Proposed Work Group: Metrics and reports for EML data package quality
September 17, 2009 (ASM, Estes Park)
Margaret O'Brien (SBC), Emery Boose (HFR), Dan Bahauddin (CDR), James Brunt, (LNO), Mark Servilla (LNO), Duane Costa (LNO), Mark Shildhauer (NCEAS), Ben Leinfelder (NCEAS), Matt Jones (NCEAS), Jing Tao (NCEAS)

The primary quality standard for EML documents is XML schema compliance and the EML parser. Schema compliance is usually enforced by the editor used to create the document or checked by the EML parser, which also checks that rules for EML ids and references have been met. Experience using EML metadata contributed to the LTER NIS to automatically read and make use of data entities indicates that a significant fraction do not have metadata of sufficient quality for this use. The primary contribution from LTER sites to the NIS is data sets, which are intended to be used in cross-site synthesis projects. Clearly, any automated use of EML in the NIS will require a higher level of metadata and data quality.

The EML data manager library was created to read and parse EML metadata documents, then to download the data entities and store them as tables a relational database. It can also query those tables using SQL-like constructs. For a table to be ingested, its metadata must be accurate (not simply valid EML) and its format must be clean, consistent and match the metadata precisely. So the data manager library can be used to create the next level of quality control checks for EML datasets and their tables.

The goals for this group are to:
1. Establish a set of secondary metrics for LTER EML data package quality,
2. Recommend content for a report on data package quality (metadata and data) to be produced by the EML data manager library, and
3. Consider implementation strategies. These might include a quality report as another choice on the EML parser HTML page, or a shell script similar to that included with the EML parser.

Initially, the quality reports can be used to
1. Inform the dataset contributor about the content of the data package, and indicate whether data are of sufficient quality to be machine-readable. XML repositories have no quality standards beyond basic XML schema and (in the case of Metacat) EML compliance, so a data package that fails these secondary quality metrics can still be uploaded or harvested. However, a dataset contributor should be aware that the usefulness of the dataset will be limited.

2. In the LTER context, collating reports can produce a list of failure modes for LTER metadata and data entities. Such a list could provide input for the design of specific tools for data providers, or help identify gaps in an LTER Site's IM system. A Site requesting supplemental funding for its IMS could use the reports as part of the proposal justification.

Additional details about the dataset may be reported at some later date, e.g., basic stats, ranges, frequency distributions, and also may be compared to metadata values. A preliminary figure showing some possible failure modes is attached.

Figure 1. Possible failure modes and reports on data quality or content for an EML data package being read by the data manager library. The boxes labeled “why not?” might indicate places where a site’s IM system could be insufficient to produce NIS-ready EML.