Published on *LTER Information Management* (http://im.lternet.edu)

EML Congruence Checker

Membership:
LTER: Margaret O'Brien (chair), Corinna Gries, Emery Boose, Dan Bahauddin, Gastil Gastil-Buhl, Jason Downing, Sven Bohm, James Brunt, Mark Servilla, Duane Costa. Associates include Matt Jones, Mark Shildhauer, Ben Leinfleder, Matt Jones and Jing Tao, from the NCEAS Ecoinformatics programming group for DML-related issues. This IMC working group was formed at the All Scientists Meeting in 2009 at Estes Park.
Mail list: eml-congruencychecker-at-lternet.edu [1]
Meeting notes: /projects/eml_congruency_checker/meeting_notes [2]

For comments on the Checker and/or checks, LOG IN and use the "Add new comment" link at the bottom of this page. You will be asked to preview before saving. You may also email the working group with their mail list, above. Comments are especially encouraged during August and September 2012, as the group prepares for acceptance of the checks, software and update plan at the annual IMC meeting.

Documentation for the Data Package Quality Engine:
https://nis.lternet.edu:8443/display/pasta/Data+Package+Quality+Engine

Upcoming events:

when done, apply wired-group's SOP for deprecating pages

2013 Events:
2013 July: ECC working group proposes process for adding and staging new checks

2012 Events:
2012 Sept: IMC annual meeting: established new working group for defining reports with checker (Package Reporting WG area [3])
2012 August 6/7: IMC Water cooler VTC for discussion of report: Water cooler link, with Presentation [4]
2012 July 30: report due to IMC on a) March workshop list of checks and b) PASTA Quality Engine behavior: Report - PDF [5] | Workbook with checks - Excel [6]
2012 March: Workshop to complete and define checks: Workshop Proposal [7]
2012 February: IMC Water cooler VTC, preparation for March workshop: Workshop Proposal [8]

2011 Events:
2011 September: IMC annual meeting: Data availability and v0.1 checker discussions (reports redacted) [9]
2011 September: EIMC Birds of a Feather session: Functional Requirements for the EML Dataset Congruency Checker
2011 August: Update to IMC via VTC (water cooler): Water cooler link, with Presentation [10]

EML Congruence Checker - 2011 Plan

At the annual meeting in 2010, the IMC was introduced to the EML congruence checker (ECC) project, and the development of a tool for reporting on EML datasets using metrics that are being established by the Information Manager's Committee. "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 first iteration of the checker is being developed as part of the NIS Data Manager Web Services, which wrapped up its testing phase in June. Please understand that 2011 represents just the first iteration of the reporting components. During 2011, we plan to use the report web service to generate a baseline report for all dataset currently in the NIS. During 2012, the working group will continue to review the dataset-checks with input from the IMC.

Here is the planned timeline for the 2011 reports:

Late July/Aug	access data URLs in EML metadata currently in the NIS Metacat catalog
Sept 1	draft baseline reports sent to both the Network and sites
Sept (tentative)	At the IMC annual meeting, report and/or break out session for feedback or discussion
Dec 31	final baseline report sent to network and sites

The report web service currently has five checks which can produce a basic report on data availability and a rough estimate of the amount of data. The checks are:

1. The content of the EML path //dataTable/physical/distribution/online/url returns content (of any kind)
2. The URL data can be read into a database from its metadata, and the first data record can be returned (excluding the header)
3. Data is displayed from the URL (information only).
4. The table can be loaded into a relational database
5. The number of rows in a table that were successfully loaded is returned, and compared to the value found in metadata.

The Data Manager Tiger Team has established a Google document which is accumulating dataset features to be reported on. The initial content of the list was contributed by the IMC at breakouts during the 2010 annual meeting (KBS), and by the EML Metrics working group. The Data Manager Tiger Team also proposed an XML format for report results, which can be transformed into a variety of formats. Further work on report tools will coincide with work on the Metadata Manager and Data Package Manager NIS modules.

Before the IMC annual meeting, each site will receive a report of their datasets and associated data currently contributed to the NIS (i.e, in Metacat). If sites use NIS Data Access Server (DAS) URLs and proxies, that system will pass the data URL to the dataset checker. Likewise, the web services are also able to read a distribution URL that returns data without human participation (i.e., a form). Currently however, the ECC cannot read data enclosed by <inline> elements (as of mid-June).

Following is a brief description of the anticipated reports. The work will be carried out by Margaret O'Brien using funds from a "NIS IM buy-out" during the latter part of 2011. Contact her for more information.

1. A Network-wide report by site showing the number of positive responses for each of the 3 checks.

Site	# dataset entities	# live data URLs	# URLs w/data	# tables attempted	# tables loaded	Total # records loaded	Notes
[acronym]	[Count of data entities in site's scope]	[count of check 1 positive responses]	[count of check 2 positive responses]	[count of check 3 positive responses]	[count of check 4 positive responses]	[sum of check 5 data records counted]	As needed

2. For each site, we anticipate a summary similar to the one above, plus the results from each data table. Remember that this is an early iteration of the checker, and not all planned functions are available. The exact format of the report is still in development, and samples will be sent intermittently for feedback.

Initial definition of "PASTA-prototype online data"

We now have an initial, strawman definition of "data online". The word document was circulated and discussed by IMExec during their Dec 1 VTC, and verbally approved. It was sent it to NISAC but they did not have time to consider it during their late 2011 call.
This definition is close to what the PASTA developers have used as their starting point, and consistent with the recently Metacat search or browse results display. This proposal is only a first step, but it represents a minimal standard, and dataset features that can actually be quantified.

Proposed definition (word-doc version attached below):
With this definition, the following examples would NOT be considered online:
1. EML metadata in the Network catalog with a data URL located at any XPath (other than the one spec'd in the definition)
2. Data that do not have EML metadata in the Network catalog (no matter where else metadata or data may reside)
3. Data that the public must specifically request from an individual (e.g., "Type II" according to http://www.lternet.edu/data/netpolicy.html [11])

Note that this means that the old "Discovery Level EML" would not be considered "online data". This is reasonable because discovery level is only online metadata. Generally, publishing only metadata advertises that certain research is occurring. It is likely that the network will still want to house it, but those policies are not part of the current discussion.

Once there is a basic definition of "online data", it can be refined further:

"Human-accessible online data" is the most basic, because a human can almost always interpret or guess how to use what s/he is given, for example a human can fill out a web form. Systems that have intervening forms (but eventually produce data via a web browser) are "human-accessible online data".

"Machine-accessible online data" would be a data package in which data can be directly accessed with no other intervention. We know that this can be complex -- not all machines can automatically access all URLs. And when URLs include web forms, we know that use-tracking systems can be designed so that they don't impede machine access. This request does not include use-tracking policies.

Attachment	Size
Request to NISAC: online_data_definition_2011.docx [12]	154.01 KB

NIS Production workshop: Defining Checks to Ensure High Quality LTER Data Packages

Workshop to be held in Santa Barbara, early March 2012

Proposal attached below.

Notes to be gathered here.

Attachment	Size
Workshop proposal: O'Brien and Downing: quality_checks_workshop_3.pdf [13]	51.98 KB

This material is based upon work supported by the National Science Foundation under Cooperative Agreement #DEB-0236154. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Please contact us with questions, comments, or for technical assistance regarding this web site.

Source URL: http://im.lternet.edu/projects/eml_congruency_checker

Links:
[1] mailto:eml-congruencychecker@lternet.edu
[2] http://im.lternet.edu/projects/eml_congruency_checker/meeting_notes
[3] http://im.lternet.edu/project/PackageReporting
[4] http://im.lternet.edu/node/1064
[5] http://im.lternet.edu/sites/im.lternet.edu/files/Data_package_quality_checks_Report_july2012.pdf
[6] http://im.lternet.edu/sites/im.lternet.edu/files/MetadataQualityReportChecks_July2012.xls
[7] http://intranet2.lternet.edu/content/defining-checks-ensure-high-quality-lter-data-packages
[8] http://im.lternet.edu/node/980
[9] http://im.lternet.edu/meetings/2011/breakout1
[10] http://im.lternet.edu/node/912
[11] http://www.lternet.edu/data/netpolicy.html
[12] http://im.lternet.edu/sites/im.lternet.edu/files/online_data_definition_2011.docx
[13] http://im.lternet.edu/sites/im.lternet.edu/files/2011_ECC_requirements_workshop_3.pdf