Evaluating a Thesaurus for Discovery of Ecological Data

Author: John H. Porter[a]

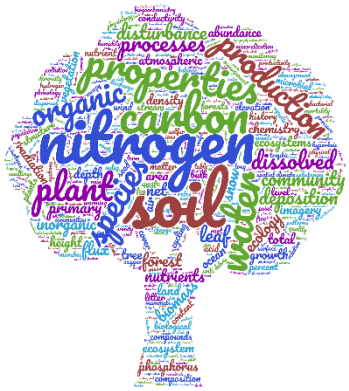
[a] Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia, USA 22904-4123, jhp7e@virginia.edu

Corresponding Author: John H. Porter, jhp7e@virginia.edu

Abstract

The increasing availability of data has driven a need for improved capabilities to discover data. Use of keywords drawn from a thesaurus is one way to enable browse-based discovery and to enhance searching. To assess the effect of using a thesaurus on the discoverability of ecological data, use of 81,415 keywords derived from 6,132 data packages drawn from 28 ecological research projects in the U.S. Long-term Ecological Research Network were examined. The vast majority (95%) of data packages included at least one keyword drawn from the thesaurus, thus enabling their discovery using a hierarchical browse interface. For searching, keywords derived from the thesaurus would reveal 17 times more data packages than *ad hoc* keywords not in the thesaurus. Additionally, searches using keywords derived from the thesaurus returned data from a median of four different projects, whereas *ad hoc* search terms would typically yield data from only a single project. Of the search terms that yielded more than five data packages across two or more projects, 78% were found in the thesaurus. Use of keywords drawn from the thesaurus increased when compared to their use prior to establishment of the thesaurus, indicating that terms from the thesaurus are being actively added to

metadata. A questionnaire assessed the process by which keywords were selected and indicated that information management personnel played an important role in assigning keywords drawn from the thesaurus. These results support the idea that adoption of a thesaurus can be an effective way to enhance the discoverability of ecological data, and that keywording practices play an important role in supporting that enhancement.



Graphical Abstract:

Highlights

- Searching of metadata is much more effective, both with respect to the expected number of "hits" and the number of projects from which data comes, when terms drawn from a thesaurus were used.
- In a case study, the vast majority (95%) of data packages could be discovered using a browse interface based on a thesaurus
- Adoption of a thesaurus or other lexical resource can aid in scientific data discovery

Keywords

Data Discovery, Thesaurus, Long-Term Ecological Research, Keywords

1. Introduction

The past decade has seen an increasing emphasis on the sharing of scientific data (Campbell, 2009; Reichman et al., 2011; Roche et al., 2015). Funding agencies and journals have adopted policies that seek to ensure that scientific data, once collected, will remain available in the future (Bloom et al., 2014; Hanson et al., 2011; Whitlock et al., 2010). As systems for sharing data have continued to evolve there remain sociological, ethical and technical challenges (Costello et al., 2013; Duke and Porter, 2013; Hampton et al., 2013; Michener, 2015; Nelson, 2009; Reichman et al., 2011; Roche et al., 2015). Here I examine how a thesaurus has been used to enhance data discoverability across a group of ecological research projects.

The increasing volume of data and the number and diversity of datasets pose challenges for the researcher seeking to use the data (Porter et al., 2012; Vanderbilt et al., 2017; Vanderbilt et al., 2010). The foremost of these challenges is data discovery. No researcher can use data that they can't find. Unfortunately, search tools using free text (e.g., Google) typically work poorly for locating data, whose metadata make up only a miniscule fraction of online text. The Schema.org initiative has helped improve searching by Internet search engines, but still depends on there being sufficient semantic content in metadata (Mika, 2015). Building on the long-standing work of the library community, a variety of approaches are available for helping to improve the efficiency and precision of browsing and searching, including controlled vocabularies, taxonomies, thesauri and ontologies (Lambe, 2014; Madin et al., 2008; National Information Standards Organization, 2005; Rosati et al., 2017). Controlled vocabularies are lists of terms to be used as keywords, with no attempt to define relationships among terms. Synonym rings, also known as synsets, allow addition of synonyms for preferred terms. Taxonomies organize a controlled vocabulary into one (a taxonomy) or more (a polytaxonomy) hierarchies that define parent-child relationships that allow terms to be linked to broader or narrower

concepts. A thesaurus broadens a polytaxonomy to allow related terms to be linked across different hierarchies. Finally, ontologies provide a more flexible mechanism for linking terms using a variety of relationships that are not limited to simple parent-child hierarchies and provide more nuanced ways to express relationships. For example, relationships of "is a," "part of," "has unit," and "derives from" might be a subset of the relationship types used within an ontology. Ontologies are thus formal, explicit descriptions of concepts that enable machine-based inference (Kless et al., 2015).

The various lexical technologies require different levels of effort to develop (Lambe, 2014; National Information Standards Organization, 2005). A controlled vocabulary is easily assembled from terms commonly used in a field that may be altered using rule-based transformations to make them more consistent (e.g., use plural form for things that can be counted, singular form for amounts). Thesauri require more effort because parent-child and cross-hierarchy relationships between terms need to be defined. Ontologies are the most demanding because of the formal logic that needs to be applied to relationships between concepts. Moreover, the wider array of relationships poses additional challenges with respect to locating where terms fall within the ontology. Here I will focus on use of a thesaurus because they are commonly used, require less effort to create than an ontology but provide a reasonable level of functionality for browsing and searching. Additionally, a creation of a thesaurus is a reasonable starting point for the ultimate development of an ontology, should one be needed (Cardillo et al., 2014; Kless et al., 2012; Kless et al., 2015; Wielinga et al., 2001).

There is a rich history of using thesauri to facilitate searches for documents (Aitchison et al., 2003; Schwartz, 2008). However, use of lexical technologies, such as thesauri, for searching for data, rather than documents, is a relatively recent development (Caracciolo et al., 2013; Garnier et al., 2017; Rosati et al., 2017), in part because most scientific data have not been available (Hampton et al., 2013; Nelson, 2009; Tenopir et al., 2011). Searching for data poses a challenge. The data themselves are

typically simply numbers (Wallis et al., 2013). Searching for "12.2" is not useful. Instead, metadata describing data must be used to enable data discovery. Often this metadata is brief, providing a relatively limited array of concepts that serve as search targets (Parker et al., 2016; Roche et al., 2015). Studies of the effectiveness of thesauri are dominated by thesauri targeted at documents (Shiri and Revie, 2005, 2006), not of thesauri aimed at data.

The creation of a thesaurus does not guarantee that data can be more effectively discovered. For that, several additional conditions must be met. First the thesaurus needs to contain the appropriate terms needed to characterize datasets. Each dataset needs to have at least one keyword drawn from the thesaurus to be discoverable via a hierarchical browse interface, otherwise no link to the dataset will be available. Second, there need to be a limited number of terms, because large numbers of terms in a thesaurus complicates both application of terms to datasets and searching or browsing. Finally, for a thesaurus to be successfully used, terms from the thesaurus must be employed as keywords or other searchable entities within metadata.

One example of a project-specific thesaurus for data is the U.S. Long-Term Ecological Research (LTER) network thesaurus. The LTER Network consists of 28 research sites funded by the National Science Foundation that study a variety of ecosystems – from arctic oceans, to old-growth forest, to grasslands, to cities and to deserts. Each site is a separate project with its own research and data objectives within some broad LTER-wide guidelines. Since 1994 the LTER Network has had a mandate to share their data and currently over 6,000 individual data packages are available (Long-Term Ecological Research Network, 2019b). Prior to 2010 almost all keywords used in LTER data packages were *ad hoc*, selected by the researcher with no guidance (a few sites implemented their own controlled vocabularies for keywords). The result was that, based on a 2006 analysis, the majority of keywords were used for only a single data package, and only 3% of the keywords were used across five or more LTER sites

(Vanderbilt et al., 2017). Starting in 2005, a working group of LTER Information Managers evaluated existing thesauri such as the National Biological Information Infrastructure Thesaurus, the GEMET Environmental Thesaurus, and the Global Change Master Directory Keyword List relative to the keywords already used in LTER data packages, but found none that had a high enough overlap to be useful (Porter 2010). Therefore they pursued development of a thesaurus, with version 1 completed in 2011 (Vanderbilt et al., 2017). The thesaurus focused on thematic terms and excluded specific taxonomic terms and place names, with the rationale that there were other resources (e.g., taxonomic naming systems and gazeteers) that would be better employed for them. The thesaurus was further expanded in 2013 to a total of 702 preferred terms. The thesaurus was made available via a browsable and searchable web site using the Tematres (Gonzales-Aguilar et al., 2012) thesaurus management system (Long-Term Ecological Research Network, 2019a), and a variety of web services were produced that allowed suggested words to be incorporated into interfaces.

Here I will examine how successful efforts at including terms from the thesaurus have been at the level of the individual data package (dataset), at the level of the research site, and the mechanisms used to promote addition of keywords drawn from the thesaurus. Specifically I examine:

- How are keywords used by ecological research projects and how do practices vary between projects?
- How discoverable are data using hierarchical browse interfaces?
- Are terms in the thesaurus more effective for browsing and searching than uncontrolled terms?
- How has use of keywords changed following introduction of the thesaurus?
- What are the mechanisms used for including terms from the thesaurus in metadata documents?

2. Methods

Catalog information on 6,132 data packages was downloaded using a SOLR query of the Environmental Data Initiative's Provenance Aware Synthesis Tracking Architecture (PASTA) on May 16, 2018. Data were drawn from 28 LTER sites, which includes network data packages not associated with any specific site and two defunded sites that are no longer part of the network, but does not include three recently-added sites that do not currently have any data publications. Among other catalog information, there was an identifier for each data package (data packages consist of one or more data files and associated metadata) and a list of keywords derived from the original metadata. Keywords were extracted and changed to lower case for consistency. Raw and summarized data used in this study are available in the Environmental Data Initiative Data Portal (Porter, 2018).

A copy of the LTER Thesaurus was downloaded from a TemaTres database as a Moodle-formatted exchange file. Terms were extracted as either "preferred" terms (702 terms) or "use for" terms (196 terms). "Use for" terms are synonyms, or alternative descriptions, of a concept described by a preferred term. All terms were converted to lower case for consistency.

To permit comparisons with earlier uses of keywords, a dataset of keywords in use in 2006 created by Porter and Costa (2006) was downloaded. The keyword list was changed to lower case, for consistency, and limited to terms in the "keyword" section of metadata documents drawn from the LTER Metacat system. Servilla (2006) reported that there were 5,296 data packages in the LTER Metacat system in 2006.

Counts of the number of keywords, number of terms found in the thesaurus (including both preferred terms and "use for" terms), and number of preferred terms were tabulated for both data

packages and for individual sites using the R statistical package, version 3.5. Counts of the number of data packages and sites using each of the keywords were also tabulated.

To discover who is primarily responsible for adding keywords to metadata documents, a brief survey was sent via electronic mail to Information Managers at the LTER sites. The survey first asked for a ranking of individuals (Researcher, Information Manager, Technician and Other) responsible for adding keywords to metadata documents. The second question allowed multiple responses and focused on the process for assigning keywords to data packages.

3. Results

3.1 Overall use of Keywords

The word "term" is used here to refer to words that describe a concept without reference to any specific data package. The word "keyword" refers to terms that have been used within data package metadata and formally identified there as keywords.

Metadata records were obtained for 6,132 data packages incorporating 81,415 keywords. The number of keywords per data package (i.e., dataset) varies from 0 to 295 with a median of 10. The distribution is highly skewed, with only 25% of the data packages including more than 16 keywords (Figure 1). Only four LTER sites had data packages containing more than 70 keywords. These data packages were characterized by long lists of taxonomic names or chemical constituents included as keywords, or with repeated, alternative forms of the same keyword (e.g., CO2 vs Carbon dioxide).

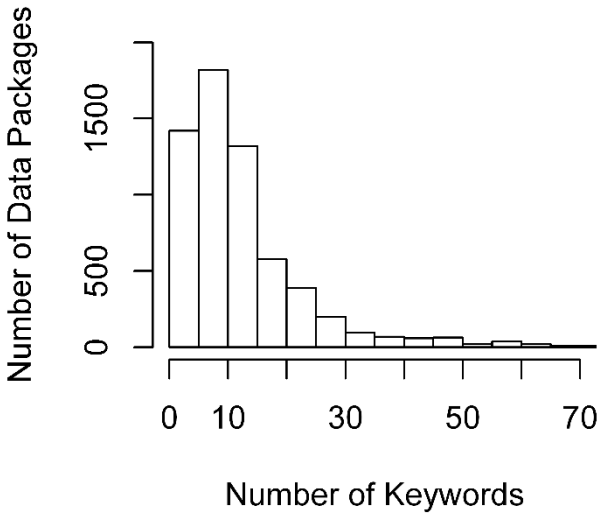


Figure 1 Number of keywords per data package. Excluded from the graph are 42 data packages that had greater than 70 keywords.

There was substantial reuse of keywords across data packages. Of the 81,415 total keywords in data packages, there are only 6,022 distinct keywords used. Some of the keywords are widely used, with a single keyword ("disturbance") used in 1,390 data packages across 25 different sites. However, this was exceptional; only four keywords were used in more than 1,000 data packages. The typical number of uses was far lower. The median number of datasets that used a keyword was 2, and the median number of sites was 1 when all keywords were considered. The most useful keywords are likely to be those that are used in multiple data packages across multiple sites. There were 824 keywords used both in five or more data packages and across more than one LTER site.

The number of keywords per data package varied widely both within and between LTER sites (Figure 2). The median number of keywords per data package ranged between 4 and 27 across sites, with 75% of the data packages having 6 or more keywords. There is no apparent relationship between the number of data packages per site and the median number of keywords per data package

(Spearman's rho =0.071 $p > 0.71$), nor are there any apparent patterns based on the type of ecosystem being studied (forest vs grassland vs coastal vs arctic). Examination of metadata for sites with exceptionally high numbers of keywords found large numbers of places, taxa or multiple alternative forms of the same keywords. Terms at the top of the thesaurus hierarchy (the broadest terms) tended to be used more frequently (a median of 36.5 data packages per term) than other levels in the hierarchy (levels 2 through 5 all had medians between 19 and 21 data packages per term).

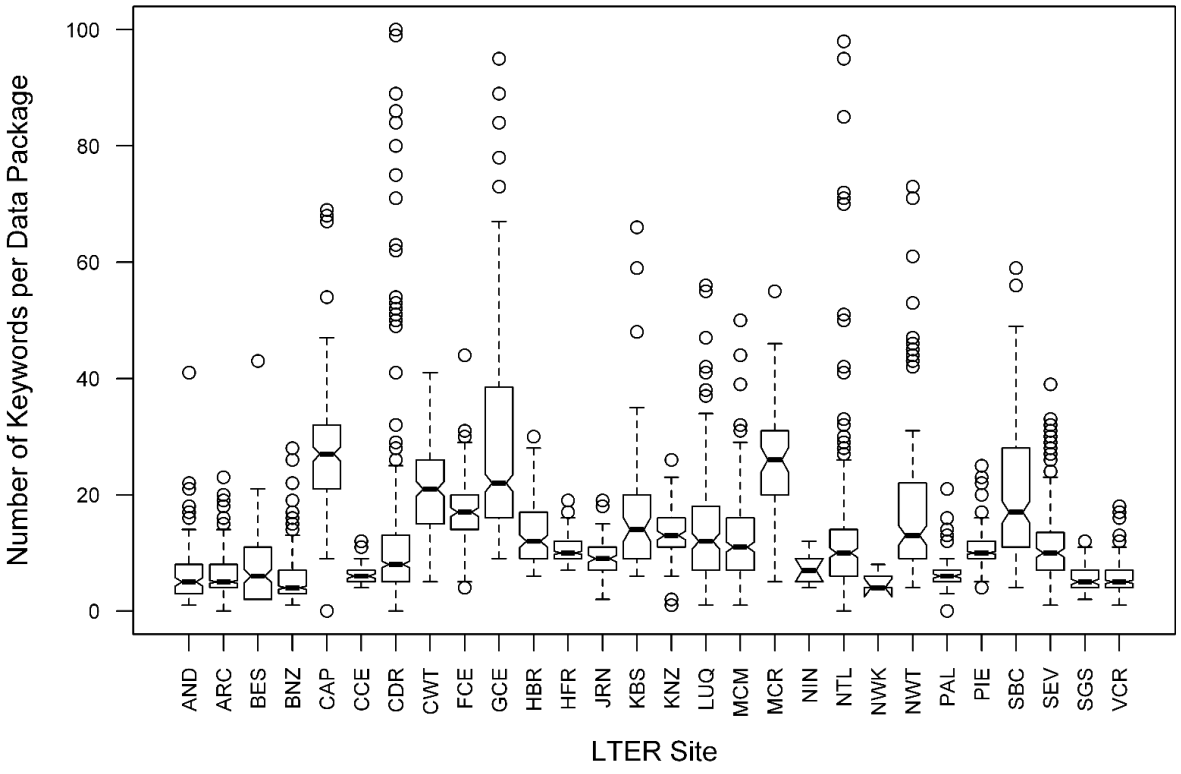


Figure 2 Number of keywords per data package by LTER site. Not shown are 20 outlier packages with more than 100 keywords. LTER sites are indicated by three letter acronyms available from https://www.lternet.edu.

3.2 Use of Terms in the LTER Thesaurus

Although the overall patterns of keyword application are of interest, our focus here is on how terms drawn from the LTER Thesaurus have been employed. When used in a "browse" interface to traverse from more general to more specific concepts, any data package that does not have one or more keywords that are found in the thesaurus will be undiscoverable. For the 6,132 LTER data packages, 95% of data packages could be located using a browse interface using only preferred terms in the thesaurus. Eighty of the packages that could not be located via a browse interface had no keywords at all. Overall, the non-browsable data packages had a median of only 2 keywords, versus a median of 10 keywords for data packages that could be located via a browse interface.

Terms in the thesaurus were more frequently used as keywords than terms not in the thesaurus. The median number of data packages which used terms absent from the thesaurus was 1 whereas for terms in the thesaurus, the median was 17 (Figure 3). Ideally a search should result in a manageable set of data packages, with neither too few nor too many matches. Most searches using terms in the thesaurus would return between 7 and 44 datasets, whereas for keywords not in the thesaurus one would expect to only find between 1 and 4 data packages.

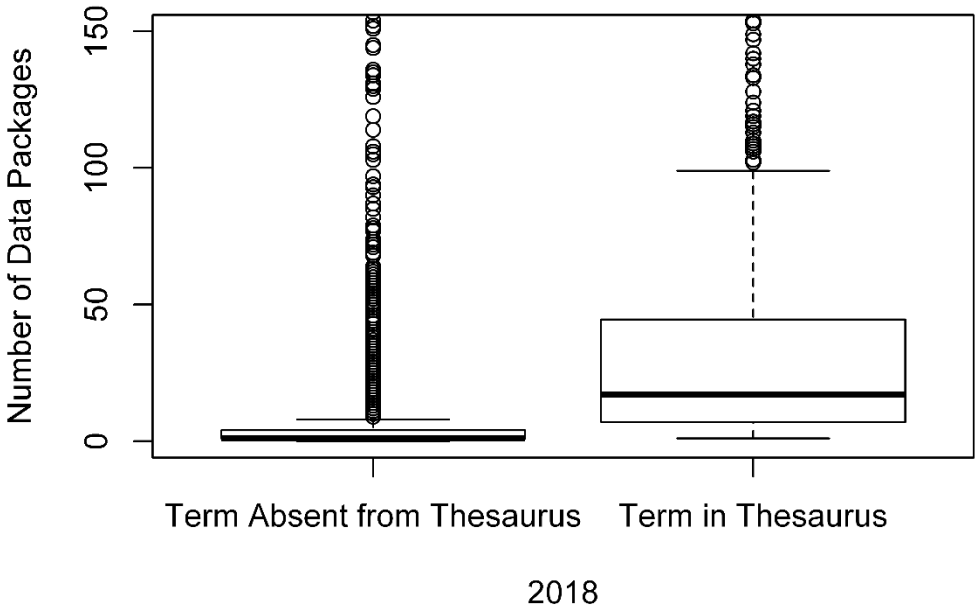


Figure 3 Number of data packages returned by searching on a term in 2018. Terms found in the thesaurus were much more likely to locate a large number of data packages. Not shown are 108 outliers which would return more than 150 data packages.

Terms in the thesaurus were much more frequently used at multiple sites than terms absent from the thesaurus. Most searches using terms in the thesaurus would return data packages from between 2 and 8 sites (median=4). Searches using terms absent from the thesaurus would be expected to return data from only a single site (median=1). Terms in the thesaurus made up the majority of all keywords used by 4 or more sites. No term absent from the thesaurus was used by more than 12 sites, whereas up to 27 sites used terms in the thesaurus (Figure 4). Of the 824 keywords that were used in five or more data packages across two or more sites, 645 (78%) were included in the thesaurus.

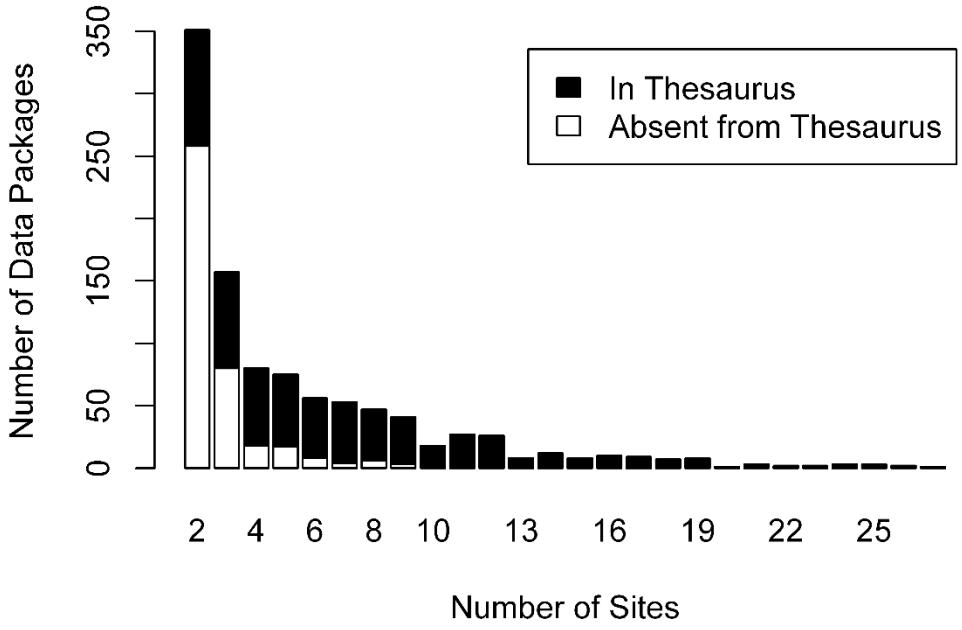


Figure 4 Number of sites where a keyword was used vs the number of data packages. Only terms included in the thesaurus were used at more than 12 sites. Not shown are 5,012 data packages where a term was used at only a single site.

3.3 Changes Since Adoption of the Thesaurus

To examine how the use of keywords has changed following the introduction of the thesaurus, statistics from keyword usage in 2006 (Porter and Costa 2006) were examined. At that time there were approximately 5,296 documented data packages drawn from 18 LTER sites (Servilla, 2006). The number of data packages for 2006 is approximate because the date of the report and date of the harvesting of keywords for analysis are not precisely the same. The total number of keywords in 2006 was 29,398 and the number of distinct keywords used was 3,206. Both of these are roughly one-half the number of keywords in 2018, reflecting an overall increase in the use of keywords in metadata since 2006.

Although the patterns are similar to the 2018 use of keywords, some trends have been amplified following adoption of the thesaurus. Comparison of Figure 5 and Figure 3 shows that keywords that were subsequently incorporated into the thesaurus matched more data packages than those that remained absent from the thesaurus . However, the magnitude of the difference between median values went from 3 (1 vs 4) in 2006 to 16 (1 vs 17) in 2018 (Figure 3), indicating a strong increase in the utility of terms in the thesaurus for searching for data. A similar pattern was observed when the number of sites using a keyword was considered. In 2006, terms in the thesaurus were used at a median of 2 sites, whereas those absent from the thesaurus were used at a median of only 1 site. In contrast by 2018, while the keywords absent from the thesaurus remained at a median of 1, keywords in the thesaurus were used at a median of 4 sites, double what was observed in 2006.

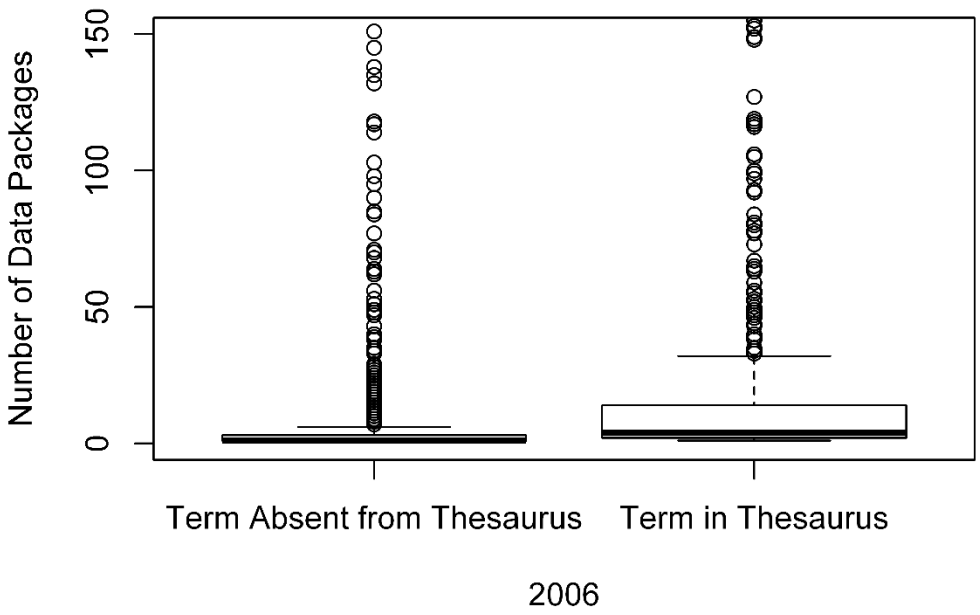


Figure 5 Number of data packages returned by searching on a term in 2006. Terms later added to the thesaurus found more data packages, but not as many as in 2018 (Figure 3). Not shown are 38 outliers which would return more than 150 data packages.

3.4 Incorporating Keywords into Metadata

The increased utility of keywords drawn from the thesaurus is not a function of the thesaurus itself, but rather of efforts to use the thesaurus as part of the metadata creation process. To explore this process a questionnaire was sent out to each LTER site's Information Manager. Responses were received from 21 sites. When asked to rank the relative contribution of different roles to the assignment of keywords to data packages, 77% responded that the site information management team played the most important role in assigning keywords, with the researcher providing the data playing the most important role at the other 23% of sites.

The questionnaire also asked sites to provide information regarding what tools and techniques were used to populate keyword metadata. The dominant responses were: "The Information Management Staff provides keywords from the Controlled Vocabulary based on a reading of the title, abstract and data attributes" and "Researchers provide preliminary keywords that the Information Management Staff uses to select final keywords using the Controlled Vocabulary" with 14 and 13 sites, respectively, using those techniques. Three additional techniques were used at 8 sites each: "A person identifies keywords using the http://vocab.lternet.edu web site hierarchy," "We have our own site listing of the terms that are most likely to be used at our site and draw keywords from that," and "We use DEIMS or another system that automatically incorporates the Controlled Vocabulary." No other tool or technique was used by more than two sites.

4. Discussion

I compared the use of keywords in 6,132 data packages across 28 research projects to determine how keywords are used by ecological research projects and how practices vary between projects. Keyword use exhibits some general properties, and some exceptions. Most data packages

incorporate 10 or more keywords, with most including between 6 and 16 keywords. However, some projects and data packages use very large number of keywords, often incorporating long lists of species, chemical names or locations. Comparison of research projects shows some variation, but differences appear to be esoteric and unrelated to the ecosystem of study or the number of data packages managed by the project.

For data packages to be discoverable using a hierarchical browse interface, at least one preferred term from the thesaurus must be used as keyword. Almost all (95%) of the data packages were discoverable because they included at least one keyword found in the thesaurus. Data packages that remained undiscoverable typically had very low numbers of keywords (in 80 instances, no keywords). This suggests that the solution to this problem lies less in expanding the thesaurus than in augmenting the number of keywords used in some data packages.

Terms in the thesaurus would be more effective for browsing and searching than uncontrolled terms, typically yielding useful numbers of data packages derived from multiple sites. Terms in the thesaurus are used as keywords over 6 times as often as terms not in the thesaurus. Indeed, terms in the thesaurus make up only 15% of the distinct keywords used, but comprise over half (52.5%) of the keywords used overall. Thus there is increased power for researchers in using search terms that are derived from the thesaurus. A search for a typical term in the thesaurus would return a useful number of datasets (between 7 and 44) whereas for keywords absent from the thesaurus a typical search would return fewer than 5 datasets. Similarly, terms in the thesaurus return data from a larger number of research projects (1 vs 4 sites).

Apart from searches for single terms, the “Advanced Search” function of the EDI and LTER Data Portals includes the ability to search on terms based on their relationships within the thesaurus. Thus searches are available that allow a researcher to search on a term and all its synonyms and narrower

terms underneath it in the hierarchy. So a search on “Forests” also returns data with the keywords “boreal forests”, “clearcut”, “clearcuts”, “forest”,” old growth” and “old growth forests.” Searches are also available for including related terms or both narrower and related terms. Such searches that leverage the structure of the thesaurus will inevitably be more productive (i.e., produce more hits) than searches using individual terms. However, keywords not in the thesaurus cannot be similarly augmented, even to the degree of including synonyms. Unfortunately, neither portal records information on the searches that were actually used by researchers, so it is not currently possible to assess how researchers interact with the search capabilities of the portal.

The dominance of terms derived from the thesaurus as keywords could result from two processes, one passive and one active. Creating a thesaurus of terms that were already widely used as keywords could lead to apparent dominance. Alternatively, an active process of purposefully selecting keywords drawn from the thesaurus could be used. The data from the LTER Network shows both effects. Comparison of keyword use prior to, and subsequent to, creation of the thesaurus found that terms later added to the thesaurus were already more widely used in 2006 than terms not used in the thesaurus. To some degree this is inevitable, because criteria for selecting terms for the thesaurus included how many data packages and the number of sites that used the term (Vanderbilt et al., 2017). However, changes from 2006 to 2018 display a near doubling in the use of keywords drawn from the thesaurus, indicating an active process that took place subsequent to establishment of the thesaurus.

The mechanism for including terms drawn from the thesaurus in metadata is largely the result of human effort. The survey of information managers confirms that for new data packages inclusion of keywords drawn from the LTER Thesaurus is an active process, not a passive one. There have been efforts to create semi-automated tools for assisting in application of keywords, such as online forms which “autocomplete” with preferred forms from the thesaurus or software which uses semantic

analysis to process text in the metadata (e.g., title, abstract, variable labels etc.) to suggest potential keywords. However, the diversity of methods by which metadata is generated (varying from metadatabases, to metadata editors, to translators that convert text or spreadsheets into needed forms, to direct text editing) makes it difficult to develop a single tool that can directly facilitate addition of keywords. This makes an interface that allows keyworders to rapidly browse or search for appropriate keywords especially important. The Tematres software used to maintain the LTER Thesaurus provides efficient browse and search capabilities, along with web services that can be used to support semi-automated keywording.

There are several challenges when adding keywords to datasets. One is that researchers often want to search for data relevant to specific uses, but that data itself is agnostic relative to use. For example, a dataset might include measurements of air temperature. Thus "air temperature" is a term that can be directly applied. But additional terms could be added that relate not to air temperature per-se, but to possible applications of air-temperature data. Thus, "climate," "weather," "meteorology," or even "global warming" might also be applied to that dataset. Keywords that address topics to which data might be applied are in the eye of the beholder. This results in an inevitable tension when adding keywords to datasets. Do you focus keywords entirely on dataset contents, or do you also include keywords that link datasets to particular uses? One approach that might help ameliorate the impact of this issue is to break keywords into distinct subsets, where one subset focuses exclusively on dataset contents whereas another subset focuses on potential uses, thus allowing a searcher to explicitly separate contents from uses. Use of an ontology would also allow more complex relationships such as "measured" vs "used to study" (Kless et al., 2015).

Design of the LTER Thesaurus focused on using a relatively small number of terms. The number of terms in a thesaurus needs to be large enough to discriminate among datasets, but not so large as to

be unwieldy. Just imagine trying to identify which of 100,000 possible terms should be applied to a dataset! Moreover, if a thesaurus contains too many closely-related terms use of the terms becomes increasingly esoteric. One metadata creator may use term A as a keyword, but another metadata creator may use (closely-related) term B instead, making it more difficult to successfully use keywords in search and browse interfaces. The ideal number of terms depends to some degree on the size of the corpus of datasets to which it will be applied. Researchers conducting searches for data will want to have more than one candidate dataset returned by a search, but at the same time not have so many candidate datasets returned as to overwhelm their ability to evaluate them. The desirable number of terms in a thesaurus will thus vary with the number and diversity of the underlying datasets. The goal is to identify a sufficient number of terms to yield informative searches that don't overwhelm the researcher searching for data. This can be a challenge when there are large numbers of pre-existing datasets whose keywords may not be included as terms in the thesaurus. To some degree this problem can be ameliorated by identifying synonyms or "use for" terms that might occur in metadata, but that match up with a preferred term in the thesaurus. However, unless a conscientious effort is made to use terms from the thesaurus in metadata documents, the benefits of using a thesaurus will not be realized.

A "best practices" document for adding keywords to metadata was created by the LTER Controlled Vocabulary Working Group (LTER Controlled Vocabulary Working Group, 2013). One of the recommendations is to use the most specific possible keywords. Their rationale was that, when searching or browsing, 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. However, the frequent use of keywords drawn from the top level of the hierarchy suggests that many data packages are including not only the most specific terms, but also their parent terms. Such use enables use of less-sophisticated search interfaces that do not integrate the thesaurus

structure into their searching, but also inflates the number of keywords that need to be added to each metadata document.

In conclusion, creation of a thesaurus has favorably increased the discoverability of data from a multi-project network of research sites. It enabled hierarchical browsing of data, and provided better opportunities for obtaining a reasonable number of search hits across several research projects. There remain challenges in assigning keywords to datasets and best practices need to be adopted. More sophisticated searching could be enabled using an ontology, with a concomitant increase in the keywording process. However, if an ontology is ultimately required, a thesaurus would serve as a valuable starting point in its creation.

Acknowledgements:

Thanks to the LTER site Information Managers, in particular, Margaret O'Brien and Kristin Vanderbilt, for many hours of discussion regarding keywords and their uses. Mark Servilla helped identify historical documents relating to the growth of the LTER data archive and two anonymous reviewers provided helpful comments. This material is based upon work supported by the National Science Foundation under grants: 0832652, 1237733, 1545288 and 1832221.

Literature Cited

Aitchison, J., Bawden, D., Gilchrist, A., 2003. Thesaurus construction and use: a practical manual. Routledge.

Bloom, T., Ganley, E., Winker, M., 2014. Data Access for the Open Access Literature: PLOS's Data Policy. PLOS Biology 12, e1001797. 10.1371/journal.pbio.1001797

Campbell, P., 2009. Data's shameful neglect. Nature 461, 145. 10.1038/461145a

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J., 2013. The AGROVOC linked dataset. Semantic Web 4, 341-348. 10.3233/SW-130106

Cardillo, E., Folino, A., Trunfio, R., Guarasci, R., 2014. Towards the Reuse of Standardized Thesauri Into Ontologies. WOP, 26-37. https://www.iit.cnr.it/sites/default/files/Paper_8_update.pdf (accessed 3.4.19)

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q., Bourne, P.E., 2013. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. 28, 454-461. 10.1016/j.tree.2013.05.002

Duke, C.S., Porter, J.H., 2013. The ethics of data sharing and reuse in biology. BioScience 63, 483-489. 10.1525/bio.2013.63.6.10

Garnier, E., Stahl, U., Laporte, M.A., Kattge, J., Mougenot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönisch, G., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. Journal of Ecology 105, 298-309. 10.1111/1365-2745.12698

Gonzales-Aguilar, A., Ramírez-Posada, M., Ferreyra, D., 2012. TemaTres: software para gestionar tesauros. El profesional de la información 21, 319-325. 10.3145/epi.2012.may.14

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11, 156-162. 10.1890/120103

Hanson, B., Sugden, A., Alberts, B., 2011. Making data maximally available. Science 331, 649. 10.1126/science.1203354

Kless, D., Jansen, L., Lindenthal, J., Wiebensohn, J., 2012. A method for re-engineering a thesaurus into an ontology, FOIS, pp. 133-146. 10.3233/978-1-61499-084-133

Kless, D., Milton, S., Kazmierczak, E., Lindenthal, J., 2015. Thesaurus and ontology structure: Formal and pragmatic differences and similarities. Journal of the Association for information science and technology 66, 1348-1366. 10.1002/asi.23268

Lambe, P., 2014. Organising knowledge: taxonomies, knowledge and organisational effectiveness. Elsevier.

Long-Term Ecological Research Network, 2019a, LTER Controlled Vocabulary. WWW Document https://vocab.lternet.edu (2019a) (accessed 3.6.2019)

Long-Term Ecological Research Network, 2019b, LTER Network Data Portal. WWW Document https://portal.lternet.edu (2019b) (accessed 3.6.2019)

LTER Controlled Vocabulary Working Group, 2013, Best Practices for Adding Science Keywords to LTER Metadata. WWW Document http://im.lternet.edu/sites/im.lternet.edu/files/BestPracticesforAddingScienceKeywordstoLTERMetadata.docx (2013) (accessed 3.4.19)

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends Ecol. Evol. 23, 159-168. 10.1016/j.tree.2007.11.007

Michener, W.K., 2015. Ecological data sharing. Ecological Informatics 29, 33-44. 10.1016/j.ecoinf.2015.06.010

Mika, P., 2015. On Schema. org and why it matters for the web. IEEE Internet Computing 19, 52-55. 10.1109/MIC.2015.81

National Information Standards Organization, 2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies. NISO Press.

Nelson, B., 2009. Data sharing: Empty archives. Nature News 461, 160-163. 10.1038/461160a

Parker, T.H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J.D., Chee, Y.E., Kelly, C.D., Gurevitch, J., Nakagawa, S., 2016. Transparency in ecology and evolution: real problems, real solutions. Trends Ecol. Evol. 31, 711-719. 10.1016/j.tree.2016.07.002

[dataset] Porter, J., Costa, D., 2006. Keywords and Terms from the LTER Network - 2006. Environmental Data Initiative. 10.6073/pasta/3eabff466e2552caa383eafb2ce2b343

[dataset] Porter, J.H., 2018. The US LTER Thesaurus: Contents and Keyword Use Statistics in LTER Data Packages in 2006 and 2018. Environmental Data Initiative. 10.6073/pasta/df2c74cad1319e947b3c3ffc83daa153

Porter, J.H., Hanson, P.C., Lin, C.-C., 2012. Staying afloat in the sensor data deluge. Trends Ecol. Evol. 27, 121-129. 10.1016/j.tree.2011.11.009

Reichman, O.J., Jones, M.B., Schildhauer, M.P., 2011. Challenges and opportunities of open data in ecology. Science 331, 703-705. 10.1126/science.1197962

Roche, D.G., Kruuk, L.E., Lanfear, R., Binning, S.A., 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biol 13, e1002295. 10.1371/journal.pbio.1002295

Rosati, I., Bergami, C., Stanca, E., Roselli, L., Tagliolato, P., Oggioni, A., Fiore, N., Pugnetti, A., Zingone, A., Boggero, A., 2017. A thesaurus for phytoplankton trait-based approaches: Development and applicability. Ecological Informatics 42, 129-138. 10.1016/j.ecoinf.2017.10.014

Schwartz, C., 2008. Thesauri and facets and tags, oh my! A look at three decades in subject analysis. Library Trends 56, 830-842. 10.1353/lib.0.0014

Servilla, M., 2006, LNO NIS Status Report. WWW Document https://lternet.edu/wp-content/uploads/2010/12/NISStatusReport20060515.pdf (2006) (accessed 3.4.19)

Shiri, A., Revie, C., 2005. Usability and user perceptions of a thesaurus-enhanced search interface. Journal of documentation 61, 640-656. 10.1108/00220410510625840

Shiri, A., Revie, C., 2006. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. Journal of the American Society for Information Science and Technology 57, 462-478. 10.1002/asi.20319

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M., 2011. Data sharing by scientists: practices and perceptions. PloS one 6, e21101. 10.1371/journal.pone.0021101

Vanderbilt, K., Porter, J.H., Lu, S.-S., Bertrand, N., Blankman, D., Guo, X., He, H., Henshaw, D., Jeong, K., Kim, E.-S., 2017. A prototype system for multilingual data discovery of International Long-Term Ecological Research (ILTER) Network data. Ecological Informatics 40, 93-101. 10.1016/j.ecoinf.2016.11.011

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Lin, C.-C., Lu, S.-S., Ogawa, A., Tuama, É.Ó., Schentz, H., Su, W., 2010. A multilingual metadata catalog for the ILTER: Issues and approaches. Ecological Informatics 5, 187-193. 10.1016/j.ecoinf.2010.02.002

Wallis, J.C., Rolando, E., Borgman, C.L., 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS one 8, e67332. 10.1371/journal.pone.0067332

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L., Moore, A.J., 2010. Data archiving. The American Naturalist 175, 145-146. 10.1086/650340

Wielinga, B.J., Schreiber, A.T., Wielemaker, J., Sandberg, J., 2001. From thesaurus to ontology, Proceedings of the 1st international conference on Knowledge capture. ACM, pp. 194-201. 10.1145/500737.500767