

Normalisation des échanges de données en terminologie : le cas des relations dites « conceptuelles »

Laurent Romary
Équipe Langue et Dialogue
Loria (UMR 7503)
B.P. 239
F-54506 Vandœuvre-lès-Nancy,
France
Laurent.Romary@loria.fr
www.loria.fr

Marc Van Campenhoudt
Centre de recherche Termisti
Institut supérieur de traducteurs et interprètes
34, rue J. Hazard
B-1180 Bruxelles
Belgique
marc.van.campenhoudt@euronet.be
www.termisti.refer.org

Résumé :

L'échange et la fusion de données constituent un aspect important de la production terminographique. L'Organisation internationale de normalisation (Iso) s'est de longue date souciee de faciliter de tels échanges. Après avoir longtemps défendu l'idée d'un inventaire clos des catégories de données (norme Iso 12 620), sur lequel elle a fondé sa norme SGML d'échanges négociés Martif (Iso 12 200), elle se dirige aujourd'hui vers une famille de normes d'échange plus génériques, utilisant XML et ouvertes à un enrichissement permanent des types de données. Cet aspect est particulièrement important dès lors qu'il s'agit de permettre un échange des relations dites *conceptuelles* engrangées dans les bases de connaissances terminologiques. Les auteurs présentent la nette évolution que constitue la nouvelle proposition de norme Iso CD 16 642 TMF (*Terminological Markup Framework*) qu'élabore actuellement l'Équipe Langue et dialogue du Loria, en concertation avec le comité technique 37 de l'Iso. Celle-ci est expérimentée, entre autres, par le Centre de recherche Termisti, dont les microglossaires contiennent de nombreuses relations conceptuelles.

Mots-clés :

Bases de connaissances terminologiques, relations conceptuelles, échange de données, normalisation, Iso, XML, TMF.

1 Introduction

L'échange et la fusion de données ont servi à la constitution de la majorité des grandes bases de données terminologiques (BDT) connues. Quand bien même elles n'ont pas été élaborées de manière « orthodoxe », ces bases sont des monuments incontournables dont la taille et le coût de constitution interdisent toute velléité de mise au rebut. Un fossé évident sépare les pratiques réelles des théories : un espace de collaboration doit se dessiner entre les chercheurs en linguistique appliquée qui aspirent à une plus grande modernité et ceux qui travaillent à la maintenance et à la difficile modernisation de ces BDT. La présente communication s'inscrit dans cette modeste perspective et examinera les aménagements qui peuvent être apportés aux normes d'échanges traditionnelles en terminologie pour mieux prendre en compte la description des relations de sens. Les auteurs n'ont aucune prétention d'œuvrer dans le domaine strict de l'intelligence artificielle. Ils partagent toutefois les aspirations d'un grand nombre de linguistes qui ont investi le champ de

l'étude des vocabulaires spécialisés : abandonner une démarche normalisatrice et onomasiologique pour mieux tenir compte des réalités du corpus.

Les spécialistes des réseaux et des ontologies connaissent la puissance de systèmes de représentation et d'échange comme Kif (*Knowledge Interchange Format*), CGS (*Conceptual Graph Standard*), Oil, (*Ontology Inference Layer*) ou encore XTM (*XML Topic Maps*, Iso 13250 2000)¹. Ces systèmes ne sont guère répandus, sinon connus, dans les milieux de la terminographie traditionnelle. Pour rester dans le cadre des modestes besoins de ce milieu - qui néglige trop souvent les relations conceptuelles -, le souci des auteurs sera plutôt de veiller à ce que les normes d'échange du comité technique 37 de l'Iso, traditionnellement utilisées en terminologie, donnent un plus large écho aux nombreux travaux sur les relations de sens au sein des BDT. Leur communication est nourrie par leur collaboration au projet européen Salt (*Standards-based Access to Multilingual Lexicons and Terminologies*)², qui s'intéresse précisément aux formats d'échanges de l'Iso, en concertation avec le consortium Olif (*Open Lexicon Interchange Format*)³, plus orienté vers l'échange de ressources lexicales utilisées en traduction automatique⁴.

2 Le point sur les formats d'échange de l'Iso TC37

La première norme d'échange parue, Martif (Iso 12 200 199), est une évolution du chapitre 13 de la TEI (*Text Encoding Initiative*) et de l'ancien format Micromater. Conçue pour des échanges négociés, Martif s'appuie sur la norme Iso 12 620 (1999), qui établit un inventaire des champs utilisés dans les bases de données terminologiques. Ce format est aujourd'hui complété par deux projets de normes permettant d'opérer des échanges aveugles : Geneter et DXLT. L'idée actuelle est de permettre des échanges entre ces formats de référence - ou d'autres à venir - au travers du projet de norme Iso Cd 16 642 (2001) *Terminological Markup Framework* (TMF) précisant les exigences structurelles minimales auxquelles doit répondre tout langage de représentation de données terminographiques ou TML (*Terminological Markup Language*) (figure 1). TMF devrait permettre le passage entre deux TML au travers d'une représentation abstraite intermédiaire nommée GMT (*Generic Mapping Tool*). Les principes structurels de TMF seront détaillés au point 3.

1. Kif : www.cs.umbc.edu/kse/kif/, CGS : www.cs.uah.edu/~delugach/CG/, Oil : www.ontoknowledge.org/oil/, XTM : www.topicmaps.org/.

2 Salt : www.loria.fr/projets/SALT/.

3. Olif : www.olif.net/.

4. Le consortium Salt étudie cependant de près l'intérêt que pourrait représenter un pointage du balisage vers un fichier de ressources externes conforme à XTM ou Oil.

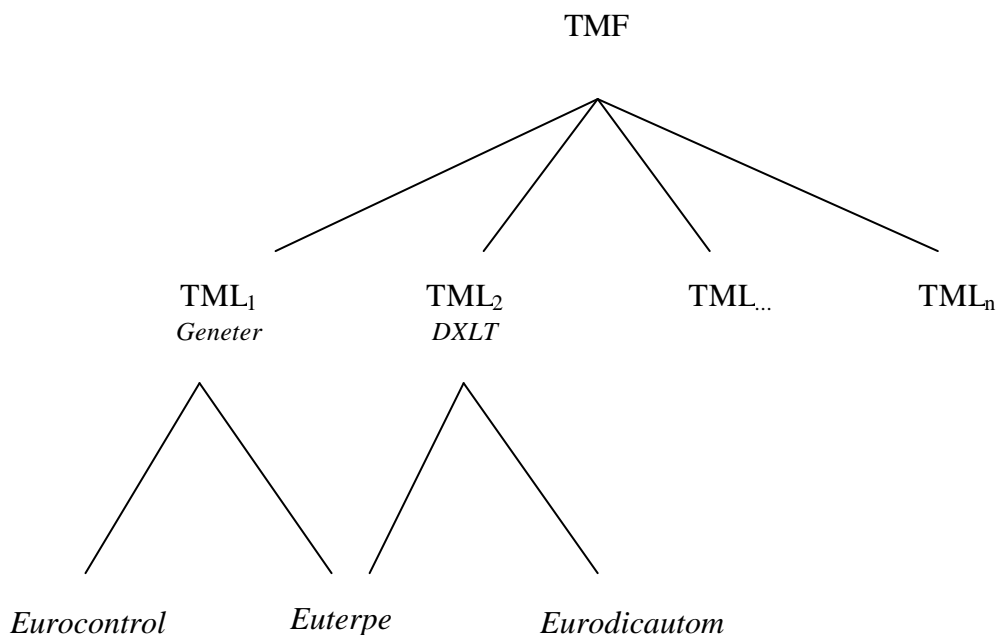


Figure 1 : vue des différents niveaux d'échange

Publiée en 1999, la norme Iso 12 620, consacrée aux catégories de données, est déjà en révision. L'idée de dresser un inventaire normalisé des champs envisageables dans la conception d'une BDT suscite toutefois diverses réserves. Les obstacles à une mise à jour régulière sont tels que les normes d'échanges fondées sur un inventaire clos risquent de ne pas satisfaire les concepteurs de BDT novatrices. Cette critique est particulièrement fondée pour ce qui concerne les relations de différents types que l'on peut identifier dans une BDT.

3 La structure typique d'une entrée terminologique

La gestion terminographique est souvent dite «conceptuelle», un adjectif dont la signification doit être nettement relativisée. Dès lors qu'elles servent à traduire un nombre élevé de langues, les bases de données terminologiques doivent miser sur la monosémie, laquelle est d'ailleurs systématiquement privilégiée en ingénierie de la langue. Le projet de norme TMF consacre un mode d'organisation systématique qui a fait ses preuves et sur l'exposé duquel nous ne reviendrons plus ici en détail. Il s'agit de distinguer des niveaux de description successifs : données dites « conceptuelles », communes à toutes les langues, données propres à une langue, données propres à un terme. Par exemple, le modèle de données retenu dans notre récent projet Dhydro⁵ (figure 2) correspond à un tel modèle, puisqu'il obéit à la structure suivante :

- 1 concept
 - décrit dans n langues
 - désigné par n termes dans chaque langue

5. www.loria.fr/projets/MLIS/DHYDRO/.

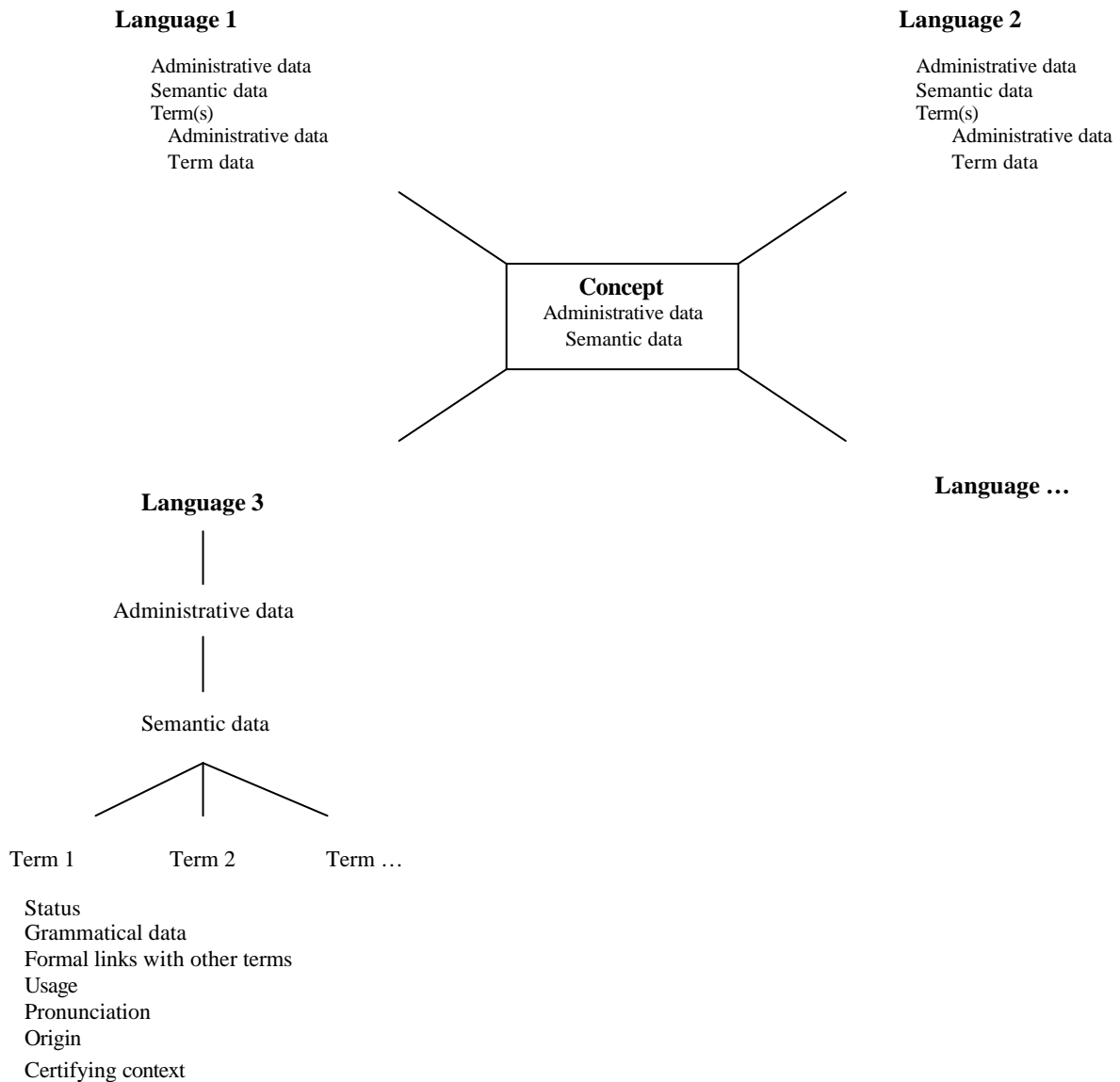


Figure 2: le modèle de donnée de l'interface Dhydro (Descotte et al., à paraître)

4 TMF : une plate-forme de description de langages de représentation de données terminologiques

La proposition de standard Iso CD 16 642 intègre la description d'un modèle permettant de spécifier les contraintes propres à un langage de description de données terminologiques informatisées (TML) exprimé en XML. Il repose sur l'hypothèse qu'un tel format peut être décrit par la combinaison de deux composantes :

- Un méta-modèle spécifiant un squelette structurel commun à tout langage de représentation de données terminologiques, et qui repose sur le modèle conceptuel classique. Ce squelette est décrit dans la figure 3, où apparaissent 7 *sites structurels*, dont certains peuvent être itérés (une entrée terminologique (TE, *Terminological Entry*) peut ainsi contenir une ou plusieurs sections associées aux langues décrites dans la base (LS, *Language Section*)) ;
- La description de contraintes de rattachement de certaines informations, ou *traits*, aux différents nœuds du modèle structurel. Chaque trait fait référence à une catégorie de données spécifiée au sein de la norme Iso 12 620 ou définie de façon spécifique par le concepteur de la base.

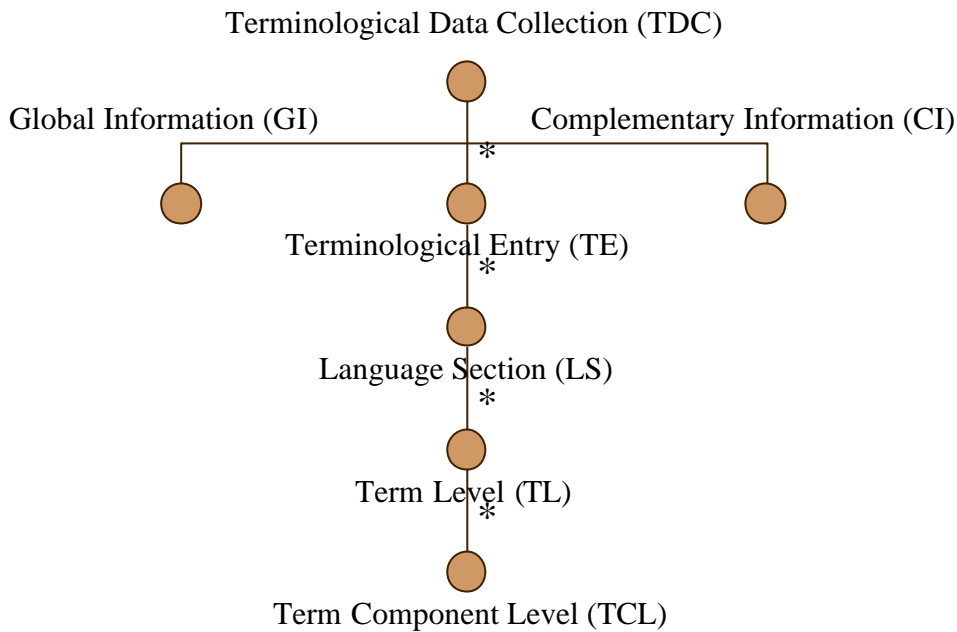


Figure 3 : le méta-modèle d'organisation d'une base de données terminologiques.

Ce modèle permet de définir un principe d'interopérabilité entre deux langages de description terminologique (deux TML) qui garantit leur équivalence dès lors qu'ils reposent sur le même ensemble de catégories de données. Ce principe permet, en particulier, de définir des filtres entre formats et spécifie dans ce cadre un outil de conversion reposant sur une représentation abstraite intermédiaire nommée GMT (*Generic Mapping Tool*). À titre d'exemple, la figure 4 représente une entrée terminologique extrêmement simple au format Martif et la représentation équivalente en GMT. On y distingue les deux éléments de base définis dans la structure XML de l'outil GMT :

- Un élément <struct> qui, combiné à un attribut 'type', permet de représenter un site du squelette structurel ;
- Un élément <feat> (*feature*) qui, lui aussi combiné à un attribut 'type', permet de spécifier un trait associé au nœud <struct> qui le contient.

En dehors de ces deux éléments, GMT comprend, d'une part, un mécanisme de regroupement de traits (élément <brack> (*bracket*)) et, d'autre part, une représentation des annotations que la valeur d'un trait pourrait contenir (élément <annot>).

<pre> <termEntry id="eid-EUTERP-94545"> <descrip type="subjectField"> health </descrip> <langSet lang="en"> <ntig> <termGrp> <term>abdominal dropsy</term> </termGrp> </ntig> <ntig> <termGrp> <term> peritoneal dropsy </term> </termGrp> <descripGrp> <descrip type="definition"> Effusion and accumulation of serous fluid in the abdominal cavity. </descrip> </descripGrp> </pre>	<pre> <struct type="TE"> <feat type="identifiant">eid-EUTERP-94545</feat> <feat type="subjectField-12620A.4"> health </feat> <struct type="LS"> <feat type="language-12620A.10.7">en</feat> <struct type="TS"> <feat type="term-12620A.1"> abdominal dropsy </feat> </struct> <struct type="TS"> <feat type="term-12620A.1">peritoneal dropsy </feat> <feat type="definition-12620A.5.1"> Effusion and accumulation of serous fluid in the abdominal cavity.</feat> </struct> </struct> </struct type="LS"> </pre>
---	--

<pre> </ntig> </langSet> <langSet lang="fr"> <ntig> <termGrp> <term> ascite </term> </termGrp> <descripGrp> <descrip type="definition"> Accumulation de liquide dans la cavite peritoneale. </descrip> </descripGrp> </ntig> <ntig> <termGrp> <term>hydroperitoine</term> </termGrp> </ntig> </langSet> </termEntry> </pre>	<pre> <feat type="language-12620A.10.7">fr</feat> <struct type="TS"> <feat type="term-12620A.1"> ascite </feat> <feat type="definition-12620A.5.1"> Accumulation de liquide dans la cavite peritoneale.</feat> </struct> <struct type="TS"> <feat type="term-12620A.1"> hydroperitoine </feat> </struct> </struct> </pre>
---	---

Figure 4 : une entrée terminologique au format Martif et sa représentation à l'aide de l'outil de conversion GMT.

4.1 Représentation des relations dans la plate-forme TMF

La notion de relation fait partie intégrante de la plate-forme TMF puisque chaque trait représentant une information attachée à un site du squelette structurel peut, si la catégorie de données correspondante l'autorise, contenir un attribut 'target' qui va alors pointer sur un autre site du squelette structurel avec lequel le site courant est en relation. Ce mécanisme permet ainsi de relier des termes entre eux (relation 'AbbreviationOf' entre deux sites TL, illustrée de façon simple dans la figure 5 ci-dessous), des concepts (relation 'BroaderConceptGeneric') ou plus généralement toute paire de sites de niveau quelconque.

```

<struct type="TE">
  <feat type="subjectField-12620A.4">terminology</feat>
  <struct type="LS">
    <feat type="language-12620A.10.7">en</feat>
    <struct type="TS" id="TS1">
      <feat type="term-12620A.1">Terminological Markup Framework</feat>
    </struct>
    <struct type="TS">
      <feat type="term-12620A.1">TMF</feat>
      <feat type="abbreviation-12620A.2.1.8.1" target="id(TS1)"/>
    </struct>
  </struct>
</struct>

```

Figure 5 : représentation d'une relation d'abréviation entre deux termes.

Bien plus, chaque trait peut aussi contenir un attribut 'source' qui contient alors une référence au site décrit par ce trait, dans le cas où l'on souhaiterait décrire certaines relations de façon externe au contenu terminologique proprement dit. Un tel mécanisme permet de rapprocher la notion de base terminologique de celle de thesaurus, en isolant la composante strictement relationnelle.

4.2 Vers une hiérarchie de catégories de données

La représentation des traits dans le modèle TMF est associée à une formalisation des catégories de données auxquelles celui-ci fait référence, et ce dans le cadre du travail de révision de la norme Iso 12 620 (*Data Categories*). Dans les dernières propositions définies au sein du comité éditorial de l'Iso 12 620, 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. Pour simplifier, RDF permet de décrire des objets (ou «ressources») à l'aide de structures propriétés-valeurs, éventuellement hiérarchiques. La figure 6 schématise ainsi, à un premier niveau, 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, comme nous le verrons, son lien éventuel avec d'autres catégories de données.

Plus précisément, la propriété 'Locus' décrit les niveaux possibles (dans le méta-modèle) auxquels peut être rattachée la catégorie de données, 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.

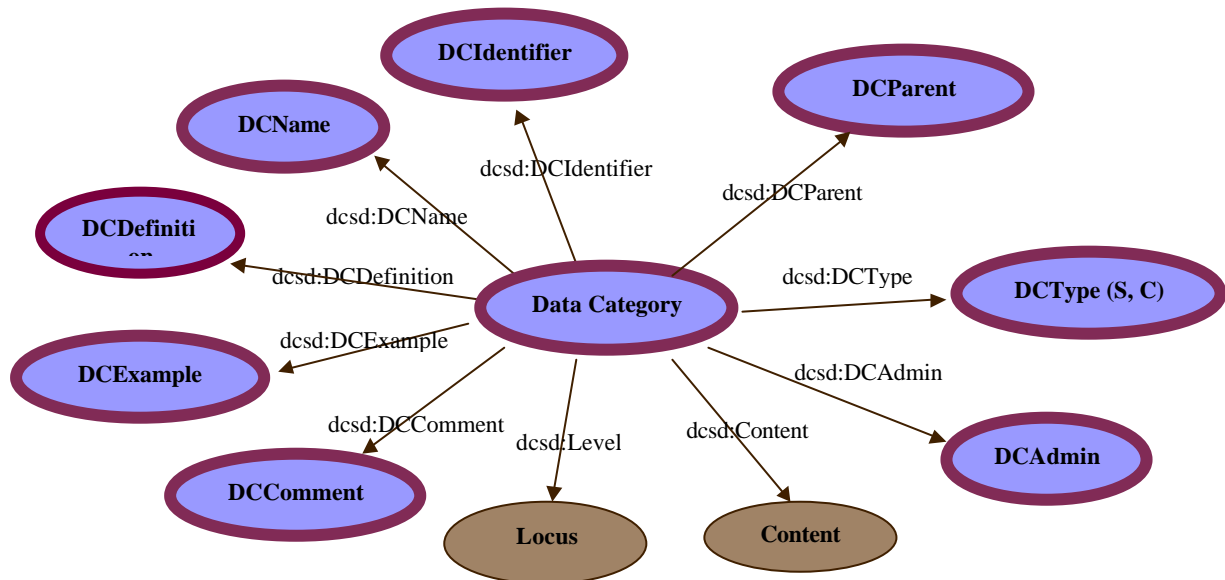


Figure 6 : premier niveau de description d'une catégorie de données.

Enfin, il est possible d'associer à une catégorie de données une propriété 'DCParent' qui va relier celle-ci avec une autre catégorie de données plus générique, dans le cadre de contraintes de cohérence entre les catégories concernées (du point de vue des sites d'ancrages possibles ou du type de leurs contenus). Cette relation permet d'aller vers une véritable ontologie des catégories de données en terminologie, en s'appuyant sur un certain nombre de catégories générales (par exemple 'relation spatiale') qui seront ensuite spécialisées, soit au sein d'un répertoire de référence tel que l'Iso 12 620, soit par spécialisation d'une catégorie de données issue d'un répertoire de référence pour un usage spécifique au sein d'une base propriétaire.

5 Les relations dans les bases de données terminologiques

Si nous partons du principe que toute BDT correctement pensée doit répondre à la structure proposée par TMF, sous peine de rencontrer de sérieux problèmes d'échanges, nous pouvons identifier au moins deux grands types de relations qui ne sont pas implicites dans une telle structure :

- des liens entre termes , comme 'abréviation de', 'troncation de', 'acronyme de', etc.
- des liens entre concepts, comme 'type de', 'cause de', 'prédécesseur de', etc.

En effet, le modèle conceptuel garanti par TMF ne suppose pas de considérer la synonymie et l'équivalence comme des liens explicites entre termes, dans la mesure où il implique que pour une entité conceptuelle donnée :

- tout terme est synonyme des autres termes relevant de la même langue ;
- tout terme est équivalent des autres termes relevant des autres langues.

Toutefois, on observera que dans le cadre de TMF, les relations dites conceptuelles peuvent être indifféremment situées au niveau du concept, de la langue ou du terme, ce qui autorise quiconque à adopter une stratégie de représentation différente.

Il convient cependant d'attirer l'attention sur le danger qu'il y a à parler de « relations conceptuelles » lorsque les liens ne sont pas établis au niveau du concept et sont clairement dépendants d'une langue donnée. Personne n'oserait prétendre détenir un inventaire de concepts universels pour un domaine donné : ce que nous nommons *concept* est en réalité une conceptualisation de la connaissance conditionnée par la langue, le domaine donné, le corpus... Rien ne prouve que les relations entre conceptualisations seront les mêmes d'une langue à l'autre, d'un domaine à l'autre, d'un corpus à l'autre. Dans une BDT multilingue, les concepts sont, en fait, des « noyaux de sens » qui sont restreints de manière à autoriser l'équivalence quelle que soit la langue source et la langue cible (Van Campenhoudt 2000). Si les relations que l'on nomme *conceptuelles* sont vérifiables pour chacune des langues envisagées, il est logique de les situer au niveau interlangue, c.-à-d. au niveau dit *conceptuel (terminological entry)*. Mais rien n'empêche d'imaginer une BDT proposant des variations du réseau dit *conceptuel* en fonction des langues. La logique voudrait alors que l'on situe les relations au niveau intermédiaire du modèle : celui des langues⁶ (*language section*).

5.1 Vers une typologie élargie et ouverte sur l'innovation

Alors qu'elle propose une très intéressante liste des relations entre termes sous son point A.2.1, la norme Iso 12 620 (1999) présente sous ses points A.6 et A.7 un inventaire de liens dont on ne peut que regretter l'indigence :

- la typologie est très élémentaire : relations espèce-genre, partie-tout, séquentielles (spatiales, temporelles, causales) et associatives ;
- la place de chaque concept dans la relation ne peut pas toujours être précisée ; tel est notamment le cas pour les relations spatiales, temporelles et causales : on ne peut préciser que tel concept est situé au-dessus de tel autre, qu'il est plus ancien que tel autre, qu'il est la cause de tel autre ...

À titre de comparaison, le projet de norme Olif2 (2000) - qui n'est pas spécifique à la terminologie - propose un inventaire déjà plus étoffé, même s'il a le tort de mêler les relations formelles (comme 'abréviation de'), sémantiques (comme 'synonyme de') et conceptuelles (comme 'fils de'). La liste intègre, par exemple, une typologie des relations méronymiques héritée des

6. On peut alors imaginer des relations conceptuelles comme 'plus proche concept générique équivalent dans l'autre langue'.

travaux de Winston, Chaffin et Herrmann (1987). Elle a, par ailleurs, le grand mérite d'envisager des types de relations conceptuelles qui déterminent la valence sémantique ou syntaxique des termes : agent, patient, action typique, instrument, etc.

Soucieux d'enrichir la typologie de la prochaine version de la norme Iso 12 620, le centre de recherche Termisti a proposé un essai d'inventaire plus large encore, prévoyant des relations hiérarchiques de parenté entre liens et des liens indirects (par transitivité) dans les cas où les concepts qui servent de points de passage dans le réseau ne seraient pas décrits dans la BDT (par exemple la relation entre *piston* et *automobile*). Dans la mesure où un tel inventaire ne sera jamais exhaustif, TMF prévoit que la référence à la norme 12 620 n'est pas obligatoire et que de nouvelles catégories peuvent être déclarées grâce au modèle RDF.

5.2 De l'information à échanger

On l'a dit : toute typologie des relations conceptuelles varie en fonction de paramètres comme le domaine, le corpus, la culture... À ce titre, tout essai d'inventaire systématique ou « définitif » est voué à l'échec. Il n'en reste pas moins vrai que si l'on souhaite exporter les informations contenues dans le réseau conceptuel d'une BDT, il convient de se mettre d'accord sur un minimum d'étiquettes. Il paraît donc important qu'un format d'échange permette d'exprimer un minimum d'informations relatives aux relations conceptuelles utilisées :

- Classification du lien, au sein d'une typologie existante ou sous la forme d'une nouvelle catégorie : p.ex. : lien de nature temporelle.
- Niveau de validité du lien : terme, langue, concept.

Au-delà de ces indications minimales, on peut souhaiter exporter ou importer une variété d'autres éléments informatifs :

- Formulation du lien dans chacune des langues envisagées : 'est au-dessus de' = 'is above' = 'ist über' = 'è sopra' = 'está arriba de'.
- Propriétés relationnelles : transitif, réflexif, bijectif...
- Restriction : fréquence, condition...
- Autre relation qui fonde la typologie : 'succède à', 'plus grand que', 'supérieur hiérarchique de'...
- Trait distinctif des opposés : taille, direction, âge...
- etc.

6 Conclusion

Il est essentiel que le système relationnel présenté ci-dessus puisse être intégré dans tout format d'échange de données terminologiques présentant un caractère un tant soit peu générique. Pour ce faire, il faut, d'une part, disposer d'une plate-forme technique qui soit à même d'exprimer de telles relations et, d'autre part, prévoir des mécanismes par lesquels on puisse à tout moment étendre un système de relations existant lorsque le besoin s'en fait sentir pour une application particulière. En effet, il n'est pas évident qu'il soit nécessaire de répertorier l'ensemble des relations utilisées dans les différentes bases terminologiques existantes, certaines étant, par exemple, trop liées à l'usage spécifique qui est fait de cette base. Par ailleurs, si l'on fait l'hypothèse que l'on dispose d'un répertoire normalisé de relations, le temps de mise à jour d'un tel répertoire dépasserait de loin, par exemple dans un cadre industriel, les contraintes d'efficacité de mise en œuvre d'une BDT particulière. Un concepteur doit alors pouvoir décrire ses propres relations en liaison étroite avec le répertoire existant, par exemple en identifiant une relation plus générique dont la relation qu'il souhaite décrire serait l'instance. Il semble que la plate-forme TMF, couplée avec les mécanismes

de déclaration dynamique de catégories de données organisées en réseau, répond exactement à ces exigences.

L'influence historique des normalisateurs en terminologie a souvent suscité de légitimes réticences chez les linguistes attentifs à la description d'un usage réel au sein d'un corpus particulier, dans des circonstances déterminées. Ils apprécieront sans aucun doute une heureuse évolution dans la manière dont semble aujourd'hui se concevoir la normalisation des échanges de données. Alors que, par le passé, on entendait définir et délimiter les seules catégories de données susceptibles d'être exportées ou importées, on se dirige à présent vers une norme ouverte à la créativité du chercheur et qui se borne à expliciter les meilleures conditions structurelles pour garantir un bon échange.

Références

CHAFFIN R., HERRMANN D.J. et WINSTON M. (1988) : « An Empirical Taxonomy of Part-whole Relations : Effects of Part-Whole Relation Type on Relation Identification », *Language and Cognitive Processes*, vol. 3, n° 1, p. 17-48.

DESCOTTE S., HUSSON J.-L., ROMARY L., VAN CAMPENHOUDT M. et VISCOGLIOSI N. (à paraître) : « Specialized lexicography by means of a conceptual data base: establishing the format for a multilingual marine dictionary », à paraître dans les actes de la *Second Conference on Maritime Terminology*, University of Turku, Finland, 11-12 May 2000.

ISO 12 200 (1999) : *Applications informatiques en terminologie – Format de transfert de données terminologiques exploitables par la machine (Martif) – Transfert négocié*, Genève, Organisation internationale de normalisation.

ISO/IEC 13250 (2000) : *Information technology - SGML Applications- Topic Maps*, Genève, Organisation internationale de normalisation.

ISO 12 620 (1999) : *Aides informatiques en terminologie – Catégories de données*, Genève, Organisation internationale de normalisation.

ISO CD 16 642 (2000) : *Computer applications in terminology - Terminological markup framework (TMF)*, 2nd working draft, 10th November 2000, www.loria.fr/projets/TMF/.

OLIF2 (2000) : *Open Lexicon Interchange Format, V.2 : Proposal for Linguistic Features/Values for OLIF2*, OLIF2 Consortium, June 2000, www.olif.net.

VAN CAMPENHOUDT M. (2000) : « De la lexicographie spécialisée à la terminographie : vers un "métadictionnaire" ? », in THOIRON PH. et BÉJOINT H., dir., *Le sens en terminologie*, Lyon, Presses universitaires de Lyon (Travaux du C.R.T.T.), p. 127-152.

WINSTON M.E., CHAFFIN R. et HERRMANN D. (1987) : « A Taxonomy of Part-Whole Relations », *Cognitive Science*, vol. 11, n° 4, p. 417-444.