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/

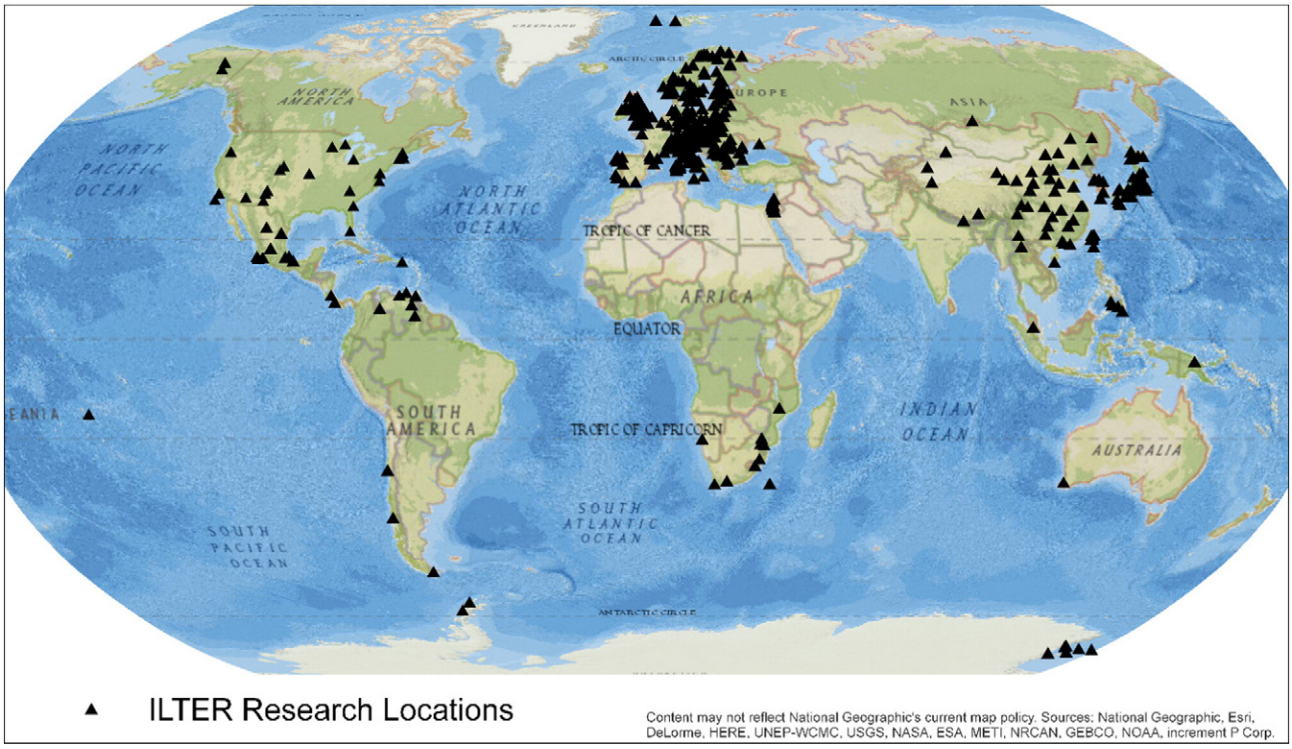

Fig. 1. International Long-Term Ecological Research (ILTER) Network research site locations.

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

concepts for "wetlands" are different, but many lower-level units (different types of wetlands such as salt marsh or mangrove) are similar. So searches on the high-level term will still find data tagged with the more specific terms.

Ideally, an ILTER end-user would query the ILTER data archive using a term in their own local language, and be able to retrieve resources tagged with that term in other languages in the database. To accomplish this query, software is needed that can use the multilingual thesaurus to find translations and then query stores of ILTER data using the translated terms. The adoption of a common software stack and metadata standard for managing data in many ILTER national networks makes this task tractable. In 2010, ILTER members agreed to use Ecological Metadata Language (EML) (Fegraus et al., 2005) as the metadata standard for the network (Vanderbilt et al., 2010). EML is implemented as a set of XML schemas that can be used in an extensible manner to document ecological data. Metacat (Berkley et al., 2001), a database for storing data packages (i.e., data + metadata) is an open-source solution for managing data and metadata and many ILTER network members use it (e.g., Lin et al., 2008; Ohte et al., 2012). Metacat stores XML documents in a relational database, from which they can be queried using a path-oriented query language. Metacat will store metadata documents in different languages. ILTER data are not stored in one centralized location, but in a distributed system of Metacats.

In 2012, the "Semantic Approaches to Discovery of Multilingual ILTER Data" workshop was held at the East China Normal University in Shanghai, China to explore how to improve data discoverability in the multilingual ILTER data archives. The results of that workshop were wide-ranging, varying from evaluation of search resources, enhancement of existing thesauri and development of a prototype distributed ecological data system. Here we describe the steps required to build a prototype web-service-based multilingual data search system, including development of base thesauri, both monolingual and multilingual, development of interfaces and search tools and subjective assessments of the prototype multilingual data search system. We also discuss lessons learned from the prototype that can be used to guide creation of a production system, and the degree to which web-based automated translations might be used to make full metadata translations available.

2. Building a monolingual thesaurus

An example of the process of building a thesaurus comes from the U.S. LTER Network. Historically, most keywords used to characterize datasets at U.S. LTER sites were uncontrolled. They were selected entirely by the data creator without reference to words used in other datasets. One of the challenges facing researchers in discovering data from LTER sites was inconsistent application of keywords. A researcher interested in carbon dioxide measurements would need to search on both "Carbon Dioxide" and "CO_2." Moreover, the existing set of keywords was highly diverse. For example, in a 2006 survey of Ecological Metadata Language (EML) documents in the LTER Data Catalog, over half the keywords (1616 of 3206) were used in only a single dataset, and only 104 (3%) of the keywords were used at five or more different LTER sites (Porter, 2006, Porter and Costa, 2006).

To address this problem, in 2005 the U.S. LTER Information Management Committee (comprised of one information manager from each of the 26 US LTER sites) established an ad hoc "Controlled Vocabulary Working Group" and charged it with studying the problem and proposing solutions. The group compiled and analyzed keywords found in LTER datasets and documents, and identified external lexicographical resources, such as controlled vocabularies, thesauri and ontologies, that might be applied to the problem (Porter, 2010). Initially the working group attempted to identify existing resources, such as the GEMET Thesaurus, the Global Change Master Directory keyword list, and the National Biological Information Infrastructure (NBII) Thesaurus (now the U.S. Geological Survey Biocomplexity Thesaurus), that LTER might be able to adopt wholesale. Unfortunately, using matches with widely-used LTER keywords as a metric, none of the external resources proved to be suitable. Too many keywords commonly used in LTER datasets were absent from the existing lexicographical resources. So, starting in 2008 the working group focused on developing an LTER-specific controlled vocabulary, ultimately identifying a list of approximately 600 keywords that were either used by two or more LTER sites, or were found in one of the external resources (NBII Thesaurus and Global Change Master Directory Keyword List). Excluded from the list were taxonomic names for species and names of geographic locations, as these were considered to be better addressed using existing taxonomic resources and gazetteers. The form of keywords were adjusted to conform to the recommendations of the international standard for controlled vocabularies (ANSI/NISO, 2010), but the original forms were preserved as synonyms or "use for" terms to facilitate searching. This draft list was then circulated to members of the U.S. LTER Information Management Committee for suggested additions and deletions, which were then voted upon (Porter, 2010).

Organization of the keywords into a polytaxonomy (i.e., multiple taxonomies) and thesaurus followed the recommendations of the "Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies" (ANSI/NISO, 2010). Members of the Controlled Vocabulary Working Group classified each keyword into one of six different types (things, properties, processes, materials, disciplines, and events). This greatly simplified the organizational process by allowing them to focus on a smaller subset of terms when organizing them into a hierarchical structure, or taxonomy. Using the Tematres online thesaurus software (http://www.vocabularyserver.com/index.html), ten taxonomies were created. Four taxonomies were of type "things" (Organisms, Ecosystems, Organizational units, and Substrates), two taxonomies were of type "processes" (Processes, Methods) and the other four taxonomies (Substances, Measurements, Events, and Disciplines) were each of one of the four remaining types. Some additional terms were added to facilitate grouping (e.g., "hydrologic properties") and some terms that were found to be too ambiguous when used alone (e.g., "aboveground") were deleted. Synonyms, abbreviations, variant spellings were added as "use for" terms to facilitate searching of existing metadata documents that had not yet been revised to incorporate preferred terms. Version 1.0 of the U.S. LTER Controlled Vocabulary Working Group Thesaurus contained 627 preferred terms and an additional 150 "use for" terms (http://vocab.lternet.edu, Porter, 2010).

3. Creating a multilingual thesaurus

EnvThes is a thesaurus developed by European projects EnvEurope (http://www.enveurope.eu/) and ExpeER (http://expeeronline.eu/). It provides a common set of defined concepts that can be used to annotate the heterogeneous data collected and managed in different ways at research sites throughout Europe. It was selected for use by the ILTER because it is the most comprehensive list of terms available to describe LTER activities of ecological monitoring, research, and experiments.

EnvThes is constructed from several existing vocabularies and thesauri, including the U.S. LTER Thesaurus. To create a comprehensive list of concepts covering the wide range of disciplines studied by the European ecological community, terms from the INSPIRE spatial data themes (http://inspire.ec.europa.eu/index.cfm/pageid/2/list/7), EUNIS habitat types (http://eunis.eea.europa.eu/habitats.jsp), and NASA units controlled vocabularies were included (Schentz et al., 2013). English was established as the main language of EnvThes, and translations of concept definitions were made to provide multilingualism. English terms may or may not have translations, depending on the resources available for translation and the degree of equivalence of the translated term. EnvThes terms are linked to definitions in existing vocabularies such as AGROVOC (http://aims.fao.org/standards/agrovoc/about), Wikipedia (http://en.wikipedia.org/wiki/Main_Page), EUROVOC (http://eurovoc.europa.eu/), GEMET (http://www.eionet.europa.eu/gemet/), and EARTh (http://uta.iia.cnr.it/earth.htm), so that EnvThes is

part of a web of interlinked thesauri. Domain scientists do the translation of the concept definitions. As was the U.S. LTER Thesaurus, EnvThes is implemented using the Simple Knowledge Organization System (SKOS; Miles and Bechhofer, 2009) which allows these links to indicate "exact match", "close match", "narrow match", "broader match" and "related match". This system of linked thesauri creates a stable semantic reference for terms in EnvThes.

EnvThes, at the time the 2012 workshop, had terms translated into fourteen European languages. Additional languages needed to be added, because thirty-three languages are used in the ILTER. During the workshop Korean, Japanese, and Chinese participants contributed to the work by translating the 627 terms from the US LTER Thesaurus that are in EnvThes. This was readily achieved during the workshop, although some of the English terms did not have exact equivalents in the target language. Regional biases for terms were noted, e.g., "typhoon" is the English word used for "hurricane" in Asia, but in EnvThes the preferred term is "hurricane" which reflects the preferred term in the US LTER Thesaurus.

4. Multilingual search prototype

4.1. Implementation

A web-services framework was adopted for development of the prototype multilingual ecological data search engine. Use of web services facilitated the use of existing systems and eliminated the need to duplicate existing functionality. EnvThes is currently implemented using the Resource Description Framework (RDF)-based TopBraid software (http://www.topquadrant.com/) which incorporates a SPARQL endpoint for remote access. Similarly, the Metacat data catalogs incorporate a web-services application programming interface (API) that allow queries to be submitted as XML documents, with the results returned as XML as well. For ease in integration with the web, web services to perform the needed functions were written in the PHP language. The ARC2 PHP module was used to facilitate communication with an EnvThes instance established to support the prototype using SPARQL.

Implementation of the prototype thesaurus-enhanced multilingual search interface required several components (Fig. 3):

- Preferred-term autocomplete – Helps guide users to existing terms;
- Thesaurus – Needed to make connections between terms;
- Search interface – Web page from which users select terms to search in their local language;
- Enhancer – Uses the underlying Thesaurus to generate an expanded list of search terms by including synonyms and narrower or related terms to the user-provided term;
- Translator – Translates the original search term and expanded terms into the language needed for the search engine;
- Search preparation – Formats the translated and enhanced terms into a query suitable for use with a specific data catalog or search engine;
- Data catalog(s) – Stores the metadata, searches based on the prepared query, and returns a list of matching datasets;
- Display engine – Translates the list of matching datasets into an attractive, user-readable form.

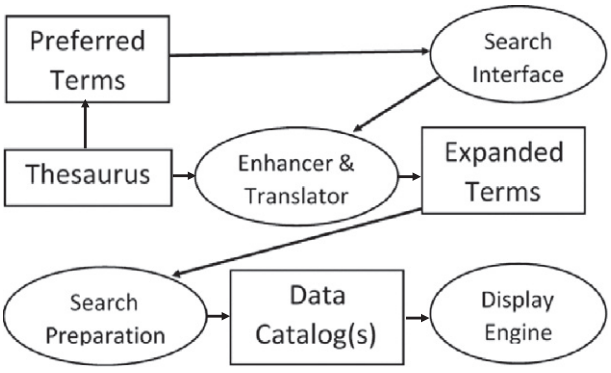

Fig. 3. Diagram of multilingual search system elements.

In the prototype, the process of identifying the initial preferred term was facilitated in the online search form by using an "autocomplete" list that suggested keywords from the thesaurus when a user started to type in a search string. This helped reduce "misses" caused by misspellings or selection of words not in the thesaurus. However, we did not implement this function for languages that use ideograms, such as Chinese, Japanese and Korean. For those languages a browse-based system may be preferable over an autocomplete system.

In our case, we were able to combine the Enhancer and the Translator functions by using the EnvThes Multilingual Thesaurus. In that thesaurus each preferred term is linked to a concept. Each concept has multiple alternative labels in different languages. Thus, a list of keywords for an enhanced multilingual search can be created by identifying a preferred term to serve as a starting point (e.g., "aquatic ecosystems"). Synonyms and narrower and/or related terms can then be identified and alternate labels in a particular language extracted (e.g., in French: "Lac", "Rivière", and "Ruisseau"). Once a list of terms has been extracted in the desired language, they need to be formatted into a form suitable for use in a search engine. In our prototype application, our targets were all Metacat data catalogs (Jones et al., 2001). These supported a web service interface that used a format known as "pathquery" which specified what should be searched and how elements of a search should be combined. In our case, we simply used a "union" operator to return datasets that had any one of the keywords in the enhanced search in it. The resulting list was reformatted using an XML stylesheet for viewing by the user.

4.2. Interface

A sample workflow shows the web pages seen by users during a search (Fig. 4). It starts with language selection, so that the proper autocomplete list of words in the user's local language can be displayed for the user and the appropriate language for the final search selected. It then proceeds to obtain the search term. The prototype takes the search term and translates and expands that term using multiple calls to the multilingual thesaurus. Multiple calls are needed because the thesaurus software typically returns only a single level of a hierarchy at a time, so burrowing down multiple levels requires multiple calls. In the prototype, resultant sets of search terms, ordered by complexity (e.g., synonyms only, synonyms and narrower terms, synonyms, narrower terms and related terms) are provided for the user to view. However, this step could be omitted for a production system and the search at a pre-determined level of complexity run automatically. The user is then queried for the data catalog to search and the search results are displayed. Each data catalog typically contains metadata only in the local language, although some may support more than one language. Translations are made into only a single language, so in the case where the Metacat contains more than one language, the user chooses which language to query.

A potential enhancement to the system would be to automate searches of multiple data catalogs, each in their preferred language. However, this would require that the enhancement/translation step be repeated multiple times and that the results returned from the different data catalogs be integrated, with any duplicates removed. The web services approach used in this prototype system would also allow non-Metacat data catalogs to be queried. For instance, the Chinese Ecosystem Research Network (CERN) does not use Metacat or EML, but could still use web services to received queries translated into Chinese from the prototype system and then return target datasets.

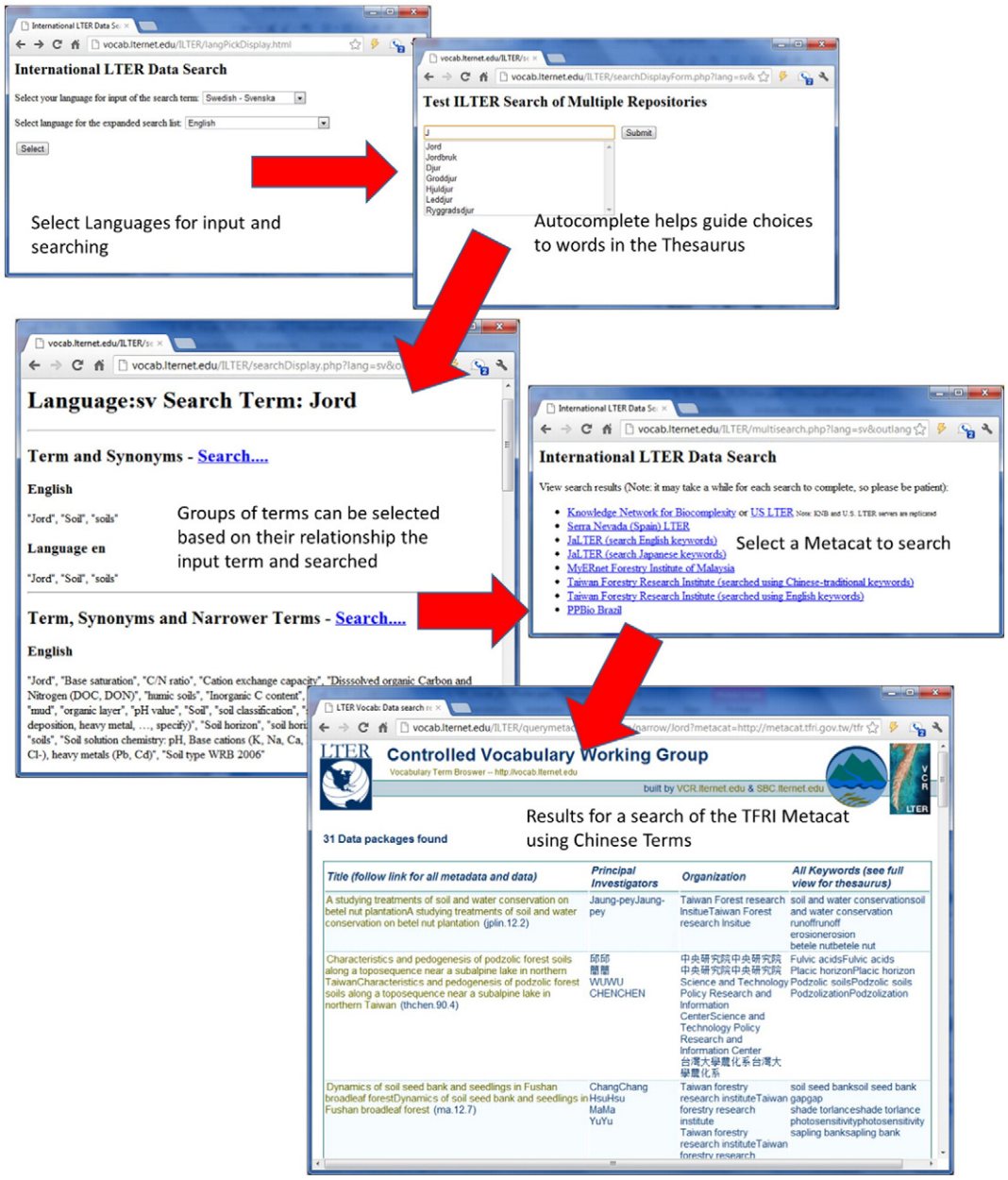

Fig. 4. A prototype system for multilingual searching of ecological data. The user selects the language they wish to use for search input (e.g., Swedish) and the Metacat to be searched (e.g., Taiwan Forestry Research Institute (TFRI) Metacat). A term autocomplete list in Swedish guides the searcher towards terms in the thesaurus. The thesaurus is used to select additional terms such as synonyms or narrower terms prior to preparing a search. The enhanced search "pathQuery" is then sent to the TFRI Metacat after being translated into the language appropriate for that particular Metacat (Chinese in this case).

4.3. Evaluation

The prototype multilingual search application was effective in improving most searches. By automatically expanding the number of search terms to include synonyms and narrower terms, most test searches were able to return relevant datasets. However, there were some limitations apparent in the prototype. First the number of translated terms varied widely across languages within the thesaurus. English had the largest number of terms in the EnvThes thesaurus used at the workshop, with over 1000 terms. In contrast many other languages were represented by 100 terms or fewer and a few, such as traditional Chinese were represented by fewer than 10. This meant that the effectiveness of the keyword enhancement for translation into those languages was relatively limited. However, effectiveness was still enhanced if the user-supplied search term could be translated into English and the resulting set of terms extracted from the thesaurus could be used. The translation of over 600 terms into Japanese, Chinese (simple), Chinese (traditional) and Korean during the workshop should dramatically help to remedy this limitation in subsequent versions of the EnvThes thesaurus, at least for Asian languages.

A second limitation was speed. The prototype multilingual data search was too slow to be used as a production implementation. Often several minutes might be required at the search enhancement and data search stages of the process. There were several bottlenecks associated with the system. First, the system was highly distributed, with the search interface in the United States, the thesaurus in Europe and the data catalogs in Taiwan, Japan, Spain, Brazil, Malaysia and the United States. Internet latency in any of the long-distance links caused the system to slow down. Secondly, the web services provided by the thesaurus were effective, but also limited in scope. As noted above, any term

could be extracted based on language and its relation to other terms within the thesaurus, but each level traversed in the thesaurus required that a separate web service query be sent for each of the members at that level. Thus, if the search term "soil" had 10 "child" terms, the keyword enhancement tool would need to perform 11 individual queries (the first on soil, then one each on each of the child terms). But if each of those child terms each had 10 additional children, the total number of required queries would be 111, each of them involving an intercontinental trip across the network. Additionally, for the purposes of testing, the prototype tool created alternative sets of search terms, using different rules regarding what should be included (e.g., synonyms only, synonyms and narrower, synonyms, narrower and related, and synonyms, narrower, related and the narrower terms of the related terms) that the user could select. This required that the entire enhancement process be repeated several times. Finally, the "pathQuery" search engine in the Metacat server is slow, especially when confronted with a large or complex query, such as those generated by the search enhancement web service. Complex searches could take up to several minutes to return a result.

Most of these performance difficulties can be relatively easily addressed in a production system. Eliminating the multiple search options (i.e., only searching for synonyms and narrower terms) is one obvious way to speed up the system. Similarly, moving the query-intensive functions closer to the server being queried would be a substantial help in reducing network latency. If necessary, applications that operate directly on the thesaurus database, rather than indirectly through a web service, could be used to reduce the time spent processing queries. If a single query sent to the thesaurus server could return the full hierarchy of needed terms instead of only a single level, speed would be substantially enhanced. Additionally, upgrades in the web service support in the component software could speed up searches. For example, recent versions of the Metacat software have added the capability to use the extremely fast Apache Solr search engine (Shahi, 2015), in place of the very general, but also very slow, "pathQuery" search. Finally, for a production system, periodic caching of either enhanced search term lists or even caching of data search results could drastically improve performance.

5. Automated translation of metadata

The multilingual search tool can help users locate suitable data, but that data will still not be useful if researchers cannot correctly interpret the metadata itself. It is beyond the scope of the prototype multilingual search system to provide full translations of the metadata. However, there are a variety of web-accessible translation engines that might be applied to this task. To evaluate the adequacy of those engines, we extracted the abstract from an EML metadata document and used several different web translation tools to transform it into English and compared that to the direct translation provided by the metadata author.

The original abstract was in Chinese (traditional):

本研究主要在建立觀霧地區大鹿林道鍬形蟲之物種名錄及族群數量等資料,自2002年4月至2003年3月間,於觀霧地區之大鹿林道設置樣站,以夜間燈光採集 法、沿路調查法、誘餌採集法和朽木採集法等四種方法進行調查。本區全年共記錄14屬28種440隻鍬形蟲,其中以高砂深山鍬形蟲(*Lucanus taiwanus*)為最優勢的種類,佔全部記錄數量的13.6%。

The translations of this paragraph were:

Google Translate Version http://translate.google.com/

This study aimed to establish Guanwu area Dalu Forest Road stag insects of the species directory and ethnic groups number and other information from April 2002 to March 2003, in view of the fog area of the large deer forest roads set the kind of station, the night light collection method along survey, bait collection and deadwood collection of four methods of investigation. This year, recorded a total of 14 genera of 28 kinds, 440, stag beetles, including high sand mountains stag beetles (*Lucanus taiwanus*) was the most dominant species, accounting for 13.6% of all records.

Bing Translate Version http://www.microsofttranslator.com/

This study mainly in view of establishing fog area lulindao rohac of list and data such as numbers of species, from April 2002 to March 2003, in view of fog district of lulindao station, along to night-time lighting collection law, investigation, decoy collection and Deadwood collection law, four kinds of methods to investigate. This area a total of 28 species of the record 14 440 rohac, mountains with high sand shovel-shaped beetle (*Lucanus taiwanus*) for most types of advantages, 13.6% percent of total recorded amount.

World Lingo Version http://www.worldlingo.com/en/products_services/worldlingo_translator.html

This research mainly establishment view fog area moose wood-road spade shape insect in materials and so on species of name list and tribal grouping tribal group quantity, from from April, 2002 to March, 2003 between, Yu Guanwu moose of wood-road establishment type station the area, by at night light gathering law, along the road the survey procedure, the bait gathering law and the rotten wood gathering law and so on four methods carried on the investigation. The home court whole year altogether records 14 to be 28 kind of 440 spade shape insect, in which take the high granulated substance remote mountain spade shape insect (*Lucanus taiwanus*) as the most superiority type, occupies records quantity completely 13.6%.

Finally, the author's original English Version was:

The purpose of this study was to draw up a namelist of stag beetles with their abundances along the Da-lu Forest Road in the Kuanwu area of northwestern Taiwan. From April 2002 to March 2003, four different methods, including light traps, transect line sampling, bait sampling, and rotten wood chopping method, were used in the investigation. In total, 28 species belonging to 14 genera of Lucanidae with 440 of stag beetles were recorded during this investigation in the Kuanwu area. The most abundant species was *Lucanus taiwanus* which accounted for 13.6% of total individuals.

Although none of the translations successfully captured a fully intelligible version of all the methods used to census stag beetles, it is at least possible to understand enough of the abstract to determine whether the dataset might be useful. It is even likely that someone familiar with the different sampling approaches used to observe stag beetles would be able to correctly discern which methods were used, even if the terminology is not typical.

We also reversed the process. When the original English abstract was translated to Chinese, co-author Sheng-Shan Lu noted that the grammar was incorrect and some characters were not in the correct order. The collection methods were also not fully understandable, just as they had been in the Chinese to English translation. He estimates that he understood about 60% of the meaning of the translation and this was sufficient to allow a determination about its usefulness to him. A similar test was done for Japanese to English and vice-versa, and the accuracy of translation was judged to be about 60% correct in both directions.

Translation success of English to Swedish and the reverse were also tested with Google Translate. As one would expect because Swedish and English are more closely related languages, automatic translations from Swedish to English or the reverse are quite good. The automatic translations were at least 90% semantically equivalent.

We did not attempt to replicate this translation experiment on all parts of the metadata. It is likely that the parts of the metadata relating to the actual structure of the underlying data tables should be even more intelligible. By combining translations of the metadata about the

column headers and descriptions with the relevant units, a more nuanced understanding of the data in a column can be achieved. For example, if a measurement of "annual nitrogen deposition rate" is mistranslated as "nitrogen in year", the underlying unit of "grams per meter squared per year" or "$g/m^2/yr$" should help to clarify any ambiguity.

6. Discussion

General ecological theories are best tested using data from widely-disparate systems (Tonn et al., 1990, Brown, 1995). Ecological processes that are important at a local scale may be quite different from those at regional and global scales (Levin, 1992, Lawton, 1999, Gross et al., 2000). A common approach is to mine the ecological literature for data, but such data are often summarized or incomplete. Access to raw and voluminous ecological data via data repositories provides new opportunities (Michener and Jones, 2012), but also demand bridging language gaps in order to locate and access ecological data from different regions or continents.

Online searching for data even in a single language can be complicated. We found that for the U.S. LTER sites, uncontrolled application of user-supplied keywords resulted in over 3000 keywords, the majority of which were used only once. The experience was much the same for the Taiwan data system. They found that almost 69% (990 of 1305) of the keywords were used only in a single dataset, and only 36 (2.8%) of the keywords were used at five or more times. Such a wealth of keyword diversity means that most searches based on a single keyword would result in only a single dataset, and that finding related datasets would require repeated searches of using different terms related to a subject. The situation is even more difficult in a multilingual context. The Taiwan Forestry Research Institute found that the English and Chinese keywords from their Metacat were used often in uncontrolled and inconsistent ways, even by the same authors. The same concepts in English keywords differed in spelling or could be singular or plural. Inconsistencies were also found in the Chinese terms used that meant the same thing. As an example, the term "wireless sensor network (WSN)" was annotated inconsistently in Chinese. They found three different translations in Chinese for WSN, 無線感測網 (five Chinese characters), 無線感測網路 (six Chinese characters), 無線感應器網絡(seven Chinese characters). The Chinese characters used vary from five to seven, but still had the same meaning.

There are two parallel approaches that can be used to help improve the reliability and efficiency of searches. One is to encourage, or require, the use of keywords drawn from a controlled vocabulary, thesaurus or ontology. Use of these preferred terms can help to remedy the "one keyword, one dataset" problem, because multiple datasets will share keywords. Moreover it addresses the unnecessary and confusing variation caused by variant spellings or the use of plural vs singular. For new metadata, autocomplete forms or keyword browsers can help guide users to preferred terms when preparing metadata. However, for existing metadata this approach can be time-consuming and expensive, because it requires going back through existing metadata to standardize selection of keywords. However, typically keywords will still be in a single language, so translation problems remain.

The other approach, that we tested here, is to increase the intelligence of the search process through the use of a multilingual thesaurus. Inclusion of synonyms or "use for" terms in the thesaurus can address the issues associated with variant spellings or use of plural vs singular terms. Additionally, the structure available in a thesaurus, where more specific terms can be linked to broader "parent" terms, allows a much more complete and reliable search to be run. For example, a search on "forests" will also return datasets that include the narrower term "trees," even though the term "forest" is never mentioned in the metadata. The tasks of translation can be moved from the metadata creation into the creation of the multilingual thesaurus, where each term need only be translated once.

These two approaches are complementary. As the quality of metadata documents is improved by the inclusion of standardized, preferred terms, so is the quality of searches provided through use of the multilingual thesaurus. However, there are also other approaches, such as free-text searches that do not depend on the identification of specific keywords. The relative brevity of many metadata documents, relative to text documents written for more general purposes, may pose a challenge for free-text searches. Moreover, if the search does not utilize the context provided by the structural descriptors that are used to define different components of the metadata, it is likely to return many false results. For example, a search for the researcher with the surname "Young" may be confounded with all manner of youthful organisms, such as "young leaves" and "young of the year"; not to mention the popular wind sensor manufactured by the "R.M. Young Company." Searches that focus on particular elements of a metadata document, such as title and keywords, are less likely to make such mistakes.

Generally, studies have shown that use of a preferred term list can improve search precision (Mackenzie-Robb, 2010) over free-text searches. One of the primary advantages the multilingual thesaurus has over free-text search is in its accuracy of results. The use of preferred terms ensures that the meaning of terms is known and assures consistency in term use that can improve search performance. Use of a thesaurus can also reduce irrelevant returns from free-term searches that are often caused by the inherent ambiguity of natural language and incompatibilities in translation of these terms. Of course, to attain this improved accuracy will exact a price. Research projects will need to adopt the thesaurus as a means for assigning terms to data resources. Limited term assignment will lead to unsatisfactory search results that miss relevant resources. Generation of multilingual thesauri is also challenging. In addition to the work required to generate a monolingual thesaurus, domain experts in each language will need to review and revise any automatic term and term definition translations to assure the term and its definition match the original concept. Additionally, terms for many languages will need to be manually translated. Beyond the work of creating a thesaurus, there are also management tasks, such as adding and defining new terms, deprecating outdated terms, and relating terms as needed.

Here we implemented enhanced searches using a single multilingual thesaurus, albeit one assembled from parts of existing thesauri. However, there is no reason that searches could not be enhanced using multiple thesauri, with synonyms and narrower terms drawn from several thesauri or ontologies. The main challenges would be developing the needed web services to harvest terms from each thesaurus and eliminating redundant terms. There would also be performance concerns, as the search could proceed only as fast as the slowest of the thesauri, and the larger number of search terms could slow down the search engines associated with data catalogs.

As discussed by Vanderbilt et al. (2010), multilingual ontologies, rather than thesauri, may offer the best long-term solution for facilitating sophisticated and accurate data discovery. Ontologies are models of concepts and their relationships within a scientific domain. Ontologies offer more relationship types (e.g., is-a, has-part, located-in) with which to capture semantic relationships between concepts (Madin et al., 2008). Work is ongoing to develop ontologies for the biological and ecological domains, in particular, the Biological Collections

Ontology, the Environment Ontology, and the Population and Community Ontology (Walls et al., 2014). Vanderbilt et al. (2010) envisioned linking multiple monolingual ontologies with a core ontology (e.g., OBOE, Madin et al., 2007; SERONTO, van der Werf et al., 2008). Such a structure would allow the wider array of relationship types available in ontologies to be exercised to further reduce the ambiguity inherent in multilingual searches. Development of thesauri pave the way for creation of ontologies (Almeida and Simoes, 2006; Rajbhandari and Keizer, 2012) but substantial additional effort to define both core ontologies and the monolingual ontologies is required. Additional effort is then required to link each of the ontologies into the core. For this reason Vanderbilt et al. (2010) recommended that thesaurus-based search be developed first, both to provide more immediate aid to searching and to help form the basis for the needed ontologies for a longer-term solution.

A more radical approach to providing access to multinational data is to retrieve actual data rather than metadata. This Linked Data approach combines computational ecology and eco-informatics (Gray, 2009) and refers to a set of best practices for publishing and interlinking structured or non-structured data on the web in a machine-readable way (Berners-Lee, 2006; Heath and Bizer, 2011; Wood et al., 2014). The Linked Data approach uses a general standard, the Resource Description Framework (RDF), to make data and metadata amenable to automated interpretation by computers. Although RDF provides a generic, graph-based data model encoding data in typed statements called triplets, it depends on the domain ontology or controlled vocabulary to specify the concepts and the relationships between concepts in the triplets (Bizer et al., 2009). Linked Open Data has been tested for ecological research in Taiwan using a monolingual data catalog (Mai et al., 2011) and it was found to be an effective way of sharing a wide variety of ecological data. The multilingual thesaurus developed by the ILTER is a first step towards building the sorts of ontologies needed to support linked data approaches in the future.

7. Conclusions

We found that a multilingual thesaurus-based data search capability could be developed using a web-service-based approach. There remain some issues regarding completeness of the thesaurus and performance, but those are issues that are readily addressed. We also found that processing metadata documents using existing online automated translation services, although far from completely accurate, provided enough information in most cases to support decisions regarding the suitability for use of specific datasets. There are additional and more sophisticated approaches that could also be used to support data discovery, but the thesaurus framework used here can serve as a useful stepping-stone, as well as a useful interim tool for researchers.

Acknowledgements

Funding: Travel support for US participants to the workshop in China was provided by a supplement grant to NSF DEB grant #021774. Local costs for all participants were generously covered by the Chinese Ecological Research Network (CERN) central office.

References

Almeida, J.J., Simoes, A., 2006. T_2O – Recycling thesauri into a multilingual ontology. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., Tapias, D. (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, May 22–28, 2006, pp. 1466–1471.

ANSI/NISO Z39.19–2005, 2010. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. NISO, Baltimore, Maryland.

Berkley, C., Jones, M., Bojilova, J., Higgins, D., 2001. Metacat: a schema-independent XML database system. Proceedings of the 13th International Conference on Scientific and Statistical Database Management, George Mason University, Virginia, July18–20, 2001. IEEE Computer Society, Washington, D.C., USA, pp. 171–179.

Berners-Lee, T., 2006. Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html (retrieved March 17, 2015).

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked data-the story so far. Int. J. Semantic Web Inf. Syst. 5, 1–22.

Broughton, V., 2006. Essential Thesaurus Construction. Facet Publishing, London.

Brown, J.H., 1995. Macroecology. University of Chicago Press.

Clarke, S.G.D., 2001. Thesaural Relationships. In: Bean, C.A., Green, R. (Eds.), Relationships in the Organization of Knowledge. Kluwer Academic Publishers, Boston, pp. 37–52.

Fegraus, E., Andelman, S., Jones, M.B., Schildhauer, M.P., 2005. Maximizing the value of ecological data with structured metadata: an introduction to the ecological metadata language (EML) and principles for metadata creation. Bull. Ecol. Soc. Am. 86, 158–168.

Gray, J., 2009. Jim Gray on eScience: A Transformed Scientific Method. In: Hey, T., Tansley, S., Tolle, K. (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington, USA, pp. 1–16.

Gross, K.L., Willig, M.R., Gough, L., Inouye, R., Cox, S.B., 2000. Patterns of species density and productivity at different spatial scales in herbaceous plant communities. Oikos 89, 417–427.

Haase, P., Frenzel, M., Klotz, S., Musche, M., Stoll, S., 2016. The long-term ecological research (LTER) network: relevance, current status, future perspective and examples from marine, freshwater and terrestrial long-term observation. Ecol. Indic. 65, 1–3.

Heath, T., Bizer, C., 2011. Linked Data: Evolving the Web Into a Global Data Space. MC Publishers.

IFLA Working Group on Guidelines for Multilingual Thesauri, 2009,. Guidelines for Multilingual Thesauri. IFLA Professional Reports 115The Hague, International Federation of Library Association and Institutions (IFLA) Headquarters (30pp.).

Jones, M.B., Berkley, C., Bojiova, J., Schildhauer, M., 2001. Managing scientific metadata. IEEE Internet Comput. 5, 59–68.

Jorna, K., Davies, S., 2001. Multilingual thesauri for the modern world - no ideal solution? J. Doc. 57:284–295. http://dx.doi.org/10.1108/EUM0000000007103.

Lawton, J.H., 1999. Are there general laws in ecology? Oikos 84, 177–192.

Levin, S.A., 1992. The problem of pattern and scale in ecology. Ecology 73, 1943–1967.

Lin, C.-C., Porter, J.H., Lu, S.-S., 2008. A metadata-based framework for multilingual ecological information management. Taiwan J. For. Sci. 21, 1–6.

Mackenzie-Robb, L., 2010. Controlled vocabularies vs. full text indexing. http://www.vantaggio-learn.com/White%20papers/vocabularies%20vs.%20text%20indexing.pdf.

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends Ecol. Evol. 23, 159–168.

Madin, J.S., Bowers, S., Schildhauer, M.P., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Eco. Inform. 2, 279–296.

Mai, G.S., Wang, Y.H., Hsia, Y.Y., Lu, S.S., Lin, C.C., 2011. Linked open data of ecology (LODE): a new approach for ecological data sharing. Taiwan J. For. Sci. 26, 417–424.

Michener, W.K., Jones, W.B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol. Evol. 27, 85–93.

Miles, A., Bechhofer, S., 2009. SKOS simple knowledge organization system reference. http://www.w3.org/TR/skos-reference (accessed 2016-04-03).

Ohte, N., Nakaoka, M., Shibata, H., 2012. ILTER and JaLTER: their missions and linkage to database development in the Asia-Pacific region. In: Nakano, S., Yahara, T., Nakashizuka (Eds.), The Biodiversity Observation Network in the Asia-Pacific Region. Springer, pp. 206–216.

Porter, J., 2006. Improving data queries through use of a controlled vocabulary. LTER databits. http://databits.lternet.edu/spring-2006/improving-data-queries-through-use-controlled-vocabulary (accessed 2016-11-20).

Porter, J., 2010. A controlled vocabulary for LTER datasets. LTER databits. http://databits.lternet.edu/spring-2010/controlled-vocabulary-lter-datasets (accessed 2016-04-28).

Porter, J., Costa, D., 2006. Keywords and terms from the LTER Network - 2006. Long Term Ecological Research Network http://dx.doi.org/10.6073/pasta/270a615ebbecf90aebc72134a1bda355.

Rajbhandari, S., Keizer, J., 2012. The AGROVOC concept scheme-a walkthrough. J. Integr. Agric. 11, 694–699.

Schentz, H., Peterseil, J., Bertrand, N., 2013. EnvThes – interlinked thesaurus for long term ecological research, monitoring, and experiments. In: von Bernd, P., Fleischer, A.G., Gobel, J., Wohlgemuth, V. (Eds.), EnviroInfo 2013: Environmental Informatics and Renewable Energies, 27th International Conference on Informatics for Environmental Protection – I and II. Shaker-Verlag, pp. 824–832.

Shahi, D., 2015. Apache Solr: An Introduction. Apache Solr. Apress, pp. 1–9.

Tonn, W.M., Magnuson, J.J., Rask, M., Toivonen, J., 1990. Intercontinental comparison of small-lake fish assemblages: the balance between local and regional processes. American Naturalist.]–>Am. Nat. 136, 345–375.

van der Werf, D.C., Adamescu, M., Ayromlou, M., Bertrand, N., Borovec, J., Boussard, H., Cazacu, C., Van Daele, T., Datcu, S., Frenzel, M., Hammen, V., 2008. In: Weitzman, A., Belbin, L. (Eds.), SERONTO, A Socio-Ecological Research And Observation Ontology: The Core Ontology. Proceedings of TDWG, October 2008, 19–24, Freemantle, Australia, pp. 17–25.

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Lin, C.C., Lu, S.S., Ogawa, A., Ó Tuama, É., Schentz, H., Su, W., 2010. A multilingual metadata catalog for the ILTER: issues and approaches. Eco. Inform. 5, 187–193.

Vanderbilt, K.L., Lin, C.-C., Lu, S.-S., Kassim, A.R., He, H., Guo, X., San Gil, I., Blankman, C., Porter, J.H., 2015. Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network. Ecosphere 6, 1–18.

Vihervaara, P., D'Amato, D., Forsius, M., Angelstam, P., Baessler, C., Balvanera, P., Boldgiv, B., Bourgeron, P., Dick, J., Kanka, R., Klotz, S., Maass, M., Melecis, V., Petřík, P.,

Shibata, H., Tang, J., Thompson, J., Zacharias, S., 2013. Using long-term ecosystem service and biodiversity data to study the impacts and adaptation options in response to climate change: insights from the global ILTER sites network. Curr. Opin. Environ. Sustain. 5, 53–66.

Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L., Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., Ó Tuama, É., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wieczorek, J., Whitacre, J., Wooley, J., 2014. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS One 9 (3):e89606. http://dx.doi.org/10.1371/journal.pone.0089606.

Wood, D., Zaidman, M., Ruth, L., Hausenblas, M., 2014. Linked Data. Manning Publications Co.