Un modèle abstrait pour la représentation de terminologies multilingues informatisées

Laurent Romary

Un modèle abstrait pour la représentation de terminologies multilingues informatisées TMF - Terminological Mark-up Framework

Laurent Romary

LORIA
Campus Scientifique
BP 239
F-54506 Vandœuvre-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 la structure générale d'une base terminologique et les informations (catégories de donnée) qui servent à décrire les différents niveaux de cette structure.

Abstract. *We are introducing an abstract model for representing computerized multilingual terminologies. This model has been developped in XML by Technical Committee 37 of ISO. It relies on a methodology which makes an essential distinction between the general structure of a terminological database and the information units (data categories) that are used to describe the various levels of this structure.*

1. Introduction

L'utilisation de données terminologiques multilingues représente une composante essentielle de nombreuses activités liées à la rédaction de documents techniques, à la traduction humaine ou encore à la traduction automatique spécialisée. Il existe actuellement de nombreuses bases terminologiques gérées par des entités publiques et privées. Il est parfois difficile d'en d'appréhender la cohérence et la compatibilité, car la plupart reposent sur des formats informatiques hétérogènes et plus ou moins bien documentés. Pour ne

mentionner que l'Union Européenne, dont on connaît les besoins en matière de traduction, les différents services concernés s'appuient sur un ensemble de plusieurs bases non interconnectées, sans parler des données spécifiques utilisées par le système de traduction automatique SYSTRAN adopté par la commission.

La représentation de données terminologiques multilingues pose de manière générale deux types de problèmes. D'une part il faut identifier l'organisation générale de ces données en grandes composantes. Le modèle le plus couramment adopté, d'inspiration Wüstérienne (Eugen Wüster est l'un des pionniers de la terminologie contemporaine), décompose une base terminologique sous la forme de niveaux hiérarchiques correspondant aux concepts, langues et termes. À l'opposé, les systèmes de traduction automatique tendent plutôt à s'appuyer sur une description plus lexicographique centrée sur une langue spécifique (dite *langue source*), où à chaque mot ou expression sont associés, en plus des informations morpho-syntaxiques, syntaxiques et sémantiques nécessaires, les équivalents possibles dans la langue cible de traduction. D'autre part, il faut déterminer les différents types d'informations (par exemple, /catégorie grammaticale/, /définition/, /restriction géographique d'usage/ etc.), qui sont utilisées dans une base donnée afin de pouvoir mettre en correspondance les informations qu'elle contient avec celles issues d'autres bases.

Il apparaît donc nécessaire de définir des standards de représentation de ces données et c'est ce à quoi a travaillé le comité technique 37 (TC37) de l'ISO (International Standardizing Organization) depuis plusieurs années.

Une première norme, MARTIF (ISO 12200), fruit de l'évolution des réflexions menées initialement dans le cadre de la TEI (Text Encoding Initiative - `http://www.tei-c.org`), proposa une DTD[1] SGML (Standard Generalized Markup Language, l'ancêtre de XML) permettant de couvrir l'essentiel des usages dans le domaine de la terminologie. Cependant, cette norme possédait quelques limitations qui rendaient sa révision nécessaire, et se voyait par ailleurs concurrencée par d'autres formats tels que Geneter qui, bien qu'exprimant plus ou moins les mêmes phénomènes, adoptaient des *styles* de représentation différents (plus grande finesse des contrôles sur les données au prix d'une plus grande complexité de la DTD).

Par ailleurs, de nombreux projets ayant à manipuler des données terminologiques et en particulier à unifier des bases hétérogènes qu'ils pouvaient posséder (la commission européenne est exactement dans ce processus en ce moment, cf. `http://www.unilat.org/dtil/etis/actasTDCnet/macphail.htm`),

1. DTD : Document Type Definition, correspond à la définition de la syntaxe d'un document par le biais de la description des éléments et des attributs utilisables pour baliser celui-ci, ainsi que leurs possibilités de combinaison.

il est exigé plus de souplesse de la part des standards proposés par l'ISO afin de pouvoir mieux exprimer les contraintes qui leur sont propres.

C'est pour toutes ces raisons que le comité technique 37 de l'ISO a décidé de travailler à une nouvelle norme qui définit, plutôt qu'un nouveau format de représentation des données terminologiques, un cadre générique permettant de définir de tels formats, une sorte de méta-modèle. Cette norme doit devenir la future ISO16642[2] , alias TMF (Terminological Markup Framework). Elle repose sur le principe simple qu'un format de description de terminologies particulier (un TML - Terminological Markup Language) repose sur la description de trois éléments :

– Un squelette structurel abstrait qui est commun à toute description terminologique ;
– Un ensemble de catégories de données correspondant aux informations que ce format veut représenter ;
– Les modes de réalisation de ce squelette structurel et de ces catégories de données dans un langage particulier pour définir un *TML concret*, sous la forme par exemple d'un schéma XML.

La norme TMF cherche à montrer qu'il est possible, d'une part, de couvrir une majorité des possibilités expressives des anciens formats MARTIF et Geneter, et d'autre part, de générer automatiquement des filtres de transfert des formats ainsi décrits vers une représentation abstraite, GMT (cf. *infra*), qui puisse servir d'intermédiaire de transformation d'un format donné vers un autre.

2. Une plateforme abstraite de définition de structures de documents

2.1. Organisation générale

L'objectif principal de TMF est de définir des mécanismes qui permettent de décrire les contraintes propres à une représentation donnée indépendamment d'un choix explicite d'une implémentation de cette structure sous la forme par exemple d'une DTD XML. De la sorte, l'ensemble des formats (ou TML - Terminological Markup Language) compatibles avec la plate-forme TMF forment une famille dont on sait définir de façon rigoureuse les conditions d'interopérabilité.

2. Le développement de cette norme a bénéficié du soutien du projet HLT/SALT (`http://www.loria.fr/projets/SALT`), dont l'objectif est de définir des outils pour la gestion et la diffusion de données terminologiques.

Pour illustrer la démarche, on peut considérer une entrée terminologique typique exprimée au format MARTIF (cf. ci-dessous). On peut y distinguer d'une part (en souligné) un certain nombre d'éléments XML qui organisent l'entrée dans un découpage en langues (`<langSec>`) et termes (`<tig>`) et d'autre part des informations qui qualifient ces différents niveaux. Du point de vue de leur représentation en XML, ces informations apparaissent sous différentes formes, soit comme des attributs (`id='ID67'`), des éléments (`<term>alpha smoothing factor</term>`) ou encore des éléments plus abstraits typés (`<descrip type='definition'>A value between 0 and 1 used in ...</descrip>`).

```
<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 ...</term>
        </tig>
    </langSet>
</termEntry>
```

Cette analyse conduit à voir une telle entrée terminologique comme une structure abstraite d'un arbre (le squelette structurel) décoré par des structures de trait correspondant aux informations identifiées à chaque niveau.

2.2. Un outil abstrait de réprésentation : GMT

La structure abstraite présentée ci-dessus peut elle-même être représentée en XML en utilisant le format GMT (Generic Mapping Tool) ainsi que l'on peut le voir ci dessous :

```
<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 ...</feat>
        </struct>
    </struct>
</struct>
```

Chaque niveau du squelette structure est ainsi exprimé à l'aide d'un seul élément récursif `<struct>` et chaque trait à l'aide d'un élément `<feat>`, ces deux éléments étant typé pour les relier respectivement à un niveau d'un métamodèle abstrait et de catégories de données définies par ailleurs (cf. section suivante).

Bien que le modèle et le format GMT qui l'accompagne, soient légèrement plus complexe que ce que nous présentons ici, on identifie facilement ici une classe particulière de documents XML analysable sur la base de la méthodologie employée.

GMT, par son caractère abstrait, apparaît en définitive comme un intermédiaire idéal entre deux TML particuliers, notamment quand il s'agit de définir des filtres de l'un vers l'autre. Les travaux récents menés au sein du projet SALT sur la rétro-conversion de bases existantes montrent que c'est aussi un outil important d'analyse de formats exogène à XML, par exemple des modèles entités-relations.

2.3. Représentation des catégories de données à l'aide de RDF

La représentation des traits dans le modèle TMF est associée à une formalisation des catégories de données auxquelles celui-ci fait référence. Chaque catégorie de données est modélisée par un ensemble de propriétés décrites à

l'aide du modèle RDF (*Resource Description Framework*) proposé par le consortium W3C. RDF permet de décrire des objets (ou « ressources ») à l'aide de structures propriétés-valeurs, éventuellement hiérarchiques. La figure 1 représente ainsi, à un premier niveau[3], le modèle de description proposé pour une catégorie de données, qui repose sur un ensemble de propriétés élémentaires permettant de lui affecter un identificateur unique (*DCIdentifier*), un nom (*DCName*), une définition (*DCDefinition*) etc., ainsi que des propriétés plus complexes déterminant les conditions d'utilisation de la relation ou encore son lien éventuel avec d'autres catégories de données (*DCParent*).

Plus précisément, la propriété « Locus » décrit les niveaux possibles (dans le méta-modèle) auxquels peut être rattachée la catégorie de donnée et « Content » donne le type du contenu de cette catégorie. Dans le cas d'une catégorie relationnelle, on peut en particulier décrire le domaine du deuxième argument de la relation ; en d'autres termes, le niveau vers lequel cette relation peut pointer.

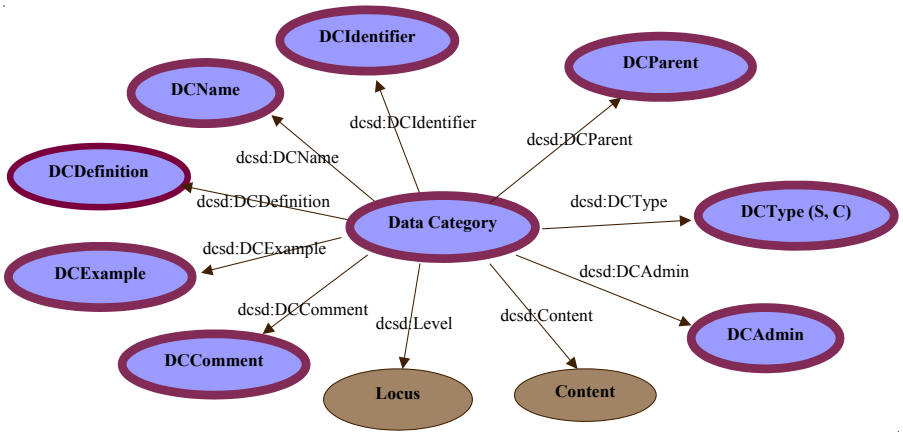

FIGURE 1 – Premier niveau de description d'une catégorie de données.

Enfin, le modèle RDF complet associé à une catégorie de donnée contient des propriétés spécifiques permettant de lui associer les éléments nécessaires à sa réalisation comme objet XML, à savoir un style (la catégorie de donnée doit s'exprimer comme un élément, un attribut etc.) et le vocabulaire nécessaire à

3. Le modèle complet décompose en particulier les ressources « Locus » et « Content » (cf. `http://www.loria.fr/projets/TMF`)

cette réalisation effective (par exemple la catégorie de donnée /définition/ se réalise sous la forme d'un élément `<def>`).

3. Outils associés

L'implication du projet européen SALT dans la définition de TMF a donné l'occasion, en parallèle à la définition du standard TMF, de mettre au point différents outils d'édition, de visualisation et de validation des formats concernés. On peut mentionner ainsi :

- Un éditeur de catégories de données reposant sur la représentation RDF présentée ci-dessus ;
- Un outil d'accès en ligne à des répertoires de catégories de données ;
- Des outils de génération automatique, à partir des spécification d'un TML, de schémas XML permettant de valider celui-ci, ainsi que de filtre XSL entre le TML et le format GMT.

Ces différents outils couplés à la définition d'API standard de manipulation de formats conforme à TMF devraient permettre une rapide diffusion et utilisation de la norme ISO 16642.

4. En guise de perspectives

Le travail mené au sein du projet SALT et du comité technique 37 de l'ISO représente une première étape dans la définition de structures abstraites pour la représentation de données structurées. L'expérience que nous avons menée porte de fait sur une classe de problèmes tout à fait particuliers (la description de données terminologiques) où, par essence, on sait disposer de cette double description en termes de structure abstraite et de traits associés. Il reste que plusieurs travaux récents montrent ([1], [2]) qu'il est possible d'étendre le champ d'application de ces concepts respectivement à la représentation de structures syntaxiques et surtout à tout le champ de la lexicographie. Ce point est essentiel si l'on veut aboutir par exemple à une comparaison systématique des nombreux dictionnaires informatisés qui voient le jour à la fois d'un point de vue institutionnel et commercial. L'idée étant à terme de pouvoir mettre en œuvre des librairies logicielles d'édition et de consultation qui puissent être indépendantes des formats manipulés.

Pour les classes de problèmes qui ne permettent pas d'identifier une organisation structurelle parfaitement reproductible, il nous semble que la notion de catégorie de donnée permet malgré tout de fournir un cadre générique de comparaison des informations manipulées par des formats différents au sein

d'un même domaine. C'est dans ce sens que l'ISO doit offrir un cadre ouvert de mise en œuvre de répertoire de catégorie de donnée où tout concepteur de format XML par exemple pourra puiser des références.

Ne rêvons pas. Nous sommes loin d'une situation où il sera facile de décrire les conditions d'interopérabilité universelles pour toute classe d'application. Nous espérons simplement que le travail mené au sein d'un domaine particulier puisse inspirer d'autres initiatives similaires.

Bibliographie

[1] N. Ide, A. Kilgarriff & L. Romary, «A formal model of dictionary structure and content», *in Proceedings of EURALEX 2000*, p. 113–126 (Stuttgart, 2000).

[2] N. Ide & L. Romary, «Standards and formats for treebanks», *in Treebanks* (Kluwer academic publishers), (à paraître).

[3] «ISO 12200 : Applications informatiques en terminologie – Format de transfert de données terminologiques exploitables par la machine (MARTIF)», Transfert négocié, Genève, Organisation internationale de normalisation (1999).

[4] «ISO 12620 : Aides informatiques en terminologie – Catégories de données», Genève, Organisation internationale de normalisation (1999).

[5] L. Romary, «Computer applications in terminology – Terminological markup framework (TMF), 2nd working draft», (2000).
`www.loria.fr/projets/TMF/`