Experiences with Multilingual Modeling in the Development of the International Classification of Traditional Medicine Ontology

Csongor Nyulas, Tania Tudorache, Samson Tu, Mark A. Musen

Stanford Center for Biomedical Informatics Research, Stanford University, US
{nyulas, tudorache, swt, musen}@stanford.edu

Abstract. The World Health Organization (WHO) in collaboration with several international stakeholders have started recently the work on the International Classification of Traditional Medicine (ICTM), which will provide a standardized system for encoding and collecting health statistics data related to Traditional Medicine practice throughout the world. ICTM is represented in OWL, and is developed by Traditional Medicine experts in a collaborative Semantic Web platform, called iCAT-TM. The content of ICTM is developed simultaneously in four languages (English, Chinese, Japanese and Korean). In this paper, we describe how we modeled the multilingual content, the Web platform used for editing, and some of the challenges we have encountered related to the multilingual aspects of the model and use of the platform.

1 The International Classification of Traditional Medicine (ICTM)

The World Health Organization (WHO) in collaboration with a large group of international stakeholders is developing the International Classification of Traditional Medicine (ICTM).[1] ICTM will provide a standardized international system for classifying Traditional Medicine (TM) related health concepts, such as disorder names, disease patterns, signs and symptoms, causal factors, and interventions [9]. One of the goals of the project is to be able to unify the data collection and monitoring for Traditional Medicine systems with those of the conventional (i.e., "Western") medicine, which will be realized by integrating a relevant part of ICTM as Chapter 23 of the 11th revision of the International Classification of Diseases (ICD).[2] ICD is an essential classification used in the United Nation countries for compiling basic health statistics, billing, and clinical documentation [8].

The content of ICTM is based on classifications of Traditional Medicine from three countries, China, Japan and Korea. Even if these classifications have a common root, they have diverged significantly over the years. The role of ICTM is to harmonize these different efforts and come to a consensus classification that can be used in health systems around the world.

With the information age revolution, WHO has changed significantly the way they build classifications. To make them ready for electronic health records and enable easy

[1] https://sites.google.com/site/whoictm/

[2] http://www.who.int/classifications/icd11/browse/f/en

cross-linking between them, the classifications have now a formal underpinning. ICTM, similarly to ICD-11, is represented as an OWL ontology and is developed using Semantic Web technologies.

Given the international nature of ICTM, tackling the multilinguality problem is one of the main challenges in the project. Domain experts from the three countries and the project coordinators in Geneva, Switzerland, are developing the content of ICTM simultaneously in four languages: English, Chinese, Japanese, and Korean. Our group has provided the ontology modeling support and the Web platform infrastructure used for editing ICTM. In this paper, we describe our experiences in supporting multiple languages in the ICTM ontology, including the model and tooling, and the challenges we encountered.

The rest of the paper is organized as follows: Section 2 describes the related work, in Section 3, we describe how we modeled the multilingual content in ICTM, Section 4 presents the collaborative Semantic Web platform used by the domain experts to edit ICTM, and finally, Section 5 presents the challenges and some lessons learned in the project, and gives an overview of the future work.

2 Related Work

As the Semantic Web matures, there is an increasing body of research on localizing ontologies. For example, the SKOS-XL extension [1] treats labels as first order resources, thus enabling the definition of explicit links between labels associated to the same concept. Montiel-Ponsoda *et al.* [4] try to overcome some of the limitations of the SKOS-XL representation and propose a module for lemon [3] that supports different types of translation relations and metadata, such as provenance and reliability scores. Extensive work on ontology localization [2] has also been done in the NEON project[3] that proposes guidelines and a tool to support this process.

Silva *et al.* present conceptME [5], a collaboration framework that supports ontology localization starting early in the conceptualization phase. Providing terminological support so early in the development process proved to enhance the conceptualization of the domain. conceptME has also support for sharing conceptual models, for content negotiation and discussion.

In this work we did not use any of the related approaches, as one of the main requirements in the project (see Section 3) was to use and/or extend the ICD-11 ontology to ensure that these two ontologies will be easily integrated at a later stage. We plan to investigate the related approaches (such as SKOS-XL and the extensions to lemon) to see if they would fit the requirements for ICTM, and if so, we will refactor our ontology accordingly.

3 Multilingual Modeling in ICTM

As we mentioned before, one of the main requirements for ICTM is that it should follow similar modeling patterns to the ICD-11 ontology [6], so that these two can be easily integrated.[4] In addition, all ICTM textual content should be available in four languages: English, Chinese, Japanese and Korean, which Traditional Medicine experts

[3] `http://www.neon-project.org`

[4] As we mentioned before, part of ICTM will be available as a separate chapter in ICD-11.

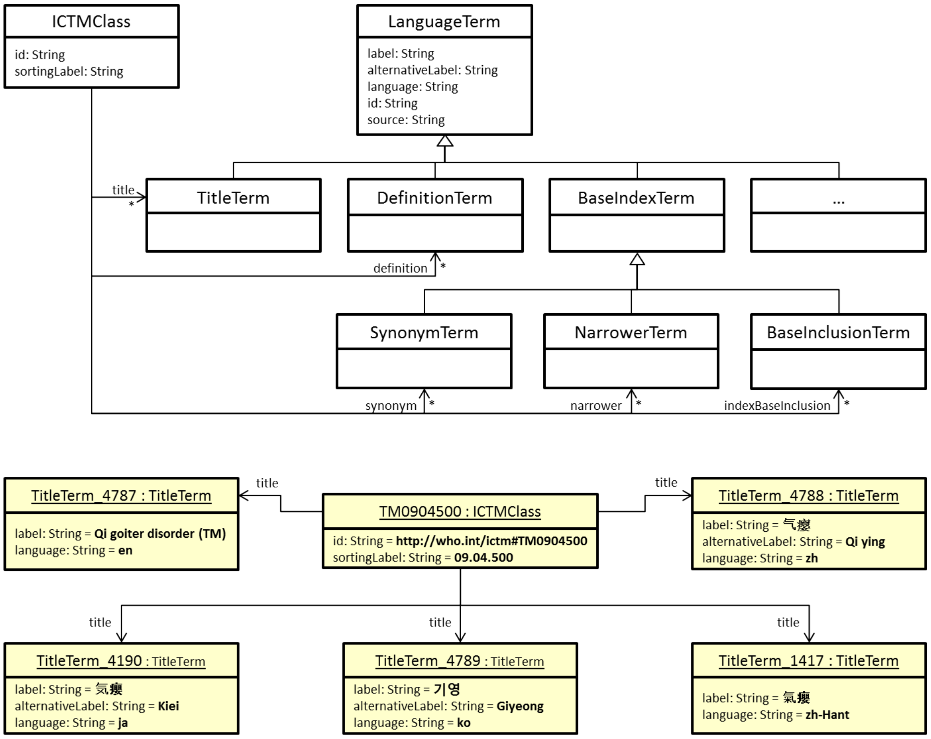


Fig. 1. Excerpt from the ICTM ontology. Language terms are modeled as instances of the reified class *LanguageTerm*. Subclasses of *LanguageTerm* represent different linguistic terms (title, definition, synonym, and so on). The subclasses may have additional properties that represent different metadata of the language term. The class level is shown in boxes with white background. We also show an example instantiation for the title terms for the *Qi goiter disorder* disease class using the boxes with darker background.

from different countries will input during development. A further requirement, which came later in the project, was to support transliteration of titles, i.e., converting the Chinese, Japanese and Korean scripts into Latin script. For example, a common transliteration for converting Chinese characters into Latin script is Pinyin. Figures 1 and 2 show the transliteration of the simplified Chinese disease title 气瘿 (meaning, *Qi goiter disorder*) into Pinyin as *Qi ying*. Other metadata will be attached to the label of a term into a specific language, such as the source of the label (e.g., the Traditional Medicine classification where the label originates from), and an internal id that is used by other WHO software.

We modeled ICTM in OWL 1.0. We used a reified class, *LanguageTerm*, to represent all linguistic terms in the ontology. We have created a taxonomy of language terms as subclasses of the *LanguageTerm* class, as some of the term types have additional properties attached to them. For example, the *SynonymTerm* has additional properties that describe if and how it will be included in an electronic index for the classification.

In the current version of the model, there are eight subclasses of *LanguageTerm* that represent among others, the title, the fully specified title, the definition, other external definitions, the synonyms, and so on. The actual value for a language term is an instance of a subclass of *LanguageTerm*.

Figure 1 shows an excerpt of the class level modeling for language terms and how different properties of a disease class (title, definition, synonym, etc.) have been reified. The figure also shows an example for modeling the five title terms for the *Qi goiter disorder* disease class. The *Qi goiter disorder* class has a property *title* that has as values five instances of the class *TitleTerm* that correspond to the titles in 5 languages (English, Japanese, Korean, simplified Chinese and traditional Chinese). Some *TitleTerm* instances (e.g., *TitleTerm_4788* for simplified Chinese) has in addition to the *label* and *language* properties, also another property, *alternativeLabel*, to represent the transliteration of the Chinese script to Latin characters. Other language terms (not shown in the figure), such as the *ExternalDefinitionTerm*—used to reference textual definitions from external resources — have additional properties that specify the source of the definition in greater detail (e.g., the ontology name, the IRI of the source ontology entity, the URL for the source ontology, etc.).

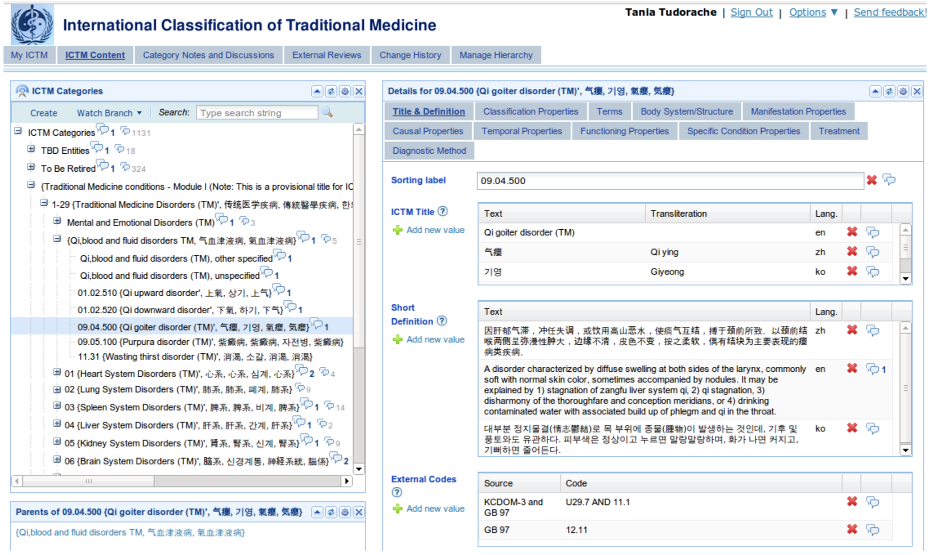
As there might be confusion about the difference between synonyms and transliterations, we would like to clarify this issue. A synonym is a term that has a similar meaning to a another term (in our case, the title term). In ICTM, as in ICD-11, synonyms are also used to store alternative titles for a disease, that are either found in scientific literature, or have minor linguistic variations, or are used in the colloquial language (e.g, the synonym for *Roseola infantum* is *Sixth disease*). The synonyms apply for terms in the same language (e.g., an English title may have other English synonyms). A transliteration, on the other hand, represents exactly the same term in the same language, but in a different script. A term may have several transliterations (Korean has 4 different transliterations).

4 The iCAT-TM Platform

Traditional Medicine experts around the world are editing ICTM using the collaborative iCAT-TM Web platform. iCAT-TM is a customization of the generic WebProtégé ontology editor [7]. The user interface of iCAT-TM is tailored for domain experts, who are not knowledgeable about ontologies or knowledge representation. iCAT-TM presents a form-based interface shown in Figure 2 that is is less intimidating for the experts than a generic ontology editor would be. The experts can edit the class taxonomy in the left panel of the *ICTM Content Tab*, and the class details, including the language terms, in the right panel.

iCAT-TM has many collaboration features inherited from WebProtégé, such as the support for simultaneous editing, change history of users' actions, and notes and discussions attached to any entity in the ontology.

We have created a generic widget that displays the content of reified individuals, and we have reused it for displaying and editing the different language terms. In Figure 2, the *ICTM Title* uses this widget to display the values of the *TitleTerm* individuals associated to the *title* property of a disease. A row in the widget table corresponds to one of the reified individual values, and the columns display the properties of the respective row individual value. The same widget is also used for displaying the short definition of a disease (the *transliteration* column has been hidden from view).


International Classification of Traditional Medicine
Tania Tudorache | Sign Out | Options ▼ | Send feedback!
My ICTM
ICTM Content
Category Notes and Discussions
External Reviews
Change History
Manage Hierarchy
ICTM Categories
Create
Watch Branch ▾
Search:
Type search string
Details for 09.04.500 {Qi goiter disorder (TM)', 气瘿, 기영, 氣癭, 気癭}
Title & Definition
Classification Properties
Terms
Body System/Structure
Manifestation Properties
Causal Properties
Temporal Properties
Functioning Properties
Specific Condition Properties
Treatment
Diagnostic Method
Add new value
Parents of 09.04.500 {Qi goiter disorder (TM)', 气瘿, 기영, 氣癭, 気癭}

- ICTM Categories 1 1131
  - TBD Entities 1 18
  - To Be Retired 1 324
  - {Traditional Medicine conditions - Module I (Note: This is a provisional title for IC
    - 1-29 {Traditional Medicine Disorders (TM)', 传统医学疾病, 傳統醫學疾病, 한
      - Mental and Emotional Disorders (TM) 1 3
      - {Qi,blood and fluid disorders TM, 气血津液病, 氣血津液病} 1 5
        - Qi,blood and fluid disorders (TM), other specified 1
        - Qi,blood and fluid disorders (TM), unspecified 1
        - 01.02.510 {Qi upward disorder', 上氣, 상기, 上气} 1
        - 01.02.520 {Qi downward disorder', 下氣, 하기, 下气} 1
        - 09.04.500 {Qi goiter disorder (TM)', 气瘿, 기영, 氣癭, 気癭} 1
        - 09.05.100 {Purpura disorder (TM)', 紫癜病, 紫癜病, 자전병, 紫癜病}
        - 11.31 {Wasting thirst disorder (TM)', 消渴, 소갈, 消渴, 消渴}
      - 01 {Heart System Disorders (TM)', 心系, 心系, 심계, 心系} 2 4
      - 02 {Lung System Disorders (TM)', 肺系, 肺系, 폐계, 肺系} 9
      - 03 {Spleen System Disorders (TM)', 脾系, 脾系, 비계, 脾系} 1 14
      - 04 {Liver System Disorders (TM)', 肝系, 肝系, 간계, 肝系} 1 2
      - 05 {Kidney System Disorders (TM)', 肾系, 腎系, 신계, 腎系} 1 9
      - 06 {Brain System Disorders (TM)', 脑系, 신경계통, 神経系統, 腦係} 2

{Qi,blood and fluid disorders TM, 气血津液病, 氣血津液病}

Sorting label: 09.04.500

ICTM Title

Text	Transliteration	Lang.
Qi goiter disorder (TM)		en
气瘿	Qi ying	zh
기영	Giyeong	ko

Short Definition

Text	Lang.	
因肝郁气滞，冲任失调，或饮用高山恶水，使痰气互结，搏于颈前所致。以颈前结喉两侧呈弥漫性肿大，边缘不清，皮色不变，按之柔软，偶有结块为主要表现的瘿病类疾病。	zh	
A disorder characterized by diffuse swelling at both sides of the larynx, commonly soft with normal skin color, sometimes accompanied by nodules. It may be explained by 1) stagnation of zangfu liver system qi, 2) qi stagnation, 3) disharmony of the thoroughfare and conception meridians, or 4) drinking contaminated water with associated build up of phlegm and qi in the throat.	en	1
대부분 정지울결(情志鬱結)로 목 부위에 종물(腫物)이 발생하는 것인데, 기후 및 풍토와도 유관하다. 피부색은 정상이고 누르면 말랑말랑하며, 화가 나면 커지고, 기뻐하면 줄어든다.	ko	

External Codes

Source	Code
KCDOM-3 and GB 97	U29.7 AND 11.1
GB 97	12.11

Fig. 2. The iCAT-TM platform is used by domain experts from the three countries to develop ICTM collaboratively on the Web. The panel on the left hand side shows the class tree, and the right hand side panel shows the details of the selected class (in this case, the *Qi goiter disorder*). The language terms for the title and short definition are also shown, as well as the transliteration for the title.

One "bonus" of using reified individual for language terms (which, as a consequence, have identity) is that we can attach notes and discussion threads to a particular individual. For example, in Figure 2, the second short definition of the disease has a comment attached to it (shown as the number *1* next to the comment icon on the second row). This feature enables domain experts to have focused discussions right in the context in which they are editing. The contextual discussions are particularly useful because each disease has several properties that need to be filled, in many cases by different experts, and the overview and management of notes and discussions is much easier.

5 Discussions and Future Work

The iCAT-TM has been in production use since February 2011 by 25 Traditional Medicine experts. As a result, ICTM contains now more than 1,500 classes, 15,000 reified terms, out of which, 10,000 are language terms. The users have created more than 60,000 changes in the ontology, and added more than 1,100 notes and discussions.

Since the beginning of the project, we have encountered several challenges related to the multilingual aspects in the modeling, tooling and use of the platform.

Modeling. ICTM was developed using OWL 1.0 to make it compatible with the ICD-11 ontology. For this reason, we had to use reified relations to model the language terms. Reified relations, even though they have the advantages described earlier, have several disadvantages, as well. First, the reified individuals clutter the domain ontology, and increase its size significantly (in ICTM, almost all property values are reified). Second, these anonymous individuals are used in reasoning (as part of the domain ontology) and can slow it down significantly. We plan to overcome these limitations by upgrading the ontology to OWL 2.0 (ICD-11 will also upgrade), and rather than using reified individuals, we plan to use annotations on axioms. We plan to change the modeling in other aspects, too. For example, the transliterations are currently modeled as a multiple cardinality datatype property that take string literals as values. Even if we can now add more transliterations for the same label (e.g., Korean has four different transliterations), we cannot specify to which script or alphabet a transliteration belongs to. We plan to address this issue by using nested annotations on axioms in the OWL 2.0 modeling. Additionally, we plan to investigate if other approaches for ontology localizations, such as the ones we mentioned in the Related Work section, are suitable for ICTM. If these approaches fit the requirements, we will refactor the ontology to use a more standard approach. This undertaking will, however, require significant effort, as we need to also change the modeling of ICD-11, as well as migrate all existing content of two live production system (iCAT for ICD-11, and iCAT-TM for ICTM) to the new structure.

Tooling. We had to make sure that our tooling works well with international characters. While these are not an issue for the Web application per se (Web browsers can show pages in different encodings), we had to adjust our Lucene-based search mechanism to work properly with multiple languages. One hurdle for the domain experts in using iCAT-TM is that the user interface is presented in English, and many of them are not very comfortable with it. We plan to redesign the user interface to better follow the principles of internationalization, so that we can more easily provide language specific user interfaces. We do expect that this step will involve a significant re-design effort.

Use of the platform. We had several user related challenges that are not necessarily of technical nature. For example, when we started the project, we used (wrongly) the country codes to model the languages (*ch, jp*, and *kr*). Later, in the process, we changed the language codes to the correct ones from the ISO 639-1 (*en, zh, ja, ko*), however, some of the domain experts complained that the correct language codes are less intuitive to use. Also, when we started the project, we did not anticipate that some content will be entered in simplified Chinese, while other will be entered in traditional Chinese, which created some confusion with the users. As a solution, we added also the traditional Chinese language code (*zh-Hant*), so that at a later date the Chinese content can be easier curated and harmonized (it is expected that in the official distribution only simplified Chinese will be used). Another challenge is related to the communication among the domain experts, as most of them speak only their native language, and sometimes English, too. To improve the communication among the domain experts and the WHO coordinators in Geneva, we have introduced the transliteration. Another challenge related to the language barrier is that experts do not agree on the English translation for a term, and “invent” new English translations. This fact also makes the curation and verification of the entire classification content very challenging, because finding Traditional Medicine experts who understand all languages and can verify that the terms in different languages really mean the same thing, is very difficult.

As future work, we plan to upgrade ICTM to OWL 2.0 to overcome the modeling issues we described before. We will also create linkages between ICTM classes and ICD-11 classes that will put into correspondence Traditional Medicine disorders with "Western" diseases. As the project progresses, we will also provide a peer-reviewing mechanism, in which external domain experts will review different aspects of ICTM.

The iCAT-TM platform is currently in production use, and we expect that by 2015, when the ICD-11 major revision is planned to end, the ICD-11 Chapter 23, containing a part of ICTM, will be finalized as well. Even after 2015, ICTM will continue to be developed as an independent classification that will address the needs of the Traditional Medicine practices around the world.

Acknowledgments

We thank our WHO collaborators and the ICTM project members for developing the project requirements and for the fruitful collaboration. The work presented in this paper is partly supported by the NIGMS Grant 1R01GM086587.

References

1. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant. `http://www.w3.org/TR/skos-reference/skos-xl.html`. Last accessed: August, 2012.
2. M. Espinoza, E. Montiel-Ponsoda, and A. Gómez-Pérez. Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture*, pages 33–40. ACM, 2009.
3. J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Garcia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, et al. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 2011.
4. E. Montiel-Ponsoda, J. Gracia, G. Aguado-de Cea, and A. Gómez-Pérez. Representing translations on the semantic web. *MSW 2011*, page 25, 2011.
5. M. Silva, A. Soares, and R. Costa. Supporting collaboration in multilingual ontology specification: the conceptme approach. *TKE 2012*, page 27.
6. T. Tudorache, S. Falconer, C. Nyulas, N. Noy, and M. Musen. Will Semantic Web Technologies Work for the Development of ICD-11? In *The 9th Intl. Semantic Web Conference (ISWC 2010)*, pages 257–272. Springer, 2010.
7. T. Tudorache, C. Nyulas, N. Noy, and M. Musen. Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic Web Journal*, pages 1–11, 2012.
8. World Health Organization. International Classification of Diseases (ICD). `http://www.who.int/classifications/icd/`. Last accessed: August, 2012.
9. World Health Organization. Traditional Medicine in Health Information Systems: Integrating Traditional Medicine into the WHO Family of International Classifications. `https://sites.google.com/site/whoictm/home/ICTMProjectPlan.pdf`. Last accessed: August, 2012.