Multilingual Information Framework for Handling textual data in Digital Media

Samuel Cruz-Lara, Satyendra Kumar Gupta, Javier David Fernández García, Laurent Romary

Multilingual Information Framework for Handling textual data in Digital Media

Samuel Cruz-Lara, Satyendra Gupta, Javier David Fernández García, Laurent Romary
LORIA / INRIA Lorraine
Campus Scientifique BP 239
54506 Vandœuvre-lès-Nancy
{cruzlara,gupta,fernande,romary}@loria.fr

***Abstract*: This document presents MLIF (Multi Lingual Information Framework) [1], a high-level model for describing multilingual data across a wide range of possible applications in the translation/localization process within several multimedia domains (e.g. broadcasting interactive programs within a multilingual community).**

Keywords: Multilingual Content, Interactive Multimedia Applications, Natural Language Display, Localization

I. 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.

II. BACKGROUND

The scope of research and development in localization and translation memory (TM) process development is very large. There are numerous independent groups LISA [2], OASIS [3], W3C [4], ISO [5] working on these aspects namely LISA, OASIS, W3C, ISO, etc. LISA is working for the GILT (Globalization, Internationalization, Localization and Translation) business community. It has evolved as the premiere organization for developing language-technology standards. OSCAR [6] "Open Standards for Container/Content Allowing Re-use" is LISA's special interest group (SIG) for the creation of open standards. OASIS evolved to drive the development, convergence, and adoption of structured information standards in the areas of e-business, web services, etc. OASIS is driven by various technical committees (TC) formed by its members. OASIS XML Localization Interchange File Format TC (OASIS XLIFF TC) was formed with the purpose to define, through XML vocabularies, an extensible specification for interchange of localization information. The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to bring the Web to its fullest potential. W3C organizes the work necessary for the development or evolution of Web technologies into Activities. Most of the activities under W3C are developing specifications/tools for management of language/text on the web. ISO is a network of national standards institutes from 148 countries working in partnership with international organizations, governments, industry and business and consumer representatives. ISO is a bridge between public and private sectors. It has also formed some technical committees, working in language-technology and IT applications in information, documentation and publishing.

Under the guidance of above-mentioned groups, many formats have been developed. Some of the major formats of specific interest for localization and TM are TBX [7], TMX [8] (LISA/OSCAR), XLIFF [9] (OASIS), Timed Text [10] (W3C), TMF [11] (ISO). TBX is an open XML-based standard format for terminological data. This capability will greatly facilitate the flow of terminological information throughout the information cycle both inside an organization and with outside service providers. TMX is a vendor-neutral, open standard for

storing and exchanging translation memories created by Computer Aided Translation (CAT) and localization tools. The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process. The purpose of XLIFF vocabulary is to store localizable data and carry it from one step of the localization process to the other, while allowing interoperability between tools. The Timed-Text specification covers real time subtitling of foreign-language movies, captioning for people lacking audio devices or having hearing impairments, karaoke, scrolling news items or teleprompter applications. It provides interoperability between different existing formats. ISO 16642:2003 specifies a framework designed to provide guidance on the basic principles for representing data recorded in terminological data collections. This framework includes a meta-model and methods for describing specific terminological mark-up languages (TMLs) expressed in XML. The mechanisms for implementing constraints in a TML are also defined in ISO 16642:2003. ISO 16642:2003 is designed to support the development and use of computer applications for terminological data and the exchange of such data between different applications. In the framework of digital media, MPEG4 and MPEG7 deal with multilingual data. MPEG-4 is standard for multimedia for the fixed and mobile web and MPEG-7 is standard for description and search of audio and visual content.

When we closely examine the different standards or formats developed by these groups, we find that they have many overlapping features. For example all these formats are based on XML schemas, provide extensibility, and bridge the gap between two systems or tools in different languages. There are many identical requirements for all the formats irrespective of the differences in final output. For example, all the formats aim at being user-friendly, easy-to-learn, and at reusing existing databases or knowledge. All these formats work well in the specific field they are designed for, but they lack a synergy that would make them interoperable when using one type of information in a slightly different context, giving rise to the fear of competition between them.

III. THE 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 groups listed in the previous section. 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.

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 DIS 12620-1 (in ISO/TC 37/SC 3). 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 [12] clearly show that it is not difficult to edit, store and reuse data categories. The SALT project combines two interchange formats: OLIF [13], which focuses on the interchange of data among lexbase resources from various machine translation systems, and MARTIF [14], 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.

IV. USING MLIF

Some multilingual content stuff has been integrated in several demonstrations that we have developed in the framework of ITEA "Jules Verne" project [15]. As all the Data Categories related to Digital Media have 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).

XMT [16] has been designed to provide an exchangeable format between content authors while preserving the author's intentions in a high-level textual format. In addition to providing an author-friendly abstraction of the underlying MPEG-4 technologies, another important consideration for the XMT design was to respect existing practices of content authors such as the Web3D X3D, W3C SMIL and HTML. XMT is suitable for many uses including manually authored content as well as machine-generated content using multimedia database material and templates. XMT may be encoded and stored in the exchangeable mp4 binary file or may also be encoded directly into streams and transmitted. XMT encoding and delivery hints exist to assist this process.

A. XMT Localization Round Trip using MLIF

The XMT localization process is performed in the following way: (see Fig. 1)

1. The original XMTFrench document containts linguistic information in French.

2. Transformation of XMTfrench document into MLIFfrench document.

3. Transformation of the MLIFfrench document into an XLIFFfrench document.

4. By using existing XLIFF environment, a professional translator performs French-English translation. We obtain XLIFFenglish document.

5. /6. Transformation of XLIFFenglish document into an MLIFenglish document.

7. /8. Transformation of the MLIFenglish document into XMTenglish document

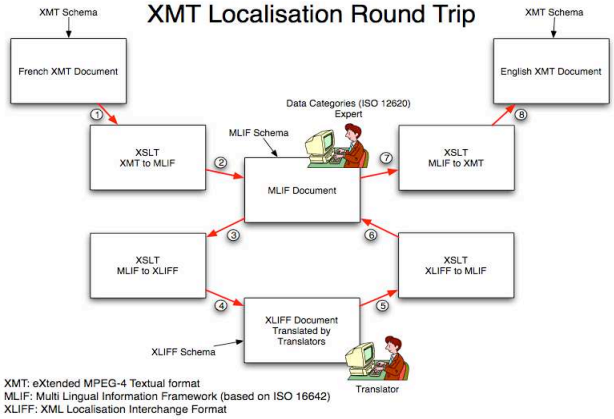

Figure 1. XMT Localization Round Trip.

Fig. 2 and Fig 3. show an interactive MPEG-4 (XMT based) application in which textual data (i.e. subtitles) is displayed dynamically, by means of user interaction, in French or in Spanish.

Figure 2. An interactive MPEG-4 presentation: textual data is in French.

Figure 3. An interactive MPEG-4 presentation: textual data is in Spanish.

The translation/localisation process that allows passing from French to Spanish follows the XMT localization round trip that has been shown in Fig. 1. It should be noted that in this example, only one single XMT document exists. This document contains French and Spanish textual data.

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

```
<string dur="800s"
    …
    textLines="&amp;quot; Presentation of MLIF;;"
    … />
```

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

- The XSLT stylesheet preserves the original XMT structure when transforming into a MLIF document,
- The XSLT stylesheet takes into account all Data Categories related to XMT original document.
- Though the task of identifying monolingual content inside an XMT 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.

C. *Historical and Geographical Textual Information Display inside MPEG-4 Applications: an example.*

In the framework of ITEA project "Jules Verne", another MPEG application has been developed[1]. 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 (again, 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.

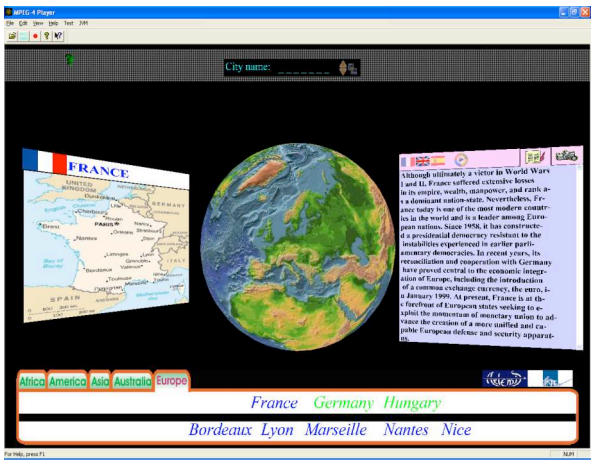

Figure 4. Textual Information (right panel) Display in English.

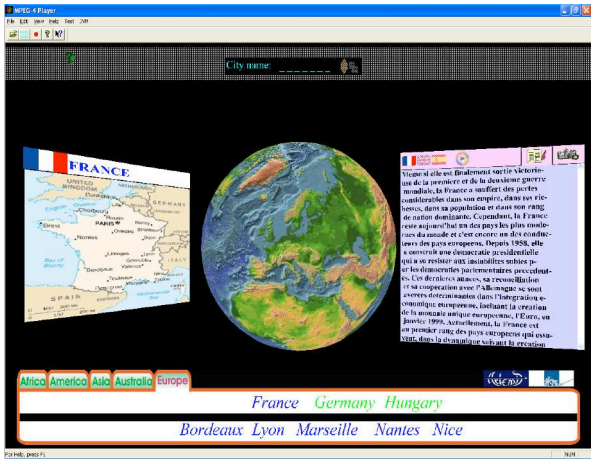

Figure 5. Textual Information (right panel) Display in French.

[1] ©2004 ARTEMIS – INT (Institut National de Télécommunications d'Evry, France)

V. Potential Implementation Areas for MLIF

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.

We have identified several potential implementations 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.

VI. Conclusion

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] S. Cruz-Lara, S. Gupta, L. Romary, 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] OSCAR. www.lisa.org/oscar

[7] TermBase eXchange Standard. www.lisa.org/tbx

[8] Translation Memory eXchange Standard. www.lisa.org/tmx

[9] XML Localization Interchange File Format. www.oasisopen.org/committees/tc_home.php?wg_abbrev=xliff

[10] Timed Text. www.w3.org/AudioVideo/TT/

[11] Terminological Markup Framework. www.loria.fr/projets/TMF/tmf.html

[12] Standards-based Access to multilingual Lexicons and Terminologies www.loria.fr/projets/SALT

[13] Open Lexicon Interchange Format. www.olif.net/

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

[15] ITEA "Jules Verne" Project. http://webserver.tudor.lu/QuickPlace/julesverne/Main.nsf/

[16] M. Kim & S. Wood. "XMT: MPEG-4 Textual Format for Cross-Standard Interoperability". IBM T.J Watson Research.