Towards a seamless multilingual Semantic Web: A study on constructing a cross-lingual ontology

Yi-Yun Cheng, Hsueh-Hua Chen
Department of Library and Information Science
National Taiwan University
1, Sec. 4, Roosevelt Rd., Taipei 10617, TAIWAN
r02126002@ntu.edu.tw; sherry@ntu.edu.tw

ABSTRACT

Cross-lingual ontology research has become a pivotal concern in the global age. Researchers worldwide try to be interoperable with ontologies written not only in English, but also in other languages. Yet, constructing a cross-lingual ontology can be difficult, and a detailed mapping method is often hard to find. This study investigates the practice for constructing a cross-lingual ontology, in the case of the Semantic Sensor Network (SSN) ontology and the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies. By adopting a three-phase research design, a cross-lingual ontology method between English and Traditional Chinese is constructed through the implementation of Protégé. The mapping results between the two languages reveal an accuracy of 80.66% on the exact-match terms, while the Chinese synonyms and related terms expressed by SKOS labels are all proven searchable in our primary evaluation. These promising results demonstrate the feasibility of the methodology proposed by this study, and further suggest that such approach is suitable to be adopted by future researchers to model their cross-lingual ontologies.

Keywords

Ontologies, cross-lingual ontology, SWEET ontology

INTRODUCTION

The emergence of big data has brought about a pressing need to build ontologies for automatic reasoning and processing. Ontologies, as the fundamental building blocks for the Semantic Web, are the highest level classification scheme in the family of knowledge organization systems (KOS) (Zeng, 2008). Not only can ontologies help express complex relationships and meanings between objects, they are also machine-readable (Berners-Lee, Hendler, & Lassila, 2001; Gruber, 1993; Noy, & McGuinness, 2001). However, nowadays, the development of ontologies is no longer limited to English. Ontologies of other natural languages have also gained considerable significance and started to be developed. To avoid building "islands of monolingual ontologies" (Gracia et al., 2012), a pool of cross-lingual ontologies mapping research has begun to thrive.

The purpose of this study is to establish a feasible practice on building cross-lingual ontologies. The study will focus on the construction of an English-Chinese ontology from an existing source ontology and a KOS source. This study will also address the synonymy and polysemy problems of the target language (Traditional Chinese).

RELATED WORK

As there are no de facto protocols for building a cross-lingual ontology, various methods have been proposed by former researchers. Though each method has its distinct approach, none seems to be solely successful (Kempf, Ritze, Eckert, & Zapilko, 2014; Shvaiko & Euzenat, 2013). For example, Fu, Brennan, & O'Sullivan (2009) and Liang et al. (2005) tried to map between the English and Chinese ontologies. The former proposed a "semantic-oriented" mapping framework by using machine translation tools, whilst the latter manually match terms between the English AGROVOC and the Chinese Agricultural Thesaurus. However, neither approach can be seen as adequate in expressing Chinese synonyms.

Accordingly, since constructing an ontology is a laborious task, the issue of "who" should be the builder often concerns researchers. Some studies take a manual processing approach, such as the abovementioned Liang et al.'s (2005) study and Albertoni, De Martino, Di Franco, De Santis, & Plini 's (2014) mapping between EARTh thesaurus and other ontologies. Alternatively, other studies use automatic matching systems like the BOAT matcher in Chua & Kim (2012), and let the machine do all the work. Still others are hoping to develop some semi-automatic tools that would accelerate the multilingual ontology building process such as Pazienza & Stellato (2005)'s OntoLing tool. Even though automatic or semi-automatic tools can effectively reduce the input of human labor, the final ontology of the studies mentioned above are usually insufficient in discerning the relationships between concepts (or classes) and oftentimes require the aid of manual input.

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

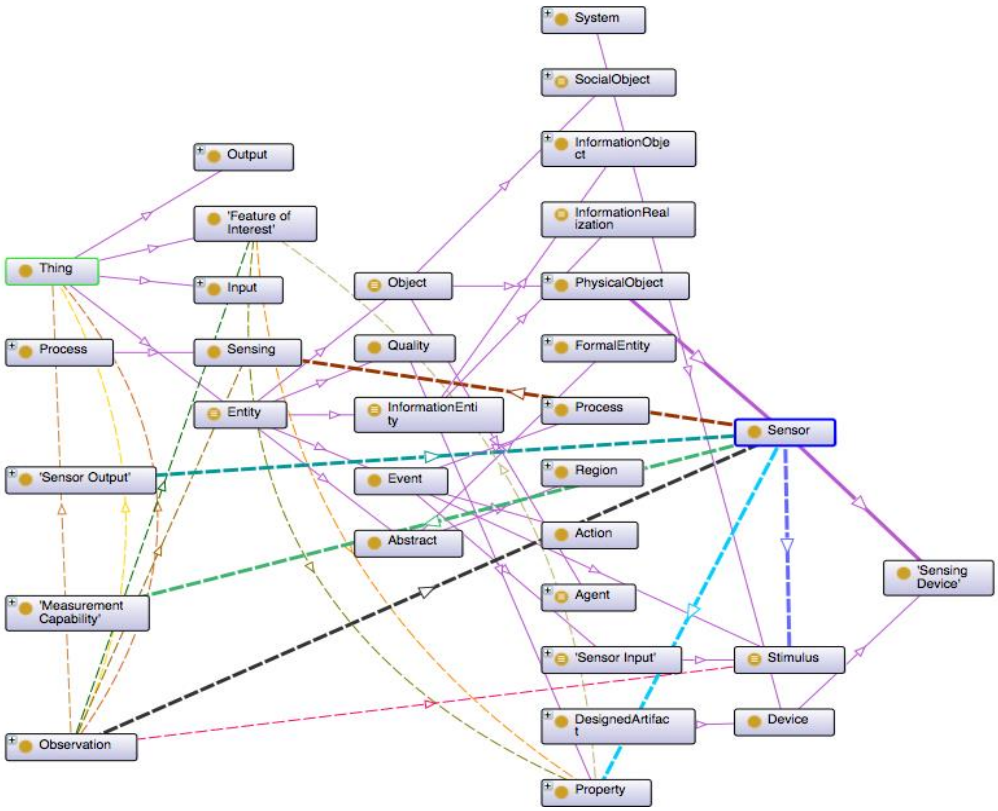

Figure 1. Visualization of SSN ontology

Through an extensive literature review, we organized the ontology construction process into three approaches. These three approaches include:

(1) Design from scratch—no existing ontologies to work with, researchers have to brainstorm all the classes and relationships. This often happens in designing task or domain ontologies such as the case in Pattuelli (2011).

(2) Knowledge organization system based (KOS-based)—use KOS (Hodge, 2000; Zeng, 2008) such as term lists like the dictionaries, classification schemes like taxonomies, or relationship lists like the thesaurus as the base of the ontological structure, borrowing the terms or relationships in KOS to form an ontology. Example of a KOS-based research can be found in Qin & Paling (2001).

(3) Use existing ontologies—this includes extending the classes and relationships in existing ontologies, or mapping between ontologies (monolingual and cross-lingual alike). Such studies can be found in Shvaiko, P.,& Euzenat, J.(2013).

Though some of the work discussed above presented workable approaches and tools, almost all of them are built upon two or more well-established, already-existing ontologies (approach 3). This study thus focuses on constructing a brand new multilingual ontology by mixing approach 2 and approach 3 by using one KOS structure and one existing ontology. Also, we try to find ways to more effectively express Traditional Chinese synonyms and polysemy, while at the same time takes a semi-automatic approach in building the ontology.

RESEARCH DESIGN

To explore the practice for building cross-lingual ontologies, we employ a three-phase research design detailed as followed:

Phase 1 is the pretest of our mapping practice on a small ontology—W3C's Semantic Sensor Network Ontology (SSN ontology).The purpose of this pretest is to ensure our mapping process is feasible. We first parse all the classes in SSN ontology by writing a SPARQL code. Then we input all the classes into spreadsheet form to map with the Chinese domain term lists provided by National Academy of Educational Research (NAER) of Taiwan.

Phase 2 is our formal mapping process. After making sure that our mapping method would work, we start building an English-Chinese ontology model by using Microsoft Access to automatically map between the large geospatial ontology—SWEET ontology—and the Chinese NAER domain term lists.

Phase 3 is the implementation of this cross-lingual ontology in Protégé, a popular open source ontology editing software, to produce an OWL ontology file with Simple Knowledge Organization System (SKOS) properties expressing the Traditional Chinese synonyms and related terms.

RESULTS

Phase 1 Pretest

The SSN ontology is a small ontology consists of 117 classes, written in English (see Figure 1). The SPARQL code we employed to parse all the classes in SSN ontology are in the following:

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?class
Where { ?class a  owl:Class .
}
```

This helped us automatically retrieve all 117 classes in plain text format in a click, so we can input them into Excel spreadsheet. By employing the classes into spreadsheet format, we then can easily map all the English SSN ontology classes with the Chinese NAER domain term lists (also in Excel spreadsheet format) in Microsoft Access. Our pretest results showed that our mapping process in Microsoft Access successfully mapped 82 of the English classes with corresponding Chinese terms (the mapping process will be explained in detail later in Phase 2). This gave us confidence in our mapping practice so we can move on to map a larger ontology—the SWEET ontologies.

Phase 2 Formal Mapping Process

According to the results shown in Phase 1, our mapping process was feasible. We used a mixed approach which combines KOS-based (approach 2) and the use of existing ontologies (approach 3). This allows us to map between the English SWEET ontologies and the Chinese NAER term lists. Figure 2 is the model we proposed for this study.

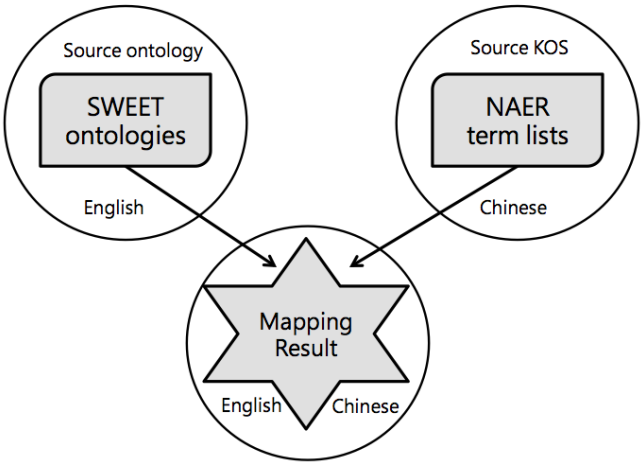

Figure 2. Mapping Model

The mapping steps are detailed below:

Step 1: Decide the ontologies to use and Parse all the classes.
In our study, our source ontology is the SWEET ontologies maintained by NASA and Caltech. The scope of SWEET ontologies covers a broad range of topics, including environment, geology, biology, etc. There are 3,770 ontology classes in total in SWEET ontologies, and we use the SPARQL code described in Phase 1 to automatically parse all the classes in SWEET ontology.

Step 2: Decide the corresponding KOS sources.
We use the term-lists provided by the NAER, a credible institution in Taiwan, as our source KOS. To match the classes in SWEET ontologies, we chose 17 domain term lists from NAER. The domains include earth science, geography, geology, atmospherics, meteorology, chemistry, agriculture,

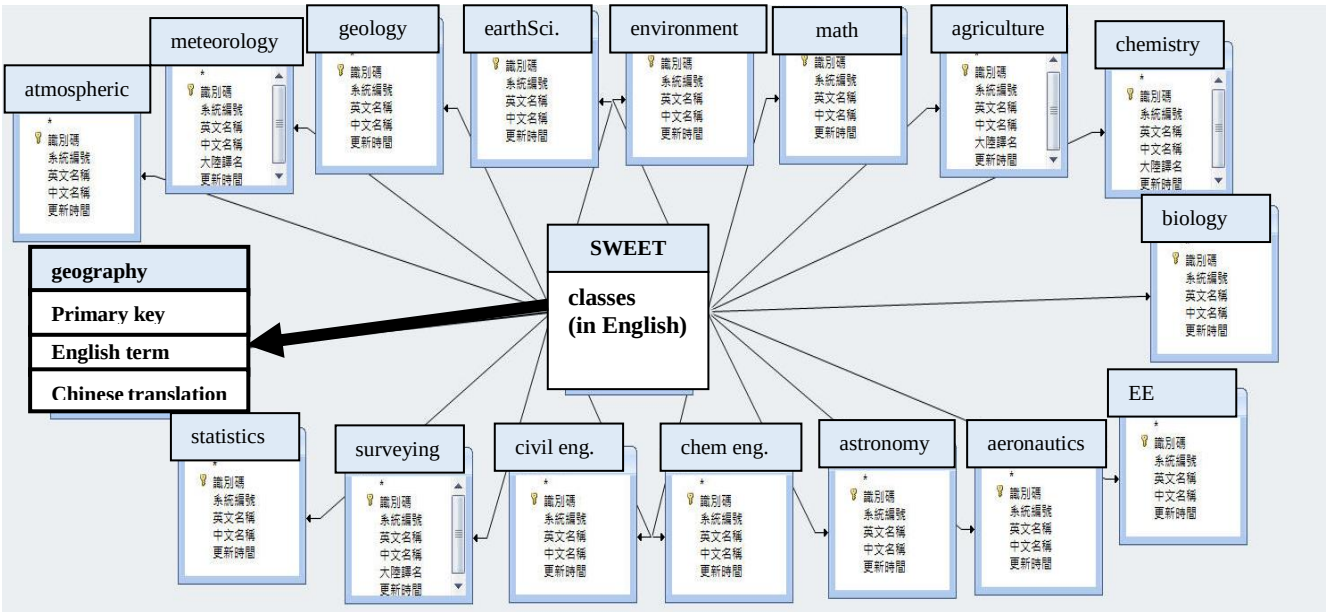

Figure 3. Mapping Process in Microsoft Access

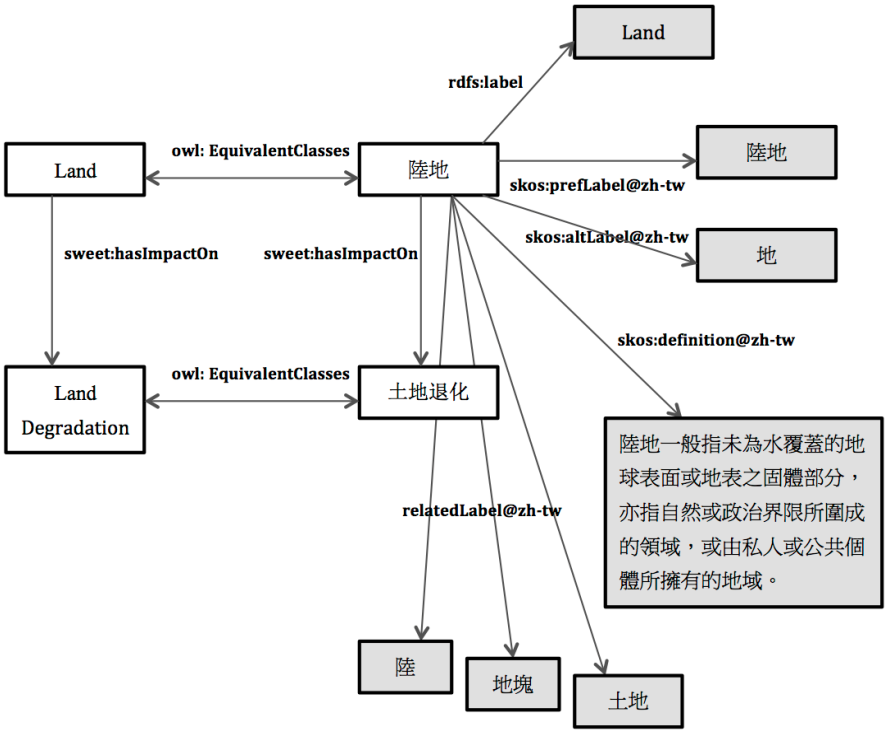

Figure 4. Example of the class 'Land' and all its corresponding Chinese classes and properties

math, statistics, etc, which consist a total of 231,264 entries (not mutually exclusive).

Step 3: Generate a cross-lingual mapping result.
Input the SWEET ontologies and NAER term-lists, both in excel spreadsheet files, into Microsoft Access (see Figure 3.). By connecting the SWEET classes (which is in English) with the NAER English terms, Access will match the exact same English terms and then automatically generated a mapping result that shows all the SWEET ontology English classes with corresponding Chinese translation in spreadsheet format. The mapping results showed that 3,041 out of 3,770 SWEET terms can be successfully mapped with exact-match Chinese terms. In other words, the mapping process demonstrated a final accuracy of 80.66%. The remaining 729 SWEET classes (19.34%) that have not been mapped are either partial-match terms, unique SWEET compound nouns, or terms that are more general so that NAER did not include it.

Phase 3 Constructing the ontology in Protégé
In phase 3, we manually input the generated mapping results in phase 2 into Protégé (ontology editing tool) to produce an OWL file. Our constructing steps are as followed:

Step 1: Choose one topic in the source ontology.
Although the total number of SWEET ontologies classes is 3,770, these classes all come from 200 distinct small ontologies in SWEET. We started by choosing one ontology in SWEET to work with—the environmental impact ontology, which consists of 48 classes. When we input the EnvirImpact.owl file into Protégé, the Protégé interface showed all the English classes of this file.

Step 2: Add Chinese classes into Protégé.
To create the same ontological structure in Chinese as the English Environmental Impact ontology, we then manually added Chinese counterpart classes into the EnvirImpact.owl on Protégé. We did not directly substitute the English classes with Chinese ones. We kept the classes in both languages for the purpose that we did not want to disarray the original EnvirImpact.owl file and that we want to make sure our Chinese classes can be easily interoperable with the English ones. Also, for future research purpose, we hope to work on ontology localization tasks, such as adding local Chinese concepts into our ontology while our ontology is connected to NASA's SWEET ontologies.

Our mapping results gave us more than one Chinese translation (across the 17 domains) for a single SWEET term. In deciding which Chinese terms should be shown as the main classes in Protégé, two rules are adopted: (1) follows the definition from SWEET ontology to locate the closest Chinese term; (2) if no definition for the SWEET terms is available, the Chinese terms used by the majority of the domains will be selected.

Step 3: Connect the equivalent classes.
Connect the English classes with the Chinese classes by using owl:EquivalentClasses properties.

Step 4: Add SKOS annotation properties.
To express Chinese synonyms and other related terms generated in our mapping results, we added SKOS annotation properties to the Chinese classes. Figure 4 shows an example of all the properties of the English class 'Land' and its Chinese main class.

- skos:prefLabel—the main Chinese classes that we decided in step 2 will also be labeled as preferred labels.
- skos:altLabel—other Chinese synonyms that contain almost the exact same meanings to the Chinese main classes will be put in alternative labels.
- relatedLabel—this is a property we created by ourselves to express the other related Chinese terms which are close in meaning to the Chinese main classes but maintain certain difference.
- skos:definition—the definition of the Chinese classes found on the NAER website.

Step 5: Search the classes and properties in Protégé.
To evaluate whether the English and Chinese classes, synonyms, related terms are searchable, we used the Protégé plug-in OntoGraf and Search Annotations bar. Our primary results demonstrate that the Chinese classes, prefLabels, altLabels, and relatedLabels can all be retrieved.

Figure 5. is the visualization of our final SWEET English and Traditional Chinese cross-lingual ontology.

CONCLUSION

This is an exploratory study for the practice in constructing a cross-lingual ontology, our mapping results proved that our method is feasible, and that the classes and labels in both language in the ontology are all searchable. One of the implications we foresee in this study is that most language cultures might not have existing ontologies ready to use, but they will at least have a KOS term list such as dictionaries or

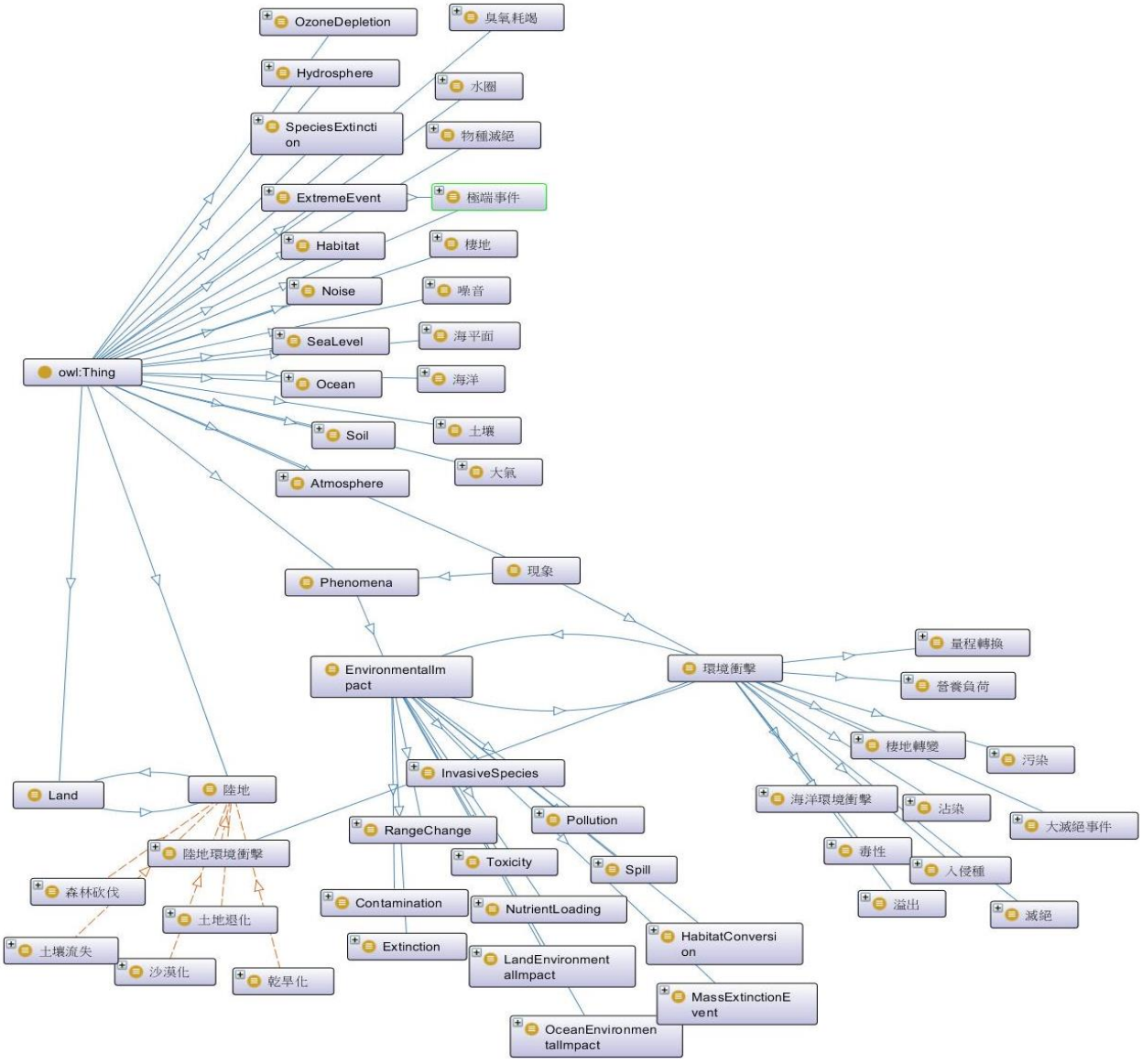

Figure 5. Visualization of the English-Chinese SWEET ontology

glossaries. By using our model, a source KOS in whichever language mapping with an existing source ontology in English (or other languages), it is possible that many language cultures may construct their own ontologies and at the same time being interoperable with other languages.

Nevertheless, the mapping results between the two languages in this study still require further improvement. In this respect, future research should take into consideration the opinions of domain experts during the mapping of the partial match terms and terms that were uniquely created by the source ontology.

The building of cross-lingual ontologies bears strong reminiscence of the Babel Tower story. If we want to break through the language barriers, challenges such as choosing the building approaches, figuring out the mapping processes, or dealing with synonyms are almost inevitable. But in the end, when all obstacles are overcome, we might actually realize the dream of a multilingual Semantic Web in which knowledge sharing among languages is seamless, and that the tower of ontologies is built safe and sound.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Science and Technology (grant no. MOST103-2627-M-002-004) of Taiwan. We thank Professor Chih-Hong Sun, Dr. Ping-Ying Tsai, and Dr. Rong-Kang Shang for their helpful suggestions. Parts of this work were conducted using the Protégé resource, which is supported by grant GM10331601 from the national Institute of General Medical Sciences of the United States National Institutes of Health. We also pay a special courtesy to NASA/JPL-Caltech's SWEET ontologies.

REFERENCES

Albertoni, R., De Martino, M., Di Franco, S., De Santis, V., & Plini, P. (2014). EARTh: an environmental application reference thesaurus in the Linked Open Data cloud. *Semantic Web*, *5*(2), 165–171.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 28-37.

Chua, W., & Kim, J. (2012). BOAT: Automatic alignment of biomedical ontologies using term informativeness and candidate selection. *Journal of Biomedical Informatics*, *45*(2), 337–349. doi:http://doi.org/10.1016/j.jbi.2011.11.010

Fu, B., Brennan, R., & O'Sullivan, D. (2009). Cross-lingual ontology mapping–an investigation of the impact of machine translation. *The Semantic Web* , 1-15. Retrieved from: http://link.springer.com/chapter/10.1007/978-3-642-10871-6_1

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *11*, 63–71. doi: http://doi.org/10.1016/j.websem.2011.09.001

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199-220.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: beyond traditional authority files.* Washington, D. C.: The Digital Library Federation. Retrieved from: http://www.clir.org/pubs/reports/pub91/pub91.pdf

Kempf, A., Ritze, D., Eckert, K., & Zapilko, B. (2014). New ways of mapping knowledge organization systems: using a semi-automatic matching procedure for building up vocabulary crosswalks. *Knowledge Organization, 41*(1), 66-75.

Liang, A., Sini, M., Chun, C., Sijing, L., Wenlin, L., Chunpei, H., & Keizer, J. (2005). *The mapping schema from Chinese Agricultural Thesaurus to AGROVOC.* Retrieved from: http://eprints.rclis.org/handle/10760/15692

Noy, N., & McGuinness, D. (2001). *Ontology development 101: a guide to creating your first ontology.* Retrieved from: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Pattuelli, M. C. (2011). Modeling a domain ontology for cultural heritage resources: A user-centered approach. Journal of the American Society for Information Science and Technology, 62(2), 314-342.

Pazienza, M. T., & Stellato, A. (2005). Linguistically motivated Ontology Mapping for the Semantic Web. In *SWAP*. Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-166/26.pdf

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: the case of GEM. *Information Research: an International Electronic Journal*, *6*(2).

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 158–176. doi: http://doi.org/10.1109/TKDE.2011.253

Zeng, M. (2008). Knowledge organization systems. *Knowledge Organization*, *35*(2). Retrieved from: http://nkos.slis.kent.edu/KOS_taxonomy.htm