Enhancement of the LTER Controlled Vocabulary to Support Data Synthesis

--John Porter (jporter@LTERNET.edu) & Margaret O'Brien (mobrien@LTERNET.edu) & the LTER Controlled Vocabulary Working Group

Goals: The credo of this working group is that: **Scientists seeking data should be able to efficiently and reliably locate LTER datasets through searching or browsing**. In 2011, the LTER Controlled Vocabulary Working Group made substantial progress towards this goal (report attached). That progress included creating a thesaurus of over 600 keywords (http://vocab.lternet.edu) and using that thesaurus to automatically enhance searches on the LTER Data Portal by automatically searching for synonyms and narrower terms. Data browsing was also enhanced by implementing a more complete browse interface. The group also produced a series of technical resources that can be used by LTER sites and researchers (http://im.lternet.edu/vocab_resources). The LTER Controlled Vocabulary was also fully integrated into EnvThes thesaurus being developed by the European ILTER. PASTA is also incorporating improved search technologies that rely on the thesaurus.

Despite these successes there remains additional work that needs to be done to maintain and further enhance the LTER Controlled Vocabulary so that LTER and other researchers searching for data will be able to locate them for use in synthesis projects.

Activities: This proposal requests funds to support a workshop, coupled with pre-workshop activities such as videoconferences and querying of data sources in the LTER Metacat and PASTA, to address the following tasks:

1. **Enhance the controlled vocabulary by adding new terms**:
    a. **Collect new keywords to be added** – Query terms currently used in LTER Datasets and see if there are additional terms that are used by more than one site that should be added to the controlled vocabulary. Additionally, requests for additional terms will be sent to each site.
    b. **Focus on adding human-related keywords** - The current vocabulary has only a few terms related to socioeconomic studies and ecosystem services, therefore we need to identify additional terms (perhaps adding one or more additional taxonomys).
    c. **Assess candidate keywords** - Assess whether new candidate keywords should be added, and if so, where in the existing taxonomys. Sites will be given the opportunity to "vote" on candidate keywords. This is also an opportunity to add needed synonyms for terms already in the vocabulary.
2. **Identify resources for additional types of keywords and plan for how to use them**
    a. **Identify potential Gazetteers** – the current controlled vocabulary excludes place names. There are large, existing, gazetteers that link place names to geographic locations and we should identify the gazetteers, in conjunction with the GEONIS working group, that might help meet this need for LTER.

b. **Identify and evaluate potential sources for taxonomic keywords** – similarly, taxonomic keywords are excluded from the existing controlled vocabulary.

c. **Create plans for implementing place-based and taxonomic Controlled Vocabularies** – Once we have identified resources, we can begin planning on how they should be used and what software will be required.

3. **Develop analytical tools for assessing keyword use in LTER Metadata** – there is currently no easy way to assess whether LTER Metadata contains keywords from the vocabulary. The tools we will develop will provide a quantitative basis for evaluating keywording for individual documents and for the network in general.

4. **Add additional relationships among terms** – the current LTER Thesaurus is primarily a polytaxonomy, with only a few related terms identified. We will take a more comprehensive approach to identifying relationships among terms that can be used to aid searching.

5. **Add definitions for all of the terms in the controlled vocabulary** – to maximize long-term utility, the standard for controlled vocabularies, NISO Z39.19, recommends that a definition be applied to each term. Currently this is done for only a few of the terms, so we propose to add the needed definitions.

6. **Complete work on the draft Best Practices for LTER Keywording document** – A draft document was produced in 2011 (http://im.lternet.edu/VocabBestPractices), but it needs additional crafting to incorporate metrics (#3) and finalize a few controversial points.

Participants: The workshop will include nine participants. Because of their familiarity with past efforts, workshop participants will be drawn primarily from the members of the LTER Controlled Vocabulary Working Group (D. Henshaw (AND), J. Jones (AND), R. Ruess (BNZ), J. Downing (BNZ), D. Costa (LNO), M. Servilla (LNO), I. San Gil (MCM), J. Brunt (LNO), E. Melendez-Colom (LUQ), T. Crowl (LUQ), C. Gries (NTL), M. O'Brien (SBC), K. Vanderbilt (SEV) and J. Porter (VCR)), which includes both information managers and Principal Investigators. Additionally we will include at least one social scientist and at least one member of the LTER GEONIS working group, to be recruited prior to the workshop.

Timing: A three-day workshop will be held at the Anheuser-Busch Coastal Research Center in Oyster, VA in early May of 2013. The ABCRC supports videoconferencing capabilities, so that remote participants can also be included. Starting in February 2013 a set of monthly or bi-weekly videoconferences will set the stage for the workshop.

Products: The workshop products will include an enhanced version of the LTER Controlled Vocabulary with additional keywords (including socioeconomic terms), definitions and relationships; metrics of keyword use in LTER metadata, and tools for collecting them; plans for how to incorporate place and taxonomic keywords into the vocabulary; and a finalized "Best Practices" document.

Budget:

Activity	Number of "in person" Participants	Airfares @$700	Travel Cost	Food & Lodging Person Days	Food & Lodging Cost @ $20+$46	Meeting Room, Van Rental, tolls	Total Cost
Workshop	9	8	$5,600	36	$2,376	$882	$8,858

Budget Justification: Lodging and food costs were estimated at $20 lodging at the ABCRC Field Station + $46 food per diem. Airfares were estimated to average $700 per attendee. Nearest airport is Norfolk International Airport, which requires a 45-minute trip across the Chesapeake Bay Bridge Tunnel ($12 toll each way). Taxis would be too expensive, so we have requested funds for rental of two mini-vans to ferry participants to the meeting and for local travel (e.g., restaurants etc.) during the meeting. Thus incidentals come to $882 which includes $150 for meeting room rental, $660 for two mini-vans, and $72 (3 round trips) in tolls.