML. In addition, PASTA's use of open source components ensures synergy with the broader ecoinformatics community into the future.

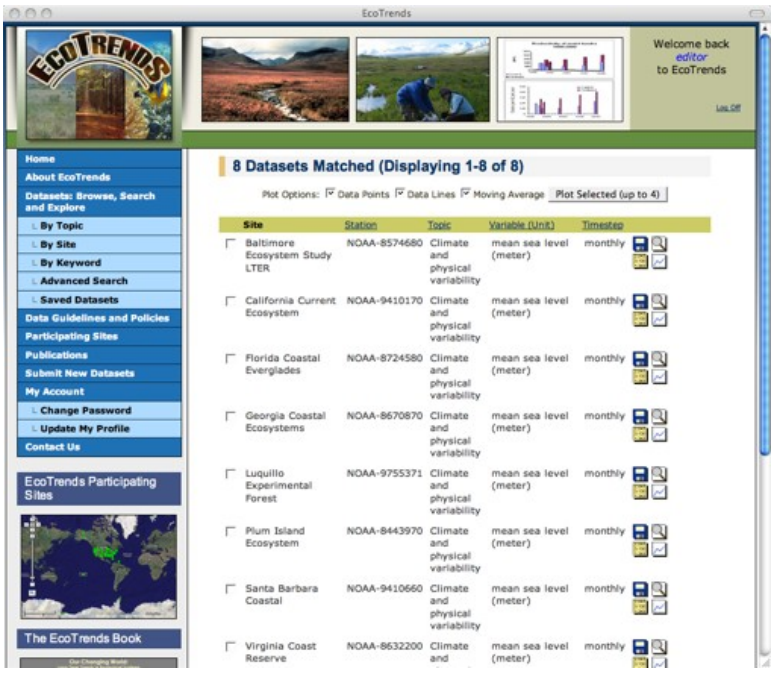

Figure 2 - Display of the search results table from the EcoTrends Web Portal.

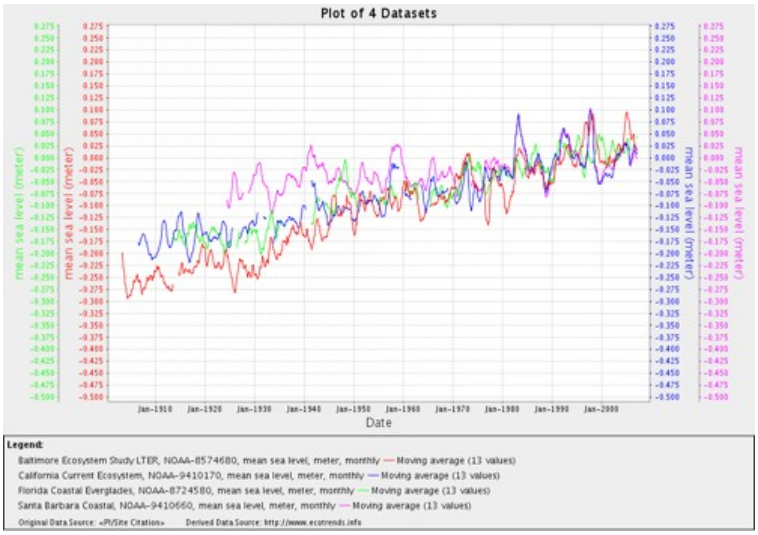

Figure 3 - Time-series plot of derived measurements from four different sites.

An equally important goal for the EcoTrends Web Portal is the ability for users to quickly discover meaningful data. The two different portal approaches for data discovery streamlines the "exploration" process of finding the right data. The "Browse Catalog" satisfies most user needs by directing their search effort to specific categories. The "pre-searched" terms provide nearly instant display of information on available data and are refreshed nightly, minimizing the probability of missing data that was recently added. When the "Browse Catalog" is too generalized, the form-based search allows the user to "fine-tune" search criteria for better control of search fidelity. To reduce the time of repeated searches, the portal's "My Data Store" can save key data sets for future use. Access to previous versions of both data and metadata is also a

noteworthy feature of the portal, since reproducibility of synthetic data products that were based on earlier versions of any derived data may be required to verify peer-reviewed studies. Perhaps, the most important feature to identify meaningful data available through the EcoTrends Web Portal is interactive plotting. Real-time plots of multiple data sets can be used to visualize and quickly interpret temporal trends of derived data, thereby focusing effort on data that is significant to the end user.

4 Acknowledgments

The authors of this paper would like to thank Dr. Debra Peters, EcoTrends Project director, and the science members of the EcoTrends Project Committee for their thoughtful review and evaluation of the EcoTrends Web Portal. We gratefully thank the technical members of the EcoTrends Project Committee, Don Henshaw, Ken Ramsey, Mark Schildhauer, Wade Sheldon, and Marshall White, for their countless hours given to discussions on the best design of EcoTrends Web Portal. The development of the EcoTrends Web Portal is supported by the National Science Foundation under Grant number DEB-0080412 and Cooperative Agreement DEB-0236154.

5 References

Berkley, C., M. Jones, J. Bojilova, and D. Higgins, 2001. Metacat: A schema-independent XML database system. 13th Intl. Conference on Scientific and Statistical Database Management: 171.

Fegraus, E.H., S. Andelman, M.B. Jones, and M. Schildhauer, 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of Ecological Society of America. 86(3): 158-168. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

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

Laney, C.M. and D.P.C. Peters, 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/)

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.

McCartney, P. and M. Jones, 2002. Using XML-encoded metadata as a basis for advanced information systems for ecological research. Proceedings of the 6th World Multiconference Systemics, Cybernetics and Informatics, (7), International Institute for Informatics and Systematics, 2002, pp. 379-384.

O'Brien, M. and C. Burt, 2007. A query interface for EML data tables. LTER DataBits Spring 2007: (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/)

Peters, D. and C. Laney, 2006. Trends in Long-Term Ecological Research projects. Jornada Trails 10(1): 2.

Servilla, M., J. Brunt, I. San Gil, and D. Costa, 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/).