Maintenance de bases de connaissances terminologiques

Daniel Beauchêne*, Christophe Roche*
Cécile Million-Rousseau**

* Equipe Condillac « Ingénierie des Connaissances » Laboratoire LISTIC
Campus Scientifique 73376 Le Bourget du Lac cedex
Daniel.beauchene@univ-savoie.fr Christophe.roche@univ-savoie.fr
http://ontology.univ-savoie.fr/
** Ontologos Corp.
178 Rte de Cran Gevrier 74650 Chavanod
Cecile.Million-Rousseau@ontologos-corp.com
http://www.ontologos-Corp.com

Résumé. L'acquisition des connaissances terminologiques de l'entreprise se fait souvent à partir des textes qu'elle utilise. Dans le cadre de ce travail, la base de connaissances terminologiques repose sur la modélisation des concepts-métier sous la forme d'une ontologie. Le problème de la maintenance de cette base et de cette ontologie doit alors être traité.
Dans cet article, après avoir donné une définition d'une base de connaissances terminologiques (BCT) et des problèmes de diachronie, nous présentons notre modèle et notre méthode d'acquisition des connaissances terminologiques de l'entreprise. Nous exposons alors notre proposition pour maintenir au cours du temps la base de connaissances terminologiques ainsi construite.
Nous illustrons ce travail sur une base de connaissance terminologique sur le cinéma d'animation en décrivant le problème de la maintenance dans une reconstitution historique de différents états de cette base lors de l'apparition des techniques numériques d'animation.

1 Introduction

L'évolution des marchés et des technologies a conduit à une modification profonde de nos sociétés. Le nouvel enjeu économique est devenu la maîtrise des informations, des connaissances et des savoir-faire (Roche 2001). Or, à travers les documents qu'elles produisent ou utilisent, les entreprises rendent explicite une partie des connaissances liées à leur savoir-faire et à leur métier (Aussenac-Gilles et Condamines 2001). Il est donc naturel de gérer ces connaissances et ces textes dans une structure commune : la base de connaissance terminologique (BCT).

Cette BCT évolue au cours du temps, au fur à mesure de la production de nouveaux textes et de l'obsolescence de certains autres. Il y a donc nécessité de maintenir à jour la BCT de l'entreprise.

Dans cet article, après avoir présenté notre vision de la BCT et de son évolution, nous précisons notre méthode de construction d'une BCT dont la sémantique des termes repose sur une modélisation ontologique (Roche et Million-Rousseau 2003). Nous décrivons ensuite notre proposition pour la maintenance de cette BCT avec un exemple de mise en œuvre.

2 Bases de connaissances terminologiques et diachronie

Le concept de « Terminological Knowledge Base » a été introduit par (Meyer et al. 1992) à partir de celui de « Terminological Data Base » et de « Knowledge Base ». Il s'agissait de regrouper dans une même structure, des informations de nature lexicale sur les termes et des connaissances expertes sur le sens des concepts dénotés par ces termes.

2.1 Bases de connaissances terminologiques

Une base de connaissances terminologique a pour ambition de rassembler, mettre à disposition les connaissances issues des textes. Pour cela, (Aussenac-Gilles et Condamines 2001) affirment qu'une BCT doit contenir une modélisation des textes qu'elle référence.

Pour nous, une base de connaissance terminologique est constituée d'un ensemble de termes, mots ou expressions du langage-métier, d'un ensemble de concepts dénotés par les termes et qui sont leur signification[1], et enfin, d'un ensemble de textes qui sont les sources d'information et caractérisent le sens des termes (signification en contexte).

Les experts du métier valident le corpus de textes ainsi que les termes et concepts retenus.

Dans nos BCT, les concepts sont organisés en une ontologie OK (Ontological Knowledge) (Roche 2000), à partir de la relation de subsomption.

2.2 Diachronie

Une BCT est construite à un moment donné à partir des textes, termes et concepts existant à ce moment. L'évolution du métier, l'apparition de nouveaux mots et concepts, la disparition d'autres, les changements de sens et de signification des termes doivent être gérés dans la BCT. Nous émettons l'hypothèse que ces évolutions sont accessibles à travers la production ou la modification des textes utilisés par l'entreprise.

Bernt Moeller a étudié les évolutions du couple (« terme » ; <concept>) (Moeller 1998). En schématisant, ce couple peut subir quatre types de modifications :

- Le terme et le concept ne changent pas
- Seul le concept change (dans sa définition)
- Seul le terme change (dans sa relation avec le concept)
- Le terme et le concept changent.

Dans notre représentation (FIG 1), nous nous intéressons à l'évolution des termes et de la relation de dénotation, des concepts et de leur définition.

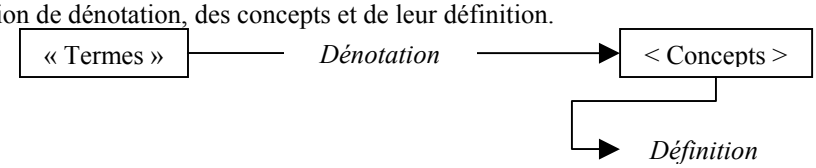

FIG. 1 - *Les éléments d'une BCT susceptibles d'évoluer*

[1] Signification considérée comme décontextualisée, à distinguer du sens, pris comme une signification actualisée en langue

Le premier cas correspond à un état de stabilité du terme et du concept. Il est donc une référence permettant de mesurer la dynamique terminologique[2] de la base documentaire.

Le deuxième correspond à une évolution de la sémantique sans que le terme n'évolue (c'est-à-dire sans que la relation de dénotation du terme au concept change). Dans notre représentation, il s'agit de l'évolution de la définition du concept. Cette situation est généralement liée aux évolutions du domaine pour lequel on étudie la terminologie. Pascaline Dury (Dury 1999), dans une étude sur les concepts dénotés par le terme « écosystème » en français et « ecosystem » en anglais, montre comment les évolutions de cette science conduisent à une évolution des concepts dénotés sans que leur désignation évolue.

L'évolution des termes qui désignent un concept, sans que ce dernier évolue est un phénomène fréquent, tout particulièrement pendant la période qui suit immédiatement l'apparition d'un nouveau concept. Valérie Bonnet (Bonnet 2003) étudie les mécanismes de formation des néologismes dans les activités scientifiques et techniques en mettant en évidence la simplification courante de désignation en partant d'une structure de type nom-préposition-nom à une structure nom-adjectif, puis à une structure de nom simple. Jacqueline Percebois (Percebois 2001) s'intéresse à une simplification encore plus grande du terme sous forme de sigle ou d'acronyme dans un processus qu'elle nomme « siglaison ».

Enfin, le dernier cas correspond, dans notre modèle, à l'apparition d'un nouveau concept et du terme qui le désigne. Il faut alors introduire ce concept dans l'ontologie en définissant le concept le plus proche qui le subsume et en identifiant la différence spécifique qui le définit. Les définitions des autres termes doivent alors être révisées en positionnant chaque terme par rapport aux nouvelles différences spécifiques.

3 Terminologies OK

Le choix de notre modèle ontologique a été guidé par le type d'applications visées et les objectifs à atteindre et en particulier les propriétés de consensus, cohérence, partage et réutilisabilité.

La dimension terminologique est, pour nous, primordiale en regard de l'importance des informations textuelles pour nos applications. Nous avons adopté une approche normative, conforme à la norme ISO 704 (NF ISO 704 2001) des signifiés lexicaux, basée sur une sémantique lexicale componentielle (traits sémantiques différentiels). Cette approche reste acceptable pour une langue technique où le lexique est propre à un groupe social défini par une activité spécifique et où l'élaboration du sens d'un mot s'appuie sur l'élaboration d'une idée.

3.1 La méthodologie OK

La première étape de la construction d'une terminologie OK consiste à définir les besoins en terme d'application et à comprendre le métier qui doit être modélisé. Ensuite vient la construction du lexique métier subdivisée en trois phases : la constitution d'un corpus textuel de référence représentatif du métier ; l'extraction automatique des candidats termes (substantifs, expressions et mots inconnus) à l'aide de notre logiciel Linguistic Craft

[2] Mesure de caractéristiques (nombre de termes qui changent, nombre de changements par terme, etc.) de l'évolution des termes et concepts dans la BCT

Workbench (LCW)[3], et la validation de l'expert ; enfin, la différentiation entre le vocabulaire appartenant aux lexiques utilisateurs ou experts qui prend toute son importance dans une application mettant en relation différents profils d'acteurs (clients, fournisseurs, experts métier, …). L'étape suivante a pour objectif de définir de façon précise les candidats termes. Dans un premier temps, l'expert parcourt le lexique afin de classer chaque terme dans une catégorie existante ou dans une nouvelle catégorie. Pour chaque catégorie, l'expert attribut à chaque terme sa nature : concept, ensemble, différence, attribut, valeur d'attribut ou relation. La dernière phase de cette étape correspond à la construction de la terminologie OK proprement dite. Cette construction s'effectue de deux manières complémentaires :

- par la définition d'une taxinomie de concepts en précisant la relation de subsomption « sorte-de » entre les concepts, puis par la différentiation des concepts associés à un même concept père pour aboutir à l'arbre de Porphyre,
- par la grille des caractères essentiels ou grille des différences dont la méthode systématique permet la définition et la comparaison des concepts. Cette grille est composée de colonnes représentant les concepts et de lignes correspondant aux différences. L'expert précise pour chaque concept s'il possède ou non la différence ou si la différence n'a pas de sens pour ce concept. La construction de l'arbre de Porphyre est alors automatisable.

4 Diachronie OK

Nous nous intéressons dans cette partie à la mise à jour de la terminologie OK lors de l'apparition d'un nouveau concept et du terme qui le désigne.

4.1 Notre approche

Lors de la maintenance de la BCT, nous analysons un nouveau corpus constitué de textes rédigés depuis la construction du référentiel métier ou la dernière maintenance. LCW élague automatiquement les lexiques générés afin d'éliminer les termes déjà présents dans la terminologie ainsi que les termes non retenus par l'expert lors de la validation des précédents lexiques[4]. L'expert valide les nouveaux lexiques afin d'en extraire les nouveaux termes. De la même manière que cela a été fait pour la construction de la terminologie, l'expert identifie ensuite à quelle catégorie appartient le nouveau terme et précise sa nature : concept ou synonyme de concept existant.

L'expert compare alors le nouveau terme avec chaque concept de sa catégorie en partant du concept le plus général (racine de l'arbre de Porphyre). Il s'agit de parcourir l'arbre en déterminant à chaque niveau quelle différence caractérise le nouveau terme. Lorsque l'expert arrive sur un concept de plus bas niveau ou lorsque aucune des différences du couple spécifiant les concepts suivants n'a de sens pour le nouveau terme, la position du nouveau terme est alors délimitée. Si le nouveau terme ne possède pas de trait différentiateur par rapport au concept sur lequel s'est arrêté l'expert, le terme sera associé comme synonyme de

[3] LCW propose sur une analyse linguistique du corpus (lemmatisation sur le principe du Tagger de Brill, définition de patterns d'expressions, génération de cooccurrences, contextualisation des termes,)

[4] Le fait d'avoir marqué les termes non retenus aux étapes précédentes est un gain de temps considérable par rapport à la construction initiale

ce concept. Dans le cas contraire, un nouveau couple de différences est créé et le terme est inséré dans l'arbre comme nouveau concept.

Chacun des concepts existant doit alors être comparé avec ce couple de différences pour situer si l'une d'entre elle le concerne (sa définition est alors modifiée) où si ces différences sont sans objet pour ce concept.

4.2 Un exemple d'application

Le Centre International du Cinéma d'Animation (CICA) référence, aujourd'hui, 17 821 films d'animation de toutes natures : courts-métrages, longs-métrages, films de publicité, films de télévision, films d'école, etc. Certains de ces films ont été candidats à la présentation au Festival international du film d'animation (FIFA) qui a lieu chaque année à Annecy. Une des caractéristiques de cette base est qu'elle est mise à jour une fois par an au moment des inscriptions des films au festival. Le CICA souhaite permettre au public de consulter en ligne les fiches techniques de ces films et, éventuellement, d'en visualiser un extrait.

La base de films est donc indexée à partir de l'ontologie du cinéma d'animation. Nous avons reconstitué, sur la catégorie des techniques d'animation, l'évolution de cette catégorie, au fil de l'apparition des techniques numériques dans le cinéma d'animation.

Nous avons eu à étudier, entre autres, les évolutions suivantes :

- Le premier film d'animation animé à l'aide de l'ordinateur apparaît en 1968. Dès le sommet de l'arbre de Porphyre, les différences spécifiques ne sont pas pertinentes pour intégrer cette nouvelle technique. Nous définissons donc un nouveau couple de différences : animation numérique vs animation traditionnelle.
- En 1974, apparaît le concept de dessin à l'aide de techniques numériques. Ce concept s'oppose aux techniques numériques d'animation utilisées jusqu'alors qui ne sont pas « figuratives ».
- En 1980, la différence spécifique entre animation de dessins et animation de volumes devient pertinente pour l'animation numérique et modifie la définition du concept de dessin numérique. Les termes 2D et 3D qui les désignent aujourd'hui, apparaissent alors sur les fiches descriptives des films.
- Enfin, en 1997, les techniques spécifiques aux films destinés à l'Internet se différencient.

L'intégration de ces nouvelles techniques dans l'ontologie (par l'intermédiaire de l'arbre de Porphyre qui représente la catégorie concernée) se fait simplement par les experts dans la mesure où peu de concepts nouveaux apparaissent chaque année et où il n'y a pas de disparition de concepts.

5 Conclusion

Le travail présenté dans cet article est le fruit d'une collaboration entre l'équipe Condillac, la société Ontologos-Corp et le CICA. Les résultats obtenus ont permis de définir une méthode de maintenance de la BCT.

Nos travaux actuels visent à automatiser au maximum la phase d'intégration des nouveaux concepts en utilisant des outils linguistiques pour alléger le travail des experts. Il

nous faut étudier la disparition de concepts et/ou de différences spécifiques qui ne sont pas pertinents dans l'application à cette base.

Enfin, nous validons actuellement cette méthode sur d'autres applications moins simples car le modèle temporel d'introduction et de modification de documents est plus complexe.

Références

N. Aussenac-Gilles, A. Condamines (2001), Entre Textes et ontologies formelles : les bases de connaissances terminologiques, Ingénierie et Capitalisation des connaissances, Eds M. Zacklad M. Grundstein. pp 153 - 175. Paris : Hermès sciences, traité IC2

Bonnet, Valérie (2003), Pour une terminologie diachronique, Travaux linguistiques du CERLICO, vol 16, G. Col & J. P. Régis eds. Morphosyntaxe du lexique-2. Rennes : Presses Universitaires de Rennes, à paraître[5].

Dury, Pascaline (1999), Etude comparative et diachronique des concepts ecosystem et écosystème, Meta, Vol XLIV, n°3, Montréal, pp 484-500.

Meyer Ingrid, Skuce Douglas, Bowker Lynne, Eck Karen (1992), Towards a new generation of terminological resources : an experiment in building a Terminological Knowledge Base, COLING-92, Proceedings of the 14th International Conference on Computational Linguistics (COLING 92), pp 956-960, 23-28 août 1992.

Moeller, Bernt (1998), A la recherche d'une terminochronie, Meta, Vol XLIII, n°3, Montréal, pp 426-438.

NF ISO 704 (Avril 2001), ISSN 0335-3931

Percebois, Jacqueline (2001), Fonctions et vie des sigles et acronymes en contexte de langues anglaise et française de spécialité, Meta, Vol XLVI, n°4, Montréal, pp 627-644

Roche, Christophe (2000), The 'Specific-Difference' Principle: a Methodology for Building Consensual and Coherent Ontologies », IC-AI'2001 2000 : Las Vegas, USA, June 25-28 2001

Roche, Christophe (2001), From Information Society to Knowledge Society : the Ontological Issue, CASYS'2001, Liège, Belgique, 13-18 août 2001

Roche, Christophe, Million-Rousseau, Cécile (2003), Construction de Terminologies Métier : l'Importance du Modèle Ontologique, Journées Francophones d'Extraction et de Gestion des Connaissances EGC 2003, Lyon, 22-24 janvier 2003

Summary

An organization's terminological knowledge is often acquired from the texts it uses. In the context of this work, the terminological knowledge base is founded on the modeling of business concepts, in the form of ontology. It is then necessary to address the maintenance issue for this base or ontology.

In this article, after defining the terminological knowledge base (TKB) and the diachrony issues, we present our model and our method for acquiring the organization's terminological knowledge. We then present our proposal for the ongoing maintenance of the terminological base that has been built.

To illustrate this work, we use a terminological knowledge base for cartoon animation and describe the maintenance issue using a historical re-creation of the various states of this base as the digital animation techniques appeared.

[5] Nous remercions l'auteur qui a autorisé cette utilisation anticipée