Meeting Notes 11 July 2011

VTC on LTER Controlled Vocabulary

Presenting: John Porter (VCR)

Participants: Don Henshaw (AND), Emery Boose (HFR), M. Gastil-Buhl (MCR), Jason Downing (BNZ), Margaret O’Brien (SBC), Yang Xia (LNO), Duane Costa (LNO), Nicole Kaplan (SGS)

Phone:

Slide 1

Since last IMC, wg has been active. Funded wg proposal for 2 wkshps march and may plus lots of VTCs. Details on WG web page.

participants

products

(see ppt)

Slide 4

Goals for initial keyword list:

Identify preferred terms to put into EML docs

“sweet spot” between too long and too short number of terms in list

slide 6

NISO standard doc 2010 update: rules for forms of words.

Nouns should be plural if something you count.

Singular if “how much” makes sense.

Slide 7

Slide 8

Structure complexity: list to ontology

Focused on reaching a taxonomy or polytaxonomy plus some related terms which would make us a thesaurs. But focus on taxonomy of parent-child.

Rules: within a taxonomy, only a certain type of thing, ie Processes or Things but not mixed.

Realized this is a smart rule. Makes it work well.

Example: primary production related to carbon. One a process and one an element so cannot go into same hierarchy. Can tie together with related terms.

Got every term into a taxonomy. Added related terms. Not same level of rigor and effort for related terms. Not fleshed in. Good stage with taxonomy, a start on thesaurus.

Now have 10 taxonomies in the polytaxonomy. 627 preferred terms pus >100 non-preferred synonomy. Carbon dioxide is preferred. CO2 is non-preferred.

When site had variant things, changed in list and tracked non-preferred terms. Miss-spellings are included as non-preferred.

Margaret is Czar on removing terms. Too ambiguous to be meaningful. “above ground” because above-ground what? Found 25 or 30 of those. Focused on science keywords. Tossed taxonomic names but left general categories; threw out Brown Trout because included Trout in general.

We will have separate keyword sets for taxonomic information. In EML that goes into taxonomicCoverage.

Slide 10

Uses of this PolyTaxonomy

Searches

Rule: a child must always refer up to the parent. Some not all. When search on general term, also search on more-specific terms and synonmys.

How to make it easy for sites to use the CV in our systems for producing EML.

HIVE

Allows read a doc and suggest kw appropriate from CV.

Web form auto-completing.

Browse to add kw.

Slide 11

TemaTres – Argintinian, or some Spanish-speaking country, developed php-based web tool for developing thesauri. Easy to create new instance of it. DataBits Article talks about this.

Customized TT to do searching.

Slide 13

Narrower terms + related terms + narrower of related = dangerous combo

Search on fire and get back data on flood.

Searching narrow and syn is safe.

Searching narrower of related can give whacky results

Slide 14

Xslt

Slide 15

This is all impl,ented with a WS framework

All it takes to run them is to put into the config file the web addr for the service interface of the tematres \_\_ you want. Type in one \_\_\_ to swtich tools.

Slide 16

Keywords Distiller

The more data you give it the better it will do.

Duane – curious did run saem text in hive to see which better?

JohnP – not tried yet. Guess hive better. Not restricted to one of these.

What comes back now is list of words which you can cut-and-paste but may like to make it easier, ie return in EML snippet, or your EML with new content added to it. Needs to be a person to check off what is right.

Slide 17

Autocomplete tool for web forms: 2 files: a php file to put on server and a sample form that has some javascript. Copy javascript into your form. Does not have to start the same. Prim finds Net Primary Productivity. Get zip.

Slide 18

WS built on top of TemaTres WS.

List of kw

Only returns preferred terms. Don’t want auto-complete with non-preferred terms.

To do a site-specific, service=your-URL then returns your list.

Slide 19

Kw expander

Used Sven’s instructions how to do good WS.

Slide 20

Next Steps

Last para of BP: what to do if need word added: propose the word, rationale, where should go in taxonomy, any synonyms. Tell where at least one dataset that already uses that word. Do not want any word to return no datasets.

Proposed words will be fwd to controlled vocab cmmttee. Conform to Z39.19 format and placement. Circulate for vote on list. Add if approved.

If you have terms specific to your site, inappropriate to all other sites, then make your own instance of TemaTres.

AND or LUQ instances of TemaTres have extra buttons, LTER vs Local.

Slide 21

Priorities for LNO. 1 week of Duane’s time to link CV to existing LTER website. Search datasets.

Slide 22

Replace metacat hierarchy of terms with LTEr CV. Limited to 2 levels displayed. May have to mush up the levels. Browse (cached searches?)

Basic search box. On metacat arrival page. Pop-up list now is not preferred terms, including all terms currently in use. Make it also search on synonyms. OR variants. Not necessary to re-keyword all EML docs as long as synonyms are in TemaTres.

Advanced Search – no autocomplete. Can use Booleans. More complicated operation. Room on page to add boxes, check-off which searches to do.

Slide 23

Slide 24

Priorities

Don – clarify. Synonyms, narrower terms, related terms. As I recall, synonyms=related.

JP – synonyms have to be almost exactly equal terms, not a variation. CO2 = carbon dioxide. In TT those are “use for” terms.

Related terms: fire related to disturbance but disturbance is a process but fire is a thing. Relation is horizontal, not parent-child.

Duane – having the WS in front of TemaTres helps, makes scope more manageable.

Emery – table on slide 24 looks great. Kw move to top, browse to 2nd. Will help on local site keywording.

JP – would be nice to provide a site of all kw in existing docs.

Duane – doable. Enter as a track ticket.

Jason – search priority most important. Priority as-is.

Keywording wont really help BNZ. KW are all in database and EML is generated out so EML snippet not useful.

JP – how should we return the data?

Could develp a sample stylesheet to convert the EML snippet to insert statements.

Jason – a list is the most useful for input to a database.

Yang – I like the search because every site can use same vocab “on same page”. How should we deal with related/similar term, such as diversity, richness, coverage.

JP – advanced search allows you to search both. Richness is not syn with Diversity but may be related. Different, related meanings. Simple search interface, given Diversity, would return that and narrower, and related terms, including Richness.

Richness would give Diversity and Diversity would give Richness.

Do some searches. Should we be doing only syn? Syn + narrower? Syn + narrower + related? Or are related too far off?

Priorities:

Searching –

browse,

simple search,

adv same as slide 24.

Kw – change priority to:

1st HIVE semi-auto

2nd Browse interface for keywording

3rd html form auto (already done

Other

Best Practices.

Basically, rec are that you try to include at least one of the words from each of the ten hierarchies. If someone searches by browsing a heir, they should find at least one of your datasets.

Do not have to have the exact term. Pick closest reasonable term. Higher imply lower. Try to get as low on tree as possible. Ie reforestation is part of landscape change.

Finally – rules for how to add additional terms.

If would like to experiment, can have own TemaTres site on your own. Will export common bibliographic formats to move to a different tool.

Nicole –

Related to govt. IMC how to move BP forward.

JP – will recommend BP to IMExec, let them carry ball from there. LTER Vocab is a creature of the IMC. Does not answer to NISAC or EB.

Margaret –

Insert kw set using xslt on eml.

LTER Core Resaerch Areas’ keywords

Parent –child: a mouse is always a rodent.

Cacti to desert plants: not all cacti are desert plants.

Children must always be under the parents.

Check to be sure these relationships are correct. Look in your area of expertise.

Aim to have everything locked down for version 1.1 of VC and Thesarus for IMC meeting.