# Best Practices for Adding Science Keywords to LTER Metadata

May 22, 2013

The goal of adding keywords to a metadata document is to assure that researchers who want to use your data will be able to locate it reliably and efficiently. Adding keywords from a controlled vocabulary means that data can be linked to other similar datasets, greatly adding to its scientific value. Before initiating keywording of datasets you should familiarize yourself with the existing keywords and taxonomys (word trees). They can be viewed at: http://vocab.lternet.edu. There are additional tools and resources available at: http://im.lternet.edu/vocab\_resources.

Here are some best practices for keywording your metadata documents:

* **Use the most specific possible keywords.** When searching or browsing, higher-level the “parents” or higher-level terms for each keyword are implied, so choosing the most specific “child” term combines the highest level of discoverability with the maximum level of discrimination. For example, rather than choosing “transects” choose the more-specific child-term “vegetation transects. ”
* **Be willing to make reasonable compromises.** By its nature keywording requires compromise. Datasets vary widely, but if that uniqueness is fully expressed in the keywords, then searching becomes virtually impossible. Therefore you may need to make reasonable compromises in order to be able to use keywords from the controlled vocabulary. For example, you may have conducted a study on the population ecology of rodents, but when you go to the controlled vocabulary, “rodents” isn’t listed, but “small mammals” is. Rather than simply adding “rodents” as an uncontrolled keyword, use the next best term (“small mammals”) instead. If you want, you can also add “rodents” as an uncontrolled keyword, but be sure to add the nearest keyword from the list as well because uncontrolled keywords don’t show up in browse-type searches.
* **Provide keywords from as many of the different taxonomys (top-level groupings) as possible.** Ideally there should be at least one keyword from each of the different taxonomys in the controlled vocabulary. However, there may be some taxonomys that are simply not applicable to a specific dataset and these may be skipped. If a user is browsing down through a taxonomy to locate data, and there is not a keyword from that taxonomy associated with the dataset, it will not be discoverable through browsing. Therefore, using a broad selection of keywords is a good idea.
* **When using keywords not in the Controlled Vocabulary, put them in the proper form**. If you really need to use a keyword not already part of the controlled vocabulary, put it in the proper form. The international standard NISO Z39.19 (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies) has recommendations the form of keywords. For example, nouns are preferred and they should be plural if they are something that is counted, but singular if they are something to which the question “how much” might be reasonably applied. See section 6 of NISO Z39.19 for details.

## Adding Keywords to the Controlled Vocabulary

If you think a keyword really should be part of the controlled vocabulary, propose that it be added. A proposal typically will consist of the

1. keyword,
2. a definition,
3. the rationale for adding the keyword,
4. a suggestion on where it should be placed in the existing taxonomys,
5. if there are any related terms to which it should be linked and non-preferred synonyms,
6. a count of the datasets already using the keyword, and
7. a count of the sites already using the keyword

The following criteria will be applied by the LTER Controlled Vocabulary Working group in considering new terms for inclusion:

### Criteria for Selection of New LTER Keywords

|  |  |  |  |
| --- | --- | --- | --- |
| **What** | **Rationale** | **Do’s** | **Problem**  **Abbreviation** |
| Keywords should be applied to a number of datasets across the LTER Network. | Data discovery is the goal, so keywords that find data are most useful. | Propose keywords that are used at several other sites, and numerous datasets | NR - not repeated in multiple datasets |
| Keywords should be used at more than one site | A goal is to enable cross-site searching | Propose keywords that are used at several other sites | A - absent from other sites |
| Avoid proposing stand-alone adjectives | Stand alone adjectives imply an “of what” question. Such as “aboveground” raises the question “aboveground what?” | Propose nouns or possibly verbs, but not stand-alone adjectives. Perferred terms can include an adjective with an object (e.g., aboveground biomass) | ADJ - stand-alone adjective |
| Be specific | Vague or ill-defined terms are hard to consistently assign | Use specific, unambiguous and well-defined terms | V - Vague |
| Avoid duplicating concepts already in the Controlled Vocabulary | Duplicative keywords lead to inconsistent keyword assignments | Avoid duplication of nearly-equivalent terms | AWE - adequate alternative word exists |
| Keywords should be well-defined | Without definition and context some technical terms may be difficult to assess or place | Provide good definitions | NC - needs clarification or better definition |
| Proposed synonyms should have exact correspondence to the preferred term | Synonyms should not refer to different concepts than the associated preferred term | Select synonyms that are exact matches for the concept described by the preferred term | NS - not a synonym |
| Keywords should be terms that users frequently search on | Keywords that are not searched for by users are not particularly useful. | Propose keywords that are frequently used in searches | NU - not used for search |

Keywords that indicate what a dataset contains are likely to be more useful than keywords that indicate scientific topics to which data might be applicable. Put another way, keywords that aid in data discovery are most frequently about what the data “contains” (e.g., air temperature) rather than what the data is “about” (e.g., climate). Although “about” (i.e., subject) keywords can be useful, they are difficult to consistently assign because to some degree they are in the eye of the beholder, and any given data might be usefully applied to many scientific topics, including those not yet identified at the time of collection. Nonetheless, there are some taxonomys within the Controlled Vocabulary (e.g., disciplines) that consist almost entirely of “about” keywords. Therefore, we recommend proposing “contains” keywords for inclusion in the LTER Controlled Vocabulary, and discourage, but do not exclude, “about”-type keywords.