EML Congruence Checker - 2011 Plan

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

EML Congruence Checker - 2011 Plan

Thu, 06/23/2011 - 9:15pm — mobrien

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.

- EML Metrics and Congruency Checker [1]

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/node/894

Links:
[1] http://im.lternet.edu/taxonomy/term/211