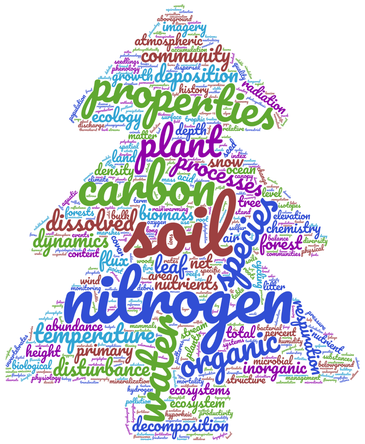


Evolution in the Use of the LTER Thesaurus in Metadata Creation

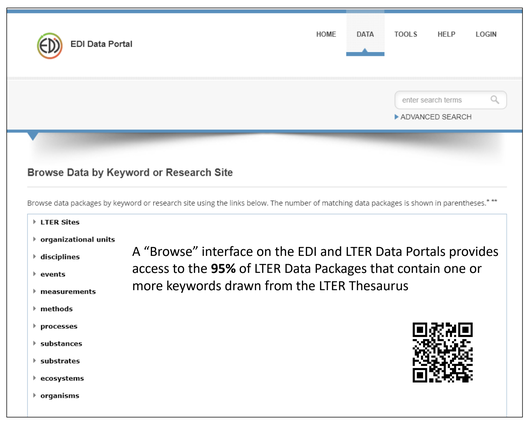
John H. Porter, VCR/LTER

LTER NETWORK

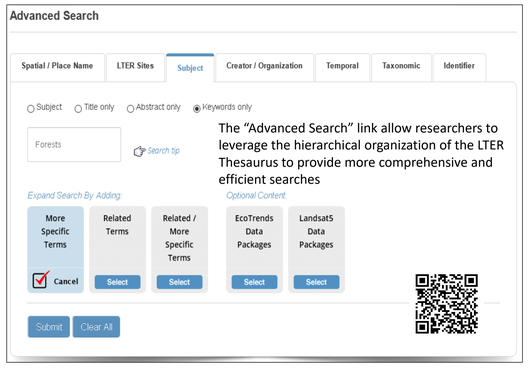
The vision of the LTER Controlled Vocabulary Working Group is that scientists seeking data should be able to efficiently and reliably locate LTER datasets through searching, browsing or following links from non-LTER systems. The purpose of the group is to help increase the efficiency and reliability of data sharing by promoting the use of controlled vocabularies that provide consistent representations of data across all LTER sites.

In 2011 version 1 of the LTER Thesaurus (also known as the Controlled Vocabulary) was created. Here we examine how LTER sites have incorporated keywords from the thesaurus into LTER metadata.

http://vocab.lternet.edu

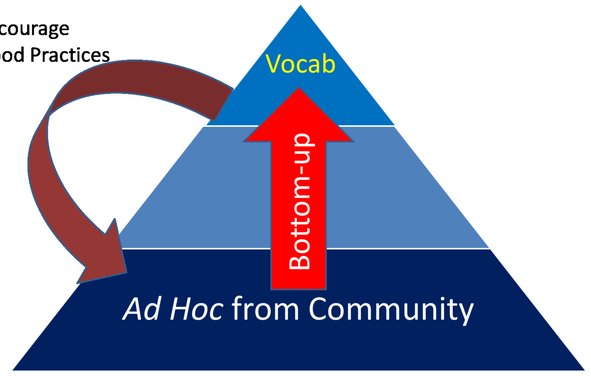


A "Browse" interface on the EDI and LTER Data Portals provides access to the **95%** of LTER Data Packages that contain one or more keywords drawn from the LTER Thesaurus



The "Advanced Search" link allow researchers to leverage the hierarchical organization of the LTER Thesaurus to provide more comprehensive and efficient searches

Encourage Good Practices



The goals of the thesaurus were to facilitate data discovery by eliminating problems with searching for variant versions of concepts (e.g., CO^2 vs Carbon Dioxide). The thesaurus was organized based on terms already used by researchers, but with synonyms and preferred forms identified. When the thesaurus was created in 2011, it was hoped that it would encourage better practices for assigning keywords in metadata documents. As the remainder of this poster shows, that feedback loop has indeed proved fruitful, with increasing and disproportionate use of terms drawn from the thesaurus in metadata documents.

2006

	Terms NOT included in the thesaurus	Terms Later Included in the thesaurus
Number of Uses in LTER Data Packages in **2006**	**14,840**	**14,550**
Median Number of Data Packages per Term in **2006**	**1** Quartiles: 1 to 1	**4** Quartiles: 2 to 14
Median Number of Sites Using a Term in **2006**	**1** Quartiles: 1 to 1	**2** Quartiles: 1 to 3

In 2006, prior to the creation of the LTER Thesaurus, most keywords in metadata were used only a single time. This meant that searching for data was very hit-or-miss. Some terms, which were used in more than one dataset across more than one LTER site, were subsequently included in the thesaurus. These, more broadly used terms, accounted for roughly 50% of all the keywords used in the metadata. Even in 2006, those terms provided substantially better search capabilities, finding between 2 and 14 data packages across 1 to 3 different LTER sites.

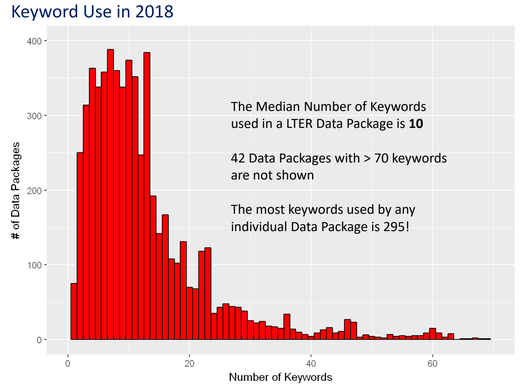
2018

	Terms NOT in the thesaurus	Terms in the thesaurus
Number of Uses in LTER Data Packages in **2018**	**38,622**	**42,795**
Median Number of Data Packages per Term in **2018**	**1** Quartiles: 1 to 4	**17** Quartiles: 7 to 44.5
Median Number of Sites Using a Term in **2018**	**1** Quartiles: 1 to 1	**4** Quartiles: 2 to 8

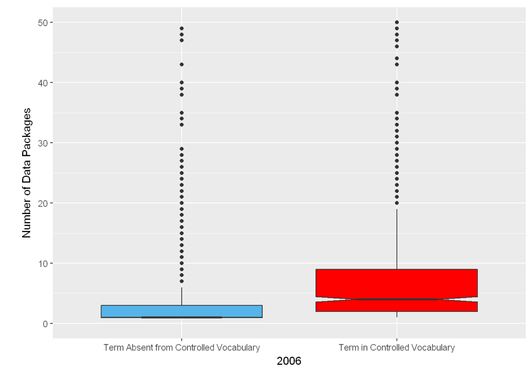
In 2018, 7 years subsequent to the establishment of the thesaurus, things remained much the same for terms not included in the thesaurus. However, those, still esoteric, terms now are in the minority, making up only 47% of the keywords used in metadata documents.

In contrast, the use of terms drawn from the thesaurus has greatly increased relative to 2006. Users searching on terms in the thesaurus can expect to receive between 7 and 44 data packages (a reasonable number to browse – neither too high nor too low), drawn from between 2 and 8 different research sites.

Keyword Use in 2018

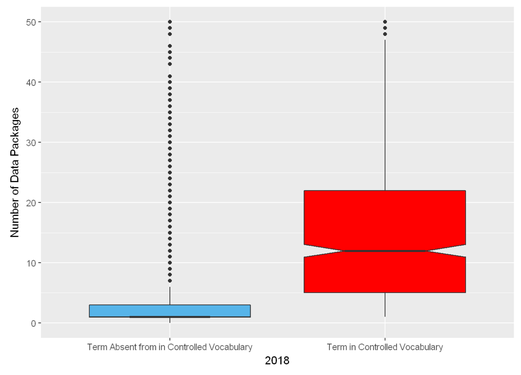


Changes over Time

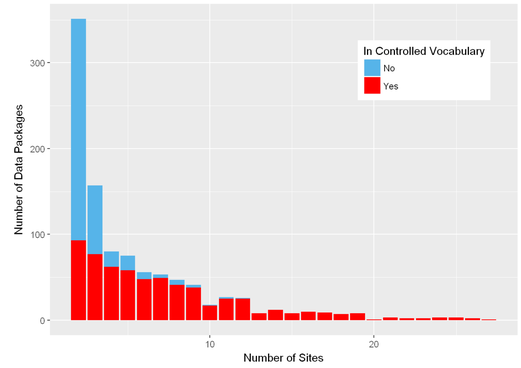


In 2006, the terms later to be used in the thesaurus were already more useful for searching than terms that were not selected for inclusion in the thesaurus. Not shown are 106 outlier terms used in > 50 data packages.

2006 data came from Porter J., D. Costa. 2006. Keywords and Terms from the LTER Network - 2006. Environmental Data Initiative. https://doi.org/10.6073/pasta/3eabff466e2552caa383eafb2ce2b343

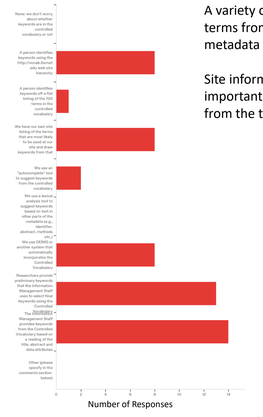


In 2018, the terms included in the thesaurus had greatly increased in utility, whereas the utility of esoteric terms had remained relatively static. In 2018 terms from the thesaurus would be expected to return a median of 17 times as many data packages (17 vs 1) as terms not in the thesaurus. Not shown are 315 outlier terms used in >50 data packages.



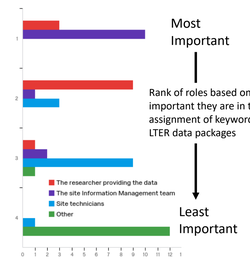
Searches using terms in the thesaurus return data packages from multiple sites much more frequently than terms not in the thesaurus. No term not in the thesaurus was used by more than 12 sites. Not shown are 5,012 terms that returned data from only a single LTER site. Of 824 terms which were used in 5 or more data packages across two or more LTER sites, 78% were found in the thesaurus. Many of the "missing" high-use terms were organization names (e.g., LTER, NSF), places (e.g., USA, Georgia, Harvard Forest) or specific taxonomic names (e.g., *Sporobolus alterniflorus*, hemlock woolly adelgid) that had been excluded from the "thematic" terms in the thesaurus.

Adding Keywords



A variety of methods are used to incorporate terms from the LTER Thesaurus into the metadata creation process.

Site information managers & researchers play important roles in adding keywords drawn from the thesaurus.



Comments on your keywording process
We have a local list of the LTER Controlled Vocabulary and of site ones (not in LTER list and place names). A lookup table is used to control spelling and terms.
Investigators suggest ad hoc keywords on data submission forms, which IM staff map to the cv and then augment based on site knowledge and reading the abstract and methods. Keywords are managed in our RDMS grouped by category/vocabulary, with the LTER CV as a primary category
Our metadata relational database parent table contains a "dictionary" of keywords originating from LTER CV, NBII, GCMD and site-specific terms. Dataset kw are constrained by foreign key.
As much as possible, *our site* aligns keywords supplied by the data provider with those in the controlled vocabulary. This is a manual process requiring looking for a word in the CV via the web interface, and altering the data-provider supplied keyword to match if appropriate. *Our site* also has a suite of keywords that it associates with each data set for internal organziation and to associate *our site* with the data set.
At *our site* I ask the researchers to use vocab.lternet.edu to assign keywords, but what I usually get back are words from the main categories, e.g. 'processes', 'ecosystems', etc. They usually make up other keywords, and I will add the LTER Vocab corrected form of the term and associate it with the LTER Controlled Vocabulary in the EML. Supplying them with a flat list might work better, now that you mention it.
Last year we replaced our custom controlled vocabulary with a modified LTER controlled vocabulary. Essentially we include every keyword from the LTER CV list and mapped all of our *site* keywords into the LTER list. Where there was not any reasonable match we added new keywords. So the list we use is a hybrid but predominantly based and hierarchically structured as the LTER CV. However, we have yet to replace most of the keywords within our EML and this will occur over time as we upload new EML files with the new LTER CV words to replace the older ones.

Thanks to the members of the LTER Controlled Vocabulary Working Group, whose hard work made this poster possible, and to the EDI/LTER Data Portal and Controlled Vocabulary Server whose web services made it possible to access the needed data.

Additional information can be found in: Porter, J.H., 2019. Evaluating a thesaurus for discovery of ecological data. Ecological Informatics 51, 151-156. 10.1016/j.ecoinf.2019.03.002