68th IFLA Council and General Conference
August 18-24, 2002

Code Number: 008-122-F
Division Number: IV
Professional Group: Classification et indexation
Joint Meeting with: -
Meeting Number: 122
Simultaneous Interpretation: -

La réalisation de l'interopérabilité entre vocabulaires d'accès matière et systèmes d'organisation de la connaissance : une analyse méthodologique

Lois Mai Chan,
School of Library and Information Science
University of Kentucky
États-Unis d'Amérique

Marcia Lei Zeng
School of Library and Information Science, Kent State University
Kent, États-Unis d'Amérique

Résumé :

L'environnement hétérogène de la recherche documentaire sur le Web a conduit à mettre en avant l'identification du besoin d'interopérabilité entre les divers systèmes. Dans la recherche par sujets, les utilisateurs rencontrent non seulement différents vocabulaires et classifications mais aussi différentes langues. En conséquence, il y a eu ces dernières années une floraison de projets cherchant à améliorer l'interopérabilité parmi des vocabulaires et des classifications, certains visant la diversité des vocabulaires et d'autres se concentrant plutôt sur les différentes langues. Cet article essaye d'analyser les méthodes employées dans ces projets. Il commence par une brève vue d'ensemble, puis examine plus particulièrement les approches et les méthodes employées dans des efforts récents.

1 INTRODUCTION

Dans l'environnement ouvert d'Internet et du Web, les ressources documentaires sont hétérogènes, et elles ont été indexées avec différents vocabulaires et organisées selon différents systèmes. Réaliser les meilleurs résultats dans la recherche multi-domaines a donc proposé un défi particulier à la profession de l'information. Dans la recherche documentaire, les utilisateurs ne sont typiquement pas avertis, ni n'auraient besoin de l'être, des mécanismes qui permettent en coulisse d'assortir les termes de leur question aux vocabulaires utilisés par les divers systèmes. L'approche idéale serait de fournir une recherche unique et sans raccord au lieu d'exiger de l'utilisateur qu'il recherche séparément dans différentes bases de données ou collections. Pour permettre une telle approche, il est important de rendre les différents systèmes d'organisation de la connaissance, tels que des vocabulaires contrôlés et classifications, interopérables au sein d'un dispositif de recherche unique.

2. EXAMEN DES PROJETS ET EXEMPLES DE VOCABULAIRES INTEROPÉRABLES

Avant d'examiner les méthodes mises en oeuvre, passons en revue brièvement un certain nombre d'efforts récents visant à réaliser l'interopérabilité entre et parmi différents vocabulaires d'accès matière (vocabulaires contrôlés et non contrôlés y compris) et systèmes d'organisation de la connaissance. Ceux-ci incluent des projets pour établir l'interopérabilité parmi des vocabulaires dans la même langue ou dans différentes langues, parmi différentes classifications, ainsi qu'entre vocabulaires contrôlés et classifications. Ces efforts ont conduit au mapping (établissement d'équivalences) et à l'intégration de systèmes existants d'organisation de la connaissance, ou à la création de nouveaux systèmes, pour partager l'information dans un environnement en réseau. Les projets ont varié à la fois dans les cibles du mapping et dans les méthodes employées pour réaliser leurs objectifs. Ces projets, organisés par similitude de tâche, sont les suivants :

2.1 Parmi des vocabulaires contrôlés dans la même langue

1. Entre les *Library of Congress subject headings* (LCSH, Vedettes-matière de la Bibliothèque du Congrès) et les *Medical subject headings* (MeSH) - Northwestern University (Olson, 2001)
2. Parmi différents vocabulaires contrôlés - H.W. Wilson Company (Kuhr, 2001)
3. Parmi différents thésaurus allemands employés pour indexer les mathématiques et la physique aussi bien que la documentation en sciences sociales - CARMEN (Content Analysis, Retrieval, Metadata: Effective Networking) (CARMEN WP12, 2000).

2.2 Parmi des vocabulaires multiples dans différentes langues et des classifications

1. Parmi des thésaurus, des classifications, des systèmes de codage, et des listes de termes contrôlés dans les domaines biomédicaux - Métathésaurus UMLS (Unified Medical Language System) (National Library of Medicine, 2001)
2. Parmi des services répartis utilisant différents vocabulaires d'indexation employés par des communautés diverses telles que des archives, les secteurs de l'enseignement supérieur et de la formation continue, des bibliothèques, des musées, la National Grid for Learning, le

Resource Discovery Network, etc. - HILT (High-Level Thesaurus Project) (HILT, 2000 ; Nicholson, Wake and Currier, 2001a)

3. Parmi des "vocabulaires d'entrée" employés par des systèmes (par exemple, les index de *BIOSIS Concept Codes*, l'*INSPEC Thesaurus*, la *U.S. Patent and Trademark Office Patent Classification,* etc.) afin de les relier aux "vocabulaires de requête" entrés dans une recherche - Projet peu connu de l'University of California Berkeley DARPA (Buckland et autres, 1999)
4. Parmi des classifications locales vers une classification commune (DDC, Classification décimale de Dewey) - Projet de Renardus (Koch, Neuroth, et Day, 2001)
5. Parmi quatre vocabulaires contrôlés et classifications : *Polish Thematic Classification* (PTC), descripteurs basés sur le *Thesaurus of Common Topics* (TCT), *CDU Universal Decimal Classification* (UDC), et *Subject-Heading Language* (SHL) de la Bibliothèque nationale de Varsovie Warsaw – Polish Project (Scibor and Tomasik-Beck, 1994)
6. Parmi les vocabulaires contrôlés employés par quatre catalogues de bibliothèques nationales dans trois langues : anglais, français et allemand - MACS (Multilingual Access to Subjects) (Freyre et Naudi, 2001)
7. Parmi des vocabulaires pour une base de données multilingue sur le patrimoine français - Mérimée (voir les statistiques rapportées dans Doerr, 2001).

2.3 Entre un vocabulaire contrôlé et un système de classification universel

1. Entre les LCSH et la *Library of Congress Classification* (LCC, Classification de la bibliothèque du Congrès) - *Classification Plus* (cédérom) et *Classification Web* (interface web en cours de développement), Library of Congress
2. entre LCSH et DDC (CDD) (Vizine-Goetz, 1996)
3. entre UDC (CDU) et GFSH (*General Finish Subject headings*) (Himanka et Vesa, 1992).

2.4 Entre systèmes de classification

1. Entre MSC (la *Mathematics Subject Classification* de l'American Mathematical Society, AMS) et la classe 510 de la DDC - State University of New York à Albany, New York (Iyer and Giguere, 1995)
2. Entre SAB (*Klassifikationssystem för svenska bibliotek*) et DDC - Bibliothèque Royale de Suède (IFLA, 2001:34)

2.5 Nouveau système pour différentes langues

Le projet HEREIN (The European information network on cultural heritage policies) a produit une interlangue, thésaurus de termes dérivés des rapports sur les politiques patrimoniales en Europe, créé sans référence directe aux terme ou à la structure d'un thésaurus préexistant. - projet HEREIN (http://www.european-heritage.net/en/index.html, cliquer sur Thesaurus).

3 MÉTHODES EMPLOYÉES POUR RÉALISER L'INTEROPÉRABILITÉ

Le souci de compatibilité des vocabulaires n'est pas nouveau. Bien avant l'avènement de l'ère électronique, bibliothécaires et professionnels de l'information avaient exploré et employé diverses méthodes pour réduire le conflit entre les différents vocabulaires qui étaient utilisés dans un même système. Les méthodes se sont d'abord presque totalement fondées sur des efforts intellectuels. Puis ont émergé des méthodes informatisées avancées pour réaliser ou améliorer l'interopérabilité, l'informatique commençant à être employée pour mieux tirer bénéfice de l'environnement géré en réseau. La section suivante énumère les méthodes conventionnelles et nouvelles qui sont devenues largement admises.

1. Dérivation/Modélisation - Un vocabulaire spécialisé ou plus simple est développé à partir d'un vocabulaire existant plus complet comme point de départ ou modèle.
2. Traduction/Adaptation - Un vocabulaire contrôlé est développé à partir des termes traduits d'un vocabulaire dans une langue différente, avec ou sans modification.
3. Équivalences (Mapping intellectuel) - Un système de mapping est développé qui consiste fondamentalement à établir des équivalents entre les termes de différents vocabulaires contrôlés ou entre des termes et des indices de classification. Un tel mapping requiert généralement beaucoup d'effort intellectuel.
4. Équivalences (Mapping assisté par ordinateur) - Un système de mapping est développé qui se fonde en partie ou fortement sur l'informatique.
5. Maillage (Linking) - Une liste de termes est développée en reliant ces termes avec d'autres termes qui ne sont pas des équivalents conceptuels mais sont étroitement liés linguistiquement. De tels liens se sont avérés aptes à augmenter les résultats de la recherche.
6. Commutation (Switching) - Un langage ou un système de commutation est développé pour servir d'intermédiaire et se déplacer parmi des termes équivalents dans différents vocabulaires.

4 MÉTHODES EMPLOYÉES DANS LE STOCKAGE ET LA GESTION DE LIEN

Une fois que les équivalences sont établies, un dispositif est nécessaire pour stocker et gérer les liens afin de contrôler le grand nombre de termes d'indexation et leurs relations complexes qui en résultent. À cette fin, plusieurs options ont été explorées et employées :

1. Contrôle d'autorité - Des champs particuliers des formats d'autorité peuvent être employés pour stocker les liens.
2. Concordances - Leur élaboration exige la désignation d'un vocabulaire/système principal et d'un ou plusieurs vocabulaires/systèmes cibles.
3 Réseau sémantique - Un réseau sémantique, également appelé web sémantique, se compose d'une structure organisée qui sert de colonne ou d'épine dorsale. Chaque unité dans le réseau représente un concept autour duquel une grappe de termes équivalents de différents vocabulaires est identifiée et stockée.

5. DISCUSSION

Un certain nombre de questions communes ont émergé dans notre analyse des méthodes employées dans les nombreux projets discutés.

5.1 Problèmes généraux du mapping

5.1.1 Mapping des vocabulaires multilingues.

Le mapping ou établissement d'équivalences est au coeur des vocabulaires contrôlés multilingues. Les relations de un à un (équivalences exactes) entre termes dans différents vocabulaires et différentes langues sont les équivalences idéales, mais sont souvent insaisissables. Les différentes expressions linguistiques pour le même concept, les degrés différents de spécificité, et les termes polysémiques sont quelques-unes des difficultés auxquelles sont confrontés ceux qui s'efforcent de relier des vocabulaires et ceux qui créent des vocabulaires multilingues ou multidisciplinaires. Les conditions et les processus complexes pour assortir les termes, qui sont souvent imprécis, ont un impact sur plusieurs aspects du mapping de vocabulaires (Koch, Neuroth, et Day, 2001) : structure de feuilletage, affichage, profondeur, classes non sujets, et équilibre entre cohérence, précision et rentabilité. Différents niveaux de mapping/lien peuvent coexister dans le même projet, comme ceux identifiés par le projet MACS : niveau terminologique (vedette matière), niveau sémantique (notice d'autorité), et niveau syntaxique (application) (Freyre et Naudi, 2001).

5.1.2 Intégration des vues de cultures différentes.

Avec pour principe que toutes les langues sont égales dans une concordance, la question est de savoir si les vues d'une culture particulière, qui sont exprimées à travers un vocabulaire contrôlé ou une classification, peuvent être convenablement transférées à celles d'une culture différente dans le processus de mapping. Hudon (1997) a noté les problèmes suivants liés aux systèmes multilingues :

1) forcer une langue à s'adapter à une structure conceptuelle étrangère au point qu'elle devient à peine reconnaissable par ses propres locuteurs ;
2) transférer une structure conceptuelle entière d'une culture à une autre, que cela soit approprié ou pas ;
3) traduire littéralement les temes de la langue source dans des expressions sans signification dans la langue cible, etc.

Elle récapitule les problèmes que cela implique tels que : problèmes de gestion, problèmes de linguistique/sémantique, et problèmes technologiques.

5.1.3 Mapping de systèmes ayant des structures différentes.

Il y a des différences de base dans les macrostructures des vocabulaires contrôlés et des systèmes de classification. Les thésaurus construits sur la base d'ISO 2788 et d'autres normes nationales assurent que la structure et la "grammaire" de tels vocabulaires demeurent cohérentes ou compatibles. Mais la construction des langages d'indexation et des classifications a, d'autre part, été guidée par les modèles ou les exemples existants. Il y a donc de fortes chances, pour dix systèmes universels différents, de trouver dix principes directeurs différents. En conséquence, les

systèmes d'organisation de la connaissance diffèrent les uns des autres dans leur structure, leur sémantiques, leur lexique, et leurs traits de notation ou d'entrée (Iyer et Giguere, 1995).
Par exemple, ils peuvent couvrir différents domaines, ou avoir une application ou une portée différente ; ils peuvent avoir des différences sémantiques causées par des variations dans leur structuration conceptuelle ; leurs niveaux de spécificité et l'utilisation de la terminologie peuvent changer ; et leurs caractéristiques syntaxiques enfin, tels que l'ordre des termes et le choix de la place des vedettes, peuvent être différents.

Ces incompatibilités ont posé des problèmes pour tout mapping dès l'origine. Par exemple, l'établissement de concordance ou de traduction entre un thésaurus et une classification ou parmi divers systèmes devient parfois impossible ou extrêmement difficile. C'est particulièrement vrai quand le système cible a un niveau plus élevé de spécificité que le système source ou d'autres systèmes en cause.

5.2 Options méthodologiques

Pour des projets visant à établir l'interopérabilité entre ou parmi certains systèmes d'organisation de la connaissance pour répondre aux nouvelles exigences de l'utilisateurs dans l'environnement des réseaux, une décision principale qui doit être prise est le choix d'une méthode appropriée. La première question complexe à laquelle il faut répondre est : intégrer, mapper, ou créer un nouveau système? Les options sont semblables à celles que propose Riesthuis (2001) pour les différentes approches possibles pour créer des thésaurus multilingues :
traduction ; fusion ; ou création à partir de zéro.

Dans chacune de ces approches les possibilités sont multiples, comme l'ont suggéré les chercheurs de HILT dans une grille bidimensionnelle (Nicholson, Wake et Currier, 2001b). Ils proposent trois options de base :

- Utiliser ou créer un système unique (LCSH, Unesco, basé sur la CDD, basé sur la CDU, entièrement nouveau) ;
- Mapper les systèmes existants (LCSH, Unesco, basé sur CDD, basé sur CDU) ;
- Mapper les systèmes existants à court terme, pour viser un système unique dans le long terme.

Sur les bases de ces options, de nouvelles considérations peuvent être appliquées :

- structure additionnelle de thésaurus ;
- nouveaux micro-thésaurus spécifiques ;
- mapping de micro-thésaurus existants par domaines ;
- capacité multilingue ;
- contrôle commun ;
- méthodes automatisées ;
- méthodes assistées par intelligence artificielle ;
- formation des utilisateurs ;
- moyens flexibles pour aider les utilisateurs ;
- cartes mentales des utilisateur ;
- uniformité dans l'application des termes grâce à la formation et au contrôle ;
- bibliothécaires qualifiés pour aider l'utilisateur à optimiser ses recherches.

Le choix de l'approche de base joint à toutes les combinaisons de ces considérations peut donner des résultats très divers et exiger des quantités variables de temps et de ressources. Toutes les méthodes et toutes les combinaisons possibles avec d'autres processus peuvent avoir du pour et du contre. Il est donc nécessaire de conduire une recherche complète et d'identifier les problèmes potentiels quand une méthode particulière est utilisée.

6 CONCLUSION

1. Qu'avons-nous appris des projets ?
2. Quels sont les problèmes encore en suspens ?
3. Que faut-il, en termes d'approches intellectuelles et techniques, pour aller de l'avant ?

A partir des exemples présentés dans cet article, nous pouvons récapituler les tendances suivantes qui forment le courant principal :

1. Le besoin d'interopérabilité entre systèmes d'organisation de la connaissance est un enjeu et un processus inévitables dans l'environnement géré en réseau d'aujourd'hui.
2. Diverses méthodes ont été employées pour réaliser l'interopérabilité parmi les systèmes d'organisation de la connaissance. Un système de commutation peut être nécessaire ou non. Il se peut ou non qu'établir une concordance entre ou parmi les vocabulaires impliqués puisse être la situation idéale. Ou il se peut aussi que l'interopérabilité puisse être plus efficacement réalisée à travers les notices d'autorité de divers systèmes en ligne.
3. Alors que mapper des vocabulaires est toujours un effort en grande partie intellectuel, l'informatique a été appliquée pour faciliter la gestion des grands fichiers de données matière comme la gestion des liens. Des niveaux plus élevés de systèmes automatisés de mapping ont également été sujets à expérimentation ou à essais. Mapping humain et mapping assisté par ordinateur coexisteront pendant un certain temps dans l'avenir.
4. Des projets nombreux de mappings multilingues et multistructures ont été lancés. Ces projets ont identifié et ont expérimenté de nombreuses méthodes. On peut à coup sûr prévoir qu'il y aura beaucoup plus de produits et services multilingues, et que bon nombre d'entre eux impliqueront de multiples systèmes structurés tels que thésaurus, classifications, vedettes-matière et index des termes dans les bases de données.

Le besoin de réconcilier différents vocabulaires d'accès matière dans l'environnement géré en réseau est incontestable. Les résultats des efforts récents pour réaliser l'interopérabilité entre des vocabulaires de différentes sortes et dans différentes langues est encourageant. Une question se pose cependant : avons-nous entièrement exploité les possibilités technologiques dans nos efforts pour améliorer l'accès matière aux myriades de ressources disponibles dans l'environnement des réseaux ?

REMERCIEMENT :

Sincères remerciements aux principaux investigateurs des projets d'interopérabilité discutés dans cet article : Traugott Koch (Suède et Danemark), Patricia Kuhr (État -Unis), Martin Kunz (Allemagne), Max Naudi (France), Dennis Nicholson (R-U), et Tony Olson (États-Unis), qui n'ont pas ménagé leur temps pour répondre à nos questions, fournir des détails concernant leurs projets, ou relire l'article avant publication.

REFERENCES

Buckland, M., et al. *(1999). Mapping entry vocabulary to unfamiliar metadata vocabularies,* D-Lib Magazine*,* 5(1). *http://www.dlib.org/dlib/january99/buckland/01buckland.html, (visité le 05-02-2002).*

CARMEN. WP12 : Cross concordances of classifications and thesauri. *http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en* (*visité le 05-02-2002*)

Doerr, Martin. (2001) Semantic problems of thesaurus mapping. *Journal of Digital information*, 1 (8). http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/#Nr.52

Freyre, Elisabeth and Max Naudi. (2001) MACS: Subject access across languages and networks. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

HILT. (2000) *HILT: High-Level Thesaurus Project Proposal*. http://hilt.cdlr.strath.ac.uk/AboutHILT/proposal.html. (visité le 05-02-2002)

Himanka, Janne and Kautto Vesa. (1992) Translation of the Finish Abridged Edition of UDC into General Finish Subject Headings. *International Classification* 19(3):131-134.

Hudon, Michele. (1997) Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge concepts. *Knowledge Organization* 24(2): 84-91.

IFLA Section on Classification and Indexing. (2001) *Newsletter* Nr.24, December 2001.

Iyer, Hermalata and Mark Giguere. (1995). Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization* 22(3/4):141-147.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001) Renardus: cross-browsing european subject gateways via a common classification system (DDC. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC. http://www.lub.lu.se/~traugott/drafts/preifla-final.html (visité le 05-02-2002)

Kuhr, Patricia S. (2001) Putting the world back together: mapping multiple vocabularies into a single thesaurus. . In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

National Library of Medicine. (2001) *Fact Sheet*: *UMLS ® Metathesaurus ®* Last updated 2001. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. (visité le 05-02-2002)

Nicholson, Dennis and Susannah Wake. (2001a) HILT: Subject retrieval in a distributed environment. . In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Nicholson, Dennis, Susannah Wake, and Sarah Currier. (2001b) High-Level Thesaurus Project: investigating the problem of subject cross-searching and browsing between communities. In *Global Digital Library Development in the New Millemnnium: fertile ground for distributed cross-disciplinary collaboration*, edited by Ching-Chih Chen. Beijing: Tsinghua University Press, 2001.

Olson, Tony. (2001) Integrating LCSH and MeSH in information systems. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Riesthuis, Gerhard J.A. (2001) Information languages and multilingual subject access. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Scibor, Eugeniusz and Joanna Tomasik-Beck. (1994) On the establishment of concordances between indexing languages of universal or interdisciplinary scope (Polish Experiences). *Knowledge Organization* 21(4):203-212.

Vizine-Goetz, Diane. (1996) Classification Research at OCLC. *Annual Review of OCLC Research*, pp. 27-33.