Ecological Informatics 40 (2017) 93–101

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

A prototype system for multilingual data discovery of International Long-Term Ecological Research (ILTER) Network data

Kristin Vanderbilt [a,*], John H. Porter [b], Sheng-Shan Lu [c], Nic Bertrand [d], David Blankman [e], Xuebing Guo [f], Honglin He [f], Don Henshaw [g], Karpjoo Jeong [h], Eun-Shik Kim [i], Chau-Chin Lin [c], Margaret O'Brien [j], Takeshi Osawa [k], Éamonn Ó Tuama [l], Wen Su [f], Haibo Yang [m]

[a] Department of Biology, MSC03 2020, University of New Mexico, Albuquerque, NM 87131, USA
[b] Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22904, USA
[c] Taiwan Forestry Research Institute, 53 Nan Hai Rd., Taipei, Taiwan
[d] Centre for Ecology & Hydrology, Lancaster Environmental Centre, Lancaster LA1 4AP, UK
[e] Jerusalem 93554, Israel
[f] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[g] U.S. Forest Service Pacific Northwest Research Station, Forestry Sciences Laboratory, 3200 SW Jefferson Way, Corvallis, OR 97331, USA
[h] Department of Internet and Multimedia Engineering, Konkuk University, Seoul 05029, Republic of Korea
[i] Kookmin University, Department of Forestry, Environment, and Systems, Seoul 02707, Republic of Korea
[j] Marine Science Institute, University of California, Santa Barbara, CA 93106, USA
[k] National Institute for Agro-Environmental Sciences, Tsukuba, Ibaraki 305-8604, Japan
[l] GBIF Secretariat, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark
[m] School of Ecological and Environmental Sciences, East China Normal University, 500 Dongchuan Rd., Shanghai 200241, China

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 21 November 2016
Accepted 21 November 2016
Available online 28 November 2016

Keywords:
Thesaurus
Ontology
Data sharing
Translation
Web services

ABSTRACT

Shared ecological data have the potential to revolutionize ecological research just as shared genetic sequence data have done for biological research. However, for ecological data to be useful, it must first be discoverable. A broad-scale research topic may require that a researcher be able to locate suitable data from a variety of global, regional and national data providers, which often use different local languages to describe their data. Thus, one of the challenges of international sharing of long-term data is facilitation of multilingual searches. Such searches are hindered by lack of equivalent terms across languages and by uneven application of keywords in ecological metadata. To test whether a thesaurus-based approach to multilingual data searching might be effective, we implemented a prototype web-services-based system for searching International Long-Term Ecological Research Network data repositories. The system builds on the use of a multilingual thesaurus to make searches more complete than would be obtained through search term-translation alone. The resulting system, when coupled to commodity online translation systems, demonstrates the possibility of achieving multilingual searches for ecological data.

1. Introduction

The International Long-Term Ecological Research (ILTER) Network, consisting of site-based research networks in 40 countries, collects long-term research and monitoring data from many ecosystems around the globe. Since its inception in 1993, this "network of networks" has collected a wide variety of data at its 633 sites (Fig. 1). The aim of the ILTER is to contribute to the understanding of international ecological and socio-economic issues through the synthesis of data at broad temporal and spatial scales that may span multiple countries (Vihervaara et al., 2013; Haase et al., 2016). One barrier to compiling datasets to explore data from more than one country is the multilingual nature of the ILTER's data archives (Vanderbilt et al., 2010, 2015). Each national network manages its data using its own local language. This poses a difficulty for scientists seeking data outside of their own national network. Successful sharing of data and information in the ILTER requires a common language that imparts understanding of what the data mean, as well as tools to do cross-language information retrieval.

One tool that can be used to help facilitate data discovery is a thesaurus. A thesaurus is a structured and organized set of terms, usually about a specific domain, that can be used to index datasets or documents so that end-users can retrieve relevant information when searching using those terms (Broughton, 2006). Thesaurus terms are cross-referenced to other terms in the thesaurus that may be equivalent (synonyms), narrower than, broader than, or related to the term (Fig. 2) (Clarke,

* Corresponding author.
E-mail address: krvander@fiu.edu (K. Vanderbilt).

http://dx.doi.org/10.1016/j.ecoinf.2016.11.011
1574-9541/

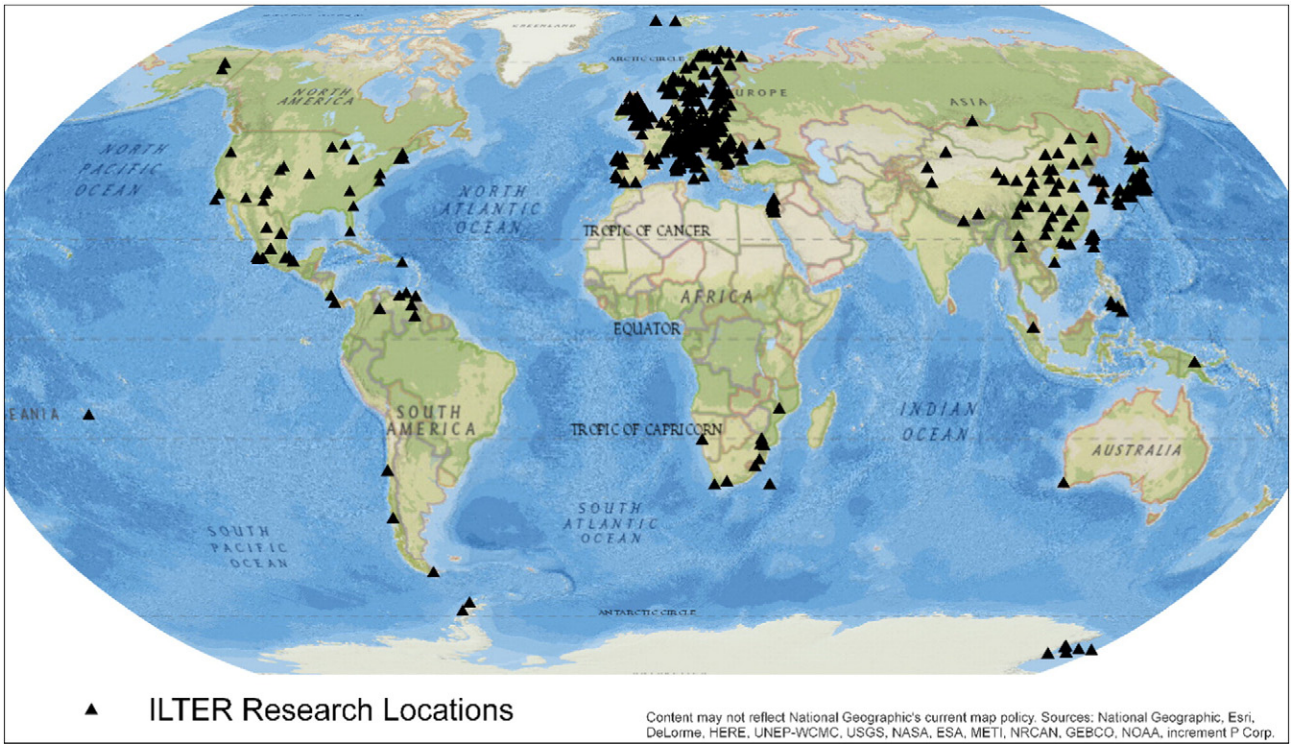

Fig. 1. International Long-Term Ecological Research (ILTER) Network research site locations.

2001). This structure serves as a navigational aid to an end-user, placing terms in a hierarchical context and alerting the user to related terms to search with. A thesaurus also constrains the terms that a data creator can choose from as they select suitable terms to describe their documents or datasets. Both the data creator and end-user benefit from having a controlled list of vocabulary terms from which to select. A monolingual thesaurus is useful within a single national LTER network, but to facilitate data discovery across the whole ILTER network, adoption of a multilingual thesaurus is needed. Several multilingual thesauri exist for the environmental domain, but they are too broad for use by the ILTER (e.g., GEMET (General Multilingual Environmental Thesaurus; http://www.eionet.europa.eu/gemet) and AGROVOC (Multilingual Agricultural Thesaurus; http://www4.fao.org/faobib/kwocinana.html)).

- Net Primary Productivity
  - BT: Primary Productivity
  - RT: Carbon Dioxide
    - UF: CO2
  - RT: Trophic Levels
    - UF: Energy Levels
    - BT: Food Chains
    - RT: Feeding Habits
    - RT: Saprophytism
    - RT: Primary Productivity
    - RT: Biological Production
    - RT: Net Primary Productivity

Fig. 2. An excerpt from AGROVOC illustrating the hierarchical nature of a thesaurus. Descriptors mean: BT: broader than; NT: narrower term; RT: related term; UF: used for. For a data creator designating keywords for a dataset, the thesaurus would tell them to use the term "carbon dioxide" instead of CO2. An end-user searching for data indexed with the term "Primary Productivity" would retrieve records tagged with "Trophic Levels" as well, if the query engine is set to return "related terms."

Even within a single monolingual LTER Network, creating a thesaurus is a challenge. Thesaurus creators must first select terms to include in the thesaurus. These will come from published lists, dictionaries, databases, or the collection of items that will be indexed by the thesaurus (Broughton, 2006). Then, the preferred term must be selected from synonyms or spelling variants (e.g., color vs. colour), and the terms organized into a hierarchical structure. Related terms are then organized into a hierarchical structure specifying "broader than," "narrower than," "related to," and "use for" relationships between terms (ANSI/NISO, 2010).

Methods for creating a multilingual thesaurus include merging existing monolingual thesauri, starting with a new thesaurus and considering multiple languages from the outset, or translating an existing thesaurus into multiple languages (IFLA, 2009). No matter the approach taken, term equivalence and structural challenges will likely be encountered (Jorna and Davies, 2001). In the context of a multilingual thesaurus, equivalent terms should be both semantically (i.e., the terms have the same meaning) and culturally equivalent (IFLA, 2009). Partial equivalence may arise when a term in one language has a somewhat broader or narrower meaning than a term in another language, or the translated term may have a different cultural connotation. The terms "loud" and "noisy", for instance, both mean "easily audible", but are only partially equivalent because "noisy" has a more negative connotation than "loud". An equivalent term in one language may not exist for a particular concept in another, and two terms in one language may be required to capture the meaning of the preferred term in the other. Semantic and cultural differences in the use of terms may result in non-symmetrical hierarchies of terms in different languages. However, one advantage to using a multilingual thesaurus, rather than a simple list of translated words, is that concepts that may be ambiguous or difficult at one level may be direct translations at another level in the hierarchy. For example, Vanderbilt et al. (2010) showed how the Japanese and English