Protégé Tool and Development of Multilingual Ontology for Architectural Corpus, Design Process, Shortcomings

Elham Andaroodi[1], Frederic Andres[2], Kinji Ono[3], Pierre Lebigre [4]
[1]Ph.D. student,[2]Associate Professor,[3]Professor Emeritus, National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo, 101-8430, Japan
elham@grad.nii.ac.jp ,{ andres, ono}@nii.ac.jp
[4] Professor, Ecole d'Architecture Paris Val de Seine, 14, rue Bonaparte 75006 Paris
lebigre@club-internet.fr

1 Introduction

This presentation introduces an innovative process of ontology design specifically for capturing lexical semantic of entities, in multilingual ontology knowledge model for historic architectural corpora. The specification of a unified lexical model for covering semantic of multilingual terminology developed in protégé is presented as far as shortcoming of this tool for multilingual and remote collaborative support.

2 Design Process

For the selected domain of this research[1] -architectural relics of Silk Roads- Ontology knowledge model starts with completing terminology-set with lexical attributes designed by protégé tool [1] in RDFS[2]. As part of aims of the research -to cover the needs of multilingual end-users of Silk Roads- The term-sets are gathered in a collaborative work[3] in different languages in Protégé tool, starting from version 1.9 and updating now to version 3 built 107.

2.1 Multilingual Terminology with Lexical Attributes

Each tem-set as part of terminology of this ontology knowledge model is designed language dependant and for each language one separated class is defined as shown in Fig1. Lexical attributes of term-sets are gathered using dictionaries of different languages as far as thesauruses.

For input of lexical specifications in protégé this research has designed a lexical model. Initially lexical attributes are separated in 2 main groups. First one a set of attributes designed as slots which are language dependant and for each class -as domain in protégé RDFS interface- are created separately, such as the slots *word* and *word_plural* with value type string, and *hierarchical_structure* of the word with value type class with allowed super class of mono lingual thesaurus.

The second group is language independent set of attributes which are created as one slot shared between term-set classes. These slots are *morphology_of* with value type class and allowed super class morphology, *Phonology* (pronunciation) with value type string, and the slot *semantic of* with value type instance.

Completing lexical data input for the latest slot *semantic_of* was the main challenge of lexical model of entities in this ontology. This research has designed ID based language independent unified lexical model in order to give semantic to the term-set in different language. The architecture of this model is shown in graph1.

4 main lexical attributes of terms such as *description*, *synonym*, *etymology* and *reference* are created as slots to collect semantic information of each term, independent of the language. These slots are given proper value types in order to be filled out with a piece of single value data. For example definition of each word is given in slot *definition* with value type string, accompanied by slot *reference* with value type instance. This slot connects a set of slots with value type string with information about reference, such as *name*, *author*, *publisher*, *place*, *ISBN*, etc. Fig 2 shows how semantic attributes model of this research -designed in protégé- is gathered for multilingual term-set for one example of term (related to corpus of historical caravanserais).

2.2 Entity-set

The definition of ontology as part of knowledge model of this research focuses on entities and their relationships on the selected corpus. Here each entity of caravanserais knowledge model is a component

[1] The research is conducted in National Institute of Informatics (NII) in Japan in cooperation with the architecture school of Paris Val de Seine (EAPVS) in France under the "Digital Silk Roads Initiative Framework" (DSRIF) in cooperation with UNESCO, as part of a PhD. study of the SOKENDAI university which is supported by the scholarship.
[2] Resource Description Framework Schema
[3] The cooperation of the experts of UNESCO is appreciated here for extending the term-set to Russian, Azeri, Italian and Chinese and for checking Arabic and French terms.

represented by a lexeme. Multi lingual terminology set defined above provides a complete set of terms with related lexical attributes presenting a components of the selected corpus. For representing each entity this research defined an ID based approach in which IDentification numbers connect multilingual equivalents of each component, regardless of ambiguity of synonym in different languages[4]. Fig 3 shows a snapshot of entity-set graph in protégé tool using TGviz plug-in. This approach helps to define and implement each entity independent of a language as a bridge between others.

3 Shortcomings of Protégé Tool and Possible Solutions

In design and data input of the above described process of ontology development, protégé tool shows some shortcoming for our selected corpus in which we present some examples in this abstract. At first although protégé uses Unicode[5] characters to cover multilingual alphabets, but representations of characters vary in different platforms like windows XP with windows 2000 with specific font supports (As for Persian term-set). Besides, data input of some languages like Chinese is possible after modifications in the file run-protégé.bat for Language (to change to Chinese). Another important challenge ahead of this research is to input standard unified phonetics characters for pronunciation[6] of terms in phonology slot in which we can not process in protégé tool until this moment.

Besides RDFS language of protégé confronts with some shortcoming in design of semantic specification of term-set, mainly in define of inverse slot for synonyms. In this model synonyms are given in slots with value type string through several interconnecting slots with value type instance. Such kind of complicated interrelated slots can not be easily inverse of each other as inverse slot is can be defined in RDFS between 2 simple slots with values type instance directly belonging to 2 classes.

Finally completing multilingual terminology in this research is part of a collaborating research between networks of experts with different languages in the selected corpus of Silk Roads. Therefore the developed ontology model using protégé V3 beta built 107 is put on server for remote access and data input. The bug of tool for failing to save the remote data input is a considerable problem we confront. One solution might be database management server with ontology file updated in V 3.1 build 173.[7]

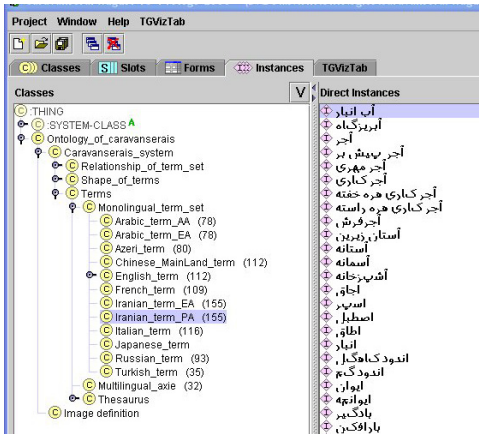
5 Conclusion and Future Trend

This abstract discussed about issues related to innovative design process and data input of multilingual lexical knowledge model as part of developing ontology for a subset of architectural heritage using protégé tool and the constraints caused by shortcoming of this knowledge representation tool. As the next step this research is dealing with design and defining of relationship between entities in order to cover the spatial attributes of architectural space [2]. For this purpose this research has moved from RDFS to OWL [3] which better supports define of rule for relationship links between instances as entities. As components in architectural corpora are not described only by lexemes but with shape, this research has planed to extend the knowledge model to cover attributes of shape using shape grammar for selected corpus of caravanserais.

References:

[1] Fridman Noy,N, Fergerson,R.W. and Musen,M.A., (2000), The knowledge model of protégé-2000: combining interoperability and flexibility, In knowledge Engineering and knowledge management: 12th International Conference EKAW2000, Juan-les-Pins, volume 1937 of lecture notes in Artificial Intelligence, PP 17-32, Berline/Heidelberg. Also as: Technical Report Stanford University, School of Medicine, SMI-2000-0830
http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2000-0830.pdf
[2] Andaroodi, E, Andres, F, Ono,K, Lebigre,P, (December 2004) Developing a Visual Lexical Model for Semantic Management of Architectural Visual Data, Design of Spatial Ontology for Caravanserais of Silk Roads, Journal of Digital Information Management (JDIM), VOL.2, NO.4, , PP 151-160, ISSN: 0972 7272
[3] Holger Knublauch, Fergerson,R.W., Noy,N.F., Musen,M.A.
The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications, third International Semantic Web Conference - ISWC 2004, Hiroshima, Japan (2004) - An architectural overview for developers and decision-makers
[4] ISO 5964, (1985), Documentation -- Guidelines for the Development and Establishment of Multilingual Thesauri

[4] The main target of this research for multilingual equivalence is to reach to a consensus for naming a component. Issues related to different types of equivalence (like exact, inexact, partial, single to multiple, etc.[4]) specially in historical architecture multilingual terminology needs more research outside the scope of this study.
[5] http://www.unicode.org/
[6] http://www2.arts.gla.ac.uk/IPA/fullchart.html
[7] The team of ontology development of this research is cooperating with Protégé tool development team to solve the problem of remote data input and save using latest version



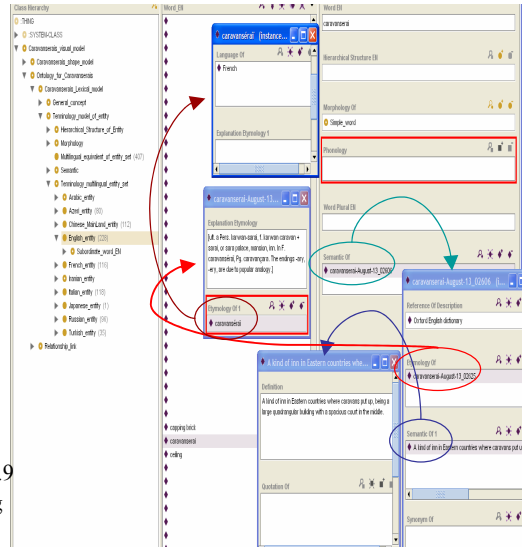


Fig1: Multilingual Term-set as classes designed in protégé V1.9
Fig2: Unified semantic attributes of multilingual tem-set using protégé v3 built107

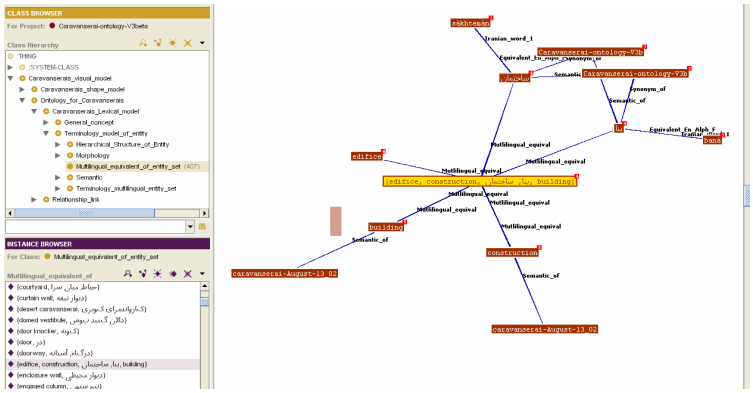
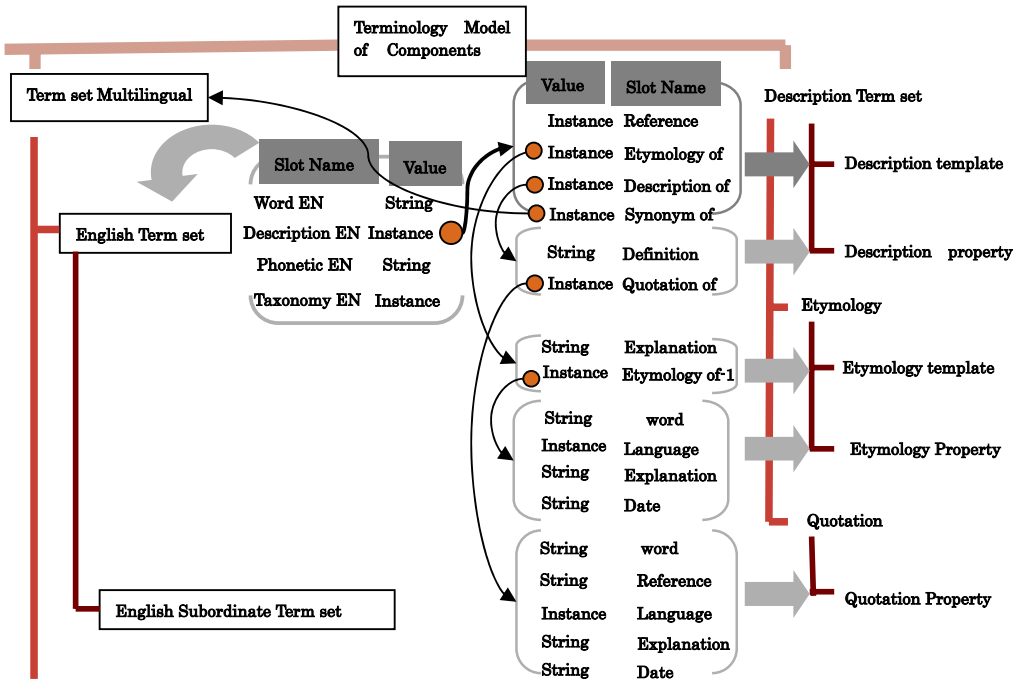


Fig 3: Graph of entities or IDs connecting multilingual equivalents of term-sets using protégé tool



Graph1: Design of unified language independent semantic lexical specifications of multilingual terminology