A "Pivot" XML-Based Architecture for Multilingual, Multiversion Documents: Parallel Monolingual Documents Aligned Through a Central Correspondence Descriptor and Possible Use of UNL

Najeh Hajlaoui, Christian Boitet

équipe GETA, laboratoire CLIPS,
385 rue de la bibliothèque - BP 53, 38041 Grenoble Cedex 9 - France
Christian.Boitet@imag.fr

Abstract. We propose a structure for multilingual, multiversion documents, built on the model of the web-oriented, cooperative lexical multilingual data base PAPILLON: a document is represented by a collection of monolingual XML "volumes" interlinked by a central volume of "interlingual links". Here, the links relate subdocuments (XML trees) corresponding to each other in monolingual "volumes". We are developing a Java application to enable direct editing of a multilingual document through the web, at the level of monolingual volumes as well as through bilingual or trilingual interfaces inspired by those of commercial "translation workbenches". Another goal is easy integration with machine translation and multilingual generation tools. For this, we add a special UNL volume. In a first stage, we split the UNL-xml document in several monolingual documents, again represented by XML files. Each document contains the text in a particular language, plus the corresponding UNL graphs, and can be modified independently. The interface is easy to build, but realigning the documents after a series of such modifications is a very difficult task.

1 Introduction

Due to Internet, the number of available documents grows dramatically. There is a strategic need for companies to control information written in more than 30 languages (HP, IBM, MS, Caterpillar). This requires the installation of powerful and effective management tools of multilingual "synchronized" documents.
There are techniques of large-grained linking (on the level of HTML pages). However, there are no techniques for structuring multilingual documents so as to allow fine-grained synchronization (at paragraph or sentence level) and even less permitting editability through the Web.

The interest to synchronize at least on the level of the sentences is double:

– for the translation and human revision with the assistance of techniques of HTHM (Human Translation Helped by Machine) and in particular of translation memory.

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 309–325.