Contents lists available at ScienceDirect

Ecological Informatics

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

A multilingual metadata catalog for the ILTER: Issues and approaches

Kristin L. Vanderbilt [a,*], David Blankman [b], Xuebing Guo [c], Honglin He [c], Chau-Chin Lin [d], Sheng-Shan Lu [d], Akiko Ogawa [e], Éamonn Ó Tuama [f], Herbert Schentz [g], Wen Su [c]

[a] Sevilleta LTER, University of New Mexico, Albuquerque, New Mexico 87131 USA
[b] LTER-Israel, Ben Gurion University, Midreshet Ben Gurion, Israel
[c] Chinese Ecological Research Network, Chinese Academy of Sciences, Beijing, China
[d] Taiwan Ecological Research Network, Taiwan Forest Research Institute, Taipei, Taiwan
[e] JaLTER, National Institute for Environmental Studies, Tokyo, Japan
[f] GBIF Secretariat, Copenhagen, Denmark
[g] Umweltbundesamt GmbH, Vienna, Austria

ARTICLE INFO

Keywords:
Challenges
Language
Translation
Ecology
Ontology

ABSTRACT

The International Long-Term Ecological Research (ILTER) Network's strategic plan calls for widespread data exchange among member networks to support broad scale synthetic studies of ecological systems. However, natural language differences are common among ILTER country networks and seriously inhibit the exchange, interpretation and proper use of ecological data. As a first step toward building a multilingual metadata catalog, the ILTER has adopted Ecological Metadata Language (EML) as its standard, and ILTER members are asked to share discovery level metadata in English. Presently, the burden of translation is on the data providers, who frequently have few resources for information management. Tools to assist with metadata capture and translation, such as localized metadata editors and a multilingual environmental thesaurus, are needed and will be developed in the near future. In the longer term, ILTER will cooperate with other communities to develop ontologies that may be used to automate the process of translation and will produce the most linguistically and semantically accurate metadata translations.

1. Introduction

The International Long Term Ecological Research (ILTER) Network (http://www.ilternet.edu) consists of 34 member countries that support long-term data collection and analysis in order to detect, interpret, and understand environmental changes. The strategic plan for the ILTER Network includes these goals:

- Foster and promote collaboration and coordination among ecological researchers and research networks at local, regional, and global scales
- Improve comparability of long-term ecological data from sites around the world, and facilitate exchange and preservation of this data.

In order to achieve these goals and conduct international research projects, the ILTER must develop a multilingual information management system that allows scientists to discover and use data gathered throughout the network. Data sharing across communities each having its own natural language and cultural and historical background poses many semantic and technical challenges.

The ILTER has adopted Ecological Metadata Language (EML) as its metadata standard. To develop a common metadata cache, member networks of the ILTER have agreed to provide at least a subset of EML in English, in addition to complete metadata in the contributing country's native language (Vanderbilt et al., 2008). This subset of discovery level metadata includes identifier, title, abstract, creator and contact. Other elements, such as keywords, are highly recommended. Even generating this seemingly small amount of metadata in EML and then translating it has resulted in significant challenges with respect to both translation accuracy and software development in a network where information management resources are often limited. It is apparent that the process to develop tools to facilitate translation and the creation of an ILTER metadata catalog will have several steps.

In this paper, we review the expected trajectory of the evolution of a multilingual ILTER metadata catalog from a system relying on free text translation to one that relies on a multilingual thesaurus and, in the long-term, ontologies. We give examples of how ILTER networks are confronting the issue of software localization in order to generate EML and how they intend to address the issue of translating metadata content. We discuss how a multilingual ecological thesaurus needs to be developed in order to lessen the translation burden on individual networks. The potential of ontologies to further resolve semantic translation issues is also reviewed.

[*] Corresponding author. Tel.: +1 505 277 2109; fax: +1 505 277 5355.
E-mail address: vanderbi@sevilleta.unm.edu (K.L. Vanderbilt).

1574-9541/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2010.02.002

2. Understanding the metadata: translation challenges

From the perspective of the usage of the ILTER network metadata catalog, it is important that a data consumer be able to understand the metadata terms in his/her language when the consumer does not have any understanding of the native language of the provider. Symmetrically, the provider may not have an understanding of the language of the consumer. English will be the language used as the bridge between providers and users, but translation to English is often imperfect.

One of the key issues causing ambiguity in multilingual and multicultural terminology is equivalence. Equivalence means that the target language contains a term that is identical in meaning and scope to the term in the source language. Units of measure are examples of equivalent concepts. Terms selected from more than one natural language vary in the extent to which they represent the same concept, and a continuum of equivalence ranging from exact equivalence to non-equivalence is recognized (Fig. 1) (Doerr, 2001). Inexact equivalence, for instance, refers to a term in the target language that expresses the same general concept as the source language term, although the meanings of the terms are not exactly identical. Partial equivalence means that a term in one language has a broader meaning than the same term in another language.

Inexact equivalence and partial equivalence are illustrated by an example from a European forest fire monitoring project. Requests for data were made to each country in its native language, and countries were asked to provide data on the number of forest fires during a specific time interval. In an early stage of analysis of this pan-European data set, it appeared that the frequency of forest fires completely changed at the border between two neighboring countries. In addition, some countries, known for their forest fires, reported only a few fires. The mysterious differences were found to be due in part to different concepts of "forest" and "forest fire". The latter concept, an example of inexact equivalence, differs between countries based on continuity of fire, area impacted and completeness of destruction. The concept of "forest" is partially equivalent, because in the language of an arid country it includes both sparsely and densely wooded areas, while a sparsely wooded area in a more mesic country is referred to as "savanna".

Single to many equivalence is illustrated by the "wetlands" concept in US English and Japanese. Both English and Japanese concepts of "wetlands" include saltwater/freshwater marshes and mangrove forests, but otherwise the concepts are comprised of elements that don't overlap (Table 1).

Further translation problems arise when the same term exists in two languages, but refers to different concepts. For example, the term "maquis" is used in France to refer to a specific habitat. The same term was translated to Italian as "macchia", but the term refers to a different habitat in Italy than its original use in France. Israel and other Mediterranean countries also use the term "maquis" to describe a particular habitat, but the actual species found vary from country to country.

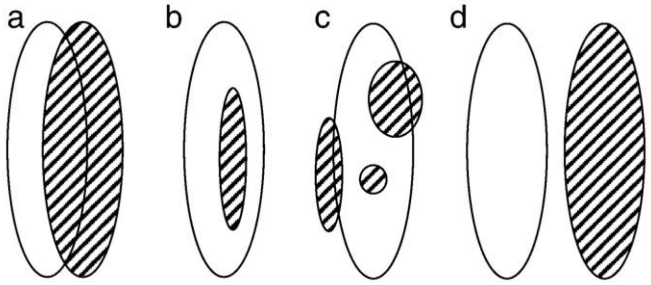


Fig. 1. Terms selected from more than one natural language vary in the extent to which they represent the same concepts. These variations can be seen as forming a continuum that ranges from exact equivalence in meaning through a) inexact equivalence, b) partial equivalence, c) single to many equivalence, to d) non-equivalence.

Table 1
Comparison of the US English and Japanese "wetlands" concept.

Wetlands[a]	湿地 ("Shicchi" ≈ wetlands)[b]
● Marshes ○ Tidal ○ Nontidal ■ Wet meadows ■ Prairie potholes ■ Vernal pools ■ Playa lakes ● Swamps ○ Forested swamps ■ Bottomland hardwoods ○ Shrub swamps ■ Mangrove swamps ● Bogs ○ Northern bogs ○ Pocosins ● Fens	● 湿原・塩性湿地 (Mires/salt marshes) ● 河川・湖沼 (Rivers/lakes and freshwater marshes) ● 干潟・マングローブ林 (Tidal flats/mangrove forests) ● 藻場 (Seaweed/seagrass beds) ● サンゴ礁 (Coral reefs)

[a] Classification by U.S. Environmental Protection Agency (EPA) (EPA, 2010).
[b] Types of wetlands in Japan (Biodiversity Center of Japan, Ministry of Environment, 2010a, b).

Even countries speaking the same language may use the same word to mean different things. For example, in northern Scotland researchers use the term "dry" to describe habitats that are not bogs, marshes or other aquatic habitats. For most other English speakers, "dry" would refer to a low precipitation region such as a desert.

3. ILTER solutions: a trajectory approach

As a first step toward developing an ILTER information management system, ILTER will use EML as its metadata standard and asks members to contribute discovery level EML in English to a common metadata cache. EML is an XML-based metadata specification developed for ecologists (Fegraus et al., 2005). The experiences of the Taiwanese and Chinese ILTER networks with respect to 1) localizing software in order to create EML, and 2) doing the translation from Chinese to English demonstrate that the costs of this effort are high and that tools are needed to assist with translation.

3.1. Localizing software for metadata capture: an example from TERN

In 2004, the Taiwanese Ecological Research Network (TERN) was the first non-English-speaking member of the ILTER to embrace the use of EML (Lin et al., 2006, 2008). In order to create EML in Chinese, TERN found that they had to localize the tools (Morpho, an EML editor, and Metacat, the database that stores EML documents) that had been developed for editing and managing EML documents by the National Center for Ecological Analysis and Synthesis (NCEAS) in the United States. The many changes to the source code that the TERN scientists had to make to get it to work with Chinese characters illustrates the effort required to localize software.

Morpho is an EML editor that was developed at NCEAS (Higgins et al., 2002). TERN started to use Morpho version 1.4.0 and 1.5.0 which provide English interfaces. TERN changed the source code of Morpho-1.6.1 to traditional Chinese characters, and now offers a complete Chinese version of Morpho, based on local scientists' request, which has a similar look and feel to the original English version. The translation to a Chinese version of Morpho was not a simple process and several problems had to be overcome. For instance, the English word strings on the user interface were replaced by Chinese characters successfully but the search function was not able to use Chinese. Morpho can be used to query a Metacat database, but

Chinese characters were displayed as squares on the results screen. TERN was able to fix both problems with the English/Chinese language by modifying the Morpho query program to let the input stream support UTF-8 string codes.

For maintaining a catalog of metadata documents, TERN adopted the Metacat system (Berkley et al., 2001) developed by NCEAS. The Metacat framework is controlled by a Java servlet that acts as the interface to any SQL-compliant relational database supporting the Java Database Connectivity (JDBC) protocol (Jones et al., 2001). Metacat includes five subsystems for storage, replication, query, validation and transformation.

When TERN first installed the Metacat-1.4.0 version in 2004, the main language problem was that the web search result page did not show up properly. The Chinese characters became question marks "??" in the web results page. This problem was caused by the Metacat query program responding with content type Unicode which could not be read by the web browser to display Chinese correctly. TERN altered the Java code in Metacat to read "response.setContentType ("text/xml; charset=UTF-8")", after which the Chinese characters displayed properly in the web page.

When Metacat was upgraded to Metacat-1.5.0, TERN found that Metacat changed the code to transfer special characters such as Chinese to numeric entities which were written into the database as "&#xxxx". This caused another problem. Chinese data are not in ASCII (American Standard Code for Information Interchange) character format and when written into Metacat are changed to html Unicode format. For example, 中文, which means "Chinese", will become "&#20013;&#25991" when saved into the database. Although the conversion code in Metacat-1.5.0 seemed to be able to solve the web page search problem, TERN still had to change the code to avoid the html Unicode format in the database.

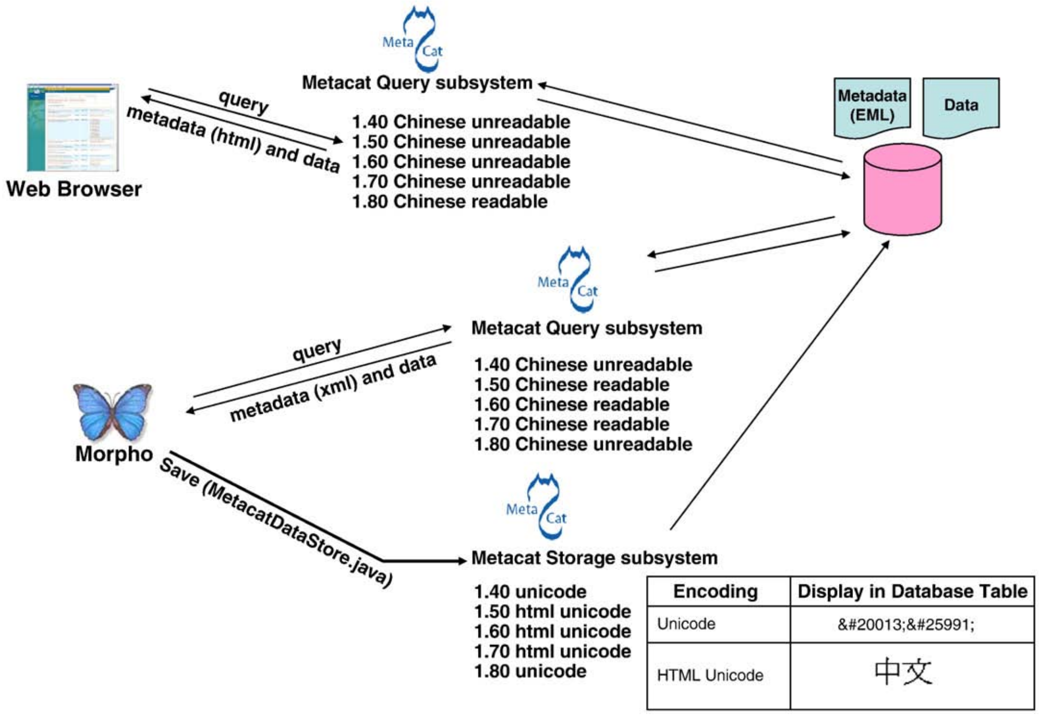
TERN used these Morpho and Metacat solutions to work smoothly with local scientists until 2008. With the release of Metacat-1.8.1, the function of character encoding conversion was eliminated and this version of Metacat fixed the language problem for the web page search. Presently, only the code in Morpho needs to change to accommodate the use of Chinese characters. Fig. 2 illustrates how the changes in Morpho and Metacat made by the developers at NCEAS influenced how Chinese characters were stored in the database and rendered as search results.

3.2. Who does the translation into English? The CERN perspective

Solving the problems of tool localization is a significant barrier to generating the EML that will feed into the ILTER metadata cache. The second major barrier is the translation of the metadata content itself. An example from the Chinese Ecological Research Network (CERN) illustrates the cost and need for tools to assist in translation.

CERN is potentially a major contributor of data to the ILTER. China has an extensive monitoring program which covers a wide variety of ecosystems. It is clearly a major resource for socio-ecological data. CERN uses a metadata standard that it developed itself (Standard of PR China GB/T 20533-2006). Data resources are physically stored and individually managed by field station sub-centers and a central synthesis center. The synthesis center harvests metadata from all field stations so that the synthesis center can provide an all-in-one search engine for all metadata for data users. The number of datasets has reached more than 5000. Having discovery level EML for the CERN data will represent a significant addition to the ILTER.

The five required elements which are defined in EML, i.e. identifier, title, abstract, creator and contact, are also included in CERN metadata. However, CERN faces problems to translate these five elements into English. The data resources belong to different organizations or field stations. The field stations' data managers may not have time to do the translation, so the question is who will take charge of converting Chinese metadata to English metadata, the centralized synthesis center or field stations? Secondly, how will the translation be done?



Encoding	Display in Database Table
Unicode	&#20013;&#25991;
HTML Unicode	中文

Fig. 2. Repeated modifications to metadata management software created for English-speaking users were needed to localize the software for use by TERN. The cylinder represents the relational database that houses data and metadata as EML. The Metacat framework allows the database to be queried from a web browser or Morpho, the application for entering metadata into EML. The Metacat query subsystem produced unreadable web browser content for metadata in Chinese characters until version 1.8.0. The Metacat storage system was fixed in version 1.5.0, by changing the storage system from Unicode to html Unicode, so that Morpho would display Chinese characters. This, however, caused the database to store characters in format "&#XXXX;", which was not useful to the Chinese database administrators.

Automated tools for translating the specialized ecological vocabulary from Chinese to English have not been found. In general, according to CERN's experience, automated translation software is very weak in accurate translation from Chinese into English and manual corrections must be done after automated translation. For instance, a word in Chinese that describes a type of landslide literally translates to English, character-for-character, as a “mud-rock flow”, a phrase that is not used in English. A query for data on a “mud-rock flow” in an English database would yield no results, since the correct English term is “debris flow” (Scott, 2004). CERN has concluded that utilizing automated translation software is not applicable and feasible, and that manual translation is unavoidable.

As far as manual translation is concerned, personnel and funding should be taken into account. Because the Chinese language, culture and thought patterns differ from English, sometimes the translation results may be “Chinglish” which is hard to understand by non-Chinese English speakers. Professionals who have the knowledge in both ecology and English must be sought and employed, so that more accurate translation can be provided. Funding is also an issue to be solved. CERN's routine funds do not cover the translation task, and other funds would have to be found.

Although it will cost much time and money, CERN is committed to translating the five required elements into English. However, the burden of translation of the complete metadata into English can not be carried by CERN. CERN has to leave the complete metadata in the local language and put the translation burden on the data user.

4. The next step: a multilingual thesaurus of ecological terms

A multilingual thesaurus from which ILTER metadata creators could select keywords would be a valuable step towards improving data discoverability and lessen the burden on metadata providers needing to translate discovery level metadata. A thesaurus is a set of concepts in a particular domain of knowledge that is organized hierarchically such that relationships between the concepts are explicit (ISO 5964:1985). Concepts are represented by a controlled list of preferred terms that are identified by the thesaurus' developers. Hierarchical relationships between concepts are indicated by “Broader Term” (a more general concept than the preferred term) and “Narrower Term” (a more specific concept than the preferred term) designations in the thesaurus. A “Related Term” is one whose scope overlaps somewhat, but not completely, with that of the preferred term. “Use For” terms are equivalent to the preferred term (synonyms). If one, for instance, queries the National Biological Information Infrastructure's Biocomplexity Thesaurus (http://www.nbii.gov/portal/community/Communities/Toolkit/Biocomplexity_Thesaurus) for “marshes”, one sees that a broader term is “wetlands” and a related term is “swamps”.

The ILTER will select 600–1000 preferred terms for the multilingual thesaurus that encompass the range of social and ecological research being done in the network. A multilingual European environmental thesaurus that will likely be a starting point for the development of an ILTER thesaurus is GEMET, the GEneral Multilingual Environmental Thesaurus (http://www.eionet.europa.eu/gemet/), which has been under development by the European Environmental Agency since 1996. GEMET is a compilation of several European multilingual vocabularies, and aims to define a common set of terms for the environment. GEMET is designed as a general thesaurus, and the current version contains 6562 terms and is available in 27 (mainly European) languages. Definitions of most concepts in GEMET are available in British English and are used to ensure the internal systematic and linguistic coherence of the thesaurus. The definitions provide a useful glossary function when the semantics of the thesaurus structure might not be completely captured.

Other environmentally-oriented multilingual thesauri that may provide terms and translations for the ILTER thesaurus include the Multilingual Thesaurus of Geosciences (http://en.gtk.fi/Geoinfo/Library/multhes.html) available in English, French, German, Italian, Spanish, Finnish and Russian; AGROVOC, the Food and Agriculture Organization of the United Nations (FAO)'s multilingual thesaurus covering agriculture, forestry, fisheries, food and related domains such as the environment (http://www4.fao.org/agrovoc/default.htm) which is available in some 20 languages; and the United States Department of Agriculture Library's Thesaurus (http://agclass.canr.msu.edu/agt.shtml), which contains terms related to natural resources, earth and the environment in English and Spanish. The United Nations Environment Program (UNEP) has also published a multilingual Thesaurus of Environmental Terms (EnVOC) in six languages — Arabic, Chinese, English, French, Russian and Spanish. Thesauri such as the European Communities' EUROVOC (http://europa.eu/eurovoc/) thesaurus, available for 22 European languages, will be evaluated for sociological terms related to the ILTER. The Multilingual Thesaurus of Geoscience of the International Union of Geological Sciences/Commission may provide terms related to hydrology and soils in Chinese, Japanese, Korean, and English (CCOP and CIFEG, 2006).

5. The long-term solution: multilingual ontologies

Having an ILTER multilingual thesaurus will facilitate translation of discovery level metadata into English. Using a thesaurus, however, can still result in vague translations when equivalent terms in both source and target languages don't exist. The ultimate goal to support translation of metadata in the ILTER is to have a multilingual environmental ontology to further semantically disambiguate concepts. An ontology is a formal model that defines “concepts and their relationships within a scientific domain such as ecology” (Madin et al., 2008). In general, a system using ontologies is a combination of a core ontology and domain ontologies. The core ontology is an extensible framework into which data from many sources can be mapped (Doerr et al., 2003). A domain ontology extends the concepts and relationships of the core ontology into specific scientific domains, such as forest vegetation or soils. An example of a core ontology is SERONTO (Socio-Ecological Research and Observation oNTOlogy) (van der Werf et al., 2008), which enables the description of data from different origins in a common conceptual manner, and is designed to be the basis for ecological domain ontologies.

5.1. A translation example using ontologies

For purposes of translation, one ontology can be mapped on to another. Suppose, for example, that ILTER develops/adopts a forest domain ontology. Suppose also that Israel develops a forest domain ontology for its own use, but wants to integrate their data with data from other ILTER countries. In this case, the ILTER ontology is referred to as the common or destination ontology, and the Israeli ontology is called a local or source ontology. Ideally the Israeli ontology is a subset of the common forest habitat domain ontology, but it is possible that it may be constructed with minor differences. If mapping can be made between the local ontology and the destination ILTER ontology, then an unambiguous translation can be made.

The following features of ontologies make them more explicit than a thesaurus:

(1) The definition of concepts is not only done by “free-text” descriptions but by relations to other concepts, thus creating formalized, man- and machine-readable sentences which explain the concepts. For instance, from Fig. 3, “Maquis HasClimateZone Mediterranean”.

(2) The mapping between the source concept and the common ILTER concept can be expressed as relations and thus all the

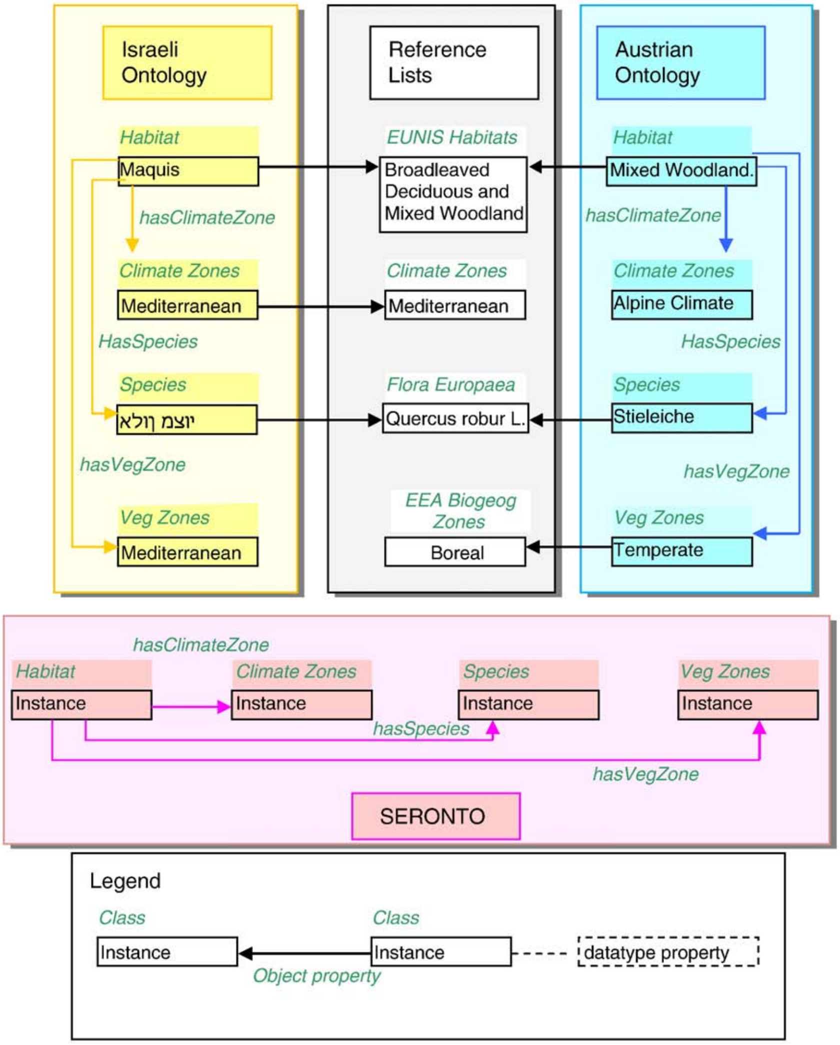


Fig. 3. This is an example of ontology mediated translation. SERONTO is a core ontology (pink) where concept classes such as habitat, species, and climate zone are related by object properties such as hasClimateZone, hasSpecies, and hasVegZone. The Israeli (yellow) and the Austrian (blue) plant community ontologies are based on SERONTO. In this example, there is a direct mapping between the Hebrew and German common names for *Querqus robur* L. using the *Flora Europaea* reference list. Even though the two habitats are instances of the EUNIS "Broadleaved Deciduous and Mixed Woodland" Habitat, they are not identical because the Israeli climate zone and the Austrian climate zone are not equivalent. Additional rules in the domain ontology would be needed to determine if the data from the two sites can be integrated.

different relations that can exist (complete equivalency, partial equivalency, etc. as shown in Fig. 1) can be realized.

(3) Source and destination concepts can be based on a common (core) ontology even when they do not follow a common structure in general. This means that ontologies allow the expression of common, agreed on global concepts of a research community and can also be extended to describe specific details, such as were explained in the example of the forest fire monitoring.

(4) Ontologies allow the concept of inheritance, which means that you can have upper concepts and lower concepts derived from the upper concepts which still can be seen under the view of the upper concepts. This allows simultaneous mapping on a higher level whereas on a lower level differing details are expressed. For instance, assume a community has defined the concept "site" but has not defined the concept "air measurement site". The concept "site" has the data type properties "latitude", "longitude", and "altitude". For the concept "air measurement site" a data type property "height above valley floor" is needed in addition. So the concept "air measurement site" can be inherited from the concept "site", thus taking on the datatype properties "latitude", "longitude" and "altitude". Any program which is designed to process "site" can process any of the individuals of the class "air measurement site."

The following example illustrates how ontologies could be used to integrate multilingual data. The example is based on the following assumptions: 1) Habitat data from Israel and Austria need to be integrated; 2) There is a common ontology, in this case SERONTO, that defines some concepts; and 3) There are some reference lists, not all commonly used, but some used individually by different countries. Ontologies from Israel and Austria have been mapped on to the reference lists where possible.

The Israeli concepts refer to common reference lists in the following way:

1) individual "Maquis" of class Habitat references to the EUNIS (http://eunis.eea.europa.eu/) Habitat "Broadleaved Deciduous and Mixed Woodland"; 2) individual "Mediterranean" of class Climate Zone references to the "Mediterranean Climate" within the assumed common list of climate zones; and individual "יוצמ ןולא" of class Species references to "*Quercus robur* L." within *Flora Europaea* (http://en.wikipedia.org/wiki/Flora_Europaea). The Austrian concepts are referring to common reference lists in the following way: 1) individual

"Mixed Woodland" of class Habitat references to EUNIS Habitat "Broadleaved Deciduous and Mixed Woodland"; 2) individual "Alpine Climate" of class Climate Zone does not have reference to a common list; 3) individual "Stieleiche" of class species references to "*Quercus robur* L." within *Flora Europaea*; and 4) individual "Temperate" of class Vegetation Zone references to "Boreal Zone" of the EEA-Biogeographic Zones list.

Ontologies provide several types of information to a scientist trying to learn if data from different sources can be integrated. For instance, although the Israeli and Austrian ontologies do not use any descriptors in common, the relations in the ontologies themselves between the habitat and species and habitat and climate zones help reveal what the terms mean. We can learn from the relations in the ontologies that, while both Israeli habitat "Maquis" and Austrian habitat "Mixed Woodland" contain species *Quercus robur* L., Israeli habitat "Maquis" has climate zone "Mediterranean" while Austrian habitat "Mixed Woodland" has climate zone "Alpine". This is more information than the scientist would have been able to glean from a thesaurus regarding the comparability of these two vegetation types. Also, because both ontologies use the same core ontology (SERONTO), the relations and classes are the same in both and the information receiver can be more clear on what he/she gets, even if the Individuals (the Instances) are not the same. Further, both Israeli and Austrian ontologies in this example are mapped to the same reference lists for Habitat and Species, and it is assumed that the two ontologies do not use the same Vegetation Zone and Climate Zone reference lists. In this situation, where the same ontology has been used to describe the non-common reference lists, we can more easily deduce the meanings of the non-common Vegetation and Climate Zones because of their relations to the common concepts.

5.2. Translation considerations

To achieve semantic interoperability throughout the ILTER, the most desirable approach would be for the ILTER community to adopt a core ontology, such as SERONTO, that uses the same concepts, individuals, and reference lists when preparing ontologies in each natural language. More likely, however, ontologies will develop with different structures in different natural languages, and mapping will then be done to the global ontology and reference lists. The multilingual ILTER thesaurus will be the starting point for multilingual ontology development.

For the common use of ontologies for translation purposes, referencing concepts (classes) individuals and relations (object properties) is crucial. Within the web-oriented formal ontology language OWL (Web Ontology Language) the referencing mechanism is based on URLs, which can be regarded as identifiers of those elements. For successful referencing it is necessary that, once established, links do not break and that the contents behind a URL are stable. There is thus a need for versioning and "permanent" possible resolution of references. A versioning system will maintain the interconnection between old contents and new contents for the documentation of the concepts' development. Possible ways to generate identifiers that are stable over time and space are the systems of European Article Number (EAN)-Barcodes or Life Science Identifiers (LSIDs) (http://lsids.sourceforge.net/).

It would be desirable for the ILTER to have a software that would map ontologies to a common ontology and perform the translation from one language to another. To date, there is much software for processing ontologies, including ontology editors, reasoners, query machines, and semantic miners. We are, however, unaware of any software that is more than experimental that allows users to see the same concepts through the lenses of different languages. Development of such software will be challenging, and should include 1) routines to execute the translations, 2) multilingual search and reasoning functions so that concepts entered in a mixture of languages can be recognized; e.g. a language mixture of English with German is common; and 3) functions that can resolve character sets (e.g. in Europe: Latin/Cyrillic; in Asia: Chinese/Korean) and interpret the meaning of the words in the different character sets.

6. Discussion

For the ILTER to achieve its goal of researching complex environmental issues across the globe, ILTER scientists must be able to discover and understand data collected at sites throughout the network, regardless of the language in which the data are documented. An ILTER ecologist in the U.S., for instance, might want to examine how long-term aboveground net primary production (ANPP) of grasslands worldwide is influenced by changing precipitation regimes. She wants to incorporate as many sites as possible, so that her study includes multiple soil types and disturbance patterns, as well as a wide gradient of precipitation. ANPP is a parameter commonly collected throughout the ILTER, but this scientist would probably discover only data from the US, South Africa, and Australia when searching current ILTER regional metadata catalogs in English. Relevant data undoubtedly exist in China, Mongolia, Hungary and elsewhere, but metadata from these sources is unlikely to have any English translation associated with them at present. The multilingual ILTER thesaurus, however, will allow Chinese metadata contributors to tag their ANPP datasets with the terms "ANPP" and "grassland" in English and Chinese, making them discoverable to the US scientist. Farther in the future, when ILTER has integrated ontologies into its information management system, intelligent query software that exploits the ontologies would also alert the scientist to data sets in English or other languages that contain variables for "biomass" and "area". The software will be "smart" enough, based on relationships defined in an ontology, to know that these variables could potentially be used to calculate ANPP (ANPP is calculated as (mass of carbon accumulated)/area/year). The multilingual ILTER metadata catalog will thus greatly increase the range of data available to ILTER researchers by facilitating cross-language queries, semantically disambiguating translated terms, and exploiting the relationships between concepts to intelligently expand queries to find relevant data sets.

Development of a multilingual metadata catalog is a significant undertaking, and ILTER can synergize with other organizations that are also examining options for developing information management systems that accommodate users of different languages. The Global Biodiversity Information Facility (GBIF), for instance, is developing a dedicated site (http://vocabularies.gbif.org) for community supported vocabularies. Mapping tools allow users to provide natural language translations of terms such as country names and vernacular names of species and to link them to relevant international standards provided by, e.g., ISO or TDWG. Sites such as this offer the possibility for communities of practice to collaborate on vocabulary term definitions and their representation in multiple natural languages. In addition, GBIF has developed the Integrated Publishing Toolkit (IPT) (http://www.gbif.org/informatics/infrastructure/publishing/) to enable publishing of data and metadata to the GBIF network. The IPT provides an integrated metadata editor that supports writing of metadata in a GBIF profile of EML. The IPT is designed to support internationalization of the interface and is already available in French, Spanish, and English. TaiBIF (Taiwan Biodiversity Information Facility) is developing an interface for the IPT in Chinese. Future versions of the editor will support writing of the recommended subset of metadata elements in multiple natural languages.

Multilinguality is a challenge for the INSPIRE (Infrastructure for Spatial Information in Europe) initiative (Annoni et al., 2008), which must support users from throughout the polyglot European Union. INSPIRE will be built on top of existing country-level spatial data infrastructures, and grapples with the same issues that ILTER does with respect to supporting discovery of data resources generated in many natural languages. INSPIRE has the additional challenge of different place names in different languages, for which multilingual gazetteers will be used. One such service is the EuroGeoNames (EGN) project (Sievers and Zaccheddu, 2005) which provides official geographical names information in 25 languages. ILTER can capitalize on the web service for this gazetteer as more ILTER spatial data is acquired, and can extend the service to place names in non-European languages.

The ILTER metadata catalog will need to integrate many ontologies, and ontologies can be expensive to construct. Pettman (2006) estimated that, given the costs of domain experts and knowledge engineers to extract concepts, agree on relationships between them, and encode them in description logic, an ontology containing 1000 concepts would cost $20,000 US in western European wages. The cost to generate an ontology could be quite large, depending on its complexity. ILTER will join other biodiversity/ecology communities developing ontologies under the auspices of standards-making organizations like the Biodiversity Information Standards (TDWG) group (http://www.tdwg.org) to avoid duplication of efforts and also to engage in the process of generating standard ontologies. In addition, some ontologies already exist that ILTER may be able to take advantage of. The Environmental Data Exchange for Inland Water (EDEN_IW) project, for instance, has developed an ontology for the surface water domain (Stjernholm et al., 2007) in English, French, Danish, and Italian. FAO is restructuring the multilingual thesaurus AGROVOC from a term-based system to a concept-based system as a step towards creating an Agricultural Ontology Service (AOS). Ontological concepts can be extracted from the AOS and used to build domain specific ontologies (Lauser et al., 2002), such as the FAO fisheries ontology (Gangemi et al., 2002).

7. Summary

Development of an ILTER Network metadata catalog faces obstacles resulting from natural language differences between member countries. The currently agreed upon approach for ILTER members, which is to provide a subset of EML in English to the ILTER metadata cache, is likely to be difficult for countries with few information management resources to support translation. This method relies on expertise of the translator for the semantic accuracy of the translation. A more standardized approach to translation should lessen the burden on individual networks and also improve data discoverability. The development of a multilingual environmental thesaurus, from which selected terms in multiple languages can be obtained, is a priority for the network. In the long-term, the ILTER will capitalize on the development of ontologies in many languages to provide the means to explicitly disambiguate translated terms and thereby improve the discovery of comparable data documented in different languages. Many other organizations around the world are also investigating methods for building multilingual thesauri and ontologies, and ILTER will seek partnerships with them to advance the development of multilingual tools for the whole ecological community.

Acknowledgements

Ideas for this paper were generated during an ILTER Informatics Workshop hosted by the Chinese Ecological Research Network (CERN) at Lake Taihu Field Station, China. Workshop participants received travel support from the ILTER Network in their own country or from GBIF. The comments of two anonymous reviewers improved the manuscript.

References

Annoni, A., Friis-Christensen, A., Lucchi, R., Lutz, M., 2008. Requirements and challenges for building a European spatial information infrastructure: INSPIRE. In: van Oosterom, P., Zlatanova, S. (Eds.), Creating Spatial Information Infrastructures: Towards the Spatial Semantic Web. CRC Press, New York, pp. 1–18.

Berkley, C., Jones, M.B., Bojilova, J., Higgins, D., 2001. Metacat: a schema independent XML database system. Proceedings of the 13th International Conference on Scientific and Statistical Database Management. IEEE Computer Society.

Biodiversity Center of Japan, Ministry of Environment. 2010a. 500 Important wetlands in Japan (in Japanese). http://www.sizenken.biodic.go.jp/wetland/, Accessed January 29, 2010.

Biodiversity Center of Japan, Ministry of Environment. 2010b. 500 Important wetlands in Japan (in English). http://www.sizenken.biodic.go.jp/pc/wet_en/, Accessed January 29, 2010.

CCOP and CIFEG, 2006. Asian multilingual thesaurus of geosciences. http://www.ccop.or.th/download/pub/AMTG_2006.pdf, accessed March 15, 2009.

Doerr, M., 2001. Semantic problems of thesaurus mapping. Journal of Digital Information 1 http://journals.tdl.org/jodi, Accessed November 23, 2008.

Doerr, M., Hunter, J., Lagoze, C., 2003. Toward a core ontology for information integration. Journal of Digital Information 4 http://journals.tdl.org/jodi/, Accessed April 13, 2009.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86, 158–168.

Gangemi, A., Fisseha, F., Pettman, I., Keizer, J., 2002. Building an integrated formal ontology for semantic interoperability in the fishery domain. Agricultural Information and Knowledge Management Papers (FAO). FAO, Rome. ftp://ftp.fao.org/docrep/fao/008/af242e/af242e00.pdf, Accessed March 11, 2009.

Higgins, D., Berkley, C., Jones, M.B., 2002. Managing heterogeneous ecological data using Morpho. Proc. of the 14th Intl. Conf. on Scientific and Statistical Database Management.

ISO 5964:1985. Documentation — guidelines for the establishment and development of multilingual thesauri. See: http://www.iso.org/iso/home.htm.

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

Lauser, B., Wildemann, T., Poulos, A., Fisseha, F., Keizer, J., Katz, S., 2002. A comprehensive framework for building multilingual domain ontologies: creating a prototype biosecurity ontology. Proceedings of International Conference on Dublin Core and Metadata for e-Communities 2002. Firenze University Press, pp. 113–123. http://www.bncf.net/dc2002/program/ft/paper13.pdf.

Lin, C.C., Porter, J.H., Lu, S.S., 2006. A metadata-based framework for multilingual ecological information management. Taiwan Journal of Forest Science 21, 377–382.

Lin, C.C., Porter, J.H., Lu, S.S., Jeng, M.R., Hsiao, C.W., 2008. Using structured metadata to manage forestry research information: a new approach. Taiwan Journal of Forest Science 23, 133–143.

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23, 159–168.

Pettman, I., 2006. Knowledge retrieval in aquatic ecology and fisheries—do we need (and can we afford) ontologies? In: Anderson, K.L., Thiery, C. (Eds.), Information for Responsible Fisheries: Libraries as Mediators: Proceedings of the 31st Annual Conference: Rome, Italy, October 10–14, 2005, pp. 25–38.

Scott, S., 2004. A Journey into Chinese–English environmental translation. Translation Journal 8 http://accurapid.com/journal/27environ.htm, Accessed November 25, 2008.

Sievers, J., Zaccheddu, P.-G., 2005. EuroGeoNames: The Vision of Integrated Geographical Names Data within a European SDI. Eighth United Nations Regional Cartographic Conference for the Americas: New York, June 27–July 1, 2005, Accessed January 29, 2010.

Stjernholm, M., Poslad, S., Zuo, L., Sortkjaer, O., Huang, X., 2007. An ontology-based approach for enhancing inland water information retrieval from heterogeneous databases. In: Haastrup, P., Wurtz, J. (Eds.), Environmental Data Exchange Network for Inland Water. Elsevier, New York, pp. 123–144.

United States Environmental Protection Agency. 2010. Wetland types. http://www.epa.gov/owow/wetlands/types/, accessed January 29, 2010.

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., Karasti, H., Kertesz, M., Kuitunen, P., Lane, M., Lieskovsky, J., Magagna, B., Peterseil, J., Rennie, S., Schentz, H., Schleidt, K., Tuominen, L., 2008. SERONTO, a socio-ecological research and observation oNTOlogy. In: Weitzman, A.L., Belbin, L. (Eds.), Proceedings of TDWG 2008, October 17–25, 2008, Fremantle, Australia, p. 24.

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Li, J., Lin, C.-C., Lu, S.-S., Ko, B., Ogawa, A., O'Tuama, E., Schentz, H., Wen, S., van der Werf, B., 2008. Building an information management system for global data sharing: a strategy for the International Long Term Ecological Research (ILTER) Network. In: Gries, C., Jones, M.B. (Eds.), Proceedings of the Environmental Information Management Conference 2008, September 10–11, 2008. University of New Mexico, Albuquerque, NM, pp. 156–165.