AGU, Dec 2013, IN53C-1576

Ensuring the Quality of Data Packages in the LTER Network Data Management System

Mark Servilla[1], Margaret O'Brien[2], Duane Costa[1]

[1]LTER Network Office, University of New Mexico; [2]Marine Science Institute, University of California Santa Barbara

Summary

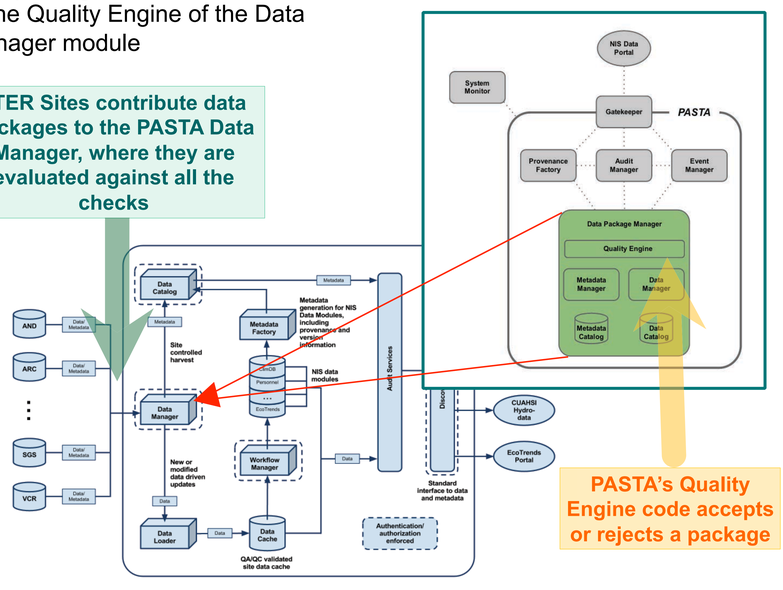
Considerable data analyses use automated workflows to ingest data from public repositories, and rely on data packages of high structural quality. The Long Term Ecological Research (LTER) Network now screens all packages entering its long-term archive to ensure completeness and quality, and requires that metadata and data are structurally congruent.

1. **Component of the LTER Provenance Aware Synthesis Tracking Architecture (PASTA)**
2. **Operates on data packages described with Ecological Metadata Language (EML), using the EML Data Manager Library (DML), written in Java**
3. **Checking is extensible for other data-types and customizable via a template**

1. LTER PASTA Architecture

The quality checks are executed as data packages are inserted into PASTA, in the Quality Engine of the Data Manager module

LTER Sites contribute data packages to the PASTA Data Manager, where they are evaluated against all the checks



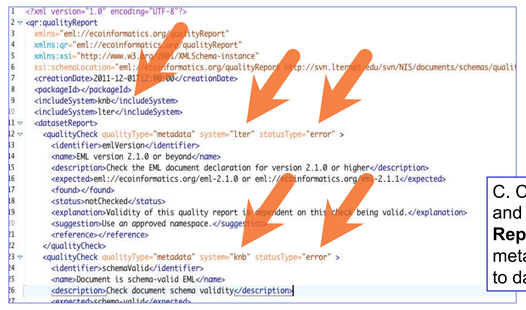
PASTA's Quality Engine code accepts or rejects a package

2. Current Implementation

- The EML metadata specification is widely used to describe environmental data
- The DML code reads EML-described data tables into a relational database for analysis. It also checks features of certain important metadata elements such as title, abstract and data URL.
- Checks are designed to evaluate metadata/data for general use (i.e., other than in a database) and the system is extensible for data types other than tables.

B. Configuration template:

In the template XML, an operator configures the System (scope) and Response status. See <includeSystem> qualityCheck/@system and qualityCheck/@statusType



A. Check classification

Scope: *the domain of the check*
- knb, lter, (or other communities)

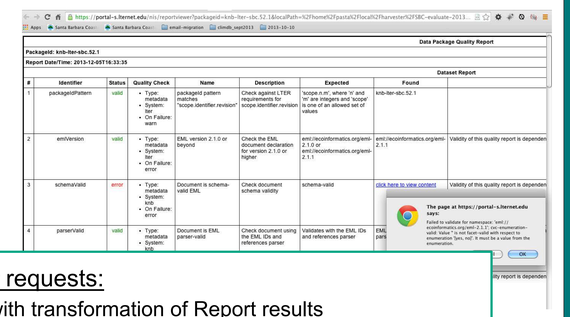
Type: *package component(s) being checked*
- metadata, data, metadata:data congruency

Response status:
- ***info*** for information only, does not affect acceptance by system
- Or one which controls system behavior:
  - ***valid*** all check-criteria were met
  - ***warn*** some problem may be present, but data package is acceptable to the system (eg, PASTA)
  - ***error*** data package cannot be accepted

C. Code populates appropriate fields (e.g. <packageID>, <found> and <status>) with results and creates an XML file called a "**Quality Report**", using the same schema. Reports are included with metadata and data in the package's resource map, and are available to data package users.

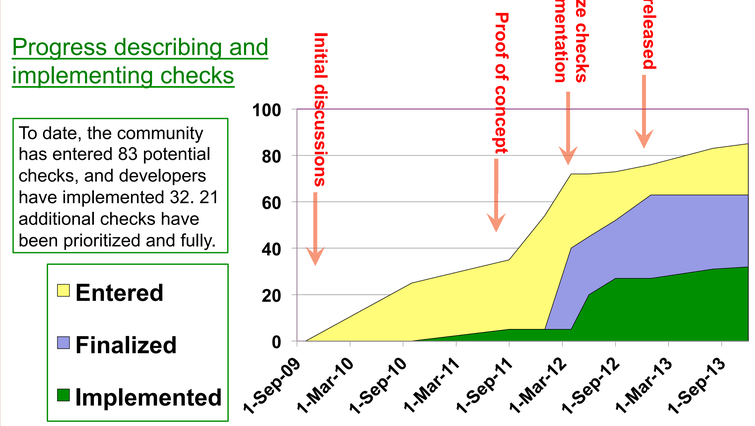
3. Community Involvement

Contributors
LTER Network Office programmers
LTER Site information managers
NCEAS programmers



Progress describing and implementing checks

To date, the community has entered 83 potential checks, and developers have implemented 32. 21 additional checks have been prioritized and fully.



User community requests:
- HTML interface with transformation of Report results
- Checks which return "error" should be implemented first
  Submitters will know which are most important issues to address
- Code should operate in two modes
  - Evaluate
    Checking continues after a failure so that a submitter sees as many problems as possible all at once
  - Harvest
    Checking stops on the first error and the data package is rejected

Contact: mservilla@lternet.edu, dcosta@lternet.edu, margaret.obrien@ucsb.edu

This research was supported by the US National Science Foundation's LTER program and the National Center for Ecological Analysis and Synthesis (NCEAS)