Les problèmes de variations terminologiques dans l'indexation de références bibliographiques

Emmanuel Nauer

Les problèmes de variations terminologiques dans l'indexation de références bibliographiques

Emmauel Nauer

Laboratoire Lorrain en Recherche en Informatique et ses Applications
UMR LORIA 7503 - 615 rue du Jardin Botanique - BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex

Introduction

Une activité de la terminologie est de spécifier le vocabulaire d'un domaine, en associant aux notions du domaine, les unités terminologiques les plus adéquates. On pourrait croire que le vocabulaire associé à un domaine est un ensemble limité de termes. Or, ce n'est pas le cas, car chaque acteur du domaine utilise son propre vocabulaire. C'est notamment le cas des fournisseurs d'informations -les producteurs de bases de données bibliographiques notamment- qui mettent chacun en place un système d'indexation des références propre. Les variations concernent tant par l'étendue sémantique couverte par les concepts pris en compte, que le vocabulaire associé.
Consciente de ce problème, la NLM (National Library of Medecine) a entrepris le développement de l'UMLS (Unified Medical Language System) [JOLIBOIS 99], dans le but d'unifier la connaissance (et le vocabulaire) du domaine biomédical. L'UMLS compile les termes biomédicaux de classifications et vocabulaires diffférents.
Nous montrons à partir d'exemples concrets que les variations terminologiques des concepts d'un domaine sont courantes dans une même langue, et qui plus est à travers différentes indexations. Ces variations posent de nombreux problèmes [NAUER 99] : en bibliométrie, comment mener à bien une analyse, si les statistiques issues des champs d'indexation ne prennent pas en compte les variations terminologiques liées à un même concept ? De même, en recherche d'information : comment peut-on interroger de multiples sources d'information qui n'utilisent pas la même représentation pour un concept commun ?

Les variations terminologiques d'un domaine

Nous commencerons par illuster nos propos par l'UMLS en nous limitant aux variations linguistiques anglaises des termes présents dans son lexique. Le lexique regroupe pour chaque concept l'ensemble de ses représentations terminologiques. Dans l'UMLS [UMLS 98] elles se distinguent selon :

- le statut du terme [TS] (Term Status) indiquant s'il s'agit du terme préféré pour représenter le concept ou s'il s'agit d'un synonyme.
- un type associé à chaque chaîne de caractères [STT] (STtring Type) qui peut prendre les valeurs suivantes : PF pour la forme préférée d'un terme, VC pour une variation de casse (minuscules-MAJUSCULES), VW pour les variations de ponctuations ou d'ordre des mots par rapport à la forme préférée, VS (respectivement VP) s'il s'agit du singulier (respectivement pluriel) de la forme préférée, VO pour les autres variations.

Les nombres de variations associés aux concepts (allant de 1 à 63 pour l'exemple de statistiques établies sur l'UMLS de janvier 1996) montrent la dispersion plus ou moins importante du vocabulaire selon le concept étudié.

Distribution du nombre cumulé de concepts de l'UMLS selon le nombre de représentations différentes qu'ils possèdentt

nrd	1	2	4	6	8	10	15	20	30	40	50	63
nc	184394	68498	20082	10318	6093	3892	1497	670	142	36	7	1

(tableau non exhaustif, les valeurs de nrd ont été choisies pour illustration)
nrd = nombre de représentations différentes pour un concept
nc = nombre de concepts possédant au moins nrd représentations différentes

En effet, sur les 252 892 concepts que comptait l'UMLS en janvier 1996, 184 394 d'entre eux n'avaient qu'une seule forme de représentation, 64 498 en avaient au moins 2, etc. Le maximum de formes différentes était de 63 pour le concept *Atrial Premature Complexes* pour lequel il y a par exemple les variations : *Atrial Beat, Premature* / *Contraction, Premature Atrial* / *Supraventricular premature beats* | *Ectopic Beat, Atrial* | *Atrial Extrasystoles* | *etc.*
Le nombre moyen de représentations différentes associées à chaque concept est de 1,733. D'après la version 1998 du thésaurus qui compte 476 322 concepts et 1 051 903 dénominations différentes, on constate que la dispersion augmente avec l'intégration de nouvelles sources ; le taux moyen ayant passé à 2,21.

L'UMLS permet de contater que le vocabulaire utilisé pour la représentation d'un même concept varie d'une base de données à l'autre (exemple : *Antidepressive Agents* et *Antidepressants* représentent tous deux le concept d'*antidépresseurs*). Si on étudie plus en détail les variations, on peut s'apercevoir que les variations peuvent avoir lieu à l'intérieur même d'une base. Bien que le vocabulaire soit controlé, il n'est pas figé. Une évolution terminologique peut advenir lorsqu'un syntagme change morphologiquement pour donner lieu à un concept à par entière. Ainsi le concept de *Charge de travail* est passé de *Work load* à *Workload*.

Variations terminologiques dans l'indexation

Les variations dans l'indexation des références bibliographiques tiennent au choix du producteur d'information. On peut citer :

- le choix d'une terminologie précise : chaque source définit un ensemble de concepts et le vocabulaire (descripteurs) qui serviront de base à l' indexation ;
- la complexité de l'indexation : la majorité des bases utilise, en matière de description, des syntagmes. Cependant, certaines bases emploient une indexation relationnelle ou une indexation à facettes pour augmenter le pouvoir d'expressivité. Dans ce cas, les descripteurs représenteront les concepts selon un point de vue spécifique. Par exemple : *Stress, Psychological_Epidemiology* indiquera qu'il est question du *stress psychologique* mais que abordé ici d'un point de vue *épidémiologique*) ;
- l'étendue de l'indexation : chaque source représente l'information selon l'aspect qui l'intéresse particulièrement. Elle utilise donc du vocabulaire spécifique. Par exemple, la base de données Psyclit représente les documents plus sur leur aspect psychologique que ne le propose Medline.

Toutes ces variations sont observables à travers les différentes indexations que l'on peut trouver dans différentes bases pour un même document. Par exemple, le document d'Arnetz BB, *Techno-stress: A prospective psychophysiological study of the impact of a controlled stress-reduction program in advanced telecommunication systems design work,* est indexé comme suit dans différentes bases de données :

Cisilo : **TELECOMMUNICATIONS** | RELAXATION EXERCISES | STRESS FACTORS | SUBJECTIVE ASSESSMENT | MENTAL WORKLOAD | **MENTAL STRESS** | BIOLOGICAL EFFECTS | **SERUM CHANGES | BLOOD PRESSURE** | **THROMBOCYTE COUNT**

Psyclit : Psychosocial Factors | Cardiovascular Disorders | Job Characteristics | Surgeons | **Occupational Stress** | Adulthood | At Risk Populations | General Practitioners | Human | Adult

Medline: **Telecommunications** | Program Evaluation | Job Satisfaction | **Hemodynamics** | Human | Prolactin | Prospective Studies | Support, Non-U.S. Gov't | Social Support | **Occupational Diseases** | **Stress, Psychological** | Questionnaires |Workload

On peut constater que les variations d'indexation concernent :

• la portée sémantique de chaque indexation : Cisilo et Medline relèvent un aspect *Telecommunications*, alors que Psyclit non ;

• l'utilisation de différents concepts pour exprimer la même information : le terme *Hemodynamics* (i.e. étude des mouvements du sang et des forces concernées) de Medline englobe trois descripteurs de Cisilo : *SERUM CHANGES, BLOOD PRESSURE, THROMBOCYTE COUNT* ;

• l'utilisation d'une terminologie différents pour dénommer un même concept : le *Mental Stress* de Cisilo est équivalent au *Stress, Psychological* de Medline ;

• le recours à des concepts plus précis dans certains cas : *Occupationnal Stress* dans Psyclit synthétise intuitivement l'information véhiculée conjointement par *Occupational Diseases* et *Stress, Psychological*.

Intérêt de l'utilisation de l'UMLS

L'UMLS peut servir à plusieurs égards dans un contexte de traitement automatique des langues. Grâce à son lexique biomédical, il permet d'unifier le vocabulaire employé par les descripteurs des différentes bases. Les différentes formes d'un même concept pourront ainsi être mises en correspondance avec la forme préférée du terme préféré du concept. Ainsi, des réindexations peuvent être possibles. Par exemple, les notices contenant *ACTH*, *Acth Hormone*, *Adrenocorticotrophic Hormone*, *Adrenocorticotropic Hormone* relatifs à un même concept pourront être réindexées par la forme préférée de ce concept, qui est *Corticotropin*. Cette étape est primordiale pour la bibliométrie ou la recherche d'information. Elle réduit l'ensemble initial du vocabulaire d'indexation. Comme chaque concept n'a plus qu'une unique forme de représentation, la recherche d'information est facilitée. De même, les études bibliométriques, qui se basent sur plusieurs fonds sont rendues possibles puisqu'on peut désormais dénombrer les apparitions de concepts et dépasser le dénombrement de chaque variation terminologique.

Le lexique de l'UMLS rend également possible l'accès multilingue à l'information. Les équivalences terminologiques qu'il propose en 4 langues (français, allemand, espagnol, italien) pour chaque concept permettent d'établir des passerelles d'indexation d'une langue à l'autre.

L'UMLS peut donc être utilisé dans un contexte de traduction et/ou réindexation automatique.

Conclusion

Les variations terminologiques inter et/ou intra bases qui rendent compte d'un même concept entraînent des difficultés et artéfacts dès lors qu'on s'intéresse non pas aux signifiants mais au signifié.

Ces biais sont particulièrement dommageables pour ce qui concerne les champs de recherche d'information et des études bibliométriques dans un conctexte multibases.

Les ressources terminologiques du métathesaurus de l'UMLS font de l'UMLS un outil permettant de pallier ces problèmes de représentation de concepts. D'autres connaissances contenues dans l'UMLS - tel le réseau sémantique par exemple - laisse envisager de belles perspectives quant à l'utilisation de l'UMLS pour des applications bien plus complexes.

Bibliographie

[JOLIBOIS 99] Samuel Jolibois, *Une base de connaissances dans le domaine biomédical : l'Unified Medical Language System*, JILA'99.

[NAUER 99] Emmanuel Nauer, *De l'importance de la normalisation en bibliométrie*, 7ème colloque sur les systèmes d'information élaborée, 7-11 juin 1999, Ile Rousse, France.

[UMLS 98] UMLS Knowledge Sources Documentation. 9th Ed. Bethesda : US Department of Health and Human Services. National Institutes of Health. National Library of Medicine; 1998. Available from URL: http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML