Interactive Handling of Multilingual Content within Digital Media

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, Isabelle Kramer

Interactive Handling of Multilingual Content within Digital Media

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, and Isabelle Kramer
LORIA / INRIA Lorraine
Campus Scientifique – BP 239
54506 Vandoeuvre-lès-Nancy, FRANCE
{Samuel.Cruz-Lara, Nadia.Bellalem, Julien.Ducret, Isabelle.Kramer}@loria.fr

Abstract
Linguistic information plays an essential role in the management of multimedia information as it bears most of the descriptive content associated with more visual information. Depending on the context, it may be seen as the primary content, as documentary content for multimedia information, or as one among several possible information components in specific contexts such as interactive multimedia applications.

In this paper we describe a generic framework that could be integrated into multimedia content. Our main objectives are both, to propose a high-level abstract model to represent multilingual content, and to offer a high degree of interactivity allowing final users to handle multilingual content within digital media

2. Introduction.

Linguistic information can appear in various formats: spoken data in an audio or video sequence, implicit data appearing on an image or textual information that may be further presented to the user.

Dealing with multilingual information is thus crucial to adapting the content to specific user targets. It requires one to consider potential situations where the linguistic information contained in a multimedia sequence is either already conceived in such way that it can be adapted on the fly to the linguistic needs of user, or by using an additional process where content should be adapted before presenting it to the user.

The Multi Lingual Information Framework (MLIF) is being designed with the objective of providing a common abstract model being able to generate several formats used in the framework of translation and localisation such as TMX or XLIFF.

3. Dealing with Multilingual Content.

MLIF also aims to propose a platform of specification for representing multilingual contents in a whole range of applications, as the localization and translation memory process, interactive and HD TV, karaoke, subtitles, accessibility, … MLIF promotes the use of a common framework for the future development of several different formats, for example: TBX (TermBase eXchange Standard. www.lisa.org/tbx), TMX (Translation Memory eXchange Standard. www.lisa.org/tmx), XLIFF (XML Localization Interchange File Format. www.oasisopen.org/committees/tc_home.php?wg_ abbrev=xliff), Timed Text (Timed Text. www.w3.org/AudioVideo/TT/), etc. It does not create a complete new format from scratch, but suggests that the overlapping issues should be handled independently and separately. It will save time and energy for different groups and will provide synergy to work in collaboration.

Presently, all the groups (i.e. LISA, OASIS, W3C, ISO, …) are working independently and do not have any mechanism for taking advantage of each other's tools. MLIF proposes to concentrate on only those specific issues that are different from others and specific to one format only, so it will create a smaller domain for the groups' developers. It gives more time to concentrate on a subset of the problems they are currently dealing with and creates a niche that helps in providing a better solution for problems of multilingual data handling and translation issues.

In MLIF, we deal with the issue of overlap between the existing formats. MLIF involves the development of an API through which all these formats will be integrated into the core MLIF structure. This is done through the identification and a selection of data categories as stated in ISO 12620. MLIF can be considered as a parent for all the formats that we have mentioned before. Since all these formats deal with multilingual data expressed in the form of segments or text units they can all be stored, manipulated and translated in a similar manner.

Our line of attack is largely inspired by the methodology used to develop TMF (Terminological Mark-up Framework) ISO 16642. So, we identify and select a set of data categories as stated in ISO 12620 (data category register). The way these data categories are related and associated to each other is described by a metamodel. This metamodel and the selected data categories are a high-level representation of multilingual content. From this high-level representation we are able to generate any specific format: we can thus ensure the interoperability between several multilingual content formats and their applications.

3.1 Handling Multilingual Content within Digital Media.

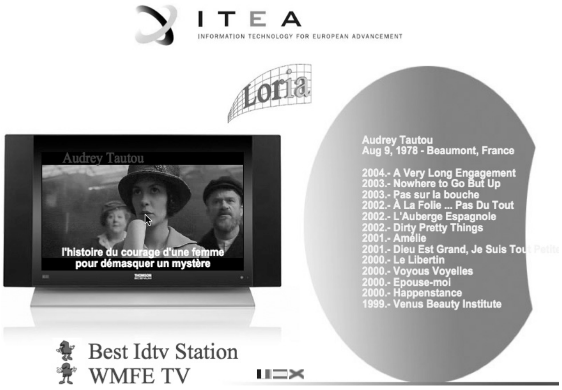
Textual data is the primary vehicle in which information (for user interfaces on the web or any digital media) is encoded. There is a pressing demand for facilitating the access to and translating/localizing of the linguistic data contained in these applications for multilingual communities. Wherever we have application serving to multilingual communities and displaying textual information, it is inevitable that the textual information is stored, handled and displayed in proper format and language of user's choice, irrespective of medium used for displaying.

Extraction of linguistic data (while keeping formatting, displaying and other information in separate tags) from multiple sources and languages (books, periodicals, newscasters, television programs, e-learning resources, etc.) and fusion into a user-chosen language requires understanding of the data and easy handling.

Within ITEA's "Jules Verne" and "Passepartout" projects, we have identified several potential implementations of MLIF. MLIF can be used in e-learning, interactive television programs and any other application having a user interface. It may be very helpful for future interactive television broadcasting. It presents ample opportunity for giving value to different languages and cultures, as is the case in Europe and Asia.

In order to have a general but rather concrete understanding about using MLIF, we may think about dealing with multilingual subtitles in digital media. The modelling of an existing multilingual database requires a specific analysis. This essential stage makes it possible to identify the available informational fields. For instance, the subtitles constitute a set of multilingual data but one may also want to retrieve, in an interactive manner, some interesting information about an actor (see Figure 1).

Starting from this analysis, one will be able to evaluate the mass of work required to satisfy the scenarios. If the multilingual data come within several distinct entities, an expert analysis of the data will have to determine which method of consultation of the base will be adapted to the nature of the data. In the case of subtitles, for instance, we wish to change dynamically the language. At present (i.e. DVD technology), subtitles are inserted into video files in bitmap format, but we need to keep them as textual content in order to provide access to the linguistic realization (language).


ITEA
INFORMATION TECHNOLOGY FOR EUROPEAN ADVANCEMENT
Loria
Audrey Tautou
l'histoire du courage d'une femme
pour démasquer un mystère
Best Idtv Station
WMFE TV

Audrey Tautou
Aug 9, 1978 - Beaumont, France

2004.- A Very Long Engagement
2003.- Nowhere to Go But Up
2003.- Pas sur la bouche
2002.- À La Folie ... Pas Du Tout
2002.- L'Auberge Espagnole
2002.- Dirty Pretty Things
2001.- Amélie
2001.- Dieu Est Grand, Je Suis Tout Petite
2000.- Le Libertin
2000.- Voyous Voyelles
2000.- Epouse-moi
2000.- Happenstance
1999.- Venus Beauty Institute

Figure 1. Multilingual Textual information associated to Digital Media.

MLIF may be associated with different multimedia applications in order be able to display multilingual information.

Currently, within the framework of ITEA "Passepartout" project, we are experimenting with some basic scenarios by using XMT ("eXtensible MPEG4 Textual format") and SMIL ("Synchronized Multimedia Integration Language").

Results of our current work are encouraging and we are also planning to use MLIF within STB (Set Top Boxes) for advanced IDTV.

References

S. Cruz-Lara, S. Gupta, & L. Romary (2004, November). *Handling Multilingual content in digital media: The Multilingual Information Framework*. Paper presented at the European Workshop on the Integration of Knowledge EWIMT 2004, Semantics and Digital Media Technology. London, UK.

S. Cruz-Lara, S. Gupta, J.D. Fernández García, & L. Romary (2005, May). *Multilingual Information Framework for Handling Textual Data in Digital Media*. IEEE AMT 2005, Paper presented at the Third International Conference on Active Media Technology. Takamatsu, Kagawa, Japan.

S. Gupta, S. Cruz-Lara & L. Romary (2005, May). *Implementing Multilingual Information Framework in Applications using Textual Display*. Paper presented at the 7th International Conference on Enterprise Information Systems, ICEIS 2005. Miami, USA.