An abstract model for the representation of multilingual terminological data: TMF - Terminological Markup Framework

Laurent Romary

An abstract model for the representation of multilingual terminological data: TMF – Terminological Markup Framework

Laurent Romary
Laboratoire LORIA
Campus Scientifique
BP 239, F-54506 Vandoeuvre-lès-Nancy
Laurent.Romary@loria.fr

Résumé. Nous présentons un modèle abstrait de représentation de terminologies multilingues informatisées en XML défini dans le cadre du comité technique 37 de l'ISO. Il repose sur une méthodologie qui distingue d'une part la structure générale d'une base terminologique, et d'autre part les informations (catégories de donnée) qui servent à décrire les différents niveaux de cette structure.

Summary. We are introducing an abstract model that has been developed by Technical Committee 37 of ISO for representing computerized multilingual terminologies. . It relies on a methodology that makes an essential distinction between the general structure of a terminological database and the elementary information units (data categories) that are used to describe this structure.

1. Introduction[1]

This paper intends to show how to describe the structure of a terminological database independently of its implementation, which would involve, for instance, a pre-defined XML format expressed as a DTD or XML Schema, or a database structure expressed as an entity-relationship model. This work has two aims. Firstly, it corresponds to the need to describe and compare existing terminological interchange formats such as MARTIF [iso12200] or Geneter [Le Meur, 1998], in terms of their informational coverage and the conditions of interoperability between these formats and hence the source data generated in them. Such an attempt should lead to more general principles and methods for analyzing existing terminological databases and mapping them onto any chosen terminological interchange format. One of the issues here is to provide a uniform way of documenting such databases considering the heterogeneity of both their formats and their descriptors. Secondly, we 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. There should be nothing to prevent a terminological project from designing a very simplified format that would only allow a terminologist to describe five types of information units expressed as XML elements and attributes (<id>, <générique>, <terme>, <catégorie>, and 'langue'), as in the following XML excerpt[2], and still be able to

[1] This work is conducted under the auspices of Technical Committee 37 of ISO and with the support of the European Union, through the co-funded HLT-Salt project (http://www.loria.fr/projets/SALT).

[2] We have limited our database to one single entry (or 'notion'). This entry does, however, make reference to the one which would express the broader concept, by means of an XPointer (see http://www.w3.org/XML/Linking) expression ("#notion[id='2']", i.e. the 'notion' element having a child element 'id' whose content is the string "2").

compare or complement them with information coming from a wider terminological database such as Eurodicautom[3] (available in MARTIF for instance).

```
<?xml version="1.0" encoding="iso-8859-1"?>
<base>
   <notion>
      <id>1</id>
      <générique target="#notion[id='2']">Insecte</générique>
      <expression langue="fr">
         <description>
            <terme>Abeille</terme>
            <catégorie>Nom</catégorie>
         </description>
      </expression>
      <expression langue="en">
         <description>
            <terme>Bee</terme>
            <catégorie>Nom</catégorie>
         </description>
      </expression>
   </notion>
</base>
```

This work is also motivated by the need to provide more connections between, on the one hand, terminological databases as handled in translation companies, major organizations like the EU or the localization industry, and on the other hand, other lexical resources dedicated, for instance, to machine translation or natural language processing, which are expressed along different modes of representation than those known in the terminological field (e.g. OLIF[4] for MT or Genelex for NLP [Antoni-Lay et alii, 1994]). We will show in this paper that such an ambitious objective is achieved by decomposing information structures into the macro-structure of the terminological database (what we will call the structural skeleton), and the elementary units of information (i.e. data categories) that can be attached to the structural skeleton. In doing so, we somehow project, in the terminological domain, a whole stream of recent works [Wright & Melby, 1999; Budin & Melby, 2000; Ide et al, 2000, Ide & Romary, 2001] attempting to devise a conceptual framework for the description of lexical resources and, more globally, linguistic structures, expressed as databases or annotations added to other resources.

2. General principles

A consequence of the above is that there is no need to propose yet another format for terminological data. Existing practices can be built on to cover a whole range (or family) of formats, which, as soon as they are shown to be compatible with the framework, become automatically comparable. Indeed, one important result of such work is the possibility to formalize the various components of a terminological database and derive, with as little human intervention as possible, model checkers for a given format, or filters from one format to another.

In this section, we describe the Terminological Markup Framework (TMF)[5], which, as a possible future ISO 16642 standard, allows one to describe a potentially infinite set of Terminological Markup Languages (TML), that can be expressed for the interchange of computerized terminological data using, for example, XML. TMF does not describe one specific format, but acts as a kind of meta-model based on the following elementary notions:

---

[3] See http://eurodic.ip.lu

[4] See http://www.olif.net

[5] See http://www.loria.fr/projets/TMF/ for further documents, samples and software.

- The meta-model: a unique information structure shared by all TMLs and which decomposes the organization of a terminological database into basic components as shown in figure 1. This model is in keeping with the traditional concept-oriented view of a terminological entry dating back to Wüster's early works [Picht & Schmitz, 2001] and widely adopted in the community;
- Information units (which we refer to as data categories): derived as a subset of a Data Category Registry (DCR, see below) as needed for a given format. This may also contain additional data categories specifically defined for the current application, which may hinder interoperability with other formats;
- Methods and representations: the means to actually implement the TML by instantiating the structural skeleton in combination with the chosen data categories, for instance by automatically generating an XML schema for the TML. This comprises the mappings between data categories and the vocabularies used to express them (e.g. as an XML element or a database field).

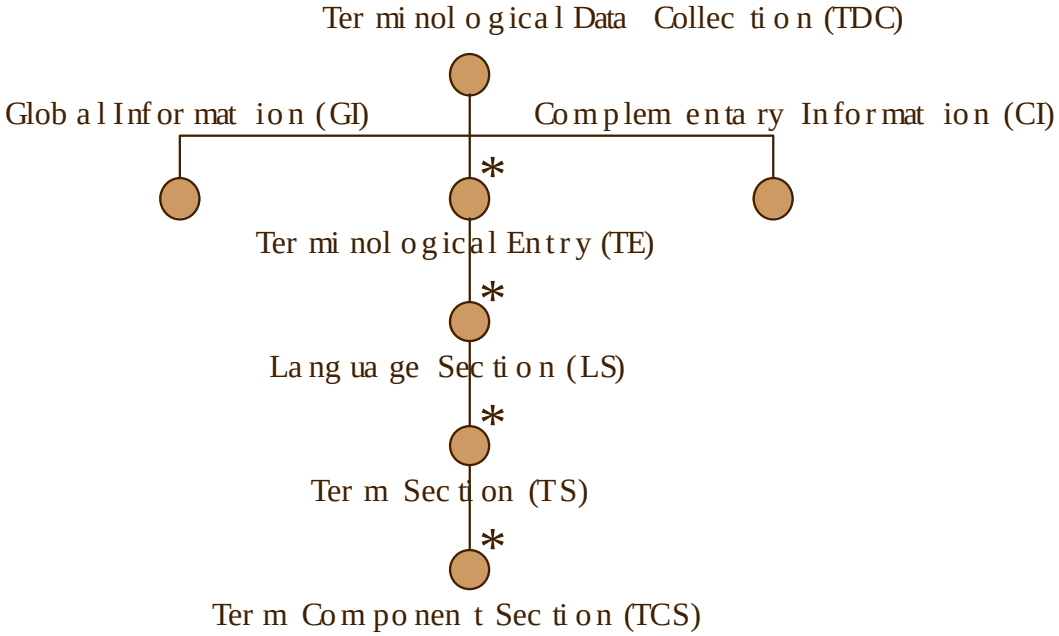

Figure 1: The meta-model of a terminological database.

The final component of TMF is the definition of a simplified XML application that can be used to map any given format, or TML, onto the abstract components of TMF. This format, also known as GMT (Generic Mapping Tool), is based on a reduced set of XML elements and attributes, which, as we shall see, serve as containers for nodes of the structural skeleton (identified by <struct> tags) and data categories (identified by <feat> tags). As shown in figure 2 below, any mapping between two TMLs can be implemented as the composition of two elementary mappings through GMT. In the same way, our experience within the SALT project has shown that GMT is an ideal tool to map a traditional relational terminological database onto a given TML.

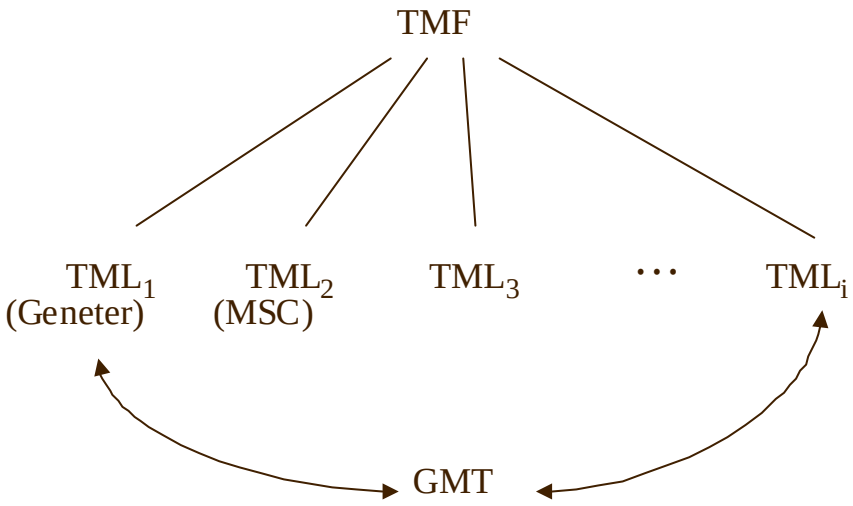

Figure 2: Mapping in the family of TML formats

3. Going through an example

To illustrate the principles of TMF, we apply its methodology to the decomposition of a typical terminological entry as expressed in MSC, a variant of ISO 12200 (MARTIF):

```
<termEntry id='ID67'>
      <descrip type='subjectField'>manufacturing</descrip>
      <descrip type='definition'>A value between 0 and 1 used in ...
      </descrip>
      <langSet lang='en'>
            <tig>
                  <term>alpha smoothing factor</term>
                  <termNote type='termType'>fullForm</termNote>
            </tig>
      </langSet>
      <langSet lang='hu'>
            <tig>
                  <term>Alfa simitisi tenyezo</term>
            </tig>
      </langSet>
</termEntry>
```

In the preceding excerpt, one can distinguish two different types of informational objects. On the one hand, we can identify (by means of an underlined bold script) three structural elements (<termEntry>, <langSet>, <tig>) which do not provide any specific information from a terminological perspective, but rather contribute to the organization of the terminological entry. The corresponding XML information structure, called *XML outline*, can easily be mapped onto the three corresponding levels of the meta-model shown above, namely: Terminological Entry (TE), Language Section (LS) and Term Section (TS).

The remaining information units, whether expressed as XML 'elements'(<term>), 'typed elements' (<descrip>, <termNote>) or as XML 'attributes' (id, lang), directly contribute to the description of the current entry and can be associated with, for instance, data categories described in ISO 12620 as shown in table 1 below.

Martif object	**Style**	**ISO 12620 Identifier**	**ISO 12620 Name**
<term>	Element	ISO12620-A01	Term

<descrip type='subjectField'>	Typed element	ISO12620-A04	Subject field
<descrip type='definition'>	Typed element	ISO12620-A0501	Definition
<termNote type='termType'>	Typed element	ISO12620-A0201	Term type
'id'	Attribute	ISO12620-A1015	Entry identifier
'lang'	Attribute	ISO 12620A100701	Language Identifier

Table 1: Mapping MARTIF XML objects onto ISO 12620 data categories

One possible representation of this decomposition is shown using the GMT format, as follows:

```
<struct type="TE">
      <feat type="id">ID67</feat>
      <feat type="subjectField">manufacturing</feat>
      <feat type="definition">A value between 0 and 1 used in
      ...</feat>
      <struct type="LS">
            <feat type="lang">en</feat>
            <struct type="TS">
                  <feat type="term">alpha smoothing factor</feat>
                  <feat type="termType">fullForm</feat>
            </struct>
      </struct>
      <struct type="LS">
            <feat type="lang">hu</feat>
            <struct type="TS">
                  <feat type="term">Alfa simitisi tenyezo</feat>
            </struct>
      </struct>
</struct>
```

In GMT, each node of the structural skeleton is expressed as an instantiation of the sole element <struct> with the type identifier used to signify the level in the meta-model, and each information unit is expressed by means of the <feat> element where the type signifies the data category name taken from ISO 12620. (see below DCName in the formal description of data categories)[6].

4. Interoperability between two terminological formats

Given two TMLs defined in accordance with TMF, eliciting their conditions of interoperability reduces to a comparison of their respective use of data categories since they share exactly the same meta-model[7]. Indeed, TMF can be used to make a precise diagnosis of the amount of data that will be preserved or lost when going from one TML to another, what we could call the *bandwidth of interoperability*.

To illustrate this we can compare the simple example presented in the introduction, corresponding to a first TML (TML_1), and the MARTIF excerpt in the preceding section (TML_2). We will make the assumption that TML_1 and TML_2 are defined as being the smallest

[6] In order to account for all the descriptive power of TMF, the GMT DTD actually comprises two other elements, namely <brack> to group co-occurring <feat>'s, and <annot>, to annotate textual content of a <feat>. It also contains additional attributes for language identification (xml:lang) and linking mechanisms ('source' and 'target')

[7] ISO/CD 16642 actually identifies more precise conditions related to data category values, data category anchoring on the levels in the meta-model etc. For instance, a definition used at the Terminological Entry level in one TML will not be received by a TML which expects definitions only at the Language Section level.

formats encompassing the examples given in this paper, in terms of the data categories being used.

As a first step, we map the XML objects of TML_1 onto data categories of ISO 12620, just as we did for the MARTIF example. The result is shown in table 2.

XML object	Style	ISO 12620 Identifier	ISO 12620 Name
<terme>	Element	ISO12620-A01	Term
<générique>	Element	ISO12620-A070201	Broader concept generic
<catégorie>	Element	ISO12620-A020201	Part of speech
<id>	Element	ISO12620-A1015	Entry identifier
'lang'	Attribute	ISO 12620A100701	Language Identifier

Table 2 Mapping XML objects from the example in section 1 (TML_1) to ISO 12620 data categories

We then align the two sets of data categories from TML_1 and TML_2 to identify the actual bandwidth of interoperability which, in this case, is limited to three types of information units.

Data Category in TML_1	**Interoperability**	**Data Category in TML_2**
Term	Interoperable	Term
	Loss from TML_2 to TML_1	*Subject field*
	Loss from TML_2 to TML_1	*Definition*
	Loss from TML_2 to TML_1	*Term type*
Entry identifier	Interoperable	Entry identifier
Language Identifier	Interoperable	Language Identifier
Broader concept generic	*Loss from TML_1 to TML_2*	
Part of speech	*Loss from TML_1 to TML_2*	

Table 3: identifying the bandwidth of interoperability between TML_1 and TML_2.

5. Styles and vocabularies

The evaluation of the bandwidth of interoperability between two TMLs is based on the data categories themselves and is completely independent of their implementation as XML objects. For instance, the data category /**Entry identifier**/ is instantiated as an attribute in TML_1 and as an element in TML_2, without preventing the corresponding information from being transferred between the TMLs. For this reason, TMF has a separate, complementary, mechanism to describe how a TML is concretely realized as an XML document, once the set of data categories it contains has been described.
The realization of an XML version of a TML is achieved by associating a *style* to each data category - selecting one possible form of XML realization - and by associating *vocabularies* to this style - selecting the names needed to realize this style.
For instance TML_1 implements the data category /**Broader concept generic**/ as follows:

Style Element
Vocabulary "générique"

which allows one to use the <générique> element to express the corresponding information unit.

TMF contains five styles (Attribute, Element, Typed Element, Valued Element, Typed Valued Element), to cover the various possibilities of realising a data category in XML.

6. A formal model for the representation of data categories

The comparison between two TMLs is only possible if there is a central repository of data categories, associated with a consistent model for these, which can act as a broker between any two formats. Such a Data Category Registry can be based upon a unified representation which will also serve for a given project to refine the characteristics of the data category it actually uses or even define its own additional data categories. In this respect, the representation of elementary features in TMF is associated to the definition of a formal description[8] of data categories that a TML will make reference to.

Each data category is described by a set of properties expressed in RDF (Resource Description Framework, see [RDF, 1999]) as proposed by the WWW consortium. Figure 3 shows the elementary properties that are used to uniquely define a data category, for example, a unique identifier (DCName), a name (DCName, which can be prefixed by a name space, in the case of multiple registries), a definition (DCDefinition) etc. as well as more complex properties determining the levels (in a structural model) at which it can be used (Level), its content type (Content), or its connection with other data categories (DCParent).

The full version of this descriptive framework allows one to assign the styles and vocabularies to each data category to describe its realization as an XML object in TMF.

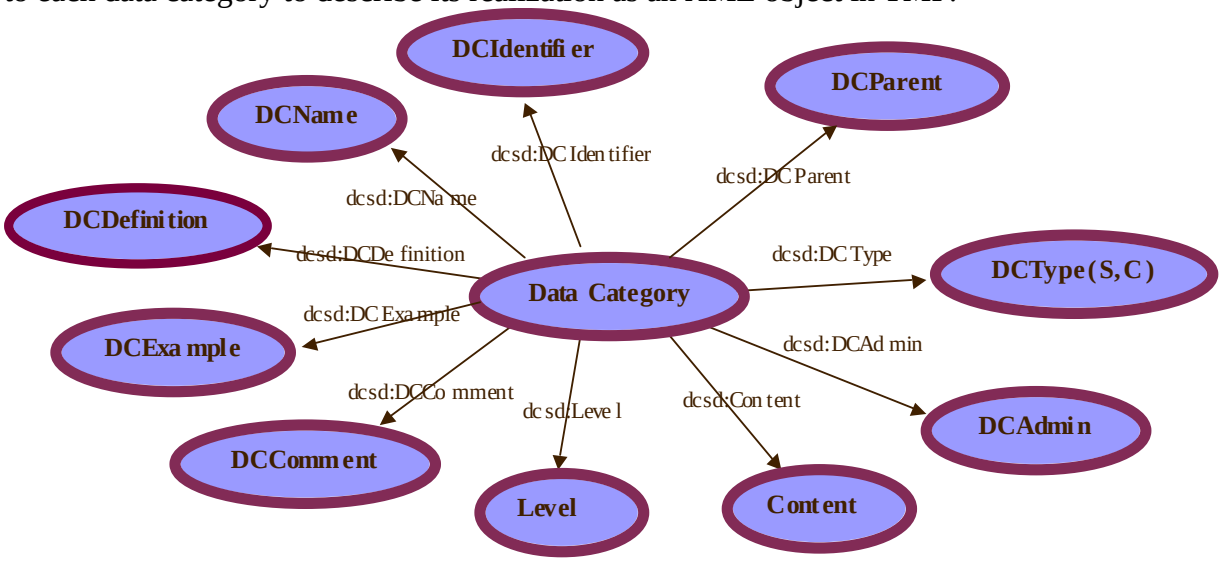

Figure 3: Core data category model expressed as RDF properties.

7. Data categories as meta-data objects

In the context of TMF, Data Category Specifications are used to specify constraints on the implementation of a TML and to provide the necessary information for the (possibly automatic) design of filters that convert from one TML to another (cf. Figure 4). From a wider perspective, it is interesting to consider data categories as abstract objects that may be used to communicate meta-data information concerning a particular (terminology) database. Indeed, any transfer of terminological data expressed in a given TML can be accompanied by an explicit reference to the DCS file that characterizes this TML, allowing the target (terminology) database to foresee the expected content of the data from an informational point of view and possibly make an *a priori* diagnosis of the compatibility of the data with its own internal format. In this respect, DCS descriptions become an essential part of the meta-data

[8] This model will be a component of part I of the proposed revision of ISO 12620.

associated with a terminological data interchange, complementary to other more traditional information such as “author”, “title” or “responsibility” for instance.

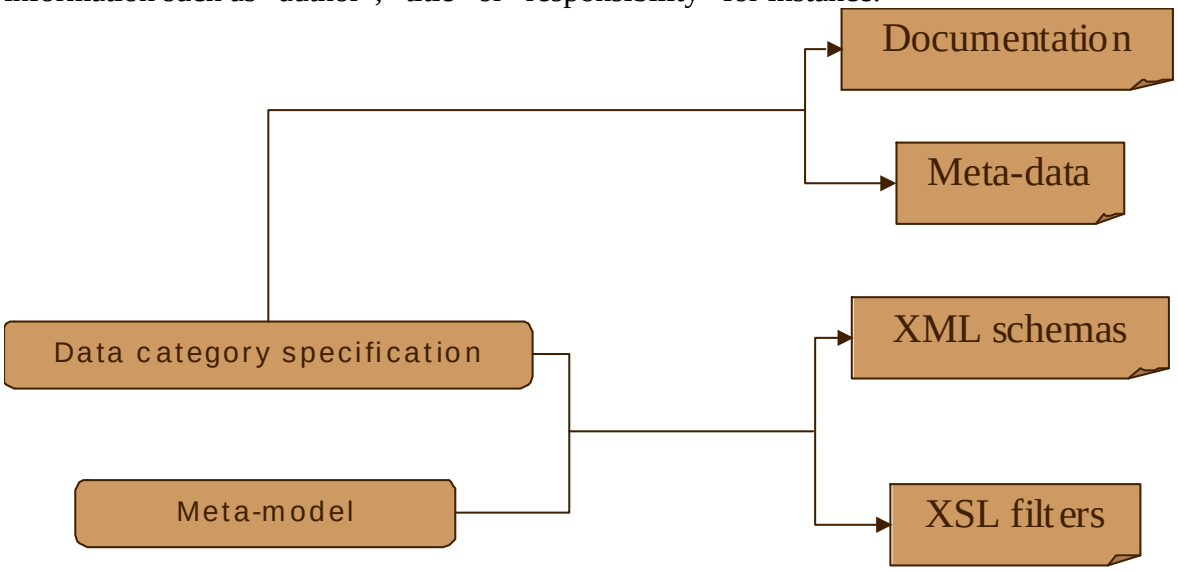

Figure 4: the various roles of data category specifications

8. Perspectives

The work conducted within the SALT project and Technical Committee 37 of ISO can be seen as a first step towards more fundamental research in the domain of implicit data-modeling. It shows how it may be possible to control classes of semi-structured documents (possibly expressed as XML applications) by means of external constraints on both their structure and content. Such an approach is particularly important in the domain of linguistic resources where a rapid consensus may not be achievable on the standardization of specific data structures, whereas it seems possible to define areas for which, for instance, it would be easy to implement data category registries. One can think in particular areas that use well-established techniques such as Part of Speech tagging, or annotation repositories (such as syntactic treebanks) for which interchange protocols are essential.

9. References

Antoni-Lay, M.-H., G. Francopoulo and L. Zaysser, 1994. A generic model for reusable lexicons: The GENELEX project, Literary and Linguistic Computing 9(1): 47-54.

Budin, Gerhard and Alan K. Melby, Accessibility of Multilingual Terminological Resources - Current Problems and Prospects for the Future, LREC’2000, Athens.

Iso12200 : Applications informatiques en terminologie - format de transfert de données terminologiques exploitables par la machine (MARTIF). Transfert négocié, Genève, Organisation internationale de normalisation, 1999.

Ide, N., A. Kilgarriff, L. Romary, 2000. A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000*, Stuttgart, 113-126.

Ide, N.and L. Romary, 2001. A Common Framework for Syntactic Annotation. *Proceedings of ACL'2001*, Toulouse, 298-305.

Iso12620 : Aides informatiques en terminologie - catégories de données. Genève, Organisation internationale de normalisation, 1999.

Ide, N., A. Kilgarriff, and L. Romary. A formal model of dictionary structure and content. In Proceedings of EURALEX 2000, pages 113–126, Stuttgart, 2000.

Le Meur, A., 1998, GENETER: a generic format for the distribution and reuse of

heterogeneous multilingual data, Proc. LREC, Grenada.
Picht H. and K.-D. Schmitz (Eds.), 2001. *Terminologie und Wissensordnung, Ausgewählte Schriften aus dem Gesamtwerk von Eugen Wüster*, TermNet Publisher.
RDF, 1999. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999 (http://www.w3.org/TR/REC-rdf-syntax/).
Wright, Sue Ellen and Alan K. Melby, 1999. Leveraging Terminological data for Use in Conjunction with Lexicographical Resources, TKE'1999