# Finding the Data: Enhancing the Utility of the LTER Controlled Vocabulary

1. *Vision and Summary:* **Scientists seeking data should be able to efficiently and reliably locate LTER datasets through searching, browsing or following links from non-LTER systems.** Unfortunately, efficient and reliable searches are not currently possible because of eclectic application of keywords in LTER metadata. To address this problem, the LTER Information Management Committee (IMC) created a working group to help create a controlled vocabulary for use in LTER datasets (<http://databits.lternet.edu/spring-2010/controlled-vocabulary-lter-datasets>). Following recommendations in ANSI/NISO Z39.19 (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies), the group created a list of 640 preferred terms and associated synonyms that maximized overlap with external vocabularies (e.g., NBII Thesaurus) and application across LTER sites. In early 2010 the LTER Executive Board committed to helping to locate some domain scientists to work with the Information Management Committee on future activities related to keywords, and endorsed the use of the list by LTER sites. Subsequently, the LNO, in association with the “Helping Interdisciplinary Vocabulary Engineering” (HIVE) project, developed some web-services and a prototype client that scans EML metadata and suggests appropriate keywords.

During the 2010 IMC meeting, the VOCAB working group identified five major actions that were needed to realize the vision of efficient and reliable identification of LTER datasets. These actions were:

* 1. Develop organizational structures and protocols for managing improvements to the LTER Controlled Vocabulary
  2. **Develop a database and associated tools for managing the controlled vocabulary**
  3. **Create tools to aid in adding keywords from the controlled vocabulary to existing and future metadata**
  4. **Create taxonomys (hierarchies) for keywords that can be used to enhance browsing,** and
  5. Develop search and browse tools that exploit the controlled vocabulary

The VOCAB product-oriented working group will focus on actions “b” through “d”. Action “a” can be accomplished without additional support, and action “e” is dependent upon the other actions and so can be deferred until more progress has been made.

1. *Scope of work:* Products of this working group will be applicable to all LTER sites and will enhance the overall functionality of the NIS (you can’t integrate data you can’t find). They will also be helpful in enhancing existing search and browse systems.
2. *Products:* Products of the proposed activities will include:
   1. **A database of keywords** that implements a data model based on “11.1.4 Term Records” in NISO/ANSI Z39.19. The database will accommodate addition of relationships needed to implement taxonomys (hierarchies) of keywords and synonym rings. Additionally, it will allow a “scope” to be applied to each term, so that individual site, as well as LTER-wide keywords can be accommodated. This database will be a critical resource for keywording and search tools, as well as facilitating management of the controlled vocabulary.
   2. **Tools for adding keywords to existing and new metadata** will be developed. In specific, we will enhance the “EML Tagger” tool developed by Duane Costa in association with the HIVE project which suggests keywords for existing documents, so that it can add keywords to EML documents, or return them to other applications. For new data, a web service and associated client to support “autocomplete” of web-based keyword entry will be created. A prototype keyword browser that uses taxonomys for locating appropriate keywords will also be developed.
   3. **Several taxonomys (hierarchies) will be developed** that place keywords into a broader context. For example, LTER core areas can be used as high-level concepts in a “research topic” taxonomy. Other taxonomys might address “biomes,” “biota” or “biogeochemistry.” In addition to fully implementing several taxonomys, we will also identify other potential taxonmys that might be created in the future.
3. *Tasks:* **Intensive Development Workshops** will be used to develop products “a” and “b.” Each workshop will consist of 4 “visiting” participants, 1 “local” participant and a variable number of “virtual” participants. Workshops will be 3 days in length with travel days before and after. The goal of the first workshop , to be held in February 2011 (probably at the LNO), will be to complete design of the controlled vocabulary database and initiate development of keywording tools and services. The second workshop will be held in late May 2011 (probably at a biological field station). It will focus on finalizing implementation of the controlled vocabulary database and completing development of the tools and services. During each workshop, time will be scheduled for VTCs to communicate ideas and results and get feedback from virtual participants.

**Development of the taxonomys** will be conducted in coordination with the May 2011 Science Council Meeting. Prior to the workshop, LTER IM’s will create several “straw man” taxonomys. An “in person” workshop involving three IM’s and two domain scientists attending the SC meeting (identified by the Executive Board) will be held on the day prior to the start of the SC meeting. The IM’s will continue work on the polytaxonomys, incorporating feedback from the designated domain scientists and other SC meeting attendees during the SC meeting, with the goal of having completed several taxonomys by the end of the SC meeting.

5. *Participants*: Participants will be drawn from membership of the VOCAB group. VOCAB chair John Porter will participate in person at both intensive workshops. Duane Costa will be a valuable participant in at least the first workshop, which focuses on building on his HIVE work, either as an in-person or virtual participant as will Kristen Vanderbilt, who volunteered during the IMC meeting. Inigo San Gil from the LNO has also expressed an interest. Additional IMs with expertise in data modeling and/or web services will be selected. IMC co-chairs Margaret O’Brien and Don Henshaw have volunteered to participate in the taxonomy development activity held in association with the SC meeting, and several other IMs have expressed an interest in this activity.

1. *Budget*:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Number of "in person" Participants** | **Airfares @$600** | **Travel Cost** | **Food & Lodging Person Days** | **Food & Lodging Cost @ $77+$46** | **Meeting Costs** | **Total Cost** |
| Database & Keywording tool Development I | 5 | 4 | $2,400 | 15 | $1845 | $100 | $4,345 |
| Polytaxonomy Meeting | 5 | 2 | $1,200 | 10 | $1230 | $0 | $2,430 |
| Database & Keywording tool Development II | 5 | 4 | $2,400 | 15 | $1845 | $100 | $4,345 |
| **Totals** | **15** | **10** | **$6,000** | **40** | **$4,920** | **$200** | **$11,120** |

1. *Budget Justification*: Airfares were estimated to average $600 per attendee. Lodging and food costs were estimated at $77 lodging + $46 food per diem. When applicable, $100 for incidental meeting costs (e.g,. the traditional caffine and twinkies needed to keep programmers operating, and meeting supplies) were requested. For the Polytaxonomy meeting, to be held in association with the SC meeting, domain scientists are assumed to be traveling to the meeting using other travel funds, but will need an extra day of lodging and food.

Submitted by John Porter on behalf of the VOCAB working group.