2011-07-06 VTC notes

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

2011-07-06 VTC notes

Wed, 07/06/2011 - 11:26am — mgastilbuhl

EML Metrics Working Group Notes from EVO meeting on 6 July 2011
aka EML Congruency Checker Working Group

Attending: Margaret O'Brien, Corinna Gries, Emery Boose, M.Gastil-Buhl, Mark Servilla, Duane Costa. (Recorded for Dan Bahauddin).

Topic:
Define "PASTA-ready" and Guidelines for sites' data to be NIS-compatible

Discussion topics:
1. Vocabulary for data packages. PASTA is code, so has precise pass/fail criteria; but Best Practices is more qualitative, so has more squishy criteria. We need a vocabulary for both. Sometimes useful to tie levels to PASTA, sometimes better not to.

2. NIS milestones, and requirements for EML data package submissions: Its possible that In 2014, only data packages with congruent data will be cataloged. What is the best way to make sure all sites can get data up to par?

3. Between now and 2014, we need to develop details for all the data package criteria (both metadata and data) and how these might need to evolve.

Here is a google doc with suggested definitions to start us off.
https://docs.google.com/present/edit?id=0AVHbE-d-kTrsZGM1NmQ0Zm1fOWd3ZDR...
[1]

These notes refer to a diagram with intersecting 3 circles inside a grey circle:
grey outer circle = any EML data package can go in Metacat. 2011 present situation
blue circle = Complete Metadata
red circle = Error-free Metadata-data congruence
green circle = Error-free data
The place where those 3 circles intersect is labeled "PASTA Ready".

The provisional criteria we have now are those of the congruency checker (ECC) as it exists currently. The ECC will evolve.

Old vocabulary to describe data package readiness we have now is old BP levels, Discovery through Integration.

2-d grid (composed at 2010 IMC) does not work as well because of intersecting issues.

Our goal for today – what we need to tell the IMC
Then... what to do in next year, next 3 years.

If this 3-circle model works, we can choose areas to focus on for criteria, policies.

4 circles if we include the current metacat data packages as an outer circle.
Where 3 intersect is "PASTA-ready".

Less than half the LTER data packages in metacat now have data.

Define "accessible" data:
"Accessible" data means a live URL direct to the data file without a form. Accessible means machine-readable. Many datasets have a form step (for a human), or html rather than the data file itself. A small percent of datasets have a url that delivers data somehow.

PASTA can either reach and read the data file, or not.

It is possible to have a URL which for human browsing is intercepted by a form but for non-browser access such as PASTA, is a direct link to the data file. A white-list or the DAS are ways of doing that.

Green data circle is new. There are features of the data that do not apply to the other 2 circles: nulls („) Previously we just discussed metadata and whether it was congruent with data. But data tables themselves can have problems.

Learned from trying to load datasets:

The Acceptance Testing for the Data Manager tiger team involved loading data packages and viewing their checker reports. We learned details about inline data, where the distribution url is in the EML doc, dateTime format, and names of entities and attributes.

Since fall 2010 the Data Manager tiger team collected a list of checks on a 'Quality Report Table' spreadsheet, categorized by data or metadata. Each check is designated as a metadata, data, or congruence check. For a congruence check, the code had to look at both the EML and data to make a decision. Duane used that spreadsheet as a blueprint to write the code for the checker.

https://spreadsheets.google.com/spreadsheet/ccc?key=0AvmNJnP7eHevdFhvNkh... [2]

NIS development timeline – gives us time to iron out difficulties.
Thru 2011+ .... any EML data package goes into metacat
2011 Aug .... First checks of congruence
2012 ..... PASTA functional prototype
2014 ..... PASTA production system and only congruent pkgs cataloged.

But how many requirements are internal to pasta, vs Best Practice?

That is for us (the EML Metrics WG) to decide.
Some things are required by the system: impossible to load a „ (two commas in a row in the data file). A missing value code is needed. Requirements of loading a file (into a db) correspond to congruence checks.
Other checks are qualitative, ie compare date coverage metadata to data.
If miss-match, is that kind of error acceptable? We have to decide and specify.

Up to us to decide what is an error vs a warning. Error not allowed into level-1.

In long run, pasta ingests as a 2-step process: first congruence check, then if no problems, then ingestion.

Mark – all data pkgs that sites produce now are harvested (none are rejected from harvests). Metacat does not care whether there is an accessible data link. Metacat does not check if can get thru to data. No restrictions or restraints. With pasta we propose a change. Based on what NSF expects of the NIS. Only data packages that actually have accessible data would be harvested as a data package itself. This oes not mean you cannot embargo the data with an access control rule. Pasta would harvest and protect data based on those rules.

In a standard harvest, pasta would go thru these checks and accept a package only if it is error-free for all three circles. The complete metadata criteria is more ambiguous, qualitative as opposed to a (code) check.

Site IMs will have the opportunity to run the checker in evaluate mode prior to harvest. To see whether complies with these rings. Can do 2-step process interactively or automatically.

Corinna –
How will you determine the data is correct?

Mark -
Use the EML as prescriptive checks. Number of records mismatch – cannot determine whether eml or data is wrong.

Number of columns vs unexpected end-of-line, jagged dataTable. Error so this data not ingested into pasta.

Functional prototype to production will be one long shake-down cruise to ...
Refinement and fine-tuning over time.

Margaret – this is a formidable task. 1.5 years until the end of 2012.

IMC does not have a concept what "PASTA ready" means. Right now it is easy to be NIS compatible (if NIS means Metacat).

Corinna –
From a first look thru EML, what everybody needs to do is get that data link in. Is that the message?

Margaret –that is a good place to start. At least can start checking a broader set of data.
Right now can check packages from about 10 sites. Data at end of link that lets code go thru. Sites using the DAS are accessible because no form attached for non-browser access.
Some sites may be using a system that cannot let pasta thru.

One option is a white-list that lets the IP addresses for PASTA through. But not all sites' systems can easily implement that tweek.
(Another option is the DAS. Mark was hesitant to involve the DAS in this discussion.)

(Making data accessible is a)
Good project for supplement. People will do it.
Margaret tried to say this in email announcement 2 weeks ago.
Did not hear many responses. Some sites are aware, read and understood it.
If supplement RFP came 3 months later, would have information (the ECC reports) to address the problem.

Margaret – some sites do not see the difference between the link pointing to an html file and a csv.

Mark –
Statistics will incur peer pressure.

Corinna –
Just with metacat you can see a lot about which sites' data is accessible.

Margaret – yes I have a script that sees what is at the end of the URLs in data packages from all sites.

Corinna – put into words
"has to be a link that if you click on it, it has to be a data file, as described."

Margaret –
Up to Mark and Duane how to handle inline.

Should we post a tentative timeline?
Corinna – we could write some text outlining this
"to be pasta-ready, have a url that will give nothing but the data"

Mark –
I agree. I do not want to be too constraining. Bottom line is machine-accessible. Whether thru URI like an ftp site (we use for LNO data). Can be http or https. Currently cannot work with inline but will in future as machine-accessible dataset.

Do not bring the DAS into it. Not clear how DAS integrates into this right now.

Margaret –
But it is true that sites that use DAS proxies can have their data read thru. Because the DAS uses the user-agent field. Does not show form if user is not a browser.

Back to the 3-ring diagram...

Mark –
The complete metadata ring is ambiguous currently

Duane -
Not ambiguous if use that spreadsheet.
https://spreadsheets.google.com/spreadsheet/ccc?key=0AvmNJnP7eHevdFhvNkh...
[3]

Margaret –
would not reject dataset for not having keywords. Not all checks are y/n binary decision.

Duane –
Yes. Metadata Manager will also do checks. Can be systematized. Have to make a call, a judgment. What is a serious enough error for a dataset to be rejected from pasta.

Mark –
Some checks are binary like data url. Some are policy issues like keywords, more squishy.

Corinna –
What are we trying to accomplish here in this meeting.
We could work on that spreadsheet... so central, needs a lot of work.
Are we trying to come up with recommendations for sites supplement proposal now or EML Metrics.

Margaret –
The Quality Report Table spreadsheet will evolve but not our task today.
Figure out if something we can tell sites now for them to use in their supplement.

To work on the Quality Report Table spreadsheet,...
This WG can start to have these regular meetings.

Would like to divvy up the WG tasks to 1 or 2 people per specific task:
Each of those checks requires a high level of knowledge about a narrow field.
Task people: how datetimes should be interpreted, for example.
So many tiny little details.
Figure out how move fwd with spreadsheet.

Can EML Congruence Checker recommend to supplement writers?

Corinna –
Steps to bring site closer to middle PASTA ready.
A few ideas already. Clearly the eventual goal is to have the data congruent with the metadata, whatever that means in detail. Preliminary steps: right?

Emery –
Today short list of items for all sites to address.
1. link to data stream

Servilla –
Be careful not to focus only on dataTables because PASTA will support other kinds of data in future. Right now first focus on tabular. Data tables are most prevalent data format. Don't know if should focus on tabular data in this supplement?
For PASTA – to be pasta-ready, depending on type, must be congruent. Don't know how to congruent check other types of data than dataTables. DML specified to tables. Overall, pasta hast to be able to harvest all types of data. Definition of pasta-ready has to be more generic. Not go into levels of potential errors.

Corinna –
True but cannot resolve that in next 10 days. Need to come up with recommendations right now for sites saying what it is, get started with one.

For example: same number of columns in EML and data table.

Anything we can do to help sites achieve this? There needs to be help, not just requirements. In past we have just told sites what to produce, not how to do it. No help out there.

Creating EML snippets.

Margaret –
Right.

Have supplements written in such a way as to re-task when refine scope later?
If people want to upgrade to a better metadata model, maybe more sites want to adopt a metadata relational database (such as Metabase.)

Also, EML 2.0.1 schema is invalid, and most sites are still using it. Upgrading to 2.1.0or at least patched 2.0.1.

First recommend:
1. Upgrade to EML 210

2. Deliver data
3. Column number test

EML upgrade path not same for everyone.
There is an xml stylesheet for 201 to 210. Export, run thru xslt, re-harvest.

Can LNO endorse these three recommendations? Suggest wording.
JamesB is sort-of on vacation, on the road.

Mark Servilla –
Would endorse that data should be accessible, described, congruent.
Column count seems too simple.
The defect in eml 201 affected fraction of docs.
Good to be at current standard so yes recommend to 210.

Mark
Data need to be machine-readable.
Well documented in eml, congruent with description.
Some sites, like MCR, have a "firewall" requiring users to sign-in. Have to white-list pasta so it can access.

Margaret –
Can suggest how to let pasta thru.

At IMExec-NISAC do we have agreement for all sites to post data, yes?
Individual site policies are not a problem?

The plan is posted on the IMC website, the plan for ECC

Mark –
There is possibility for sites to collaborate. If there is a site-crossing problem, may be able to hire programmer to address problem.

Corinna –
Huge problems with money-moving. Each site can only apply for certain amount. Cannot shuffle money from one site to another. So fi something can be done piecemeal, contribute to bigger whole, that works better. Was only possible last year because NTL had larger main budget.

Mark –
Unfortunate because some tasks scale better.

Corinna –
Supplements are limited to a certain percentage of the site funding.

Gastil –
Isnt that what the LNO or NIS dev does?

Mark –
LNO is funded to develop the NIS, as identified in the OP? Actually still have not ever received formal approval on the OP.
Our job is to focus on the NIS. We do support site requirements on as-needed

basis, as fits into schedule. Takes lower priority than NIS itself. Duane's time allocated partly to CV because so niche-oriented but small one week.

Corinna –
Idea is that legacy data initiative/ synthesis data initiative. Will fill niche of helping sites getting their data ready. That is not the NIS dev.

Mark –
If there are synergies in dev of the NIS and what site requires, then can capitalize on synergies. DML is syn with some site needs. AS we extend the DataManager, some aspects the sites can capitalize on. But if orthogonal to our mission, not likely to happen.

Margaret –
There is code in the NIS that is useful. Can we extend that?

Back to data synthesis project, these things in RFP.
Don't know status of that now. Never went to whole data synth committee. Not sure what is going on. Need to say central team to help sites, have to be familiar with NIS, be able to take code form NIS and apply to sites as useful.

EB is meeting this afternoon. Will be update.

Mark –
Will have better feel what re-useable over next 6 months as more core services shake out, become more stable. One thing is federated identity concept, not sure how could be applied at site level. Not having to come up with own authorization/access control.

Corinna-
Can we suggest a project collaborative that would help the cause.

4 goals for now

4. Encourage sites to look at Bob's list of potential projects,
look as broadly as can

- EML Metrics and Congruency Checker [4]

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/897

Links:
[1] https://docs.google.com/present/edit?id=0AVHbE-d-kTrsZGM1NmQ0Zm1fOWd3ZDR3eGdm&hl=en_

[2] https://spreadsheets.google.com/spreadsheet/ccc?key=0AvmNJnP7eHevdFhvNkh4VG1GR0JETUtfeE1EbFI
[3] https://spreadsheets.google.com/spreadsheet/ccc?key=0AvmNJnP7eHevdFhvNkh4VG1GR0JETUtfeE1EbFI
[4] http://im.lternet.edu/taxonomy/term/211