Data Synthesis Workflow Development and Testing

Wade Sheldon, John Porter and Corinna Gries

Abstract

The NIS PASTA framework is nearing its first production release in December 2012, and LTER sites have made great progress over the past several years in getting EML-described core data online. LTER site Information Managers and others in the eco-informatics community have also made progress in developing or adapting software to perform analytical workflows using EML-described data. In many ways this is a watershed moment in LTER, when we are poised to capitalize on these long-term investments and realize the vision of computer-mediated data synthesis based on structured metadata.

We therefore propose to convene a workshop to develop and test workflows that leverage the PASTA framework, EML-described data and EML-based software to generate derived data products for one or more network synthesis projects (e.g. ClimDB, Veg-DB and Cross-Site Coastal Water Quality). This workshop will not only provide documented workflows that can be used by these synthesis efforts, but will also provide crucial feedback to workflow software developers, EML providers and NIS developers on any problems encountered to improve these resources. This workshop will also provide practical, real-world experience that will inform views on best practices for EML metadata content, EML-data congruency and site data package management, helping to guide site practices as we move from the era of human-readable to machine-readable metadata and data in LTER.

Introduction

The LTER 30-year review noted "the LTER network as a whole must invest in making LTER data comparable across sites and more readily available to those interested in network-wide analyses." However, developing such "comparable" datasets is a non-trivial task because most existing LTER datasets are at a relatively low-level of processing and integration. These "raw" data are a critical element of the provenance of scientific analyses (e.g. you can always aggregate raw data into derived data products, but you cannot disaggregate derived data products to reproduce the raw data), but to achieve maximum scientific advances the raw data alone are not sufficient.

Production of higher-level data products is a complicated process that requires detailed understanding of the data themselves, of algorithms for transforming and aggregating data and of techniques for developing workflows. Leaving integration of raw data to individual data users is undesirable, because few users will have the in-depth expertise to successfully integrate different types of data (e.g., a climatologist may be great at integrating meteorological data, but completely lost when integrating biological data). The large number of choices that need to be made during the transformation and integration process virtually guarantees that no two researchers will independently come up with the same final product, and too often their work cannot be reproduced because processing steps are not documented. In contrast, creating an integrated data product using a workflow approach in consultation with domain experts provides a higher quality, consistent dataset that can be used by researchers from a variety of fields. Workflows can also be re-run when new raw data are available to

produce new or longer-term integrated data sets, and can be documented and archived with the derived data as part of the analytical provenance.

Need

Prior attempts to use workflow-based approaches for large-scale synthesis projects in LTER (e.g. EcoTrends) have not been entirely successful. Variability in raw data formats and data distribution systems across sites, and changes in data formats and URLs over time made it difficult to develop consistent, reusable workflows and necessitated frequent manual intervention, nullifying most of the benefits. However, use of PASTA and EML-based workflow tools can potentially eliminate most of these barriers. PASTA provides consistent access to versioned data objects from a central location, insulating workflows from changes in site data distribution practices. New EML-based workflows tools, including Kepler (www.kepler-project.org), VCR's EML2STAT web service for R, SAS, SPSS and MATLAB (http://www.vcrlter.virginia.edu/data/eml2/) and the GCE Data Toolbox for MATLAB (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/), use EML metadata information to download and parse data tables automatically. These tools therefore provide consistently formatted data within a desired analytical environment, insulating workflows from format variations and potentially from changes in number and positions of columns. Used together, PASTA and EML-based workflow tools could dramatically simplify developing work-flow based approaches for LTER synthesis projects; however, more testing with real-world use cases and data is needed to harden these tools and more documentation is needed to facilitate broader use.

Approach

We will assemble a working group of 10 participants, including both LTER Information Managers and domain scientists, with collective expertise in data management, workflow programming and data synthesis. The working group leaders will contact leaders of targeted synthesis projects to identify their data needs, analytical approaches, and desired synthesis products and to solicit participants. Additional members will then be recruited from the broader LTER community via email, with input from IM-Exec and NISAC if necessary. The working group will meet at LNO with NIS development staff for a week-long workshop, during which participants will divide into 3-4 person teams to design, code, test and document workflows using various workflow environments and produce derived products using EML-described source data uploaded to PASTA. Teams will work in parallel, with periodic group discussions to identify problems and solutions and to work on reports and other products.

Products

Workflow source code and associated documentation will be archived in a Subversion repository at LNO (https://svn.lternet.edu/websvn/) to provide code versioning and persistent access. General information and links will also be provided on the LTER IMC website (Drupal) to direct users to these resources. Briefing documents will be prepared for the synthesis working groups that participate, describing how the workflows can be leveraged and describing any outstanding issues that need to be resolved. Reports on any challenges encountered and best practice recommendations that emerge will be prepared for the appropriate IMC working groups, and also provided to NISAC, the EB and LTER sites that provided

data. Follow-up activities to develop training modules for LTER IMs, scientists and students will also be planned.

Budget

Participants	Days	Travel	Lodging	Per diem	Total
	5	$700	$120	$56	$1,580
10					$15,800

Budget Justification

We budgeted travel for 10 people to Albuquerque to attend a 5-day workshop, based on cost estimates provided by LNO staff.