Topics in Language Resources for Translation and Localisation (Chapter : Standardising the Management and the Representation of Multilingual Data : the Multi Lingual Information Framework)

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

Standardizing the management and the representation of multilingual data: the Multi Lingual Information Framework

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

LORIA / INRIA Nancy Grand Est - Projet TALARIS
Nancy-Université
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy
FRANCE

{Samuel.Cruz-Lara, Nadia.Bellalem, Julien.Ducret, Isabelle.Kramer}@loria.fr

Abstract

Due to the critical role that normalization plays during the translation and localization processes, we propose here to analyze some standards, as well as the related software tools that are used by professional translators and by several automatic translating services. We will first point out the importance of normalization within the translation and localization activities. Next, we will introduce a methodology of standardization, whose objective is to harmonize the management and the representation of multilingual data. Without a doubt, the control of the interoperability between the industrial standards currently used for localization [XLIFF], translation memory [TMX], or with some recent initiatives such as the internationalization tag set [ITS], constitutes a major objective for a coherent and global management of multilingual data. The Multi Lingual Information Framework MLIF [ISO AWI 24616] is based on a methodology of standardization resulting from the ISO (sub-committees TC37/SC3 "Computer Applications for Terminology" and SC4 "Language Resources Management"). MLIF aims at proposing a high-level abstract specification platform for a computer-oriented representation of multilingual data within a large variety of applications such as translation memories, localization, computer-aided translation, multimedia, or electronic document management. MLIF should be considered as a unified conceptual representation of multilingual content and is not intended to substitute or to compete with any existing standard. MLIF is being designed with the objective of providing a common conceptual model and a platform allowing interoperability among several translation and localization standards, and by extension, their related tools. The major benefit of MLIF is interoperability because it allows experts to gather, under the same conceptual model, various tools and representations related to multilingual data. In addition, MLIF should also make it possible to evaluate and to compare these multilingual resources and tools.

1. Introduction

The extremely fast evolution of the technological development in the sector of Communication and Information Technologies, and in particular, in the field of natural language processing, makes particularly acute the question of standardization. The issues related to this standardization are of an industrial, economic and cultural nature. Nowadays, an increasing number of standards are frequently being used within most scientific and technical domains. Translation and localization activities just cannot remain isolated from this important and novel situation. The advantages of normalization are currently fully recognized by most professional translators: using standards means working with a high level of quality, performance, and reliability within a very important market that is becoming more and more global and thus more and more challenging. Indeed, the standards combine simplicity and economy by reducing the planning and production costs, and by unifying several kinds of terminology (i.e. validated vocabulary) and several kinds of products. At the national and international levels, standards stimulate cooperation between different communities of trade while ensuring interoperability within information exchanges and reliability of all generated results by using standardized methods and procedures; this is why normative information has became fundamental.

The scope of research and development within the localization and translation memory process development is very large. Several industrial standards have been developed: TMX, XLIFF, OLIF, etc. However, when we closely examine these different standards or formats by subject field, we find that they have many overlapping features. 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 the synergy that would make them interoperable when using one type of information in a slightly different context. Modelization corresponds to the need to describe and compare existing interchange formats in terms of their informational coverage and the conditions of interoperability between these formats, and hence the source data generated in them. One of the issues here is to explain how a uniform way of documenting such data that takes into account the heterogeneity of both their formats and their descriptors.

We also seek to answer the demand for more flexibility in the definition of interchange formats so that any new project may define its own data organization without losing interoperability with existing standards or practices. Such an attempt should lead to more general principles and methods for analyzing existing multilingual databases and mapping them onto any chosen multilingual interchange format.

2. Normalization: a key issue for translation.

The translator is the most important element of the translation process: because of his experience and knowledge he ensures that the translated document is accurate with respect to the original one [GOM 01]. A good translation does not only need linguistic awareness but it also needs a good knowledge of the technical or scientific field of the documents that have to be translated. Most texts are addressed to specialists and ignorance of the specialized expressions can justifiably cause rejection by the reader. In the same way, English technical terms - whose equivalents however exists in the target language - are often used just as they are in French, for example. Obviously, a technical translation is not limited to the data processing or computer science translation. A technical translator must have, in addition to his knowledge, a high-quality set of documents. Even if his level of knowledge is high, he must continuously seek advice from technical documents (i.e. journals, specialized dictionaries, magazines, data bases, etc.). Somehow these technical documents correspond to a set of essential tools allowing a translator to analyze the information located on the covered subject. So, a translator must evolve constantly by acquiring new information and new experiences, because this way he will obtain additional linguistic and non-linguistic knowledge related to the domains he has to deal with [HUR 96].

Standards constitute a fundamental tool for translators, because they will provide, as high-level models for technical specifications (i.e. symbols, definitions, codes, figures, methodology, etc), abundant, exact, and about all, interoperable and reliable information. Unfortunately, several fields – and specially translation - have numerous and often non-compatible standards. This requires a parallel activity of normalization of these standards in order to ensure, at least, a minimal degree of compatibility. This activity constitute the main issue of the "Documentation on standards" or the "Structured set of standards" [PIN 93]. As the texts, objects of translation, have a great diversity of subjects, the documentary activities applied to the standards can also guide and direct the translator in the search for standards relative to a given field. The production of standards became sometimes so prolific that it is quite difficult to understand exactly what method or procedure has to be used.

The documentary techniques bring essential assistance. The services of information and dissemination of the national and international organizations of standardization give easy access to all this information. The standards worked out by the organizations of standardization (i.e. ISO, W3C, LISA, ...) are more and more accepted by the translation services and the translators with an aim of guaranteeing a high-level quality in their services and their products. Standardization thus becomes a synonym of quality for the customers who wish the best possible results. In addition, standards represent also an essential tool for translators, as they aim at creating normalized terminological and methodological linguistic resources, in order to improve the national and international exchanges as well as cooperation within all possible fields. It is also necessary to point out the important efforts of standardization of ISO and W3C in the field of information technologies, especially those related to the computerized processing of multilingual data and the Internet.

Standardization is present during the whole process of translation. So, the access to the normative information represents a stage impossible to circumvent within the activity of translators. Within this task (i.e. standardization) translators are assisted not only by resource centers charged to disseminate the information of the standards worked out by the national and international organizations, but also by other specialized private agencies (i.e. PRODOC, http://www.prodoc.de) whose main objective is to advise the customers with regards to standards. In the same way, the Internet allows access to thousands of web sites (i.e. terminology trade, data

bases, etc…) that provide important information related to using standards, as well as access to several interesting research projects in progress whose objectives are the development of standards and recommendations in the field of the "industry of the language".

2.1. Translation memories and translators.

Among various documentations available for translators, translation memories (TM) occupy a dominating place. Translation memories are built from already translated texts and constitute an assistance for translating repetitive texts, as well as for performing searching operations on terminological data bases. However, there does not exist yet a standard for the development and the management of translation memories, but rather some standards related to techniques and methods belonging to the field of document management and indexing. That is the reason why it is necessary to take into account the standards related to human indexing: the ISO standard 5963-85 [ISO 5963] encourages the services of indexing and other resource centers to unify their practices. One can divide translation memories into two great classes: memories built starting from an indexing of complete sentences, and memories built starting from an indexing of all words. The method of textual searching will thus be determined by the technique of indexing having been used.

For the evaluation of the quality of software related to computer-aided translation (CAT), we have the ISO/IEC standard 9126: 1991 [ISO 9126]. This ISO/IEC standard defines the requirements for quality, the methods of evaluation, and the application of the procedures of evaluation. Within this context, it is suitable to point out the work of the Expert Advisory Group on Language Engineering Standards (EAGLES). The activities of this group are related to multilingual electronic dictionaries, multilingual electronic thesaurus, terminological data bases management systems, and translation memories. However, the EAGLES group does not aim at producing international standards but rather to present the needs and the requirements of operational applications and to accelerate the process of standardization in this matter.

There are no specific standards for automatic machine translation either. However, the systems of CAT were subjected to many evaluations thus making it possible to gradually improve the methodologies used for these evaluations. This is the reason why some of these evaluations can be considered as being standards de facto for the future evaluation of CAT technologies.

On its side, the LISA organization proposes a recommendation called Translation Memory Exchange (TMX) that aims at facilitating the exchange of the data related to translation memories between tools and software CAT systems. Although TM Tools are based on the same basic idea, we must note that for the same sentence each tool proposes rather different ways to implement the required formatting information: on the one hand, formatting is applied to the source and target texts of a translation unit and this formatting is not exported to the corresponding TMX file; on the other hand, formatting is sometimes exported to the TMX file. In the following table (see Figure 1), the sample sentence "`the` **`sentence`** `contains` *`different`* `formatting information`" is represented in TMX by using several tools [ZER 05]. Some of these tools use external files to store formatting information (i.e. Déjà Vu and SDLX), but all of them use different ways of encoding that information.

TRADOS 6.5	DÉJÀ VU	SDLX
<seg> This <ut>{\b </ut>sentence<ut>}</ut> contains <ut>{\i </ut>different<ut>}</ut> <ut>{\ul </ut>formatting information<ut>}</ut>. </seg>	<seg> <ph x="1">{1}</ph>This <ph x="2">{2}</ph> sentence <ph x="3">{3}</ph> contains <ph x="4">{4}</ph>different <ph x="5">{5}</ph><ph x="6">{6}</ph>formatting information <ph x="7">{7}</ph>. </seg>	<seg>This <bpt i="1"x="1"> <1></bpt>sentence <epti="1"></1></ept> contains <bpt i="2"x="2"> <2></bpt>different <epti="2"> </2></ept> <bpt i="3"x="3"><3></bpt> formatting information<epti="3"></3></ept> . </seg>

Table 1. Comparison of formatting across tools

In addition, the segmentation rules used by TM tools are not compatible: each tool applies his own rule to split the text into various segments. In a same sentence some tools consider various separators. For example the semi-colon is considered as a separator for Déjà vu, but not for SDLX. Segmentation organizes and structures the data. If every one uses his own rules, the exchange is no more possible; that's why SRX for several years tries to normalize segmentation rules. SRX guidelines are useful to evaluate translation memory qualities and ensure interoperability of multilingual data.

2.2. Standards: proliferation and necessity.

As we have previously mentioned, we have to deal with the growing number of standards. Succession in standardization is usually a problem [EGY 02]. The advantages of improvements are weighed against those of compatibility. This evolution could be easily explained because the priorities in standardization could change, so the rules for developing standards are revised. Standards could be updated or become obsolete. This is part of the dynamics of standardization, irrespective of the area of interest.

A number of critical problems in the field of Information and Communication Technology (ICT) occur because many standards have functional equivalents. That is, they address the same problem and offer similar functionalities. Sometimes competition between them leads to "standards wars".

Completeness has been identified as an important design criterion for interchange formats but less attention has been paid to the sequential relations between standards, that is, the way that previous standards (i.e. predecessors) are revised and succeeded by new standards (i.e. successors). Succession in standardization implies change and renewal. Renewal comes in various shapes: new editions, revisions (i.e. new versions, technical corrigenda, amendments, annexes etc.) and new standards. The successor addresses the same area, and is an improvement on its predecessor. It is designed to succeed and thus take over the predecessor's role. New entrants in the market (standards users) naturally prefer and implement the successor.

Those who standardize the successor may or may not seek compatibility with the predecessor. They usually do, and need to have good reasons not to seek compatibility (e.g. technically impossible or a change in the product). There are many kinds of compatible successors. The most common one is the downward compatible successor, which replaces the more elaborated original standard.

If the successor standard is compatible, compliant technologies should be able to work together with products that interoperated with its predecessor. Such is typically the aim when the successor is a new edition or a minor revision of a standard. Examples are incremental innovations: the improvements made are part of normal problem solving. Dilemmas regarding compatible succession are often of a mixed socio-technical nature (i.e. technical, implementation, esthetic, etc.). A characteristic of dilemmas is that the conflicting arguments are both persuasive.

Within the framework of the management of multilingual content, some standards as TMX - related to translation memories - and XLIFF – related to the activity of localization - have right now some dedicated software, as well as several resources respecting their respective recommendations. Although they are not at all out of date, these standards however cannot satisfy the needs being born from new information technologies.

Within ISO's TC34/SC4 "Linguistic Resources Management", a group of experts is currently working on the specification of a new standard aiming at, on the one hand, covering the whole functionalities of the above mentioned standards, and on the other hand, satisfying the linguistic enrichment and the interoperability of multilingual data: the "Multi Lingual Information Framework" (MLIF) is currently being developed. MLIF is an ISO's "Approved Work Item" [AWI 24616] from TC37/SC4, working group WG3 "Multilingual Text Representation" (see figure 2).

	TMX	XLIFF	MLIF
Related Domains	Translation, Computer Assisted Translation (CAT)	Localization, Computer Assisted Translation (CAT), word processing program, terminology management systems, multilingual dictionary, or even raw machine translation output	Localization, Computer Assisted Translation (CAT) tool, word processing program, terminology management systems, multilingual dictionary, or even raw machine translation output, e-learning, multimedia applications, …
Global Information (ex : date, author, …)	Available on the head and on the meaning units.	Global	Available on the head and at the non-terminal levels of the model
Multilingual data	Multilingual	Bilingual	Multilingual
Possibility to use additional linguistic information	No	No	Yes. Terminological data, Lexical data, Morphological data, …
Segmentation	Textual segments	Blocks, paragraphs, sentences, or phrases	Blocks, paragraphs, sentences, or phrases
Internal or external references	External	External	Internal (ex: anaphoric references, ellipse, …) External (ex: data bases, terminology, …)
Missing translation	Ignored	Ignored	Indicated

Table 2. Global comparison of TMX, XLIFF and MILF

3. Contribution of standards

As we previously discussed, the life cycle of standards is conditioned by new needs, adaptations to technologies, or new trades. It is important to determine the fields concerned, as well as the concerned people and their work practices. The work practices make it possible to determine the minimum lattice of information to represent, as well as the set of features needed to specify for rendering this information relevant within a given framework. A multilingual software product should aim at supporting document indexing, automatic and/or manual computer-aided translation, information retrieval, subtitle handling for multimedia documents, etc. Dealing with multilingual data is a three steps process: production, maintenance (i.e. update, validation, correction) and consumption (i.e. use). To each one of these steps correspond a specific user group, and a few specific scenarios. It is important to draw up a typology of the potential users and scenarios of multilingual data by considering the various points of view: production, maintenance, and consumption of these data. Indeed, we are not just trying to develop a new standard, nor we are aiming at competing with any existing standard. Rather, we are trying to specify a high-level model allowing to represent and to integrate the whole set of actors of the translation and localization community. This is the reason why the participation of these actors to our work is a fundamental issue in the aim of the creation of a successful new standard.

The development of scenarios considers the possible limits of a multilingual product, thus the adaptations required. Normalization will also allow the emergence of new needs (e.g. addition of linguistic data like grammatical information). Scenarios help to detect useless or superseded features that may not be necessary to implement within standardized software applications. These scenarios must also be based on well “on work practices” while also making it possible to envisage some possible extensions. Normalization will facilitate the dissemination (i.e. export multilingual data) as well as the integration of data (i.e. import of multilingual data from external databases).

Providing normalized multilingual products and data must be considered as a way, for a scientific community, to be well known, to be acknowledged.

4. Terminology and methodology of normalization

In this section, we will introduce our methodology of normalization as well as the terminology related to our standardization activities. Like any other technical field, standardization has its own terminology and its own specific rules. As with the "Terminological Markup Framework" TMF [ISO 16642] in terminology, MLIF will introduce a structural skeleton (metamodel) in combination with chosen "Data Categories", as a means of ensuring interoperability between several multilingual applications and corpora. Each type of standard structure is described by means of a three-tiered information structure that describes:

- a metamodel, which represents a hierarchy of structural nodes which are relevant for linguistic description;
- several specific information units that can be associated with each structural node of the metamodel;
- several relevant annotations that can be used to qualify some part of the value associated with a given information unit.

4.1. What is a metamodel?

A metamodel does not describe one specific format, but acts as a high level mechanism based on the following elementary notions: structure, information and methodology. The metamodel can be defined as a generic structure shared by all other formats and which decomposes the organization of a specific standard into basic components. A metamodel should be a generic mechanism for representing content within a specific context. Actually, a metamodel summarizes the organization of data.

The structuring elements of the metamodel are called "components" and they may be "decorated" with information units. A metamodel should also comprise a flexible specification platform for elementary units. This specification platform should be coupled to a reference set of descriptors that should be used to parameterize specific applications dealing with content.

4.2. What is a data category?

A metamodel contains several information units related to a given format, which we refer to as "Data Categories". A selection of data categories can be derived as a subset of a Data Category Registry (DCR) [ISO 12620]. The DCR defines a set of data categories accepted by an ISO committee. The overall goal of the DCR is not to impose a specific set of data categories, but rather to ensure that the semantics of these data categories is well defined and understood.

A data category is the generic term that references a concept. There is one and only one identifier for a data category in a DCR. All data categories are represented by a unique set of descriptors. For example, the data category */languageIdentifier/* indicates the name of a language which is described by 2 [ISO 639-1] or 3 [ISO 639-2] digits. A Data category Selection (DCS) is needed in order to define, in combination with a metamodel, the various constraints that apply to a given domain-specific information structure or interchange format. A DCS and a metamodel can represent the organization of an individual application, the organization of a specific domain.

4.3. Methods and representation

The way to actually implement a standard is to instantiate the metamodel in combination with a set of chosen data categories (DCS). This includes mappings between data categories and the vocabularies used to express them (e.g. as an XML element or a database field). Data category specifications are, firstly used to specify constraints on the implementation of a metamodel instantiation, and secondly to provide the necessary information for implementing filters that convert one instantiation to another. If the specification also contains styles and vocabularies for each data category, the DCS then contributes to the definition of a full XML

information model which can either be made explicit through a schema representation (e.g. a W3C XML schema), or by means of filters allowing the production of a "Generic Mapping Tool" (GMT) representation. The architecture of the metamodel, whatever the standard we want to specify, remains unchanged. What are variable are the data categories selected for a specific application. Indeed, the metamodel can be considered in an atomic way, in the sense that starting from a stable core, a multitude of data can be worked out for plural activities and needs.

5. Specifying the Multi Lingual Information Framework

Linguistic structures exist in a wide variety of formats ranging from highly organized data (e.g. translation memory) to loosely structured information. The representation of multilingual data is based on the expression of multiple views representing various levels of linguistic information, usually pointing to primary data (e.g. part of speech tagging) and sometimes to one another (e.g. references, annotations). The following model identifies a class of document structures that could be used to cover a wide range of multilingual formats, and provides a framework that can be applied using XML.

All multilingual standards have a rather similar hierarchical structure but they have, for example, different terms and methods of storing metadata relevant to them. MLIF is being designed in order to provide a generic structure that can establish a basic foundation for all these standards. From this high-level representation we are able to generate, for example, any specific XML-based format: we can thus ensure the interoperability between several standards and their committed applications.

5.1. Description of MLIF[1]

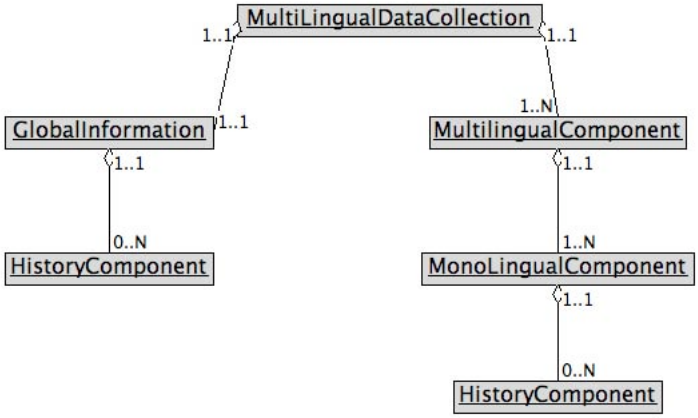

Figure 1. Hierarchical representation of MLIF

The MLIF metamodel is constituted by the following components:

Multi Lingual Data Collection
Represents a collection of data containing global information and several multilingual units.

Global Information
Represents technical and administrative information applying to the entire data collection. Example: title of the data collection, revision history, …

Multi Lingual Component
This component represents a unique multilingual entry.

[1] This presentation of MLIF is based on the "New Work Item Proposal" (NWIP) submitted to ISO TC37/SC4 "Linguistic Resources Management". This NWIP was approved after an international ballot process in August 2006. MLIF is now an ISO's "Approved Work Item" [AWI 24616].

Mono Lingual Component
Part of a multilingual component containing information related to one language.

History Component
This generic component allows modifications to be traced on the component it is anchored to (i.e. versioning).

In order to provide a larger description of the linguistic content, the MLIF metamodel allows anchoring of other metamodels, such as MAF (morphological description), SynAF (syntactical annotation), TMF (terminological description), or any other metamodel based on ISO 12620:2003.

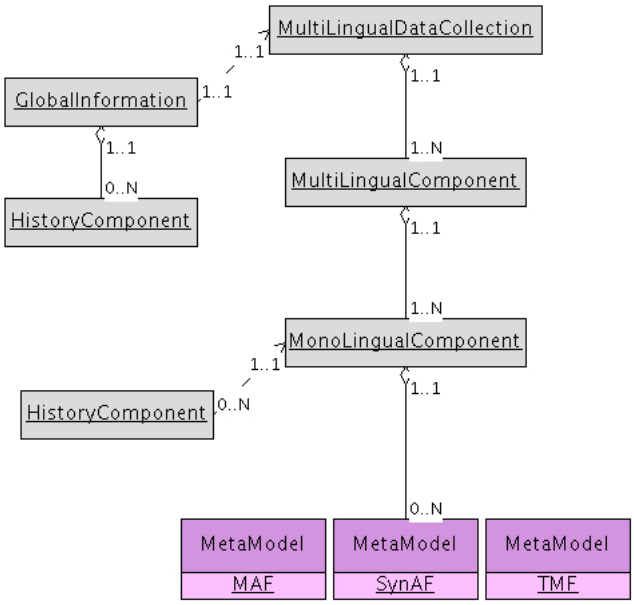

Figure 2. MLIF metamodel

All the different models have very similar hierarchical structure but they have different terms and methods of storing metadata relevant to them in particular. MLIF provides a generic structure that can establish basic foundation for all other models. This model will provide flexibility to make any element or attribute of this format to be defined explicitly or not. If the value is not defined explicitly it will take default value. Most of the models will also define their own elements and attributes those will fit into this using extensibility that is one of the basic requirements of MLIF model.

5.2. Data Categories

Global Information
/source/
- "A complete citation of the bibliographic information pertaining to a document or other resource".
- Reference to a resource from which the present resource is derived.

/sourceType/
- "In multilingual and translation-oriented language resource or terminology management, the kind of text used to document the selection of lexical or terminological equivalents, collocations, and the like".
    - "Both parallel and background texts serve as sources for information used in documenting multilingual terminology entries. "[ISO12620]

/sourceLanguage/
- "In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written".

/projectSubset/
- An identifier assigned to a specific project indicating that it is associated with a term, record or entry.

/subjectField/
- "A field of special knowledge."

Multilingual Component
/identifier/
- A unique name [source:IMDI_Source_Tag]
    - Dublin Core equivalent: DC:Identifier [source:IMDI_Source_Tag]

Monolingual Component
/languageIdentifier/
- A unique identifier in a language resource entry that indicates the name of a language.

/primaryText/
- Linguistic material which is the object of study.

/sourceLanguage/ :
- "In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written".
    - The identifiers specified in ISO 639 should be used:
        - en = English
        - fr = French
        - es = Spanish (Español)
        - …

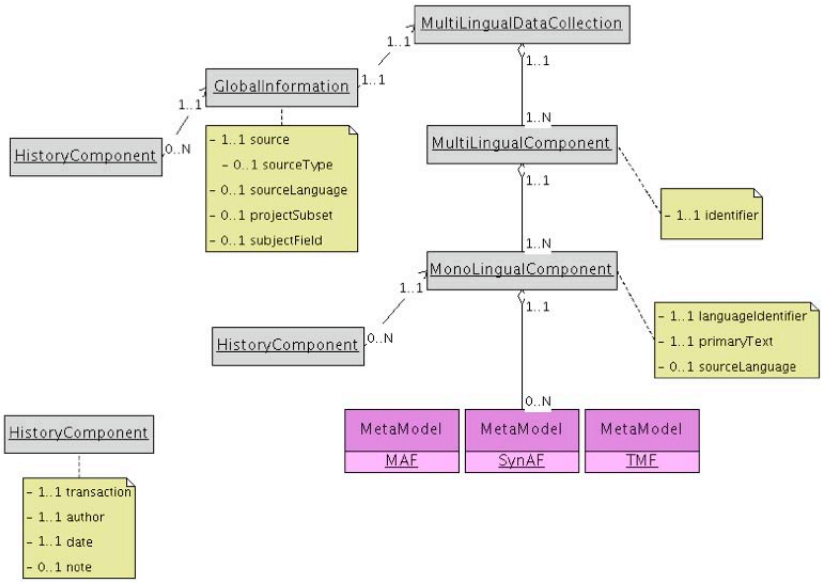

Figure 3. MLIF metamodel with selected "data categories"

5.3. Introduction to GMT

GMT "Generic Mapping Tool" can be considered as an XML canonical representation of the metamodel. The hierarchical organization of the metamodel and the qualification of each structural level can be realized in XML by instantiating the abstract structure shown above (see Figure 3) and associating information units to this structure. The metamodel can be represented by means of a generic element <struct> (for structure) which can recursively express the embedding of the various representation levels of a MLIF instance. Each structural node in the metamodel shall be identified by means of a type attribute associated with the <struct> element. The possible values of the type attribute shall be the identifiers of the levels in the metamodel (i.e., Multilingual Data Collection, Global Information, Multilingual Component, Monolingual Component, Linguistic Element).

Basic information units associated with a structural skeleton can be represented using the <feat> (for feature) element. Compound information units can be represented using the <brack> (for bracket) element, which can itself contain a <feat> element followed by any combination of <feat> elements and <brack> elements. Each information unit must be qualified with a type attribute, which shall take as its value the name of a standard data category [ISO 12620] or that of a user-defined data category.

5.4. A practical example: MLIF and TMX

Now, we will use a very simple TMX example (see Figure 6) for the purpose of showing how MLIF can be mapped to other formats. As we discuss further details about MLIF, it will be clear that all features can be identified and mapped through data categories.

```
<tmx version="1.3">
    <header
        segtype="sentence"
        creationdate="19970101T163812Z"
    2   creationid="ThomasJ"
        changedate="19970314T023401Z"
        changeid="Amity">
        <note>This is an example of TMX</note>
    </header>
    <body>
        <tu tuid="0001">
            <note>Text of a note at the TU level.</note>
            <prop type="x-Domain">Translation</prop>
            <prop type="x-Project">LORIA</prop>
1           <tuv
                xml:lang="EN"
            4   creationdate="19970212T153400Z"
                creationid="BobW">
                5 <seg>The little cat is dead.</seg>
            3   </tuv>
            <tuv
                xml:lang="FR"
                creationdate="19970309T021145Z"
                creationid="BobW"
            4'  changedate="19970314T023401Z"
                Changeid="ManonD">
                5' <seg>Le petit chat est mort.</seg>
            </tuv>
        </tu>
    </body>
</tmx>
```

Figure 4. Part of a TMX document

In Figure 6, we found structural elements of TMX : 1 represents the <tmx> root element, 2 the <header> element, 3 represents a <tu> element, 4 and 4' represent respectively the English and French <tuv> element. Next, we will match these structural elements of TMX with the metamodel of MLIF (see Figure 7).

TMX structure	MLIF component
1 <tmx>	Multilingual Data Collection
2 <header>	Global Information
3 <tu>	Multilingual Component
4 <tuv>	Monolingual Component

Table 3. Matching TMX with MLIF components

Then, we will tag each element descriptor of TMX into 3 types: attribute, element or typed element. All these descriptors will be standardized into a MLIF descriptor element (i.e. a data category). For example the TMX "xml:lang" attribute will be next matched with the data category named /languageIdentifier/.

TMX descriptor	Type	Data Categories
<note>	*element*	*/note/*
<prop type= ''x-project''>	*typed element*	*/projectSubset/*
xml:lang	*attribute*	*/languageIdentifier/*
tuid	*attribute*	*/identifier/*
<seg>	*element*	*/primaryText/*

Table 4. Typing of descriptor elements and matching with data categories

Finally, the mapping of TMX elements into MLIF elements is represented in the following GMT file (see Figure 9). Note that this GMT file is nothing but a canonical representation of a MLIF document.

6. Interoperability, adaptation, and some other important issues

The principles of TMF[ISO 16642] and, by extension those of MLIF, can be translated in the form of formal constraints making it possible to precisely specify the informational cover of a given model and, by doing this, to compare two models with the objective to determine the precise conditions of interoperability. Next, we must be able to translate these constraints into XML-related structures, in order to provide true data models, in particular those which will be used later for the construction of software tools being able to handle and, to exchange multilingual data. This is the stage that we present in this last section, while seeking again to define a general enough framework that would be able to cover a broad variety of applications.

6.1. Multiple utilizations and adaptation.

Roughly, as one single format for representing multilingual data is likely always to be regarded either as too complex, or like not answering exactly such or such particular need [ROM 06]. We actually wish to show that it is possible to consider a family of formats, within a sufficiently accessible framework of representation, that many categories of users can easily adapt the groundwork suggested to their own needs. We are thus positioned in the continuity of thoughts carried out by the actors of standardization themselves [ROM 01] [BAU 04] which consider that the proposal of a framework of standardization is not incompatible with the identification of operations of adaptation of the standards by extension or restriction of a data model. Actually, it is a question of transposing, within the field of the representation of data, the concept of "subsumption" of models. The objective is to arrive to the definition of a true platform of specification of multilingual data which is capable of guaranteeing that the same element of data (i.e. utilization or reference to data categories)

will be represented in an identical way in two different applications and thus to avoid the trap which locks up the standard in a yoke of too specific applications.

The choices that we present here were guided by another important concern, namely the need for foreseeing the potential integration of multilingual data within a broader framework of representation of textual documents. Indeed, we think that multilingual data must not be dissociated from the documents where they are used. Concretely, that means that within multilingual textual documents, we must be able for example, to annotate and to connect all multilingual terms being used, or be able to establish links with the entries of some terminological database.

From this point of view, one can of course mention the official standardization documents of ISO which integrates, as a mandatory section, the whole set of the terms and definitions used within the body of these documents.

```
<struct type="Multilingual Data Collection">
      <struct type="Global Information">
            <brack>
                  <feat type="transaction">creation</feat>
                  <feat type="date">19970101T163812Z</feat>
                  <feat type="author">ThomasJ</feat>
            </brack>
            <brack>
                  <feat type="transaction">modification</feat>
                  <feat type="date">19970314T023401Z</feat>
                  <feat type="author">Amity</feat>
            </brack>
            <feat type="note"> This is an example of TMX</feat>
      </struct>
      <struct type="Multilingual Component">
            <feat type="identifier">0001</feat>
            <feat type="note"> It's just an example</feat>
            <feat type="subjectField">Translation</feat>
            <feat type="projectSubset">LORIA</feat>
            <struct type="Monolingual Component">
                  <feat type="languageIdentifier">EN</feat>
                  <brack>
                        <feat type="transaction">creation</feat>
                        <feat type="date">19970212T153400Z</feat>
                        <feat type="author">BobW</feat>
                  <brack>
                  <struct type="Lingustic Segment">
                        <feat type="sentence">The little cat is dead.</feat>
                  </struct>
            </struct>
            <struct type="Monolingual Component">
                  <feat type="languageIdentifier">FR</feat>
                  <brack>
                        <feat type="transaction">creation</feat>
                        <feat type="date">19970309T021145Z </feat>
                        <feat type="author">BobW</feat>
                  <brack>
                  <brack>
                        <feat type="transaction">modification</feat>
                        <feat type="date">19970314T023401Z </feat>
                        <feat type="author">ManonD</feat>
                  </brack>
                  <struct type="Lingustic Segment">
                        <feat type="sentence">Le petit chat est mort.</feat>
                  </struct>
            </struct>
      </struct>
</struct>
```

Figure 5. GMT representation

Within another different context (i.e., captions within DVD movies), multilingual textual information may also need to be structured in different ways (i.e. paragraphs, sentences, but also surface annotations) that those related to the field of the translation memories or localization. It is thus important, since that seems to be

possible, to offer a representation of multilingual data which is integrated within a broad framework of representation of textual information.

6.2. Working within the scope of the TEI.

The TEI (Text Encoding Initiative - http://www.tei-c.org), is an international initiative which, since 1987, gathers together most of the actors who have to manage great projects of textual data. The TEI covers many applicative domains as, prose, poetry, theatre, manuscripts, dictionaries and is strongly concerned with multilinguality issues. Today, the TEI offers a platform of specification, ODD (One Document Does it all), which is an ideal framework to implement the approach which we defend here, namely the definition of a family of compatible models.

ODD [BEAM 04] is a language of data specification which is based on the principle of "literate programming" which combines descriptive elements with formal elements in order to provide, starting from a single source, at the same time: a diagram allowing to control the effective syntax of a document, and a documentation providing to a user the fine semantics of the objects defined in the specification. Without going here into too technical details, we show here the two essential characteristics of ODD, namely the concepts of modules and classes, for providing next, some indications on the specification of objects XML themselves.

The ODD platform makes it possible to organize any documentary structure as a combination of one or more modules joining together a coherent unit of elements and classes. The directives of the TEI propose modules thus making it possible to represent the heading of a document, the common elements (e.g. divisions) to all types of documents, the elements specific to theatre, poetry, etc. A user can thus decide to use the basic modules allowing to represent simple textual data and to associate it to a terminological module, in order to insert descriptions of terms in the body text.

Two principal types of objects are described inside a module, elements and classes. The classes allow to gather elements having a syntactic behavior or a similar semantics. Thus, all elements giving any morpho syntactic indication within a dictionary or a terminology database, belong to the class "model.morphLike". This way, if one wishes to integrate all these elements within a model of contents, it is enough to refer to the related class. In a complementary way, if one wish to add a morpho syntactic descriptor, it is enough to add an element to the class.

For the definition of the models of contents, the TEI is based on elementary fragments of RelaxNg diagrams which are then combined to generate complete RelaxNg diagrams, but also DTD's XML, or W3C XML Schemas.

7. Perspectives

A first implementation of MLIF within multimedia applications has been used within several prototypes developed in the framework of the ITEA Passepartout project (ITEA 04017). Within these prototypes some basic scenarios have been implemented: MLIF has been associated to XMT (eXtended MPEG-4 Textual format) and to SMIL (Synchronized Multimedia Integration Language). Our main objective in this project has been to associate MLIF to multimedia standards (e.g. MPEG-4, MPEG-7, and SMIL) in order to be able, within multimedia products, to represent and to handle multilingual content in an efficient, rigorous and interactive manner (see Figure 10).

At present, we are also working on the issue of proposing several compatibility-related filters with ODD. Within a more practical framework, we are also developing a PHP multilingual gateway: all multilingual textual information is directly encoded by using MLIF.

8. Conclusion

In this chapter, we have analyzed why normalization is a key issue within translation and localization activities. Within this context, we have also shown that it is possible to define, in a coherent way, the various phases of designing a general normalized framework for the handling and representation of multilingual textual data within localization and translation activities. The MLIF "Multi Lingual Information Framework" ISO [AWI 24616] is being developed this way. As we have clearly indicated, MLIF must be considered as a unified conceptual representation of multilingual content and is not intended to substitute or to compete with

any existing standard. MLIF is being designed with the objective of providing a high-level common conceptual model and a platform allowing interoperability among several translation and localization standards, and by extension, their committed tools. We think that this platform is a continuum between a truly linguistic work of collecting multilingual data and the development of a data-processing software environment intended to accommodate such data.

Figure 6. Dynamic and Interactive displaying of multilingual subtitles and multilingual textual information

MLIF continues to evolve and within the next months an ISO's "Committee Draft" (CD) should be published. This CD will reflect comments and remarks from the MLIF's Experts Committee so the metamodel and related data categories will certainly be modified. Also, as we have mentioned, our current research tends to prove that the specification of a format of representation such as MLIF can be elegantly associated with a broader normative approach, such as the TEI.

Last but not least, it is important to point out once again that MLIF has been successfully associated to multimedia standards such as XMT and SMIL. In our opinion, text must no longer be considered as the "ugly duckling" of multimedia.

9. References

[BAU 04] BAUMAN S., FLANDERS J., « Odd Customizations », *Extreme Markup Languages*, 2-6 août 2004 Montréal, Canada, 2004.

[CRU 04] S. Cruz-Lara, S. Gupta, & L. Romary (2004) *Handling Multilingual content in digital media: The Multilingual Information Framework.* EWIMT-2004 European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. London, UK.

[EGY 02] Egyedi, T.M. & A.G.A.J. Loeffen (2002b). 'Succession in standardization: grafting XML onto SGML', Computer Standards & Interfaces, 24, pp.279-290.

[GOM 01] Gómez, Carmen et María Pinto (2001): *La normalisation au service du traducteur*, , in META, XLVI, no 3, p. 564.
[HUR 96] Hurtado Albir, A. (ed.) (1996), « La enseñanza de la traducción », Castellón : Universidad Jaume I.
[ISO 12620] ISO 12620 (1999) : Computer applications in terminology -- Data categories,
[ISO 16642] TMF. ISO 16642 (2003) *Computer applications in terminology -- Terminological markup framework,* Genèva, International Organization for Standardization
[ISO 5963] ISO 5963 (1985) Documentation -- Methods for examining documents, determining their subjects, and selecting indexing terms, Geneva, International Organization for Standardization
[ISO 5964] ISO 5964 : 1985. *Documentation. Principes directeurs pour l'établissement et le développement de thésaurus bilingues*, Geneva, International Organization for Standardization
[ISO 639-1] ISO 639-1 (2002) *Code for the representation of names of languages – Part 1: Alpha-2 code,* Geneva, International Organization for Standardization
[ISO 639-2] ISO 639-2 (1998) *Code for the representation of names of languages – Part 2: Alpha-3 code,* Geneva, International Organization for Standardization
[ISO 9126] ISO/IEC 9126 : 1991. *Technologies de l'information. Évaluation des produits logiciels. Caractéristiques de qualité et directives d'utilisation.*
[ISO WD 24611] MAF: ISO/TC37/SC4/WG2 WD 24611 Morphosyntactic Annotation Framework.
[ISO WD 24615] SynAF: ISO/TC37/SC4/WG2 WD 24615 Syntactical Annotation Framework.
[ITEA 04017] ITEA "Information Technology for European Advancement". Passepartout project "Exploitation of advanced AV content protocols (MPEG 4/7)" ITEA 04017.
[ITS] ITS. W3C (2003) Internationalization Tag Set (i18n). http://www.w3.org/TR/its/
[PIN 93] Pinto, M. (1993). *Análisis documental. Fundamentos y procedimientos*, 2ª ed. rev. y auhm., Madrid, Eudema.
[ROM 01] ROMARY L., Un modèle abstrait pour la représentation de terminologies multilingues informatisées È, *Cahiers GUTenberg*, 39-40, p. 81-88, mai 2001.
[ROM 06] Romary L., Salmon-Alt S. Kramer I., Roumier J.(2006), Gestion de données terminologiques: principes, méthodes, modèles. In: Terminologie et accès à l'information. Hermes, collection Techniques et traités des sciences et techniques de l'information, Paris.
[SMIL] Synchronized Multimedia Integration Language (SMIL 2.0) . World Wide Web Consortium. http://www.w3.org/TR/smil20/
[SRX] SRX. Segmentation Rules eXchange. SRX 1.0 Specification. Oscar Recommendation 20 April, 2004. http://www.lisa.org/standards/srx/srx.html .
[TMX] Oscar / Lisa (2000) Translation Memory eXchange, http://www.lisa.org/tmx.
[XLIFF] XLIFF. (2003). XML Localisation Interchange File Format.http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff.
[XMT] XMT. extended MPEG-4 Textual format. ISO/IEC FCD 14496-11, Information technology -- Coding of audio-visual objects -- Part 11: Scene description and application engine; ISO/IEC 14496-11/Amd 4, XMT & MPEG-J extensions.
[ZER 05] TMX and SRX Exchanging TM Data. Angelika Zerfass, Consultant and Trainer for Translation Tools. LRC-X Conference, University Of Limerick, Ireland. 13-14 September 2005.