LTER IMC 2018-07-09

overview and logistics

Date: Monday, July 9th at 12 pm PDT/3 pm EDT

Topic: Working Group updates

Connection: https://ucsb.zoom.us/j/8058932071

In preparation for on annual IMC meeting on September 30th, we would like to share updates from working groups and other IM activities. We plan to hold these VWC up through September (July 9, Aug 13, Sep 10). For this upcoming July VWC, we will hear from several groups, depending on time:

John Porter – Vocab

Yang Xia – DEIMS

Gastil-Bulh – ECC

attendees

Emery Boose, Stevan Earl (notes, Part I), Don Henshaw, John Porter, Yang Xia, Duane Costa, Kris Hall, Gastil Buhl, Corinna Gries, Dan Bahauddin, Kristin Vanderbilt, Renee Brown, Ken Ramsey, Margaret O'Brien, Jim, Chris Turner, Stace, Suzanne Remillard (notes, Part II)

Part I: notes taken by Stevan Earl

John Porter reporting for controlled vocabulary group

Idea of controlled vocab (aka thesaurus) was to make it easier for people to locate data as there would be agreed-upon keywords. Focused on terms that were applicable across LTER sites. Looking for a sweet spot of terms so that there not too many such that there are splitting-hairs-types differences but broad enough to cover a wide array of topics.

Ran statistics on usage in 2013:

- 96% of LTER packages contain one or more terms found in thesaurus.
- a user is 9-times more likely to find a data set if they use a term in the thesaurus, and are likely to find data from about five sites.
- maximum number of keywords for any data set is 295
- most data sets have at least 5 keywords

Recent activities:

- statistical analyses of keyword use
- conducted a survey that noted that IMs were instrumental in adding or choosing keywords to or for data sets

- considering new terms to add to the thesaurus
-

A thesaurus improves searching and browsing owing in part to the parent-child relationships (as opposed to just a list). Term relationships are incorporated into PASTA.

The LTER thesaurus is technically a thesaurus but because there are few links may be considered more of a polytaxonomy. Need to consider whether we should translate the thesaurus into an ontology, which would feature a broader array of relationship among terms (beyond parent-child, synonyms, etc.).

Issues to address at the ASM:

- do we need to move to an ontology or another lexical structure?
- should we abandon the LTER controlled vocabulary in favor of another, existing resource?
- if we decide to continue with the LTER controlled vocabulary, what updates are needed (note the software has not been upgraded since 2013)?
- how do we deal with place names and taxonomic names as keywords?

There are several scheduled sessions during the ASM that are related to this topic.

Margaret commenting that the annotation feature of EML 2.2 does not require an ontology (but that it may perform better if so).

Gastil asking logistical questions that may be of interest primarily to new IMs, such as is it case sensitive, and do plurals match. John noted there is an accessible process for adding new keywords. Case sensitivity is not an issue. Plurals depend on what is being described. Generally, though, there are documents on the IM page and the controlled vocabulary that describe these resources.

Don inquiring about an earlier request about licensing for the controlled vocabulary. John seems to think that we generally resolved to CC0 but could not recall if this was officially resolved.

Jim inquiring if the controlled vocabulary is being effectively used - is there a return on this effort? John noting that we can see what people viewed but not what they used, and that this is generally difficult to evaluate quantitatively - would love suggestions on how to address this? Jim suggesting that we could discuss this issue with investigators at the ASM.

John posted the following resources in the chat window:

http://im.lternet.edu/vocab_resources is the URL for the resources related to the controlled vocabulary, including the Best Practices Document
http://vocab.lternet.edu is the link to the controlled vocabulary itself.

Yang Xia providing an update regarding the DEIMS working group

Eda is leading the DEIMS group but was unable to attend the VTC this day.

This group last met in March. Frustrations and approaches to dealing with issues were discussed at this meeting, as well as consideration for adopting Drupal 8. Eight people (7 sites) are participating. Participating sites include (we think): KNZ, NTL, LUQ, PIE, ARC, JRN, VCR

The DEIMS group will host an ad hoc meeting at the ASM. At that session, they would like to consider new tools or approaches for future iterations of DEIMS, in addition to helping fellow DEIMS users with any issues.

Jim noting that the group is considering how to move DEIMS forward but with an eye toward developments that EDI is addressing so as to avoid duplicating efforts.

John inquiring whether any other groups (other than LTER) are using DEIMS and/or contributing to its development? Yang indicating that they could use help since Inigo left and with little outside help without input from the network office. Jim suggesting that other groups do use DEIMS but few. Ken noted that there are several international groups that use DEIMS but it is hard to gauge whether they are making improvements or developments that would benefit all users. Ken noting that some other implementations, such as in Europe, are structured quite differently, particularly more centralized. Ken suggests that there is wide speculation (generally disapproval) of DEIMS.

Gastil providing an update of the ECC group

Accomplishments include releases of two new checks (checkSums in fall 2017 and dateTime in spring 2018).

The group chats frequently on line and hosts virtual meeting as needed, with one in-person meeting at the annual IMC meeting. All materials and resources are available through GitHub.

The plan is to continue with biannual releases. That attribute names match header string check is scheduled for fall 2018.

Please always remember that the goal of these checks is to contribute to improved data quality, and not to create bottlenecks.

Gastil detailed the check development life cycle.

Gastil inquiring with the attendees as to whether checks that are strictly informative (i.e., those that provide feedback but would not yield a warn or error) can be implemented outside of the agreed-upon biannual release cycle? Also, is it acceptable to make cosmetic changes outside of the biannual release cycle?

The attendees overwhelmingly approved implementation of new, informative-only checks outside the biannual release cycle. The first of these will be the proposed header string check noted above.

Gastil noting that the group's highest purpose is to encourage the archiving of data with congruent metadata.

New checks are reported in the EDI newsletter.

Duane detailing the header string check that will provide the user with lists of field names provided in the metadata and the field names in the data entity so that the user can visually compare these lists of strings.

A discussion of the meaning of header row ensued to which there seems to be some different perspectives. Some feel that 'header row 10' means row 10 has the data field names with rows 1-9 being supporting information whereas Duane would think of 'header row 10' as meaning rows 1-10 are the header.

John's VCR data has 20 or so rows of information before the 'header row' with row 20 (or whatever) contains the column names of the data entity.

Suzanne asking for an example of a data set that would have more than one header row?

Margaret noting that we can now do an assessment of existing data sets once the check is available.

Ken inquiring as to whether new checks can be sent via email instead of just through newsletters. Gastil seemed to think that this would be feasible. Ken suggesting that PIs should be notified as well so that site administrators are aware also of IM requirements.

Margaret inquiring (to Dan) as to whether the EB is discussing information management as part of the reviews. Dan noted that the reviews just came out and there has been little if any discussion at this point. Corinna noting that there is not a lot of excitement or discussion among the EB about data unless there is something really pressing. A take home being that it is up to the Site IMs to remain aware of IM requirements. Ken inquiring if we could assemble a central resource for all documents and information relevant to renewal proposals that would be accessible and decipherable to site PIs (not just IMs). Ken noting that there was little instruction in the RFP except for reference to the IM website.

August water cooler:

The next water cooler will feature updates from Margaret about units, and from Suzanne about the WiRED working group.

Part II: notes taken by Suzanne Remillard

John Porter – Controlled Vocab WG Update

2013 was last version of CV release.

Goal – researchers can find data. Identified preferred terms; focused on LTER-wide searches; tried to hit a “sweet spot” for number of terms.

In 2011 & 2013, assembled list of words already in LTER metadata (EML documents); ran some statistics.

New statistics; 90% of LTER packages use one or more term in thesaurus.

Resent activities: statistical analysis of keywords; survey to sites about how keywords are assigned.

Current CV is in a thesaurus: synonyms, broader to narrower terms, hierarchical relationships. Integrated into PASTA.

New version of EML includes annotation; do we need to move towards an ontology?

What needs to be addressed at ASM:

Do we need to move to an ontology?

Should we abandon LTER CV in favor of another existing resource?

If not, what upgrades are needed:

How do we deal with place names and taxonomic names as keywords?

For adding words, see document on IMC website.

Questions:

How to add keywords?

How to judge the effort versus success?

Yang Xia – DEIMS

8 people (7 sites) participating; met in March 2018; talked about experience and bugs and needs. Talked about Drupal 8 functionality. Will use DEIMS2 for 2 more years and develop DEIMS3 for Drupal 8. Talked about new features they would like to develop. DEIMS will host ad-hoc ASM meeting (Oct 2). Discussion about where to go forward; how to integrate with EDI. Lost a lot of development capability with the loss of Inigo.

Comments: JP- DEIMS is nice because it is open source, so other groups can participate and help develop; is this happening?

Jim – some international groups are using DEIMS. Ha

Ken- EU uses; Brazil, Taiwan, etc. There are some other efforts. The level of technical support is very limited.

Gastil – ECC

Six active members of WG since last year's meeting. Twice per year release (fall 2017 – checksum); Spring 2018- datetime). Frequent communication within group. Resources are posted on Github. Plan for fall 2018 – add attribute Names; consider checks for EML2.2. Highest purpose is for data reuseability.

Long list of checks which have been cultivated. These are then designed, developed, tested, and staged, then into production.

The group would like to know if they can add 'info' checks outside of the standard release schedule. These will not affect site's harvests, they are informational only.

They would like to roll out the attribute Names / header check to make sure that attribute names in metadata and data file match.

There seems to be some differentiation of how people use and define header rows. If header row = 10; does this mean the first 10 lines or the 10th line? Should it be a best practice for attribute name to be last row in header?

Ken – last RFP provided only a link to the IMC webpage for IM requirements rather than a specific list of requirements.

This seems inappropriate. We should discuss this with NSF at our summer meeting.