Vers une construction normalisée d'une base de données terminologique multilingue

Par ***Bechir BOUDHIR***
Laboratoire Paragraphe, Université Paris8, France
Et ***Abderrazak MKADMI***
Institut supérieur de la documentation, Université de la Manouba, Tunisie

Mots clés : Base de données terminologique, multilinguisme, XML, travail collaboratif, ontologies de domaine, terminotique.

Introduction

Nous nous situons aujourd'hui dans un contexte de « mondialisation » où la communication et ses différents modes prennent une place importante. Cette communication dite « internationale », doit être interopérable et compatible aux différents moyens de communication. L'interopérabilité sollicitée nécessite d'avoir recours à une terminologie partageable et normalisée afin d'assurer une compréhension la plus large possible.

Notre travail se situe dans ce contexte et vise à présenter une méthodologie de construction collaborative d'une base de données terminologique multilingue, cette construction collaborative inclut les différentes étapes ; organisation du travail terminologique, édition de la fiche terminologique et élaboration des réseaux conceptuels.

Notre méthode de travail est basée sur un squelette structurel TMF (Terminological Mark-up Framework), il s'agit d'un méta-modèle inspiré de la norme ISO 16642 qui permet d'assurer l'interopérabilité entre langues en respectant le même cadre de représentation de données terminologiques.

Nous choisissons la technologie XML, comme métalangage de description ; de structuration et d'échange de données, le plus adapté à

notre application. En effet, XML apparaît comme ressource intéressante pour tout ce qui peut constituer de véritables réservoirs de ressources numériques, ainsi qu'un moyen d'interopérabilité sûr de développement des activités collaboratives qui peuvent exister entre les différents utilisateurs de ces ressources. Caractérisant la sémantique et la hiérarchisation d'unités textuelles, XML constitue un noyau d'une base de données terminologique sous forme d'une ontologie de domaine.

Notre travail s'articule en trois parties. La première consacrée au recueil terminologique selon la méthode onomasiologique basée sur le méta-modèle TMF. Une deuxième partie s'intéresse à l'organisation du travail terminologique ainsi que l'édition de la fiche terminologique via la technologie XML. Dans une dernière partie nous allons chercher à montrer le lien entre terminologie et ontologie et construire les ontologies de domaine, qui sont à notre avis les fruits d'une terminologie normalisée, à travers un outil informatique pour rester toujours dans un contexte d'une production informatisée.

Méthode de travail

Pour toute production terminologique, une méthode de travail claire et précise est indispensable surtout dans le cas d'une terminologie multilingue et collaborative.

En terminologie, il existe deux méthodes largement connues, la première est dite onomasiologique et la deuxième est sémasiologique. Dans le cas d'une méthode sémasiologique, on part du signe jusqu'au concept. C'est une étude centrée sur les signes, leur formation, leur évolution, leurs agencements, leur variabilité dans le corpus. Cette démarche est celle du linguiste qui s'occupe du terme en tant que signe, utilisé dans un système linguistique bien spécifique à une société, une vision du monde, une culture.

Par opposition à cette méthode, au cours d'une démarche onomasiologique, le point de départ est le concept, il peut s'agir d'une notion technique, d'un objet scientifique, d'un comportement et même d'une pratique dans un domaine bien particulier, pour explorer les différentes réalisations du terme dans les différentes langues. C'est une approche « notionnelle » dans laquelle on trouve qu'une notion correspond à un terme et inversement.

Le choix de la méthode pour le travail terminologique est fixé par les objectifs et les besoins du terminologue. Notre objectif, dans ce travail, est de construire une base de données terminologique (BDT) partageable, cohérente et exploitable par les outils informatiques. Ceci dépend du système notionnel. C'est pour cela que nous optons pour une démarche onomasiologique. Cette démarche nous semble la plus adaptée à nos besoins du fait qu'elle s'occupe de l'étude des notions (concepts) et des mots ou expressions (les termes) qui les représentent.

Les deux démarches terminologiques (onomasiologique et sémasiologique) sont en opposition, mais elles restent synergiques. La synergie est le plus souvent la règle si on considère l'aspect humain de l'activité. Les différentes terminologies s'élaborent en exploitant la liaison synergique entre les deux méthodes. Cependant les terminologues informaticiens ont défini une norme (onomasiologique) pour plusieurs raisons :

« Une banque terminologique sémasiologique connaît une explosion des liens de relations, surtout si elle est multilingue. »

Construire des banques multilingues en partant des termes est quasi impossible parce que l'on rencontre les plus grandes difficultés à apparier in fine les structures conceptuelles.

L'interopérabilité des terminologies dans le web-sémantique exige un choix normatif et celui du comité technique de l'ISO_1 (ISO/TC37) (TMF : Terminological Markup Framework) permet cette interopérabilité et sert de base à la construction d'ontologies.

Les mécanismes du TMF, et la normalisation des catégories de données permettent précisément d'amorcer la réalisation modulaire de ces ontologies elles-mêmes en cours de standardisation par le W3C (langage OWL) » (HUDRISIER. 2005).

Structuration des données terminologiques

La structuration des données terminologiques sert à décrire les relations sémantiques ou conceptuelles qui relient les termes entre eux. Il existe divers types des relations, mais pour notre cas de travail nous nous intéressons au modèle hiérarchique. Dans ce type de relations, les concepts sont organisés en niveaux dans lesquels le concept superordonné renferme au moins un concept subordonné. Les concepts subordonnés de même niveau partagent les mêmes critères de subdivision qui sont appelés concepts coordonnés. Notons que la relation hiérarchique englobe les deux types de relations générique et partitive. Dans le cas d'une relation générique, la compréhension du concept subordonné, dit concept spécifique, suppose la compréhension préalable du concept superordonné, dit concept générique, en ajoutant au moins un caractère distinctif supplémentaire.

Dans le cas d'une relation partitive, les concepts subordonnés, surnommés aussi concepts coordonnés, forment les pièces constituantes du concept superordonné. Dans ce genre de relation, un concept superordonné est considéré comme étant un concept dit intégrant (englobant) alors que le concept subordonné est dit concept partitif (englobé).

La structuration des données terminologiques constitue une étape essentielle dans le travail terminologique. En revanche cette classification doit être bien ciblée, en effet plus une BDT est riche, plus elle est difficile à traiter et à gérer, et donc plus elle nécessite de faire appel à une méthode précise et claire pour stocker les informations. Un bon stockage des fiches terminologiques permet une bonne gestion en arrière plan, une recherche ciblée, une maintenance facile afin d'assurer :

> « La mise à jour (actualisation) et épuration de collections partielles,
> Le développement systématique d'une collection,
> L'élaboration de glossaires par domaines,
> Les échanges de données dans certains domaines. » (CST, 2014)

Dans un travail terminologique il ne suffit pas de faire un recueil de termes, mais il faut aussi et surtout les classer et organiser les données terminologiques. A notre avis une méthode de travail normalisée constitue une condition essentielle pour une meilleure production collaborative, que ce soit pour l'organisation ou l'échange des données terminologiques. Dans ce contexte nous nous proposons, dans la partie suivante, un modèle de structuration des données terminologiques qui fait référence à des normes de l'ISO-TC37 afin de préserver un certain niveau d'inter-compatibilité des ressources terminologiques.

Il s'agit ici d'un méta-modèle strictement normalisé qui permet de structurer les termes d'une façon hiérarchique grâce à leurs relations (lexicales). Ce méta-modèle est appelé TMF (Terminological Markup Framework), inspiré de la norme ISO-16642[2] et prend en compte des

recommandations de la norme ISO-12620$_3$ qui traite les catégories de données.

TMF : exemple du modèle de structuration terminologique

TMF permet d'assurer la cohérence, la compatibilité et l'interopérabilité des données terminologiques à travers une architecture de base où la collection des termes obtenue sous la désignation de collection de données terminologiques va être catégorisée et reliée (encapsulée) dans des paramètres prédéfinis. Ce méta-modèle repose sur la description de trois éléments suivants :

« 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 » (Romary. 2001).

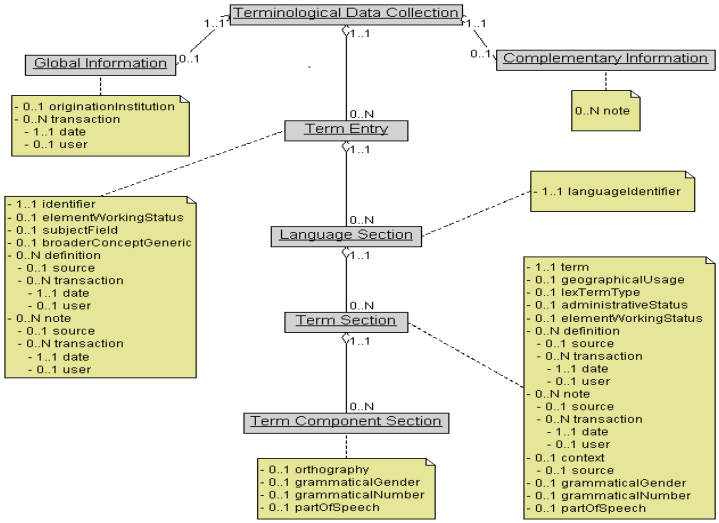

Figure 1 : le méta-modèle d'organisation d'une base de données terminologique (Kramer. 2004).

Ce modèle conceptuel se découpe en sections fondamentales, représentées de haut en bas de la manière suivante (Boudhir-Tientcheu, 2005):

TDC (Terminological Data Collection) : C'est une collection de données contenant l'information sur le concept des champs de sujet spécifiques. [ISO 1087-2.21]

GI (Globa Information) : Information technique et administrative appliquée à la totalité de la collection de données. Exemple : l'historique des versions, le titre de la collection de données.

CI (Complementary Information) : Information supplémentaire décrite dans les entrées terminologiques et partagée à travers la collection de l'entrée terminologique. Ex : Hiérarchie de domaine, description des institutions, références bibliographique, etc.

TE (Terminological Entry) : Entrée contenant une information sur des unités terminologiques (exemple : sujet spécifique, concept, terme, etc.), à relier avec nos champs qui étaient définis comme de véritables entrées terminologique

LS (Langage Section) : une partie d'une entrée terminologique contenant l'information relative à une autre langue. Une entrée terminologique doit nécessairement posséder des informations sur un ou plusieurs langues.

TS (Term Section) : Partie d'une section du langage donnant des informations relatives à un terme tel que l'usage d'un terme, éléments de terme.

TCS (Term Component Section) : Partie d'une section de terme délivrant des informations linguistiques à propos des composants d'un terme.

Le méta-modèle ci-dessus s'adapte parfaitement à une approche intégrative, utilisable pour les données terminologiques existantes et dans la conception des nouvelles terminologies qui sont typiquement utilisées dans un mode relationnel ou orientées texte dans un système de gestion. Le squelette TMF peut être décrit au moyen d'un élément générique <struct> qui décrit aussi la structure des niveaux de représentation d'une collection de donnée terminologique.

Chaque nœud structurel doit être identifié au moyen d'un attribut type associé à l'élément <struct>.

Les valeurs possibles de l'attribut type doivent être les identificateurs des niveaux du méta modèle. Exemple : GI, TE, LS, TS, TCS. Les unités de bases informationnelles associées à un nœud de la structure peuvent être représentées en utilisant l'élément <feat>.

Une unité d'information représentationnelle (paragraphe représentatif d'un programme) peut être spécifiée en utilisant l'instruction <brack>

qui peut elle-même contenir un <feat> suivi par une combinaison de <feat> et de <brack>.

Toute unité d'information doit être qualifiée par un type d'attribut prenant pour valeur le nom d'une catégorie de donnée de l'ISO 1262O ou défini par l'utilisateur.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <struct type="terminologicalDataCollection">
- <struct type="terminologicalEntry">
<feat type="conceptIdentifier">A1</feat>
- <struct type="languageSection">
<feat type="languageIdentifier">fr</feat>
- <struct type="termSection">
<feat type="term">enseignement mixte</feat>
<feat type="lexTermType">fullform</feat>
</struct>
</struct>
</struct>
```

Figure 2 : l'organisation structurelle du méta- modèle exprimée en XML

Exemple des fiches terminologiques :

Le squelette TMF nous a permis de rendre les diverses terminologies dédiées à chaque langue, interopérables entre elles. Le grand avantage de TMF est de focaliser plutôt sur le concept que le terme. En numérotant un concept, on lui associe une définition. Dans ce travail nous nous intéressons plutôt à la terminologie arabe, notre point de départ est la norme ISO- 2382[4]. Il s'agit des vocabulaires liés au domaine de l' "e-learning "). تعلّم إلكتروني

En mettons en application ladite norme TMF, nous garantissons que la fiche contient le maximum des informations dont l'utilisateur

(traducteur, linguiste, rédacteur….) a besoin. Dans notre travail nous s'intéressons à des informations dont nous semble utiles et indispensables pour tout travail terminologique tels que : définition, contexte, langue, identificateur, champs d'utilisation, contributeur, type lexical…

La richesse d'information dans une fiche terminologique va rendre aussi la banque des données terminologiques plus riche, cependant nous signalons « plus une fiche multiplie les catégories d'information, plus elle est complexe et difficile à déchiffrer, et plus la maintenance devient difficile. La multiplicité de données de nature différente peut alourdir la recherche, rendre la tâche difficile pour de nouveaux terminologues ainsi que pour les autres personnes chargées de la préparation ou terminologues ainsi que pour des autres personnes chargées de la préparation ou de la mise à jour des fiches terminologiques. Il faut donc trouver un juste milieu » (CST, 2014)

Nous listons dans ce qui suit des exemples des fiches terminologiques. En prenant comme concept générique le « e-learning » () "تعلّم إلكتروني"), il s'agit d'une entrée terminologique, section langue, section terme, composants du terme.

```
<?xml version="1.0" encoding="UTF-8"?>
<Terminological Data Collection xsi:no Name space Schema Location="bechir.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <Global Information Origination Institution="text" transaction="text" date="text" user="text"/>
  <Term Entry concept Identifier="e-learning" element Working Status="starter" subject Field="learning and training" broader Concept Generic="subordonné" definition="en : learning facilitated by information and communications technology
                                                    Fr: apprentissage facilité par l'utilisation des technologies de l'information et de la communication.
                                                    Ar : المعلومات تقنيات استخدام بواسطة ميسر تعليم
```

```
والاتصال">
        <Language Section language Identifier="en">
            <Term Section term="Computer based learning"
geographicalUsage="Information technologies" textTermType="complet"
administrativeStatus="terme préféré" elementWorkingStatus="text"
definition="text" transaction="text" date="text" user="Bechir Boudhir"
context="text" source="SC36 WG1 " grammaticalGender="neutre"
grammaticalNumber="singulier" partOfSpeech="nom">
                <TermComponentSection orthography="Computer"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
                <TermComponentSection orthography="Based"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="verbe"/>
                <TermComponentSection orthography="learning"
grammaticalGender="féminin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
            </TermSection>
        </LanguageSection>
        <LanguageSection languageIdentifier="fr">
            <TermSection term="apprentissage assisté par ordinateur"
geographicalUsage="Information technologies" textTermType="complet"
administrativeStatus="terme préféré" elementWorkingStatus="text"
definition="text" transaction="text" date="text" user="Bechir Boudhir"
context="text" source="SC36 WG1 " grammaticalGender="neutre"
grammaticalNumber="singulier" partOfSpeech="nom">
                <TermComponentSection orthography="apprentissage"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
                <TermComponentSection orthography="asisté"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="adjectif ou participe passé"/>
                <TermComponentSection orthography="par"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
                <TermComponentSection orthography="ordinateur"
```

```
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
        </TermSection>
      </LanguageSection>
      <LanguageSection languageIdentifier="ar">
        <TermSection term="تعليم باستخدامِ الحاسِبِ"
geographicalUsage="Information technologies" textTermType="complet"
administrativeStatus="terme préféré" elementWorkingStatus="text"
definition="text" transaction="text" date="text" user="Bechir Boudhir"
context="text" source="SC36 WG1 " grammaticalGender="neutre"
grammaticalNumber="singulier" partOfSpeech="nom">
          <TermComponentSection orthography="تعليم"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
          <TermComponentSection orthography="باستخدام"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="adjectif ou participe passé"/>
          <TermComponentSection orthography="الحاسوب"
grammaticalGender="masculin" grammaticalNumber="singulier"
partOfSpeech="nom"/>
        </TermSection>
      </LanguageSection>
    </TermEntr="تعليم باستخدامِ الحاسوبy>
    <ComplementaryInformation note="text"/>
</TerminologicalDataCollection>
```

Figure 3 : Exemple XML compatible avec le schéma TMF

Vers les ontologies de domaine

Pour conclure cet article nous mettons l'accent sur l'utilité d'avoir une terminologie de domaine, qui nous permet en fin de compte de construire des ontologies de domaine.

L'objectif de notre travail n'est pas de construire des ontologies de domaine, mais plutôt de montrer le lien entre l'étude terminologique et les ontologies de domaine. C'est ici que réside l'apport et

l'importance de la terminologie normalisée, c'est le passage du domaine linguistique ou terminologique au domaine de l'ontologie et l'ingénierie des connaissances, autrement dit l'espace onto-terminologique.

Au cours de l'étude terminologique, nous faisons une phase de conceptualisation et de représentation structurelle d'un domaine sous forme d'un arbre. Cette phase nous permet d'ordonner et de classer les notions dans un domaine donné selon des « classes d'objets », ce qui nous permet d'avoir des liens lexicaux entre les termes et de les classer par la suite dans des réseaux sémantiques d'une manière hiérarchique. Il s'agit des représentations partielles de sous-domaines de concepts dont chaque représentation est structurée autour d'un concept principal. Les concepts hiérarchisés forment une ontologie de domaine, « l'ontologie sert alors le plus souvent à hiérarchiser et classer les éléments composant le domaine ainsi qu'à décrire leurs relations » (Kembellec. 2008).

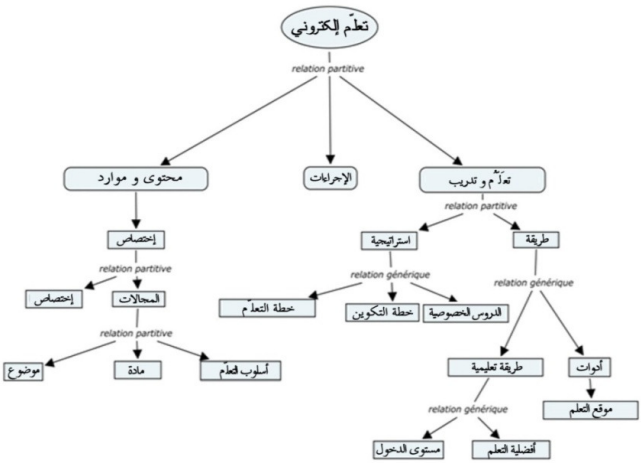

Figure 4 : représentation graphique d'une ontologie partielle de l'e-learning

Cette représentation simplifiée du domaine de connaissance de l'e-learning renferme trois sous-domaines, chaque sous-domaine peut converger ensuite vers des termes en troisième niveau qui, selon le dictionnaire terminologique PAVEL, sont « la dénomination spécialisée qui désignent un objet, concret ou abstrait, et qu'il est possible de définir de façon non équivoque dans sa lexicalisation comme une unité de connaissance dans une langue individuelle » (Pavel).

La relation hiérarchisée liant le concept et les termes de sous-domaines exprime une représentation structurée au cours de laquelle l'héritage des propriétés est aussi assuré.

Conclusion

L'étude terminologique présentée dans cet article est concentrée sur la structuration et l'organisation des données terminologiques. A notre avis la structuration des données constitue une étape indispensable dans le travail terminologique du fait qu'elle donne suite à toute tâche qui concerne la gestion sémantique.

La structuration que nous avons choisi dans ce travail est basée sur un méta-modèle normalisé connu sous le nom du TMF, ce méta-modèle nous a permis de gérer au mieux le traitement normalisé des termes ainsi que les relations entre les termes dans une phase appelée « conceptualisation » ou « dénomination » selon la méthode onomasiologique.

D'un point de vue pratique, nous avons montré dans la dernière partie de cet article l'importance de faire le lien entre la normalisation de la terminologie et la construction des ontologies de domaine. C'est dans ce passage du plan linguistique et terminologique au plan ontologique que réside l'apport de la terminologie normalisée.

Références bibliographiques

1. BOUDHIR Béchir & TIENTCHEU Joseph. novembre 2005. « Normalisation des termes de la FOAD : Proposition d'une démarche d'élaboration d'une Base de Données Terminologique ». Colloque Initiatives : La norme comme instrument de réussite d'une société de la connaissance partagée. Sommet mondial sur la société de l'information, Tunis En ligne : http://www.initiatives.refer.org/Initiatives-2005/document.php?id=269
2. CST – Conférence des Services de Traduction des États européens Groupe de travail « Terminologie et documentation » Recommandations relatives à la terminologie, Berne, 2014
3. HUDRISIER Henri. CARTAGO, Débat thématique 3, 2 mars 2007. un outil collaboratif pour le recueil à visée omni-lingue de l'expertise terminologique éducationnelle liée au développement des normes du e-enseignement. Actes du colloque Initiatives 2005 [en ligne], Disponible sur Internet : http://www.initiatives.refer.org/Initiatives-2005/document.php?id=196.
4. KRAMER Isabelle. 2004. Easy TMF : Livre blanc. Loria. 73p.
5. (Pavel) http://dictionnaire.sensagent.com/concept/fr-fr/
6. ROCHE Christophe. « Terminologie et ontologie ». In la terminologie discipline scientifique. Acte de colloque 17 octobre 2003. Paris : le savoir des mots. 2004. Pp47-56. ISBN 2-9521893-1-5.
7. ROMARY Laurent. 2001, Un modèle abstrait pour la représentation de terminologies multilingues informatisées TMF - Terminological Mark-up Framework. Cahiers GUTenberg, n° 39-40, p83-91. [En ligne]: http://hal.inria.fr/docs/00/10/05/88/PDF/Romary.pdf
8. SALEH Imad, MKADMI Abderrazak. 2008. Bibliothèque numérique et recherche d'informations. Paris : Hermès science publications. 281pages. ISBN: 2-7462-1820-8.

9. KEMBELLEC Gérald. 5et 6 juin 2008 « Ontologie franco/ anglaise du domaine informatique comme accès à un corpus de textes scientifiques ». In Actes du colloque TOTH 08 : Terminologie & Ontologie : Théorie et applications. Annecy : Institut Porphyre Savoir et Connaissance.. ISBN 978-2-9516-4539-4.

Notes :

[1] ISO/TC37 s'intéresse à la normalisation des principes, méthodes et applications relatives à la terminologie et aux autres ressources langagières et ressources de contenu dans les contextes de la communication multilingue et de la diversité culturelle.

[1] TMF (Terminological Mark-up Framework ; ISO 16642) est un standard international qui fournit un cadre pour la représentation des bases de données terminologiques en XML. Ce standard ne représente pas un format particulier mais plutôt un méta-modèle permettant de spécifier les contraintes propres à un langage de description de données terminologiques.

[1] La norme ISO 12620 fournit les catégories de données terminologiques qui seront utilisées pour représenter les unités d'information d'un langage de représentation de données terminologiques.