Gestion de données terminologiques : principes, modèles, méthodes

Laurent Romary, Isabelle Kramer, Susanne Salmon-Alt, Joseph Roumier

Chapitre 7

Gestion de données terminologiques : principes, modèles, méthodes

1.1. Terminologies - quels usages ?

Les données terminologiques multilingues représentent une composante essentielle dans de nombreux secteurs d'activité, dont la rédaction et la traduction techniques, l'indexation, la recherche d'information [WRI 97] [WRI 01]. L'élaboration de ces ressources s'avérant fort coûteuse, les échanges et la fusion de données constituent un aspect important de la constitution de la majeure partie des bases de données terminologiques connues. Dans le présent chapitre, nous essayons de montrer, en cohérence avec les grandes initiatives internationales de normalisation (ISO[1], W3C[2] et TEI[3]) comment il est possible d'apporter un certain nombre de réponses cohérentes à ces difficultés.

La perspective générale que nous adoptons ici est celle de la conception de bases de données terminologiques. Il s'agit en effet de pouvoir répondre aux interrogations que se posent tant les terminologues que les concepteurs de systèmes d'informations, à deux niveaux complémentaires :

– À un niveau macroscopique, il s'agit, d'une part, de situer le fonds terminologique dans un contexte où il est alimenté par des sources multiples et souvent hétérogènes, tout en devant répondre à de nombreux usages potentiellement différents (traduction, rédaction technique ou indexation), et, d'autre part, où il s'agit de définir des procédures de maintenance cohérentes pour l'ensemble du fonds terminologique (par l'établissement d'un comité éditorial par exemple) ;
– À un niveau microscopique, il s'agit de fournir des modèles de données qui soient en mesure de répondre aux exigences précédentes, tant du point de vue de la finesse des descriptions linguistiques associées au fonds terminologique, que de l'identification précise des opérations de gestions de ce fonds (méta-données).

De fait, notre propos sera ici centré sur le problème de la représentation fine de données terminologiques en montrant qu'il n'est pas nécessaire, voire possible, de proposer un modèle unique de représentation, mais au contraire, d'offrir les mécanismes nécessaires pour à la fois répondre au mieux aux besoins des différents projets terminologiques et garantir un haut niveau d'interopérabilité entre projets.

Le présent chapitre s'articule ainsi autour de trois parties complémentaires. Dans un premier temps nous rappelons les principaux éléments linguistiques qui caractérisent l'approche terminologique, et qui ont amené à l'organisation générique des données terminologiques proposée dans le cadre de la norme ISO 16642 (TMF). Dans un deuxième temps, nous montrons que cette norme peut conduire à une représentation formelle des modèles compatibles avec l'ISO 16642, conduisant ainsi à une vision calculable de l'interopérabilité entre

[1] Organisation Internationale de Standardisation. http://www.iso.org
[2] World Wide Web Consortium. http://www.w3c.orglogi
[3] Text Encoding Initiative. http://www.tei-c.org

Chapitre rédigé par Susanne ALT, Isabelle KRAMER, Laurent ROMARY et Joseph ROUMIER.

structures terminologiques. Enfin, nous décrivons comment il est possible d'intégrer concrètement cette démarche dans les directives de la TEI en donnant les caractéristiques principales du module terminologie de l'édition P5 de ces directives.

1.2. Modèles de données terminologiques

1.2.1. Les fondements de la terminologie moderne

La théorisation de la terminologie dans son acception moderne est généralement attribuée à Eugen Wüster (1898-1977) qui la résume dans sa *Allgemeine Terminologielehre* [WÜS 74] [WÜS 68]. Ingénieur, terminologue et membre du cercle de Vienne, Wüster aborde, à travers ses nombreux écrits depuis les années 30, des questions linguistiques fondamentales interrogeant la nature du signe linguistique, le positionnement de la sémantique linguistique vis-à-vis de la logique et la variabilité langagière. Parallèlement, il œuvre dans une perspective fortement applicative, dominée par une approche terminologique normative, considérée comme la clé d'une communication optimale dans les domaines scientifiques et techniques.

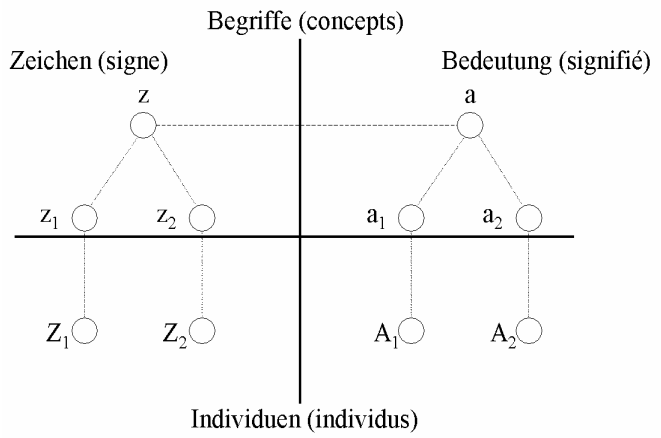

Figure 1. *Le modèle quadri-partite du mot, selon [WÜS 59/60]*

Un bref rappel de son modèle quadri-partite du mot (Figure 1) permettra de mieux situer son approche dans un contexte de positivisme logique, ainsi que les conséquences pratiques sur la terminographie. Les quatre pôles de son modèle forment, verticalement, une correspondance entre *individus* et *concepts* et horizontalement, entre *signe* et *signifié*. Le modèle retrace les trois étapes – décomposition, abstraction et association – de la « mise en mots du monde » [WÜS 59/60], i.e. de l'acte de communication dans sa perspective productive : le monde en tant qu'entité complexe est mentalement décomposé en individus (A_1 , A_2). La perception des individus conduit à la formation de concepts individuels, caractérisés par un ensemble de traits identificatoires (a_1 , a_2). Sur les concepts individuels s'opère un processus d'abstraction qui mène à la formation d'un concept générique (a), défini en tant que représentation mentale stabilisée. C'est à ce concept générique qu'est associé un signe linguistique (z), lui-même considéré sous l'angle d'une abstraction sur des réalisations individuelles (Z_1 , Z_2).

Sur ces bases, Wüster dégage les spécificités de la terminologie, dont l'objectif consiste à étudier les fondements théoriques et les outils méthodologiques pour la « représentation des relations entre individus au moyen d'un réseau de concepts et de leurs signes associés de façon univoque »[4]. En particulier, il formule les fondements de la terminographie dont on peut résumer les points principaux :

- La nécessité de la caractérisation d'un concept à travers une définition précise, afin de le situer par rapport aux concepts voisins ;
- L'ordonnancement onomasiologique, i.e. l'accès aux signes au moyen d'une classification logique des concepts, les relations conceptuelles retenues étant la relation *spécifique–générique*, la relation *tout-partie* ainsi que la relation d'*appartenance* ;

[4] Traduit et adapté d'après [FEL 01] , p. 247.

– L'introduction éventuelle de marqueurs d'autorité, permettant de gérer les contraintes de planification linguistique, i.e. le degré de normativité d'une association entre signe et concept.

1.2.2. Le courant de la terminologie textuelle

Les positions wüstériennes de la terminologie ont pu provoquer des réticences dans des courants linguistiques plus particulièrement intéressés par une approche descriptive de la variabilité des faits linguistiques en corpus, comme c'est le cas de la terminologie textuelle [BOU 00] [CON 03]. La terminologie textuelle reproche en particulier à Wüster le postulat de la pré-existence, de l'immuabilité et de l'invariabilité des concepts [RAS 95]. Or, Wüster n'a jamais présenté la monosémie d'un signe comme un dogme – mais tout au plus comme un fait idéal dans l'objectif de faciliter la communication dans des domaines techniques – ni nié la variabilité des significations d'un même signe (surtout dans ses travaux tardifs, cf. [WÜS 74]) [5]. Plus fondamentalement, la critique adressée à Wüster par la terminologie textuelle ne suffit pas à remettre en cause la nécessité de normalisation terminologique dans des contextes applicatifs, mais a le mérite d'ouvrir l'interrogation sur les modes de normalisation et le rôle de la linguistique dans le processus d'acquisition, de stabilisation et de renouvellement de connaissances terminologiques. Dans cette optique, la terminologie textuelle défend la construction de terminologies sur la seule base de faits linguistiques généralisables à partir d'usages et de régularités observés en corpus. Toutefois, le refus théorique de l'antériorité des concepts aux mots mène à deux interrogations sur la méthodologie mise en oeuvre par la terminologie textuelle : d'une part, le passage des observables aux concepts (sur quel critères et à quel moment décide-t-on de faire émerger, à partir des occurrences observées en corpus, des concepts ou « signifiés normés » ?), et d'autre part, la finalité même de la construction de terminologies (si elles émergent des corpus sans faire appel à des concepts stabilisés antérieurement, pourquoi construire, *in fine*, des terminologies ?). Enfin, et paradoxalement[6], le modèle de données sous-jacent à la mise en œuvre de la terminologie textuelle (la *Base de Connaissances Terminologique*, [CON 03] est entièrement compatible avec des modèles de données reposant sur des principes wüstériens (cf. la norme ISO 16642 présentée *infra*), à ceci près que le rôle de la définition diminue au profit d'une structuration relationnelle des concepts : l'accès au terme peut être à la fois onomasiologique et sémasiologique, et le champ *lien–concept* subsume parfaitement la fonction de marqueur d'autorité.

1.2.3. TMF, une plateforme de spécification de données terminologiques

La norme TMF [ISO 16642] s'inscrit dans la lignée des travaux d'E. Wüster et plus particulièrement d'une série de normes portant sur les méthodes [ISO 704] et modèles [ISO 12200] [ISO 12620] de représentation de données terminologiques au sein du comité technique 37 de l'ISO. Dans cette perspective, la norme vise à fournir des outils permettant de représenter de façon inter opérable et donc pérenne, toute une variété de fonds terminologiques.

Né d'une volonté de définir une plate-forme unifiée de spécification et de représentation de données terminologiques multilingues, le *Terminological Markup Framework* (TMF, [ISO 16642]) permet ainsi de décrire des méta-contraintes pour le marquage terminologique, c'est-à-dire des contraintes structurelles minimales auxquelles doit répondre tout langage de représentation de données terminographiques. Ces contraintes s'expriment par la combinaison d'un *méta-modèle* unique et de *catégories de données* qui lui sont associées.

Le méta-modèle décrit par TMF (Figure 2) reprend les grandes composantes de la plupart des bases terminologiques récentes. Il structure une base de données terminologiques (/terminologicalDataCollection/) en un ensemble de concepts, à savoir les entrées terminologiques (/terminologicalEntry/). Chaque entrée terminologique est ensuite subdivisée en un ensemble de blocs linguistiques (/languageSection/), réunissant

[5] Contrairement à ce qui est affirmé dans les travaux cités, Wüster ne postule ni des concepts universaux, ni des concepts permanents et invariables. Cf. en particulier la notion de *concept subjectif* , et la discussion du rôle de la langue en tant que *création et créatrice* (*Schöpfer und Geschöpf*) d'une communauté, « Das Worten der Welt », 1959/60. Cf. aussi les discussions concernant précisément la non-universalité des systèmes conceptuels et le besoin de multiplication des dimensions de description dans la comparaison inter-langue, « Die terminologische Sprachbehandlung », 1953 [WÜS 53]).

[6] Peut-être pas si paradoxalement, à condition de considérer un rapprochement possible du *concept* wüstérien avec le concept en tant que « *signe normé* » de Rastier et al. (1994)

l'ensemble des informations relatives à l'expression du concept dans une langue donnée[7]. Dès lors, l'une des caractéristiques importante du méta-modèle TMF est d'encapsuler systématiquement, à l'intérieur du bloc linguistique, la description relative à un terme à l'intérieur d'un bloc terminologique *autonome* (/termSection/). Ceci permet en particulier d'associer à chaque terme tout un ensemble d'informations reatives à ses caractéristiques linguistiques, son statut normatif, ou son registre d'usage (diachronique, diatopique, diastratique, ou diaphasique). Quand cela est nécessaire, les blocs terminologiques peuvent être décomposés en blocs de description des composants du terme (/termComponentSection/), afin d'associer par exemple des informations plus fines relatives à l'origine dérivationelle du terme. Des descripteurs élémentaires – ou catégories de données [ISO 12620] – s'ancrent sur les différents composants du méta-modèle et sont des éléments d'information permettant de décrire, structurer et d'enrichir les données d'une entrée terminologique.

La macro-organisation de TMF est donc d'inspiration « wüstérienne », dans la mesure où elle reflète une approche lexicographique onomasiologique, à travers un regroupement de termes par concept. À chaque entrée terminologique correspond un et un seul concept. Les termes sont les signes linguistiques représentant ce concept dans un domaine de connaissance. Plusieurs termes pour un même concept (et une même section linguistique) entretiennent une relation de parasynonymie, relation qui peut être étendue aux équivalences de traduction. La polysémie s'exprime par l'association d'un même terme à plusieurs concepts. Les autres caractéristiques de la terminologie « wüstérienne » sont gérées par des catégories de données appropriées : expression des relations entre concepts par des liens générique-spécifique et associatifs, possibilité d'associer une définition aux concepts, gestion des usages (normatifs ou descriptifs) par des marqueurs spécifiques.

Toutefois, TMF, en tant que modèle de données, ne répercute ni le postulat de l'antériorité du concept au terme, ni la contrainte d'une relation univoque entre signe et concept. En particulier, TMF ne présage en rien des procédures d'acquisition et de normalisation de données terminologiques et conceptuelles par planification linguistique ou par abstraction sur des usages effectifs. Le modèle ne fait ainsi aucune hypothèse concernant un possible usage prescriptif ou descriptif qui en sera fait. Par ailleurs, l'introduction de catégories de données spécifiquement dédiées à la description des termes en tant que signes linguistiques, telles que la description de ses composants au sein de la /termComponentSection/, la définition associée à un terme ou l'encodage de relations linguistiques entre termes, associé à l'usage possible de métadonnées pour documenter l'usage d'un terme relativement à un contexte donné sont parfaitement compatibles avec les pratiques lexicographiques modernes. D'une certaine manière, TMF et un bon compromis entre une ligne terminologique dure et la lexicographie computationnelle basée sur corpus.

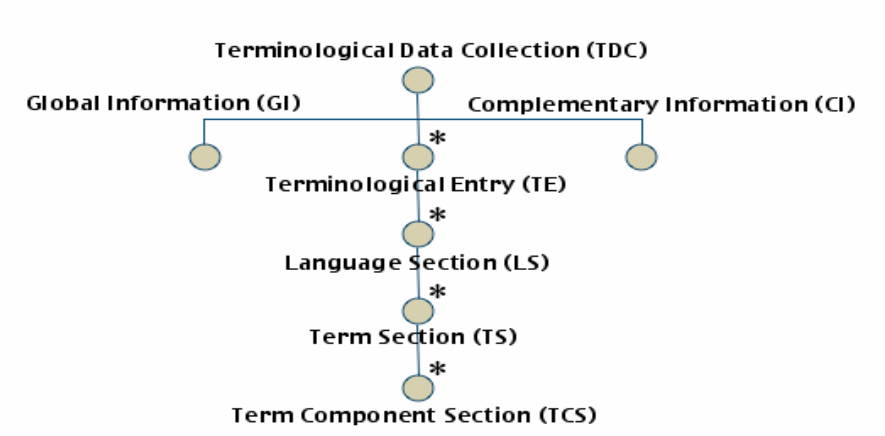

Figure 2 : *Le méta-modèle de TMF*

1.2.4. Application à la représentation de bases de données terminologiques

Dans cette section nous illustrons l'utilisation de TMF pour la modélisation de deux bases terminologiques d'origines différentes, dans la perspective de les intégrer au sein d'un même fonds terminologique de référence.

[7] On observe ici que même s'il est bien sûr applicable en situation monolingue, le modèle de l'ISO 16642 est par essence multilingue.

Ces exemples sont issus des travaux menés dans le cadre de la définition du portail TermSciences[8] [KHA 01], qui réunit un vocabulaire pluri-disciplinaire scientifique multilingue. Fusionner des données d'origines diverses nécessite donc dans un premier temps une analyse de leur similitude et de leur différence afin de proposer un format assez générique pour pouvoir conserver leurs spécificités. Effectivement, TMF propose l'instanciation d'un modèle noyau d'usage générique, en ce sens qu'il subsume les caractéristiques des terminologies récentes. Par ailleurs, ce modèle noyau peut être enrichi d'informations complémentaires (de *catégories de données*) pour des usages spécifiques.

Ainsi, le thesaurus de la BDSP[9] (Figure 3) indique que le concept [EMBRYON] a pour concept générique [STRUCTURE EMBRYONNAIRE], ce qui sera réalisé par une catégorie de données relationnelle /broaderConcept/. Dans l'autre base qui nous intéresse, éditée par l'INRA [BOU 01], les trois termes « embryon », « œuf » et « pré-embryon » lexicalisant ce même concept, ils entretiennent *de facto* une relation de parasynonymie. Toutefois, l'usage spécifique en contexte de ces termes, décrit par une note textuelle, peut être spécifié en TMF par des catégories de données appropriées, i.e. un marqueur d'usage. L'équivalence de traduction entre les termes français et anglais est implicite et peut être assimilée à un phénomène de parasynonymie inter-langues. L'exemple d' [EMBRYON] (Figure 3) montre que la description d'une donnée terminologique peut varier fortement d'une ressource à l'autre. Ainsi le thesaurus de la BDSP ne présente ni équivalent linguistique, ni définition, ni note, cependant il propose un terme générique du concept associé au terme « embryon ».

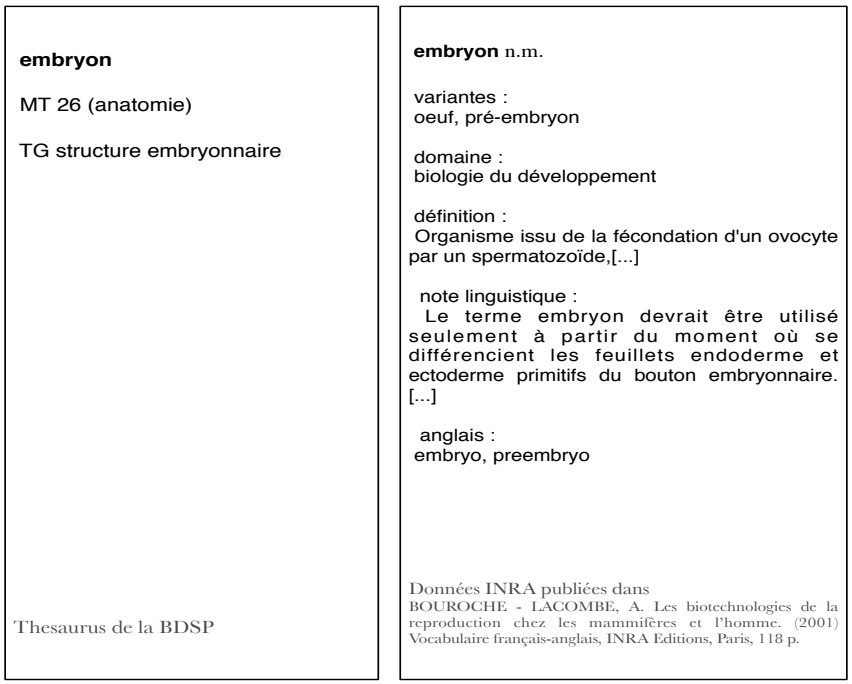

Figure 3 : *Différentes représentations du concept d'embryon (BDSP et INRA)*

Dans la ressource fusionnée, la définition proposée pourra, si elle est générique [ISO 704], être ancrée au niveau du concept (/terminologicalEntry/), ou à défaut au niveau de la langue (/languageSection/) ou du terme (/termSection/). Il en sera de même pour la note. L'ensemble des éléments extraits du thésaurus de l'INRA pourrait être accompagné de cette source. Les ressources une fois fusionnées grâce au formalisme proposé par TMF peuvent être schématisées comme dans la Figure 4 : Le concept [EMBRYON] rassemble l'ensemble des informations présentes dans les différents ressources. Chaque élément linguistique (définition, note et terme) est renseigné par une information concernant sa source. L'origine des données indique l'institution ou une référence bibliographique et peut faciliter d'autre part les mises à jour collaboratives.

[8] http://www.termsciences.fr
[9] Banque de Données Santé Publique, http://www.bdsp.tm.fr

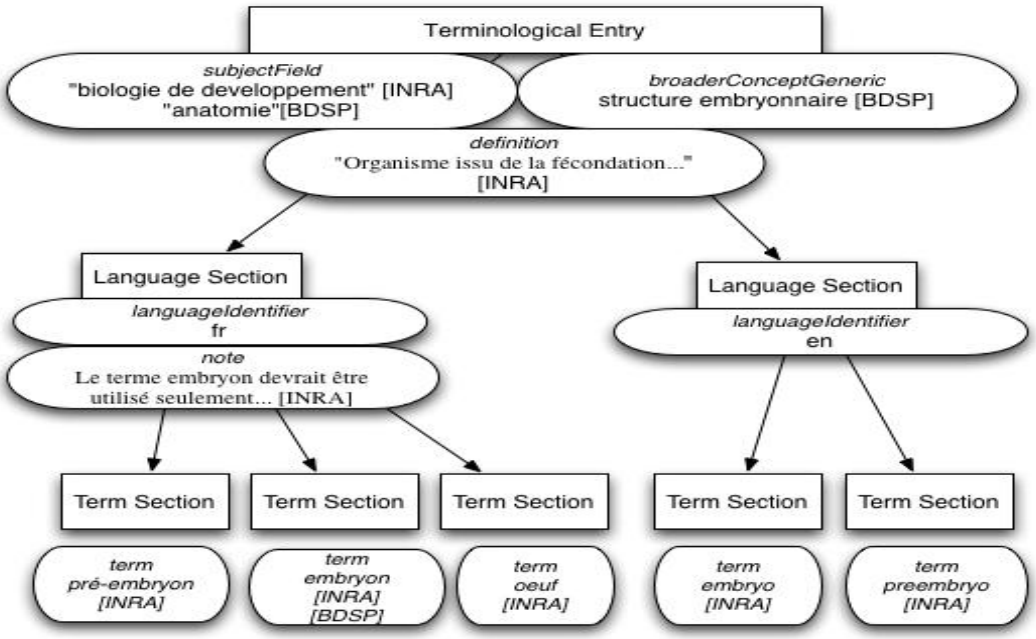

Figure 4 : *Représentation du concept d'embryon dans la ressource fusionnée*

En résumé, l'utilisation conjointe de la norme TMF [ISO 16642] et des catégories de données [ISO 12620] permet d'assurer la cohérence, la compatibilité et l'interopérabilité des données terminologiques. Toutefois, il n'existe pas à ce jour d'outil permettant de s'assurer qu'une ressource terminologique est réellement conforme à TMF, et la mise en évidence automatique de la bande passante d'interopérabilité est un vœu pieux : seule une analyse humaine rigoureuse permet jusqu'à aujourd'hui d'évaluer le degré d'interopérabilité. Or, le cadre normatif proposé par la norme ISO 16642 permet en quelque sorte à un expert de vérifier « manuellement » la compatibilité de sa représentation avec le méta-modèle pivot. Cette compatibilité de haut niveau garantit que les restrictions sur la bande passante d'interopérabilité sont uniquement liées aux catégories de données attachées aux différents niveaux du méta-modèle. L'identification de cette bande passante d'interopérabilité peut être automatisée en représentant les modèles spécifiques sous la forme d'ontologies[10] OWL [DEA 04] compatibles avec les logiques de description[11] [HOR 03]. L'expression des contraintes des modèles sous forme d'expressions logiques permet par l'utilisation de moteurs d'inférence[12] d'automatiser la tâche de classification des ressources.

Dans la suite, nous présentons les éléments fondamentaux que nous nous donnons pour la tâche de modélisation, puis leur utilisation pour la constitution d'une ontologie décrivant TMF. La section suivante présente l'utilisation de l'ontologie TMF pour construire des modèles terminologiques et les résultats pour la classification automatique de ces modèles.

1.3. Classification des modèles terminologiques

Le cadre de modélisation repose, comme TMF, sur le choix d'une séparation entre le méta-modèle et les catégories de données qui ornent le modèle. Le méta-modèle permet de structurer, d'organiser les informations de manière motivée. Le concept fondamental – dont hériteront tous les concepts associés aux composants du méta-modèle - est nommé composant (*Component*, cf. Figure 5). La relation *componentRelation* permet de lier les différents composants les uns aux autres. C'est cette relation qui assure la cohésion du méta-modèle, en exprimant, dans le cas de la norme ISO 16642 la structure hiérarchique d'une base de données terminologiques. Les catégories de données ornent les niveaux du méta-modèle. Ce sont des sous-concepts du concept *DataCategory* (correspondant aux catégories de données, cf. Figure 5). Le raffinement de ces catégories de

10 Une ontologie est une représentation conceptuelle des connaissances d'un domaine.
11 Les logiques de description sont une branche de la logique dédiée à la représentation des connaissances, largement utilisée pour la représentation des connaissances car combinant puissance d'expression et calculabilité.
12 Un moteur d'inférence explicite automatiquement des faits à partir d'une description donnée. Il peut aussi valider des propositions.

données par d'autres catégories de données est rendu possible par la relation *hasRefinement*. Et l'annotation du contenu des catégories de données est modélisée par l'attribut *hasAnnotation*. Enfin, le lien entre le méta-modèle et les catégories de données est réalisé par la relation *hasDataCategory*.

Les éléments de base pour la modélisation sont peu contraignants et en petit nombre pour favoriser l'interopérabilité entre modèles et leur réutilisation dans des cas variés. Ils s'intègrent par ailleurs dans une vision générique de la modélisation de données permettant d'étendre la méthodologie à d'autres ressources linguistiques. Dès lors, la description de l'ontologie du méta-modèle TMF importe cette ontologie fondamentale et l'étend avec ses niveaux, tous issus de *Component* : /terminologicalEntry/ (TE), /globalInformation/ (GI), /complementaryInformation/ (CI), /languageSection/ (LS), /termSection/ (TS) et /termComponentSection/ (TCS) (cf. Figure 6).

Dans un deuxième temps, les modèles terminologiques particuliers sont définis par les catégories qui ornent les différents niveaux du méta-modèle TMF. Les catégories introduites dans la modélisation sont issues de *DataCategory*. La figure 7 montre ainsi comment on peut déclarer les deux catégories /grammaticalGender/ et /definition/ comme faisant partie de l'ontologie globale.

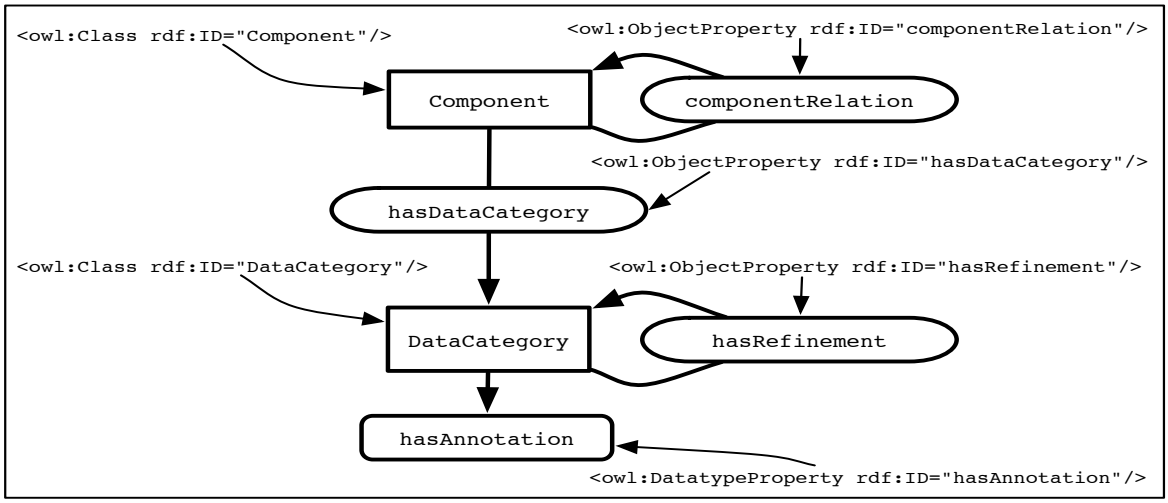

Figure 5 : *Concepts et propriétés de base*

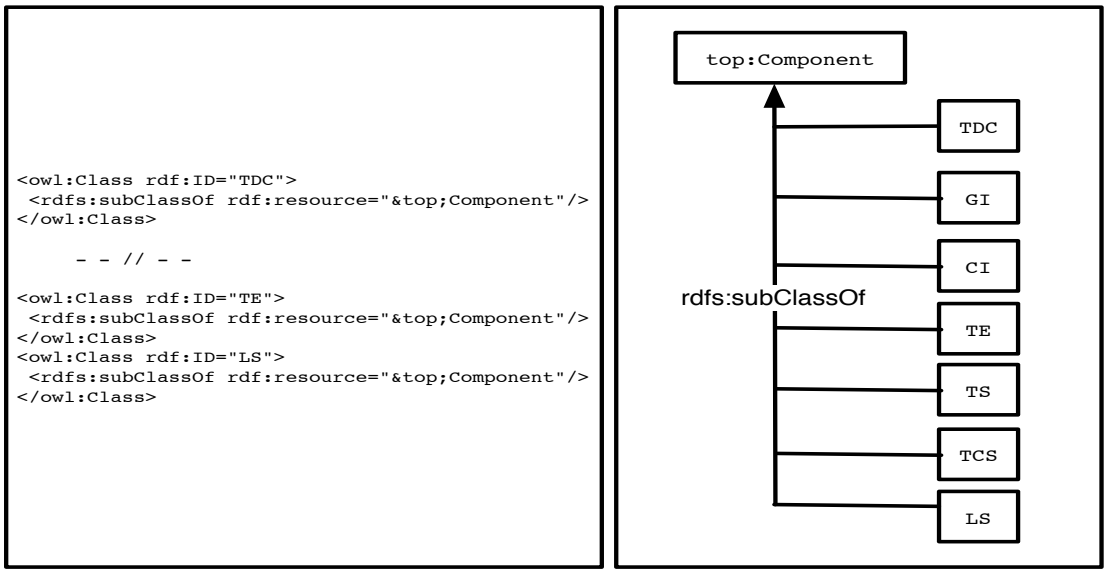

Figure 6: *Déclaration des niveaux de TMF et représentation graphique correspondante du fragment d'ontologie*

Figure 7 : *Déclaration des catégories de données*

Contraintes logiques pour les catégories de données des modèles terminologiques

En OWL-DL, tel que spécifié dans la norme du W3C, il n'est actuellement pas possible de définir pour un concept des restrictions sur la cardinalité d'une relation en fonction des concepts liés (QCR : *Qualified Cardinality Restrictions*). La demande étant forte de voir cette possibilité intégrée [KNU 05], la prochaine version de OWL la prendra certainement en compte. L'éditeur d'ontologie *Protégé* [KNU 04] associé au moteur d'inférence *Racer* [HAA 03] en a déjà produit la spécification et l'implémentation. Pour l'heure et afin d'exprimer une contrainte équivalente, nous créons pour chaque catégorie de données une relation héritant de *hasDataCategory*.

Classification des modèles

Le langage OWL-DL permet de fusionner les modèles terminologiques représentées sous forme d'ontologies. Cela consiste à déclarer l'équivalence des catégories de données et des propriétés utilisées indépendamment dans les différents modèles. Une fois ces relations d'équivalence établies, la classification automatique est réalisée en utilisant un moteur d'inférence, sur la base des différences entre ornementations des niveaux du modèle.

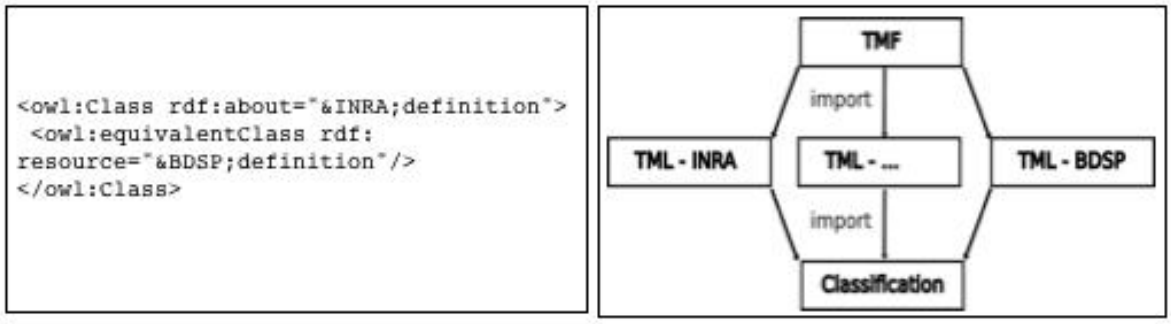

Figure 8 : *processus de classification des TMLs*

Par exemple, si le niveau **TE** du modèle de la BDSP (TML-BDSP dans la figure 8) ne comporte pas d'équivalent à la catégorie de données */definition/* présente dans le niveau **TE** du modèle INRA (TML-INRA) - les autres catégories de données présentes étant toutes équivalentes -, le moteur d'inférence le détectera et classera le niveau **TE** de TML-INRA comme étant plus spécifique que celui de TML-BDSP. D'un point de vue logique on pourra parler de subsomption entre les deux descriptions du composants TE pour l'INRA par rapport à celle de la BDSP.

En généralisant cette démarche, on voit ainsi que l'on peut, grâce à cette formalisation, identifier si un modèle terminologique donné est compatible avec un autre, ou s'il peut être attaché à une classe existante de modèles. De nouveau, il s'agit là d'une relation générale de subsomption entre modèles. La traduction en OWL-

DL des modèles terminologiques apporte ainsi au travail terminographique l'automatisation d'une partie des tâches, facilitant la détection et l'aménagement de la bande passante d'interopérabilité entre les modèles.

1.4. Représentation de données terminologiques

Dans la section précédente, nous avons vu comment les principes de la norme ISO 16642 pouvaient se traduire sous la forme de contraintes formelles permettant de spécifier précisément la couverture informationnelle d'un modèle donné et, ce faisant, de comparer deux modèles afin de fournir des conditions précises d'interopérabilité. Il reste maintenant à pouvoir traduire concrètement ces contraintes afin de fournir de véritables modèles de données, notamment sous la forme de structures XML qui vont servir de base à la réalisation d'outils de manipulation et d'échange de données terminologiques. C'est cette étape que nous présentons dans cette section, en cherchant de nouveau à nous intégrer dans un cadre suffisamment générique pour que la réponse apportée puisse couvrir une large variété d'applications possibles.

La démarche envisagée ici n'est pas de fournir un format unique de représentation de données terminologiques, sachant que suivant les usages, un tel format risque de toujours être considéré soit comme trop complexe, soit comme ne répondant pas exactement à tel ou tel besoin particulier. Nous souhaitons en fait montrer qu'il est possible d'envisager une famille de formats, dans un cadre de représentation suffisamment accessible, pour que de nombreuses catégories d'utilisateurs puissent sans trop de difficultés adapter le canevas proposé à leurs propres besoins. Nous nous situons ainsi dans la continuité des réflexions menées par les acteurs de la normalisation eux-mêmes [ROM 01] [BAU 04] qui considèrent que la proposition d'un cadre de normalisation n'est pas incompatible avec l'identification d'opérations d'adaptation des normes par extension ou restriction d'un modèle de données. De fait, il s'agit de transposer dans le domaine de la représentation de données la notion de subsumption de modèles que nous avons introduite dans la section précédente. L'objectif ultime est d'arriver à la définition d'une véritable plateforme de spécification de données terminologiques qui soit à même de garantir qu'un même élément de donnée sera représenté de façon identique dans deux applications différentes.

Les choix que nous présentons ici ont été guidés par une autre préoccupation importante, à savoir la nécessité de prévoir l'intégration potentielle de données terminologiques dans un cadre plus large de représentation de documents textuels. En effet, les données terminologiques sont indissociables, et c'est tout le sens de la démarche de la terminologie textuelle, avec les documents où elles sont utilisées. Concrètement, cela se traduit, du point de vue des textes, à pouvoir annoter les termes rencontrés et faire des références explicites aux entrées d'une base terminologique. Il s'agit aussi parfois d'intégrer dans le texte lui-même des descriptions terminologiques quand une nouvelle notion est introduite. Dans cette perspective, on peut bien sûr citer les textes de normalisation de l'ISO qui intègrent, comme une section obligatoire, l'ensemble des termes et définitions utilisés dans le corps du texte. Inversement, les données terminologiques peuvent, comme nous l'avons vu, des illustrations textuelles, dont la représentation nécessitera la représentation de structures (paragraphes, phrases, mais aussi annotations de surface) qui ne sont pas propres au domaine de la terminologie. Il est donc important, dès lors que cela semble être possible, d'offrir une représentation de données terminologiques qui s'intègre dans un cadre plus vaste de représentation de documents textuels.

La TEI (Text Encoding Initiative ; http://www.tei-c.org) répond de fait à ces exigences et ce à différents niveaux :

- Il s'agit d'une initiative internationale qui, depuis 1987, regroupe l'essentiel des acteurs ayant à gérer de grands projets de données textuelles ;
- Elle couvre de nombreux domaines d'application, prose, poésie, théatre, manuscrits, dictionnaires… et terminologie ; ce dernier domaine ayant fait l'objet d'un chapitre spécifique devenu obsolète, et dont la mise à jour était de toute façon requise ;
- Elle offre une plateforme de spécification, ODD, qui est un cadre idéal pour implémenter l'approche que nous défendons ici, à savoir la définition d'une famille de modèles compatibles.

ODD (One Document Does it all) est un langage de spécification de données qui repose sur le principe de la programmation littéraire (literate programming), c'est-à-dire qui combine des éléments descriptifs avec des éléments formels afin de fournir, à partir d'une source unique, à la fois un schéma permettant de contrôler la syntaxe effective d'un document, et une documentation fournissant à un utilisateur la sémantique fine des objets définis dans la spécification. Sans entrer ici dans les détails trop techniques, nous présentons ici les deux

caractéristiques essentielles de ODD, à savoir les notions de modules et de classes, pour ensuite fournir quelques indications sur la spécification des objets XML proprement dits. Nous montrerons ensuite comment ces principes sont appliqués à la définition d'un module terminologique intégré aux directives de la TEI.

La plateforme ODD permet d'organiser toute structure documentaire comme une combinaison d'un ou plusieurs modules réunissant un ensemble cohérent d'éléments et de classes (cf. *supra*). Les directives de la TEI proposent ainsi des modules permettant de représenter l'en-tête d'un document, les éléments communs (e.g. divisions) à tous types de documents, les éléments propres au théâtre, à la poésie, etc. Un utilisateur peut ainsi décider d'utiliser les modules de base permettant de représenter des données textuelles simples et d'y adjoindre, et c'est le cas qui nous intéresse, le module terminologique afin d'insérer des descriptions de termes dans le corps du texte.

Deux types principaux d'objets sont décrits à l'intérieur d'un module, des éléments (au sens XML du terme) et des classes. Les classes permettent de regrouper des éléments ayant un comportement syntaxique ou une sémantique similaire. Ainsi, tous les éléments donnant des indications morphosyntaxiques dans un dictionnaire ou une terminologie appartiennent à la classe *model.morphLike*. De la sorte, si l'on souhaite intégrer tous ces éléments, dans un modèle de contenu, il suffit de faire référence à la classe en question, et de façon complémentaire, pour ajouter un descripteur morphosyntaxique, il suffit d'ajouter un élément à la classe.

Pour la définition des modèles de contenu, la TEI s'appuie sur des fragments élémentaires de schémas RelaxNg qui sont ensuite combinés pour générer des schémas RelaxNg complets, mais aussi des « classiques » DTD XML, ainsi que des schémas W3C.

Dans la suite de cette section, nous présentons plus précisément les principales caractéristiques du schéma de représentation adopté pour la définition du chapitre terminologie de la TEI. Cette syntaxe s'inspire fortement du format TBX adopté par l'association LISA, lui-même issu de l'ancienne norme ISO 12200 (Martif – Machine Readable Terminological Interchange Format). Cependant, comme nous l'avons signalé plus haut, nous ne reprenons pas l'ensemble des descripteurs, mais seulement ceux qui correspondent à un scénario de base de représentation de données terminologiques multilingues.

Tout d'abord, nous avons fait le choix de n'instancier des éléments que pour les composants TE , LS et TCS du modèle TMF. Il s'agit en effet de limiter la proposition à la représentation d'entrées terminologiques proprement dits, qui peuvent ensuite s'intégrer sous différentes formes dans un document plus complexe, soit sous la forme d'un répertoire terminologique indépendant (à l'intérieur d'une division d'un document TEI) ou à l'intérieur même d'un texte, pour illustrer par exemple une notion en cours de description. Les trois composants sont instanciés par les trois éléments : <termEntry>, <langSet> et <tig>, afin de garder la compatibilité avec le vocabulaire TBX.

Le modèle de contenu de chacun de ces éléments est ensuite décrit à l'aide de deux classes regroupants les catégories de données administratives et descriptives pertinentes pour le composant correspondant. Ainsi, au niveau TS (Term Section, élément <tig>), les deux classes model.TSAdminParts et model.TSDescParts regroupent les éléments administratifs et descriptifs associés au niveau du terme.

Le tableau suivant présente ainsi les principaux éléments XML du module terminologie[13] et les classes auxquelles ils appartiennent.

Elément	broaderConcept	subjectField	definition	languageIdentifier	note	term
Classe(s)	model.TEDescParts	model.TEDescParts	model.TEDescParts model.LSDescParts	model.LSAdmParts	model.TEDescParts model.LSDescParts model.TSDescParts	N/A

Tableau 1 : *Catégories noyau du module terminologie et leurs classes d'appartenance*.

[13] Afin de ne pas alourdir le texte, les références aux catégories de données de l'ISO 12620 :1999 ne sont pas fournies. Les noms des éléments sont suffisamment explicites pour qu'il soit possible de faire le lien de façon quasi-automatique avec la norme.

Ces éléments correspondent à un modèle noyau de représentation de données terminologiques utilisable pour une classe d'applications simples de repérage de données terminologiques multilingues. Ainsi, l'exemple [Embryon] issu de [BOU 01] peut être représenté sous la forme suivante, conforme au module terminologie de la TEI :

```
<termEntry xmlns="http://www.tei-c.org/ns/1.0">
    <broaderConcept>structure embryonnaire</broaderConcept>
    <subjectField>biologie du developpement</subjectField>
    <definition>Organisme issu de la fécondation d'un ovocyte par un
    spermatozoïde, depuis la première segmentation jusqu'à la différentiation des
    tissus et des organes.</definition>
    <langSet>
        <languageIdentifier>fr</languageIdentifier>
          <note>Le terme embryon devrait être utilisé seulement à partir du
          moment où se différencient les feuillets endoderme et ectoderme
          primitifs du bouton embryonnaire. Avant ce stade, les termes corrects
          sont zygote, morula, puis blastocyste. On parle de foetus quand sont
          formés les différents organes. Le terme pré-embryon est utilisé dans
          les textes législatifs de certains pays pour désigner l'embryon humain
          depuis la fécondation jusqu'au 14° jour.</note>
        <tig><term>pré-embryon</term></tig>
        <tig><term>embryon</term></tig>
        <tig><term>oeuf</term></tig>
    </langSet>
    <langSet>
        <languageIdentifier>en</languageIdentifier>
        <tig><term>preembryo</term></tig>
        <tig><term>embryo</term></tig>
    </langSet>
</termEntry>
```

Une fois définis les éléments de base d'un format de représentation de données terminologique, la TEI offre un mécanisme naturel d'extension qui permet à tout utilisateur d'adapter la structure proposée aux besoins de sa propre application.

Le cas le plus courant d'extension qu'un utilisateur souhaite pouvoir effectuer consiste à ajouter une catégorie de données au modèle minimal. L'organisation des éléments en classes rend une telle opération particulièrement simple puisqu'il suffit de déclarer un élément (déjà disponible ou nouvellement créé) comme faisant partie de la classe correspondant au comportement syntaxique attendu. Ainsi, si l'on souhaite ajouter des marques d'usage à la description d'un terme, il suffit de rattacher l'élément <usg>, déjà disponible dans les directives de la TEI et de l'ajouter à la classe model.TSDescParts, suivant le modèle suivant :

```
<elementSpec ident="usg" mode="change">
   <classes>
      <memberOf key="model.TSDescParts"/>
   </classes>
</elementSpec>
```

Une autre possibilité offerte par les mécanismes d'extension de la TEI est de restreindre le modèle de contenu d'un attribut ou d'un élément. Par exemple, on peut vouloir limiter les domaines utilisés dans un projet à une liste contrôlée précise. Dans ce cas, il s'agit d'écrire une instruction ODD qui indique un changement par rapport au modèle de référence de l'élément correspondant (ici <subjectField>). Une telle instruction aura la forme suivante :

```
<elementSpec ident="subjectField" mode="change">
    <content>
        <valList type="closed">
            <valItem ident="psychologie"/>
            <valItem ident="linguistique"/>
            …
```

```
            </valList>
        </content>
    </elementSpec>
```

1.5. Conclusion

Dans ce chapitre, nous avons voulu montrer qu'il était possible d'envisager de façon cohérente les différentes phases de conception d'un projet de représentation de données terminologiques. Partant des principes de bases en terminologie, tels qu'énoncés par l'école de Vienne et repris dans la norme ISO 704, nous avons essayé de présenter différents points de vue associés à l'organisation de données terminologiques : méta-modèle générique de représentation (ISO 16642), formalisation des modèles (OWL) et représentation concrète (spécification XML). Ces différents points de vue sont destinés à convaincre les différents acteurs d'un projet en terminologie qu'il existe un continuum entre un travail véritablement linguistique de recueil de données terminologiques et le développement d'un environnement informatique destiné à accueillir de telles données.

Sur un plan pratique, la démarche proposée, qui reste encore à affiner, permet d'envisager des projets complexes de terminologie, qui peuvent intégrer, au sein d'une même interface, des accès à des fonds hétérogènes dans leur couverture et leur complexité informationnelles. On peut ainsi combiner, dans ce qui pourrait être taxé d'irénisme forcené, des fonds terminologiques extrêmement normatifs (le vocabulaire officiel de la DGLF-LF[14]) et des terminologies concrètes observées dans les pratiques textuelles (avec les attestations correspondantes).

D'un point de vue plus fondamental, nous avons au passage montré, d'une part, que l'approche wüstérienne était pleinement d'actualité pour qui veut représenter concrètement des données terminologiques, et, d'autre part, que la spécification d'un format de représentation pouvait se faire en bonne intelligence avec une approche normative plus large, telle que proposée par la TEI.

1.6. Index des abréviations et sigles

BDSP : *Banque de Données Santé Publique*. http://www.bdsp.tm.fr

DGLF-LG : *Délégation Générale à la Langue Française et aux Langues de France*.

INRA : *Institut National de la Recherche Agronomique*. http://www.inra.fr/

ISO : *Organisation Internationale de Standardisation*. http://www.iso.org

ODD : *One Document Does it all*

OWL : *Web Ontology Language*. http://www.w3.org/TR/owl-features/

QCR : : *Qualified Cardinality Restrictions*

TBX : *TermBase eXchange*. http://www.lisa.org/standards/tbx/

TEI : *Text Encoding Initiative*. http://www.tei-c.org

TMF : *Terminological Markup Framework* [ISO 16642]. http://www.loria.fr/projets/TMF/

TML : *Terminological Markup Language*

W3C : *World Wide Web Consortium*. http://www.w3c.orglogi

1.7. Bibliographie

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

[BOU 00] BOURIGAULT D., SLODZIAN M., « Pour une terminologie textuelle », *Terminologies Nouvelles*, 19, p. 29-32, 2000.

[BOU 01] BOUROCHE-LACOMBE A., *Les biotechnologies de la reproduction chez les mammifères et l'homme. Vocabulaire français-anglais*, INRA Editions, Paris, 118 p., ISBN 2-7380-0935-2, 2001.

[CON 03] CONDAMINES A., Sémantique et corpus spécialisés : constitution de bases de connaissances terminologiques. Habilitation à Diriger des Recherches, Université Toulouse Le Mirail, 2003.

[14] Délégation Générale à la Langue Française et aux Langues de France.

[DEA 04] DEAN M., SCHREIBER G., BECHHOFER S., VAN HARMELEN F., HENDLER J., HORROCKS I., MCGUINNESS D. L., PATEL-SCHNEIDER P. F., STEIN L. A., OWL Web Ontology Language Reference. W3C Recommendation (http://www.w3.org/TR/owl-ref/), 10 février 2004.

[FEL 01] FELBER H., *Allgemeine Terminologielehre, Wissenslehre und Wissenstechnik. theoretische Grundlagen und philosophische Betrachtungen*, TermNet Publisher, Wien, 2001.

[HAA 03] HAARSLEV V., MÖLLER R., « Racer: A Core Inference Engine for the Semantic Web », *2nd International Workshop on Evaluation of Ontology-based Tools (EON2003)*, Sanibel Island, Florida, USA, p. 27-36, 20 octobre 2003.

[HOR 03] HORROCKS I., PATEL-SCHNEIDER P. F., VAN HARMELEN F., « From SHIQ and RDF to OWL: The Making of a Web Ontology Language », *Journal of Web Semantics*, vol. 1, n° 1, p. 7-26, 2003.

[ISO 704] ISO 704, Terminology work – Principles and methods, 2000.

[ISO 12200] ISO 12200:1999, Machine Readable Terminological Interchange Format, 1999.

[ISO 12620] ISO 12620:1999, Computer applications in terminology – Data categories, 1999.

[ISO 16642] ISO 16642, Computer applications in terminology - Terminological markup framework (TMF), 2003.

[KHA 06] KHAYARI M., SCHNEIDER S., KRAMER I., ROMARY L., « Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative. », International Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine, Genoa : Italie, 2006 (http://hal.ccsd.cnrs.fr/ccsd-00022424)

[KNU 04] Knublauch H., Fergerson R. W., Noy N. F., Musen M. A., « The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications », *Third International Semantic Web Conference*, Hiroshima, Japan, 2004.

[KNU 05] KNUBLAUCH H., HORRIDGE M., MUSEN M., RECTOR A., STEVENS R., DRUMMOND N., LORD P., NOY N. F., SEIDENBERG J., WANG H., « The Protégé OWL Experience », *Workshop on OWL: Experiences and Directions*, Galway, Ireland, 2005.

[RAS 95] RASTIER F., « Le terme : entre ontologie et linguistique », *La banque des mots*, 7, p. 35-65, 1995.

[ROM 01] ROMARY L., « Un modèle abstrait pour la représentation de terminologies multilingues informatisées », *Cahiers GUTenberg*, 39-40, p. 81-88, mai 2001.

[WRI 97] WRIGHT S. E., BUDIN G. (Red.), *Handbook of Terminology Management. Volume I : Basic Aspects of Terminology Management*, Amsterdam/Philadelphia, John Benjamins, 370 p., 1997.

[WRI 01] WRIGHT S. E., BUDIN G. (Red.), *Handbook of Terminology Management. Volume II : Application-Oriented Terminology Management*, Amsterdam/Philadelphia, John Benjamins, 550 p., 2001.

[WÜS 53] WÜSTER E., « Die terminologische Sprachbehandlung », *Studium Generale*, 6/4, p. 214-219, 1953.

[WÜS 59/60] WÜSTER E., « Das Worten der Welt, schaubildlich und terminologisch dargestellt », *Sprachforum*, 3-4, p. 183-204, 1959/60.

[WÜS 74] WÜSTER E., « Die Allgemeine Terminologielehre — ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. », *Linguistics*, 119, p. 61-106, janvier 1974.

[WÜS 68] WÜSTER E.,. Dictionnaire multilingue de la machine-outil. Notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et l'ordre alphabétique. Volume de base anglais-français, Technical Press, London, 1968.