IMPLEMENTING MULTILINGUAL INFORMATION FRAMEWORK IN APPLICATIONS USING TEXTUAL DISPLAY

Samuel Cruz-Lara, Satyendra Kumar Gupta, and Laurent Romary
Loria (UMR 7503) / INRIA Lorraine, Campus Scientifique – BP239, 54506 Vandoeuvre-les-Nancy, France
Email: {cruzlara,gupta,romary}@loria.fr

Keywords: MLIF, Interactive multimedia applications, Natural language display, Localization

Abstract: This paper presents implementation of MLIF [1] (Multilingual Information Framework), a high level model for describing multilingual data across wide range of possible applications in translation/localization process within several multimedia domains (e.g. broadcasting of interactive multimedia applications), natural language interfaces, geographical information systems for multilingual communities.

1 INTRODUCTION

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 (text illustrated by pictures or videos), as documentary content for multimedia information, or as one among several possible information components in specific contexts such as interactive multimedia applications.
Linguistic information can appear in various formats: spoken data in an audio or video sequence, implicit data appearing on an image (caption, tags, etc.) or textual information that may be further presented to the user graphically or via a text to speech processor.
In this context, dealing with multilingual information is 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.

Finally, there are a wide variety of applications within which multilingual information may appear, which supports development and implementation of generic framework, MLIF, for dealing with multilingual content: subtitling of video content, dialogue prompts, menus in interactive TV, descriptive information for multimedia scenes, karaoke management, etc. Such information should be considered in the light of the experience of more specialized communities traditionally dealing with multilingual content, namely the translation and localization industry.

2 MULTILINGUAL INFORMATION FRAMEWORK

The Multi Lingual Information Framework (MLIF) is designed with the objective of providing a common platform for all the existing tools developed by the different groups (LISA [2], OASIS [3], W3C [4], ISO [5]). It promotes the use of a common framework for the future development of several different formats: TBX, TMX, XLIFF, Timed Text, TMF, 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 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.

MLIF deals with the issues of overlap between the existing formats. MLIF involves the development of an API through which all these formats are integrated into the core MLIF structure.

This is done through the identification and a selection of data categories [10] (DCs) as stated in ISO DIS 12620-11 (in ISO/TC 37/SC 32). 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. This kind of data can easily be stored in data categories and in terminological mark-up. The results of IST SALT project [6] clearly show that it is not difficult to edit, store and reuse data categories. The SALT project combines two interchange formats: OLIF [7], which focuses on the interchange of data among *lexbase* resources from various machine translation systems, and MARTIF [8], which facilitates the interchange of *termbase* resources with conceptual data models ranging from the simple to the sophisticated. It provides a graphical user interface that can be used to access or to define new data categories or modify them.

3 IMPLEMENTATION OF MLIF

Some multilingual content stuff has been integrated in the demonstrations developed. As all the Data Categories related to Digital Media has not been yet identified and defined, in both applications multilingual content has been encoded in XLIFF (which is completely interchangeable to MLIF with the help of XSLT styelsheets, we have written).

3.1 XMT-O LOCALIZATION ROUND TRIP USING MLIF

In this application, we have obtained by means of translation/localization process, a document in English (Figure 3) corresponding to an XMT document that is in French (Figure 2). Figure 1 shows how the localization process is performed.

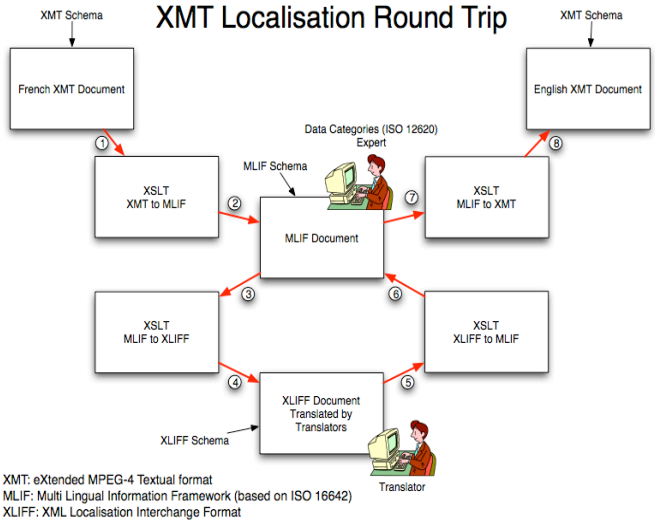

Figure 1: XMT Localisation Round Trip.

1. The original XMT_{French} document containts linguistic information in French.
2. Transformation of XMT_{french} document into $MLIF_{french}$ document.
3. Transformation of the $MLIF_{french}$ document into an $XLIFF_{french}$ document.
4. By using existing XLIFF environment, a professional translator performs French-English translation. We obtain $XLIFF_{english}$ document.
5. /6. Transformation of $XLIFF_{english}$ document into an $MLIF_{english}$ document.
7. /8. Transformation of the $MLIF_{english}$ document into $XMT_{english}$ document

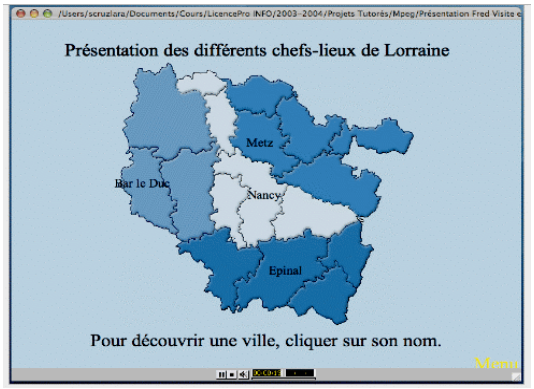

Figure 2: An Interactive MPEG-4 Presentation in French.

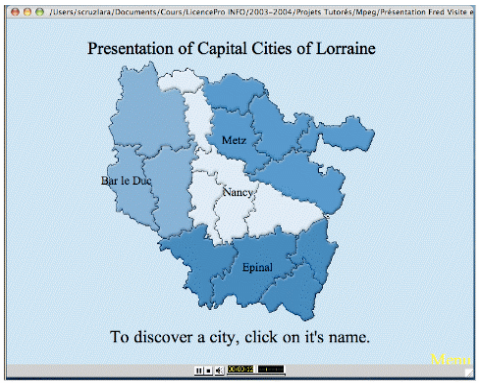

Figure 3: Interactive Presentation in English.

3.2 Identifying Monolingual Content in XMT-O

Identifying monolingual content in XMT-O document may be considered from two points of view:

1. Textual information related to metadata,
2. Textual information related to data (i.e. subtitles).

For simplicity, we will consider only textual information related to data, that is, textual information that may be associated under the form of subtitles, to a multimedia presentation.

Thorough study of XMT-O reveals that all textual information related to data (i.e. subtitles) in XMT-O document, is included in the "textLines" attributes of the <string> tag, for example:

<string dur="800s"
…
textLines="" Presentation of MLIF;;"
…/>

So, it is rather easy, by parsing a XMT-O document, to retrieve all monolingual textual information related to data. Developing a XSLT stylesheet, transforming a XMT-O document into MLIF document or vice-versa is easy task. However, we must verify that

- The XSLT stylesheet preserves the original XMT-O structure when transforming into a MLIF document,
- The XSLT stylesheet takes into account all Data Categories related to XMT-O original document.

Though the task of identifying monolingual content inside an XMT-O document is rather easy, identifying Data Categories is a complex task.

The very first step is to setup an DCS (Data Category Specification) related to Digital Media. This activity is very complex because we have to:

- Identify all existing DCs that may be used in the context of Digital Media knowing that several Data Categories may be common to several different kinds of language resources,
- Very few DCs related to Digital Media have been identified and defined. Identifying and defining DCs is complex process because
  - Digital Media experts have to be involved in identification of DC for multimedia,
  - DC experts must approve all the DC identified (and proposed) by Digital Media experts,
  - It is an official ISO normalization process that takes time.

3.1 Historical Information Display in MPEG-4 Application

In the framework of ITEA project "Jules Verne", an MPEG application has been developed. This application shows an interactive globe in the middle of the screen and two other display boards on either side of the globe. In the display board on left, it shows the map of the country chosen and on right side it shows the historical information about the country chosen. This also has some flags on the display board on right. User can click on any of the flags and the historical information is displayed in the chosen language. This information is encoded in XLIFF/MLIF document (both formats are completely interchangeable with the help of stylesheets). This information is extracted from the document, on run time, with the help of chosen parameters for language identifier and country name identifier.

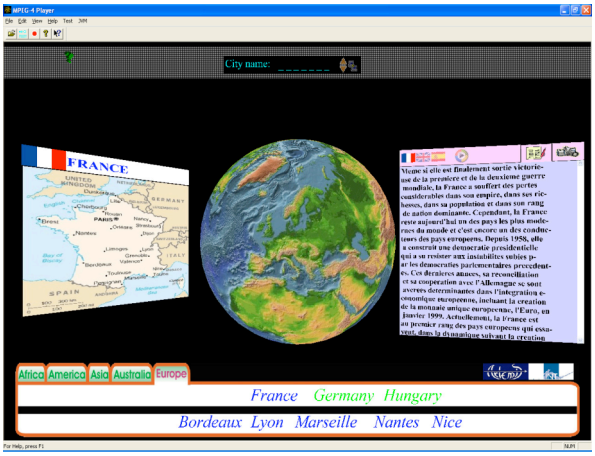

Figure 4: Information Display in French.

Figure 4 shows one screenshot of MPEG-4 Multimedia presentation (© INT-ARTEMIS http://www-artemis.int-evry.fr/). In this screenshot, the textual information (in French) appearing at the right side of the screen has been retrieved from

XLIFF/MLIF document. Figure 5 shows the same application once user has chosen to see information in English.

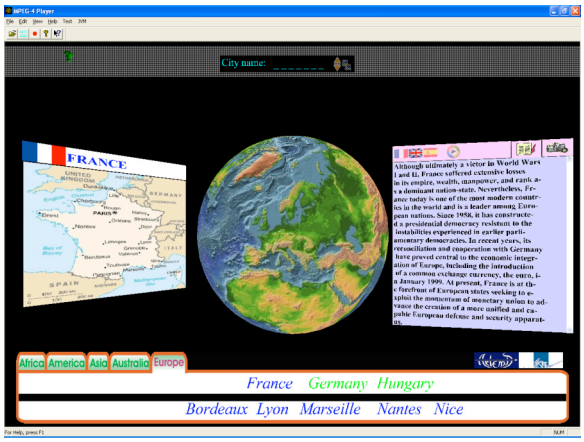

Figure 5: Information Display in English.

4 USE OF XSLT TO DEAL WITH DIFFERENT MULTILINGUAL CONTENT FORMATS

Figure 1 has shown that dealing with multilingual content means transforming one document (XMT for instance in figure 1), first into MLIF document and then into XLIFF document. The XSLT stylesheet XMTToMLIF is a very simple way to obtain MLIF documents (canonical from) from XMT-O document. It should be noted that XSLT stylesheet allows to retrieve only textual information and original structure in not preserved.

We have written many stylesheets for the purpose of all the transformations shown in figure 1. Besides these stylesheets, we have also written some more stylesheets to transform documents from/to MLIF to/from TMX, I18N.

5 POTENTIAL IMPLEMENTATION AREAS OF MLIF

Textual data is the primary vehicle in which information (for user interface on web or any digital media) is encoded. There is 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.

We have identified wide range of potential implementation of MLIF. MLIF can be used in e-learning, interactive television programs and any other application having user interface. It is 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.

6 CONCLUSIONS

In this paper, we have shown implementation of MLIF with different multimedia applications to display linguistic information. With the help of style sheets, use of data categories while transforming document to/from MLIF gives good results. Results of our current work are encouraging and we are working to use MLIF with STB (Set Top Box) for advanced IDTV.

REFERENCES

[1] Cruz-Lara, S., Gupta, S., Romary, L., 2004. Handling Multilingual content in digital media: The Multilingual Information Framework. In EWIMT-2004, *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*.

[2] LISA. *www.lisa.org*.

[3] OASIS. *www.oasis-open.org/home/index.php*.

[4] W3C. *www.w3c.org*.

[5] ISO. *www.iso.org*.

[6] SALT. Standards-based Access to multilingual Lexicons and Terminologies. *http://www.loria.fr/projets/SALT/saltsite.html*

[7] Open Lexicon Interchange Format. *www.olif.net*

[8] MAchine-Readable Terminology Interchange Format. ISO 12200: 1999

[9] XMT: MPEG-4 Textual Format for Cross-Standards Interoperability. Kim, M.; Wood, S. IBM T.J. Watson Research.

[10]Computer Applications in Terminology – Data Categories. ISO 12620. *http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=2517&ICS1=1&ICS2=20&ICS3=*