Extracting Domain Ontologies from Reference Books

Simon Carolan, Francisco Chinesta, Christine Evain, Morgan Magnin, Guillaume Moreau

Extracting domain ontologies from reference books

Simon Carolan[1], Francisco Chinesta[1], Christine Evain[1], Morgan Magnin[1] & Guillaume Moreau[1]
[1]Ecole Centrale de Nantes
[simon.carolan, francisco.chinesta, christine.evain, morgan.magnin, guillaume.moreau]@ec-nantes.fr

***Abstract*—Encyclopedic knowledge bases can be powerful tools for the acquisition of fundamental knowledge for learners. However, the structure and the very nature of these documents can impede learning processes. By extracting domain ontologies from reference books and using this same material to populate an intelligent learning system, we propose a methodology for lifelong learners.**

Keywords-ontology; knowledge management; engineering

I. INTRODUCTION

Reference books and encyclopedic knowledge bases provide learners with an important source of fundamental knowledge for given subjects. Through these mediums, they are presented with a series of concepts that are subordinate to other concepts and learners are therefore required to read accompanying documents, going from page to page to get a good understanding of the subject. However, the resulting knowledge acquisition process is hindered by the linearity of these resources. It is difficult for learners to understand the explicit and implicit ways that different concepts that they encounter are linked. Two objects, far apart in a linear model, can appear closer in another dimension and in an infinite environment all objects are linked in some way.

Knowledge is complex and can be considered as a multi-dimensional element. Recent studies have sought to define this structure, increasingly using a core-periphery model [1]. In addition, the information presented in online encyclopedic knowledge bases is not necessarily subject to a rigorous editorial process; the information presented represents the collective knowledge of contributors on given subjects [2].

Semantic tools, deployed within a larger semantic web are opening possibilities for the implementation of such models within electronic resources but we are yet to see the emergence of knowledge management (KM) systems that replicate the complexity of human learning processes within them. The integration of domain ontologies to these systems is a promising avenue. Used in conjunction with established corpora they generate new forms of intelligence [3].

Recent works demonstrate the creation of domain ontologies from both structured and unstructured corpora. [4] propose a methodology for the extraction of domain ontologies from engineering handbooks. However, the original documents are not the result of consensus and it is difficult to imagine that a consensus will be reached on the resulting ontology. [5] present a potentially interesting methodology for the extension of existing domain ontologies using unstructured data but fail to consider the flaws of the existing ontologies, representative of a given viewpoint of a given domain at a given time by a given community. In addition, whilst the works recognize the importance of online environments and produce open standard formats of their ontologies, they have stopped short of proposing applications of these ontologies in learning environments.

By using recognized reference books, conceived by the wider scientific community as source material, we can extract coherent domain ontologies that will provide improved representation of the domain and more opportunity for consensus to build around the resulting ontology. These ontologies can then be used as learning objects in virtual learning environments (VLE) that incorporate knowledge acquisition strategies. After detailing the specificities of our approach (experimental context), we will describe our methodology for the creation of domain ontologies (ontological engineering). We will then explore applications envisaged for the ontologies (perspectives), notably, the creation of domain specific VLE for use by engineering schools, industrial partners and beyond.

II. EXPERIMENTAL CONTEXT

We are intent on profiting from recent developments to propose VLE for the engineering sciences [6]. We have identified the domains of composite materials and virtual reality for experimentation. These domains are complex and contain cross-disciplinary aspects that make them interesting cases for study. These VLE will be built upon domain ontologies obtained from the electronic versions of acknowledged reference books in the given domains [7, 8]. These documents, written in French, have been selected as they gather scientific communities in their authorship, have undergone a rigorous publishing process. The methodology for the creation and development of the ontological model and the applications envisaged will be detailed and explored herein. The algorithms employed account for the particularities of the French language but the methodology itself could be applied to other languages.

III. ONTOLOGICAL ENGINEERING

For the preparation of our working documents, the bibliography required attention. The titles of the referenced texts present a concentration of keywords. This artificial concentration could bias the results of the analysis. It was necessary to remove the bibliographies and table of contents. These elements are conserved as they serve the validation of the ontology and are integrated into the VLE.

For the formal semi-automatic analysis of the reference books, we used a combination of statistical, syntactic and semantic tools and methodologies. To identify the frequency of concepts, it was necessary to identify the lemmas of the book contents in order for us to simultaneously consider all grammatical forms of a given word or word family. For lemmatization, we adopted the Carry algorithm [9]. Built

upon Porter [10], this algorithm provides greater precision for the French language. The algorithm was programmed in Java, in particular for its effective handling of accentuated characters. The stemming process effectively treated the majority of word but, on rare occasions, the algorithm was insufficient in treating highly morphological words such as *thermodur* - *thermoducissable*. We therefore completed our analysis through the comparison of our results with a lexicon [11], enabling us to complete insufficient stems.

Using our program, we then created a lexicon based on the frequency of the resulting stems, eliminating all words below the 0.05% barrier. We were also able to eliminate a certain number of words whose frequency in the given texts corresponds to their frequency in significant French corpora [12]. This process enabled us to outline the leading concepts and attributes for the population of our ontology. The final step in our initial analysis of the source documents consisted in analyzing the co-occurrence of the pre-identified concepts within the text on the level of a sentence, a paragraph, a chapter and a volume, enabling us to identify the links between the different key concepts (sentences, paragraphs), the relationship between the key concepts and themes (chapter) and the overall correspondence of the different terms to indicate the role of the authorial voice (volume).

Having ascertained the key concepts through the identification of leading domain specific lemmas and established the links between these lemmas through the analysis of the co-occurrence of these keyword, we were able to devise an ontological model. This was established in relation with an expert who indicated the relations of subordination in the existing structure. See the following website: http://somewhereinnomansland.wordpress.com/.

Once the provisional ontological model was established, we are able to validate it through elicitation. Considering the applicative nature of the aforementioned subject domains and our objective of creating ontologies for engineering students that will accompany them in lifelong learning, we have solicited experts from both academia and industry. By presenting the selected experts with a pre-established model, it was possible to optimize the ontology production process. Working from a common pre-defined base, in the majority of cases, consensus can be achieved much sooner. Following the successful completion of the elicitation process, the validated ontology was structured using Web Ontology Language. This is facilitated through the use of OWL Protégé [13] that facilitates the integration of ontologies into workspaces and provides powerful visualization tools.

IV. PERSPECTIVE APPLICATIONS

Favoring the development of learning strategies, the resulting domain ontologies allow learners, in line with Vygtskian precepts, to identify their Zone of Proximal Development. Beyond, this initial application, the validated OWL ontology can be integrated into a Semantic Mediawiki where we can re-introduce the original content, enriched with XML tags that will enable us to build the correspondence between the content and the ontological model allowing for rich text searches. External sources including the aforementioned online knowledge bases can then be integrated, to complete the environment and establish uncharted links. This information will be mapped into an evolving multi-dimensional environment where we will integrate sequential models, based upon the links established in the ontology in order to provide learners with suggested paths or routes for discovery of the domain, subject to their learning profile. We imagine these semantic "books" or VLE to accompany learners through their education and beyond, offering seemingly endless possibilities for lifelong learners. The established model will also prove to be useful for KM in industry, where companies hold a great deal of unclassified data that could be of benefit if channeled correctly. In addition, we believe that this may provide the grounding for document annotation tools. Indeed, the growing intelligence of documents through semantic tools means that one day soon, a document will know its contents well enough to be able to answer learner interrogations directly. In this way, we will explore the implementation of SILK technology in order to effectively respond to increasingly specific requests. It is also important to take into consideration the ubiquitous nature of modern learning. It is for this reason that we shall be exploring the exportation of portable personalized documents based on learner paths in the VLE.

V. REFERENCES

[1] Y. Zhang, (2011). Learning, Innovating, and an Emerging Core of Knowledge, Conference on Information Systems and Technology (CIST), Charlotte, November, 2011.

[2] A. Halavais & D. Lackaff, (2008). An Analysis of Topical Coverage of Wikipedia. Computer-Mediated Communication, 13:429-440.

[3] J. P. McCrae, (2009). Automatic extraction of logically consistent ontologies from text corpora. PhD thesis. National Institute of Informatics, Japan.

[4] S-H. Hsieh, H-T. Lin, N-W. Chi, K-W. Chou & K-Y. Lin, (2011). Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks. Advanced Engineering Informatics.Vol.25:2, Apr. 2011, pp 288–296.

[5] R. Anantharangachar, S. Ramini & S. Rajagopalan (2013). Ontology guided information extraction from unstructured text, Int. Journal of Web & Semantic Techonology, Vol. 4, No. 1, Jan. 2013.

[6] S. Carolan, F. Chinesta, C. Evain, M. Magnin & G. Moreau, "Towards augmented learning in science and engineering in higher eduaction," IEEE International Conference on Advanced Learning Technologies (ICALT 13), IEEE Press, July 2007, pp. 512-3,

[7] F. Chinesta & C. Evain. La Vie intime des matériaux composites. Editions Publibook, Paris: 2009.

[8] P. Fuchs & G. Moreau, (Eds.) Traité de la réalité virtuelle. Presses de l'Ecole des Mines, Paris: 2006-9.

[9] M. Paternostre, P. Francq, M. Saerens, J. Lamoral & D. Wartel (2002). Carry, un algorithme de désuffixation pour le français. Technical report, Université Libre de Bruxelles.

[10] M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130−137.

[11] B. New, C. Pallier, L. Ferrand & R. Matos (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, L'Année Psychologique, 101, Sep. 2001, 447-462.

[12] E. Brunet, (2002). Les 1500 mots les plus utilisés de la langue française. French Educational Authorities. http://eduscol.education.fr

[13] H. Knublauch, R. W. Ferguson, N. F. Noy and M. A. Musen (2004). The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Lecture Notes in Computer Science Volume 3298, 2004, pp 229-243.