

Enhancement of the LTER Controlled Vocabulary to Support Data Synthesis, 2013

- John Porter & Margaret O'Brien

The LTER Controlled Vocabulary Working Group obtained funding for a workshop aimed at enhancing the utility of the LTER Controlled Vocabulary. It aimed at 6 tasks: addition of new terms to the vocabulary, identification of resources for place and taxonomic resources and developing a plan for how to use them, develop quantitative tools for assessing keyword use in LTER metadata, adding additional relationships among terms, adding definitions for terms and completing work on a "Best Practices" document for LTER data keywording.

Substantial progress was made on all these tasks during the workshop and during pre- and post-workshop planning sessions.

Participants in the in-person workshop included professional librarians (Sherry Lake and Ivey Glendon from the University of Virginia) and LTER Information Managers Margaret O'Brien, Kristen Vanderbilt, Donald Henshaw and John Porter. It was held May 22-23 at the Anheuser-Busch Coastal Research Center in Oyster, VA.

TASK 1: Adding Terms

A solicitation for additional terms was sent to every LTER site. In response proposals for 230 new terms were received from four sites. Each term was discussed by workshop participants and a decision made regarding the suitability of the term for inclusion in the Controlled Vocabulary. 75 terms were selected for addition as preferred term. Several additional terms were added as "use for" terms for existing preferred terms.

Accepted terms have been added into the Controlled Vocabulary, maintained in TemaTres. Justification was provided in a Google document for each rejected term. Terms were evaluated based on whether there was repeated use at a site, use at other sites, used as a search term at the LTER portal, the uniqueness of the term (not represented by existing terms), and the clarity of the term. Definitions of many new terms were added when a good definition was provided.

Reasons for rejecting proposed terms, and abbreviations used are:

NR - only used [once] at [1] site

NU - not used as a search term (per user logs)

A - absent from other sites

ADJ - a stand-alone adjective

V - Vague: leaves a question, not a likely search term

AWE - adequate alternative word exists

NC - needs clarification/better definition

During the course of the discussion of terms, several recommendations emerged:

- All sites should explore the current list with the new terms while adding new terms to data packages.
- Sites that contributed terms for evaluation should check the evaluation spreadsheet to better understand why terms were rejected and for hints on replacement terms.
- Many of the terms were too specific as search/keyword terms and were better fitting at the attribute level.
- If we should evaluate new terms again: sites should first carefully screen potential terms versus the LTER Vocab and only suggest terms that cannot be well-represented by existing terms.
- Definitions of new terms should be required.

In addition to words formally proposed, we also examined lists of keywords used at sites and added otherwise suitable terms that were used in 10 or more datasets at two or more sites. This resulted in the addition of 25 additional terms to the Controlled Vocabulary.

Additionally we discussed ideas for enhancing social-science keywords with Ted Gragson, lead PI of the CWT LTER site, himself a social scientist, during the Science Council meeting in Las Cruces, NM. We also scrutinized lists of keywords from sites mandated to pursue social science (e.g., BES, CAP, CWT and NTL). A challenge for adding social science keywords is the relative paucity of social science datasets relative to ecological datasets in the LTER Metacat and PASTA systems.

TASK 2: Adding Definitions

Margaret O'Brien developed a web service client that queried definitions for preferred terms in the Controlled Vocabulary. The raw results required vetting to improve formatting and to eliminate meanings that are not relevant in an ecological context. Definitions for 309 terms were vetted and added to the Controlled Vocabulary, in addition to those added when new terms were added. This brings the total to 363 defined terms in the Controlled Vocabulary.

TASK 3: Complete "Best Practices" Document

The workshop jointly completed the "Best Practices for Adding Science Keywords to LTER Metadata"

(<http://im.lternet.edu/sites/im.lternet.edu/files/BestPracticesforAddingScienceKeywordstoLTERMetadata.docx>) document. It includes recommendations for selecting keywords for metadata documents and procedures for proposing new keywords for the Controlled Vocabulary.

TASK 4: Identify Alternative Thesauri for Places and Taxa

In the process of identifying alternative resources the working group concluded that the geographic coverage and taxonomic coverage elements of metadata documents were a more appropriate way to convey spatial and taxonomic information than through the use of keywords. We have subsequently been in contact with the GEONIS working group regarding spatial information and have initiated some informal discussions among Information Managers regarding development of web services to produce taxonomic elements for EML documents.

TASK 5: Metrics

M. Gastil-Buhl developed a set of queries that can be used to collect a variety of keywording information from existing documents. Duane Costa ran previously developed queries to collect information on queries from users of the LTER Data Portal.

TASK 6: Relationships

During the course of adding new preferred terms, new relationships were added to the Controlled Vocabulary. The majority of these were for synonyms (“use for” terms), but additional relationships were also added. Additional work needs to be done on better defining best practices for related terms, because existing guidance (e.g., Z39.19) is weak.

ADDITIONAL TASKS

- The Tematres software used to manage the Controlled Vocabulary was updated to the latest version.
- There was a discussion of the “vision” for how the Controlled Vocabulary should be used. It focused on the need to incorporate the capabilities provided by the Controlled Vocabulary in the new PASTA Portal.

2013 Controlled Vocabulary Statistics:

710 total preferred terms

200 synonyms (“use for” terms)

309 new definitions added

363 total definitions