Defining Checks to Ensure High Quality LTER Data Packages

NIS Product Oriented Working Group Proposal - November 2011

Margaret O'Brien (SBC) and Jason Downing (BNZ)

Summary: The "quality" of a data package is a reflection of how well it may be used for a specific purpose. The emerging LTER Network Information System (NIS) is being developed to ingest EML datasets and process them in automated workflows. To do so, a data package must include access to data, complete metadata (defined by Best Practices), and have strong agreement between the metadata and data (i.e., congruence). Tools such as the EML Congruence Checker are being developed for reporting on data packages submitted by individual LTER research sites. This proposed workshop will produce a comprehensive set of checks to assure that LTER data packages meet a standard of high quality, and that can be used by checking software.

Background: Experience using LTER Ecological Metadata Language (EML) data packages indicates that some data access, metadata content, and metadata/data congruence are incomplete or inconsistent. The Network has defined guidelines for Best Practices for EML datasets, and the infrastructure to evaluate data packages against criteria is now the focus of the nascent LTER Network Information System (NIS). A NIS Data Package Manager web service component called the "Quality Engine" codifies checks that are applied during data package analysis; however, lacking from this effort to date is a comprehensive set of checks

The EML Congruency Checker (ECC) working group, Data Manager Tiger Team, and NIS Developers have compiled and classified a set of 35 checks that address both data congruence and metadata best practices, and 5 were implemented in an early version of the Quality Engine. The goal of this proposal is to fund a workshop consisting of community experts who will extend this current set. The challenge is to determine exactly what quality checks will be required to meet the criteria of the LTER community for high quality data packages, and secondly to prioritize those checks for the greatest return of investment. Consideration must be given to the time required for actual check-implementation, and also the potential effect on sites actively writing code to produce datasets. The Quality Engine is based on the open source Data Manager Library, which itself is managed as part of the EML suite of schemas and libraries, and so involving the EML community beyond LTER is essential.

The entire network, and particularly the IMC, is promoting the development of a dataset "checking tool". Most sites are anticipating higher standards for data packages, and currently are retooling their local systems. Additionally, sites involved in the Network Synthesis Data Project (BNZ, SGS, CWT) are interested in the timely development of standards, metrics and tools to diagnose and evaluate EML completeness and congruence. We have already received broad interest from the IMC, and these sites plan to participate in evaluation of the workshop products: CWT, VCR, SGS, CAP, JRN, AND, and NTL.

Plan: The focus of this proposal is to convene a small set of community experts (both from and outside LTER) with extensive knowledge of EML and issues exhibited by existing data packages. In a workshop, they will describe and prioritize the list of quality checks. After being vetted and approved, the list will be forwarded to the lead NIS developer for scheduling and implementation. Events are planned before and after the workshop for communication and feedback from interested groups via IMC VTC sessions (Table 1). The proposed timing will also reinforce work of the Metadata Quality Tiger Team, which began in November 2011.

Workshop Products:

1. Complete list of checks; organized by types (data accessibility, metadata, congruence), status response (pass, fail, warn, info) and the criteria for each, Checks are prioritized based on a perceived return of investment for the community, with priorities justified.
2. Draft of a document describing the checks and Quality Engine behavior for comment by stakeholders and NISAC, and which can be further developed to inform those engaged in manual evaluation of datasets (requested by IMExec).

Table 1. Schedule of activities.

	Timing	Type	Who	Activity
1	Feb 2012	VTC	Workshop participants	1. Review Metadata Quality Tiger Team effort 2. Define workshop goals
2	Mar 2012	VTC	Workshop representatives, IMC metrics WG, IMC stakeholders	1. Present anticipated design and priorities to stakeholders 2. Solicit feedback on policies, especially, implications for specific EML-generation systems (i.e., site retooling)
3	Mar 2012	2 day Workshop	Workshop participants	1. Build the set of quality checks 2. Prioritize with justification
4	Apr 2012	VTC	Workshop representatives, IMC stakeholders, NISAC	Present workshop products
5	Apr-May 2012	Self-paced	IMC, IMExec, NISAC	Receive and comment on workshop products
6	Jun 2012	VTC	Workshop participants	Present final approved set of quality checks to NIS developers

Background Material:

1. Checker reports from V0.1 Data Manager (available from LNO)
2. Google-document describing checks accumulated to date: http://goo.gl/xg87p
3. Notes from IMC Breakout (26 Sep 2011): http://im.lternet.edu/meetings/2011/breakout1
4. Slides and notes from EIMC Birds of a Feather (28 Sep 2011), see link above

Budget: $5500

Justification: For this workshop, the Network-recommended average of $1100 per participant was used. We estimate that 3 to 5 people will be traveling to this workshop. It could be held in either Santa Barbara, where several expected participants reside, or in Albuquerque for LNO logistical support. In either case, the total cost is less than or equal to $5500.

Expected workshop participants:
Mark Servilla (LNO, lead NIS developer)
Duane Costa (LNO, NIS developer)
Ben Leinfelder (NCEAS Ecoinformatics group, EML DML developer)
Margaret O'Brien (SBC, EML Best Practices, Data Manager & Metadata Quality Tiger Teams)
Jason Downing (BNZ, Synthesis Data Project)
M. Gastil Buhl (MCR, EML Best Practices, NISAC, Data Package Manager and Metadata Quality Tiger Teams)