Notes 10 Jan 2013 Vocab VTC

Video: Don, Duane, JohnP, Margaret, Yang, Corinna, Gastil, JamesB, Kristin

Phone: Eda, Sven

Agenda: <http://im.lternet.edu/node/1133>

<http://im.lternet.edu/sites/im.lternet.edu/files/Enhancement%20of%20the%20LTER%20Controlled%20Vocabulary.pdf>

The pre-workshop tasks are the six mentioned in the proposal. Final determinations to be made at the workshop. Section 2. Section 3 tasks best done at workshop.

Margaret and John identified tasks to do in advance:

1. enter definitions and/or relationships for existing terms (item 2.1 in agenda)

scope notes is place to put definitions.

Right now those are only there for terms unknown to some.

2. now a polytaxonomy: parent-child but relatively few related terms, linking between different taxonomies.

Enter more related terms (relationships)

Adding additional terms

Solicit new terms from sites.

Mine Metacat to find what existing terms are being used. (orig list in 2006)

Eda – those tasks may be related to each other. When seek new candidate terms, at LUQ they have suggested some new terms for the vocabulary. Eda volunteers for 2.3 (solicit new terms from sites)

Margaret – re new terms from sites… In general, this vocab so far was not intended to encompass all terms at every site.

Eda – when there does not an existing word to accommodate what the scientists want to express with a keyword, then we should definitely consider that as a candidate for a new word.

JohnP – slow to accept words that only apply to one site. For searching LTER, not specific sites. An issue: (in 3.1) discuss whether should expand for other uses.

Assessment (item 3.2 in agenda) – how are people using the words?

Back to section 2 in agenda…

Solicit new terms from sites: have guidelines in draft of Vocab BP. Want proposals for kw that contain:

Definition

Rationale

Where to place in existing taxonomy

Synonyms

Id of dataset already using that kw

Eda – not all proposed kw will be accepted. A protocol for submission.

John P – 2.2 mine metacat/pasta for new candidate words

pathQuery to pull out kw

analytical phase to tally

Gastil – volunteer to pathQuery Metacat

Duane did the original one. Can dig up original code. Duane & Gastil on this task.

Related to that,

Tells rel btwn terms, datasets, and sites.

Don’t have anything that says dataset kw count how many in CV.

& Eda

Complete defs and rels for existing terms (item 2.1)

For current terms

For terms not currently in the Thesaurus

Kristin and Margaret volunteer for this one.

Corinna

Look into taxonomic (species) keywords also.

JohnP

A big one. Margaret and John discussed this. There are other parts of the EML doc, ie taxonomicCoverage better place than keyword for some terms.

Not highest priority.

Question, esp as get into pasta to DataONE where don’t nec have as sophisticated knowledge where to search in metadata, then may be value to put placenames and taxonomic names in kw. Maybe build something that goes into coverage sections and creates additional place/taxon keywords.

Margaret

This group has not stepped back and considered what this vocab is supposed to do, at a high level. What kind of searches do we expect it to help?

The use of the term “keyword” we use for different things. There is a <keyword> element in EML. There is the concept of a keyword as a term in the vocabulary. Those are not synonymous.

If we decide there ought to be a part of the CV that searches taxonomic coverage, then we figure how that will happen.

Eda

The BP is specific about goal of having a vocab. Assure that data is found for a search. The use of the word keyword is what we need to be clear about.

“The goal of adding kw to a met doc is to ensure that reasearchers who want to use your data will be able to locate your data reliably with \_\_\_... Your data can be linked to other datasets, adding scientific value.”

Corinna

What were (the review panel) looking for when they did not find datasets?

Eda

Mentioned Metacat, mentioned local search in Drupal. Kristin and I looked for the datasets they could not find and we found it. Something has to be more clear.

Corinna

Asking… not sure it is up to us to make decisions what vocab use should be. May need to know what people are searching for. I have no problem with having taxonomic terms in vocab if it helps people find datasets.

Our search engine does not use coverage (for kw searches) so having taxa there does not help (for that kind of search).

Margaret

We have a tree of hundreds of terms. We have datasets with terms in title, absyact, and keywords. We could only look in keywords. That is a very different task. Is that the way we want searches to work?

One way puts more burden on the site. Other puts more burden on central system to be complete with relations between terms.

John P

We do need to keep in mind our job is to provide a useable resource to people designing search systems. Not our job to design search system.

Ability to do browing thru vocabulary. Not able to browse based on title, abstract. Have to have parent-child rel for browing to work.

Margaret

Our keywords are flat text in an eml doc.

Corinna

Like idea of thinking of LTER CV as a means to browse. A wider search for terms, including title, abstract, and EML <keyword> to be more inclusive.

If too general of stuff, too many datasets… a fine line.

No problem with having all taxonomic names show up in the EML keywords sets or CV (?) Then if other search engine coming along, put same words into taxonomicCoverage.

Currently looking how my datasets look in DataONE, PASTA, and our system. I am putting stuff everywhere until each of these search systems finds my datasets the way I wan thtem to be found. Tricky when Im finding datasets from several sites. Need to make my title so recognized as I want (ie as NTL).

Margaret

I put some taxonomic words in my EML <keywords> also sometimes.

When we mined EML to get keywords, it get skewed towards sites who use many keywords. Some sites have not yet put words in their documents. We are skewing users’ perception of what LTER does. If we add the taxonomic words then we skew results to wards those sites with more taxa listed.

Corinna

That is because of the way we have gone about this process., how to construct a vocabulary.

I do not agree that this is the best method but we had to start and need to progress. We should add words so it is not skewed. Those sites with fewer words need to use appropriate words. Not 5 ways of saying “forest”.

Margaret

Expand vocab however needed to enable good searches by picking terms.

JohnP

In addition to defs and rels for existing terms, mining terms, and soliciting terms, we have another pre-workshop group aimed at:

Identify options and resources for including taxonomic an geographic keywords (terms) in our metadata. That group can explore options.

Corinna and Don.

Don

Important to think what the taxonomic words are. Placenames seems more difficult.

Corinna

Aaron (not here today) gets volunteered in absentee ☺ in GIS wg today.

JohnP

At least one social scientist: humans so far largely absent from CV. For Eda (since LUQ does soc sci).

Margaret

Need to search by taxa and placenames but not sure those should be in CV. Is it our goal that CV will be used to search datasets ONLY by terms in the <keywords> element?

JOhnP

Create resource that is useful for searching. They can use it however they want.

Duane

Example: we are currently using the existing CV in the Metacat search, both in the basic search (kw, title, abs, people). Used CV in the Advanced Search which searches the Subject field. When you have the CV… the web servies that JP provides that tap into Temetres, can e used in diff ways for diff searches. Not constrained mapping btwn cv and eml elements.

Margaret

Problem for user is … I know what people put in their eml. I would never only search pripro in abstracts or only kw.

JohnP

A dataset that has one of the terms as a kw in the dataset should probably get a higher priority in sorting (relevance) than one which uses the term in the abstract.

Q for Duane: browsing in Metacat was restricted, only 2 levels. For PASTA, will there be more extensive browsing.

Duane

(Next iteration is Knuth.) Knuth will have a much more robust browsing interface. Multi-tiered browsing would be a worthwhile goal.

Workshop (3 on agenda)

3.1 best done at workshop. Vision.

3.2 vocab in current eml

% of sites that use one or more cv kw in eml

How many kw from Disciplines Taxonomy are used versus Substances Taxonomy.

How many datasets have kw from more than one of the Taxonomies

How many datasets are using syn vs kw

Could be linked into metrics.

3.3 BP doc. Draft from 2011 on im.lternet.edu/VocabBestPractices

<http://im.lternet.edu/VocabBestPractices>

The EML BP group may have rec how to put kw into EML. The Vocab BP is how to put terms in the vocab.

Eda

I read not all datasets have core areas, true?

Margaret

The core areas are too broad and general so as to make no sense. Example: a research project concerned with meteorological events such as storms. We measure wave height. Big waves affect the community. A core area called Distubance. I think my wave height dataset may be listed under Disturbance. I do not want when someone looks for Disturbance that they get my wave height data.

JohnP

If a CoreArea kw is included, put in separate <keywordThesaurus> not along with other kw.

Margaret

That will only work in a search that knows that policy. In DataONE that wont be respected.

Meteorological data should not show up under Disturbance or Primary Production. (Ed: I differ)

JohnP

At workshop: get reports from pre-workshop groups.

I can set up a sandbox for adding relationships. We would be approving a 2.0 version.

Schedule VTC to track tasks.

Kristin – defs and rels

Gastil – mining metacat

Eda – new terms

Corinna – tax and geo terms

Those people sched sub-group vtcs.

Sched main wkshop –

Adv to sooner but adv to later to get more pre- work done and weather better ☺

Coast research center in Virginia. Doc, harbor, ocean, not delux.

JohnP

One addl person interested is Todd Crowl who had parcipated in earlier workshops.

Eda

We need a scientist.

Margaret

If one task is to vet and be sure defs are correct, might need broad expertise in ecological concepts.

JohnP

Could have some PIs on standby for email questions during workshop.

Roger Ruse, Julia jones, Todd crowl participated in orig vocab.

Cc doodles to vocab@lternet

Margaret

Schedule to track progress on tasks.

JohnP

Sub-groups can report back via email.

Margaret

DataONE does faceted searches. They are mining vocabs. They are putting terms into ontologies. They will present in Feb.