NIS Product Oriented Working Group Proposal
Requirements for the EML Dataset Congruency Checker (ECC)

November 2011

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

Summary:

"Dataset congruence" is the agreement between a data entity and its EML metadata, and reflects the degree to which EML-described data can be automatically loaded and used, e.g., by a workflow. The “EML Congruency Checker”, or ECC, is being developed for reporting on datasets submitted to the NIS PASTA framework. The first iteration of the ECC (Version 0.1) became available in June 2011 with a minimal number of checks implemented, and this workshop is proposed to further define the requirements that will enhance the capabilities for the ECC given knowledge gained from this first iteration.

Background:

Current experience using EML metadata contributed to the LTER NIS indicates that a significant fraction is not of sufficient quality for automated use. However, only data packages that can be ingested into PASTA will be usable for data synthesis within the NIS. The LTER Network SIP and potential collaborations with other networks (e.g., NEON through the Synthesis Data Pilot Project) will rely on accurate assessments of current site capabilities in order to plan new funding and data-support activities. Consequently, a tool with which sites and the network can assess dataset usability is being developed. This tool also has uses beyond the LTER, as it is built on code that is part of the EML Data Manager Library, and so is applicable to any data package described with EML.

In 2010 - 2011, the IMC EML Congruency Checker working group, the Data Manager Tiger Team, and the NIS Developers jointly began a list of features of EML datasets that could be examined, and classified these according to several criteria. Version 0.1 of the ECC implements five checks from this list, and reports testing all LTER datasets (i.e., in the Metacat metadata catalog) against these five checks are now being constructed.

In September 2011, general feedback on the ECC status and plans was solicited from the broader community during a “Birds of a Feather” session at the Environmental Information Management Conference. The session was attended by ~30 people, evenly split between LTER and the community at large. The consensus was that every effort should be made to continue development of the ECC. Additionally, during the 2011 IMC meeting a breakout discussion focused on the use of reports. Most IMs consider the ECC to be a tool for IMs to check quality of data package, but like all technological advances, there are other ramifications to consider. These issues surrounding the uses and audiences for reports should be addressed in the near future, possibly beginning at this workshop.

Workshop Products:

1. Criteria for “PASTA-readiness” We should expect some input from NISAC on the starting point for this definition. Possibly these criteria will be organized in a hierarchical manner similar to the old “levels of completeness for metadata”
2. Elaborate the list of checks for the ECC; organized by types (data, metadata, congruence), status response (pass, fail, warn, info) and the criteria for each. Delineate which checks are appropriate for externally configured responses and/or criteria.
3. Options for storage and maintenance of checker configuration (e.g., YAML or RDB)

4. Time line for implementation of checks for 2012- 2014, that dovetails with PASTA development and the activities of the Synthesis Data Pilot project.
5. Outline a user-friendly form that accesses the ECC web services that will make the software as accessible as is practical and lower the barrier for ECC use
6. Review of possible HTML reports (if time permits)
7. Schema for XML report format (if time permits)

Tasks:

The initial portion of the workshop will outline and define the ECC behavior. After a brief review of the current checker status and recent discussions concerning its development, we will refine the list of data package checks and further clarify what constitutes "PASTA-ready". This will also include time consider management of checker configuration.

The second portion will recommend a work schedule, with priorities evaluated against the time lines of PASTA development and milestones in 2012 and 2014, and anticipated schedule of the Synthesis Data Pilot Project.

Third, and as time permits, we will consider and refine report structures, and suggest a list of policies that should be developed. Various ECC report structures are likely to be of interest to different target audiences, and these should be defined and forwarded to the appropriate LTER governance bodies. If the participants feel it is within their purview, the schema for report XML itself will be reviewed and finalized.

Participants:

The work will require 6-8 participants in a 2- or 3-day workshop. We plan to include representation from several Network groups, including

- NIS Developers
- Tiger Team members: Data Manager Suite, Metadata Quality Suite
- IMC working groups: EML Best Practices, EML Congruence Checker
- Pilot sites for the Network Synthesis Data Project (BNZ, SGS, CWT)
- NISAC members
- Sites who generate their EML in similar ways, e.g., the DEIMS, Metabase and Excel-to-EML groups

We have already received broad interest from the IMC, and in addition to the organizers, representatives from these sites are interested in attending: MCR, CWT, VTC, SGS, CAP, JRN, AND, and NTL. The NCEAS Ecoinformatics programming group is planning to send Ben Leinfelder, who is one of the authors of the original code.

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) are available as an upload to the IMC website: http://im.lternet.edu/meetings/2011/breakout1
4. MS PowerPoint slides and notes from EIMC Birds of a Feather (28 Sep 2011) are also available from the page above

Budget and Justification:

For this workshop, the Network-recommended average of $1100 per participant was used. We estimate that 6 or 7 people will be traveling to this workshop, and if held in Albuquerque where several participants reside, the total cost is $7700. If this workshop can be held in conjunction with another meeting, such as the IMExec winter meeting, the cost could be significantly less.