A Controlled Vocabulary for LTER Datasets

This document lays out an operations plan for enhancing data discovery of LTER datasets. Approval is requested from the Executive Board and/or Science Committee to move forward with the plan.

In Brief:

The Problem: Existing science keywords used to describe LTER datasets are esoteric and disjunct, making data catalogs difficult to search and reducing the reliability of search results.

The Proposed Solution: Provide a consistent controlled vocabulary of LTER keywords for use by sites in publishing datasets. This controlled vocabulary would supplement, not replace, site-specific keyword lists, and would be periodically reviewed and modified by the LTER Information Management Committee (IMC), subject to the authorization of the Executive Board and/or Science Council.

In Full:

Background:

Currently the keywords used to characterize datasets at most LTER sites are uncontrolled, meaning that they are selected entirely by the data creator. One of the challenges facing LTER and external researchers in discovering data from LTER sites is inconsistent application of keywords. A researcher interested in carbon dioxide measurements must search on both "Carbon Dioxide and "CO2." Moreover, the existing set of keywords is highly diverse. For example, in a 2006 survey of EML documents in the LTER Data Catalog, over half (1,616 of 3,206) the keywords were used in only a single dataset, and only 104 (3%) of the keywords were used at five or more different LTER sites (Porter 2006).

To address this problem, in 2005 the LTER Information Management Committee established an *ad hoc* "Controlled Vocabulary Working Group" and charged it with researching the problem and proposing solutions. To that end the group compiled and analyzed keywords found in LTER datasets and documents, and identified external lexographic resources, such as controlled vocabularies, thesauri and ontologies, that might be applied to the problem (Porter, 2006). Initially the working group attempted to identify existing resources, such as the National Biological Information Infrastructure (NBII) Thesaurus, that LTER might be able to adopt wholesale. Unfortunately, using widely-used LTER keywords as a metric, none of the external resources proved to be suitable. Too many keywords commonly used in LTER datasets were absent from the existing lexographic resources. So, starting in 2008 the working group focused on developing a LTER-specific controlled vocabulary, ultimately identifying a list of ~600 keywords that were either used by two or more LTER sites, or were found in one of the external resources (NBII Thesaurus and Global Change Master Directory Keyword List), and conformed to the recommendations of the international standard for controlled vocabularies (NISO 2005) . This draft list was then circulated to members of the LTER Information Management for suggested additions and deletions, which were then voted upon (Porter, 2009). The final list consists of 640 keywords (http://intranet.lternet.edu/im/files/im/LTER_Keywords_V0.9.xls).

The final list was presented to the Information Management Committee (IMC) at the 2009 All-Scientists' Meeting. The sense of the meeting the keyword list was sufficiently evolved to form the basis of an LTER Controlled Vocabulary, but that adoption of an official LTER controlled vocabulary was beyond the powers of the IMC, and that a system of procedures needed to be developed for managing LTER-specific lexographic resources. Therefore, the Controlled Vocabulary Working Group was charged with preparing a proposal to the Executive Board for adoption. This document is that proposal.

Proposal:

We propose that the Executive Board and/or Science Council delegate to the LTER Information Management Committee Executive Committee (IMEXEC), or their designee, the task of adopting and maintaining an official list of LTER Keywords that should be used by sites in their dataset metadata (along with whatever additional "local" keywords sites wish to use), and that the LTER Network Office (LNO), in consultation with the Network Information System Advisory Committee (NISAC), should assist in providing databases and tools needed for managing and using the list. Procedures established by IMEXEC for managing the controlled vocabulary should include periodic review and revision of keyword lists.

References:

NISO. 2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. ANSI/NISO Z39.19. http://www.niso.org/kst/reports/standards/kfile_download?id%3Austring%3Aiso-8859-1=Z39-19-2005.pdf&pt=RkGKiXzW643YeUaYUqZ1BFwDhIG4-24RJbcZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ymGsj_IKVa86hjP37r_hONsJghRDv2N-zj4TZCh8Dp01rZbmK3O-8vcVjh4hezP

Porter, J. H. 2006. Improving Data Queries through use of a Controlled Vocabulary. LTER Databits Spring 2006. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/#4fa

Porter, J.H. 2009. Developing a Controlled Vocabulary for LTER Data. LTER Databits Fall 2009. http://databits.lternet.edu/node/70