Date : 02/08/2008

RELU PAR LE CFI

Concordances: la cartographie terminologique et son efficacité dans la recherche d'information

Philipp Mayr & Vivien Petras
GESIS Social Science Information Centre (GESIS-IZ)
Bonn, Germany
philipp.mayr|vivien.petras@gesis.org

Traduit de l'anglais par *Bernadette Patte, SCD Paris 4,*
Bernadette.patte@paris-sorbonne.fr
& Anne-Céline Lambotte, Bibliothèque nationale de France
Anne-celine.lambotte@bnf.fr
Juillet 2008

Meeting: **129. Classification and Indexing**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Résumé *: Le Ministère allemand fédéral de l'éducation et de la recherche a financé une initiative importante dans le domaine de la cartographie terminologique qui s'est achevée en 2007.*
Le projet était d'organiser, de créer et de gérer des concordances entre des vocabulaires contrôlés (thesaurii, classifications, listes de vedettes matière) a priori dans le domaine des sciences sociales mais il a rapidement évolué vers d'autres champs thématiques.
64 passerelles contenant plus de 500 000 relations ont été ainsi établies. Pendant la phase finale du projet, un important effort d'évaluation a été mené afin de tester et de mesurer l'efficacité d'une cartographie terminologique dans un système d'information.
Cette communication porte sur le travail de concordance et sur l'évaluation des résultats.

1. Introduction

En Allemagne, le portail Vascoda[1], ambitieux projet concernant les recherches d'information en une seule étape, est un projet mené conjointement par le Ministère de l'éducation et de la recherche et la Fondation allemande pour la recherche. Vascoda permet de rechercher, à partir d'une interface commune, dans de nombreuses bases de données disciplinaires et interdisciplinaires (services d'indexation et de résumé, catalogues de bibliothèques, bases de données, articles en texte intégral,…) ainsi que dans des collections de ressources

[1] http://www.vascoda.de

électroniques (voir Depping, 2007, pour une présentation générale). Depuis 2007, Vascoda est partenaire du portail WorldWideScience.org[2].

Le concept clé du portail Vascoda est de structurer et d'intégrer des sources de haute qualité informationnelle à partir de plus de 40 fournisseurs dans un espace unique de recherche (approximativement 81 millions de documents).
L'espace de recherche est organisé et réparti dans des portails disciplinaires (bibliothèques thématiques virtuelles) et chacune des collections intégrées est agrégée dans des groupes thématiques Vascoda (physique et sciences de l'ingénieur, médecine et sciences de la vie, droit, économie et sciences sociales, sciences humaines, régions et zones culturelles, collections multidisciplinaires).

Le portail Vascoda contient de nombreuses collections de ressources qui sont développées et structurées méticuleusement. Elles comportent des métadonnées sujet réparties en schémas élaborés (vedettes matière, thesauri ou classifications) permettant de décrire et d'organiser le contenu des documents au niveau individuel des collections. L'interface générale de recherche permet seulement une recherche plein texte dans tous les champs des métadonnées sans prendre en compte les outils d'accès sujet précis qui ont été établis pour ces collections de ressources. En raison de la nature des collections, de leur répartition et des schémas d'accès sujet nombreux et diversifiés, il est techniquement difficile d'intégrer et de gérer toutes ces sources d'information avec les mêmes outils d'accès sujet, efficaces et uniquement déclinés dans une seule interface de recherche.

Parallèlement, des applications de recherche à spectre très large ajoutent aussi bien de nouvelles collections et des accès sujet pour une recherche avancée à leur répertoire (par exemple clustering et recherche par vidéo dans Flickr). Les applications du web sémantique[3] qui dérivent des fonctions intelligentes à partir d'ontologies et d'autres données sémantiques constituent également un exemple significatif. Alors que des tentatives d'organisation de l'information comme le web sémantique s'efforcent de structurer par des liens sémantiques renforcés le contenu de l'information, comment des interfaces appliquées à des bibliothèques électroniques peuvent-elles réduire leur champ de possibilités?

En 2004, le Ministère allemand de l'éducation et de la recherche a financé une initiative importante dans le domaine de la cartographie terminologique (projet KoMoHe[4]) au centre d'information en sciences sociales GESIS à Bonn (GESIS-IZ) qui remit ses conclusions à la fin de l'année 2007. Un des aspects de ce projet était d'organiser, de créer et de gérer des concordances (alignements terminologiques) entre les plus importants vocabulaires contrôlés dans le domaine des sciences sociales, dans un premier temps, mais qui s'est bien vite étendu à d'autres domaines. L'objectif principal était d'établir, d'implémenter et d'évaluer un réseau terminologique permettant l'intégration de ressources hétérogènes afin de mener des recherches d'information dans l'environnement de bibliothèques numériques types.

L'objectif de l'intégration sémantique est de connecter différents systèmes d'information par l'intermédiaire des métadonnées-sujet afin de permettre une recherche dans plusieurs systèmes d'information grâce aux outils de recherche avancée par sujet, accessibles dans les différentes bases de données. A partir de l'alignement des différentes terminologies relatives aux sujets, un « accord sémantique » est établi pour l'ensemble des termes susceptibles d'être

[2] http://worldwidescience.org
[3] voir http://www.w3.org/2001/sw/
[4] http://www.gesis.org/en/research/information_technology/komohe.htm

cherchés. La cartographie terminologique, l'alignement terminologique de mots et de phrases issus d'un vocabulaire contrôlé à un autre permet de passer de façon transparente d'une recherche à partir d'une base de données à divers scenarii de recherche dans l'environnement des bibliothèques numériques.

Cette communication décrit le projet de cartographie terminologique KoMoHe, les vocabulaires et les bases de données impliqués, ainsi que l'implémentation des concordances dans le processus de recherche d'information. Ce projet étudie également les découvertes et les résultats d'une évaluation de la recherche d'information approfondie analysant ainsi l'impact de la cartographie terminologique en termes de taux de rappel et de précision.

2. Hétérogénéité sémantique

En général, il y deux principales approches pour traiter l'hétérogénéité sémantique dans les bibliothèques numériques : l'approche intellectuelle et l'approche automatisée. Accepter l'idée d'une divergence structurelle entre les différentes terminologies est essentiel en ce qui concerne les tentatives dans le domaine de la cartographie terminologique. Aucune de ces approches ne peut à elle seule, endosser la responsabilité du transfert entre des collections hétérogènes, principalement pour des raisons de qualité et de coût. Il est important de noter que les approches opèrent bilatéralement au niveau de la base de données. Selon Krause (2003) les différentes approches devraient se compléter mutuellement et fonctionner en synergie.

- Concordances entre des vocabulaires contrôlés : les différents systèmes sont analysés dans un contexte orienté vers l'utilisateur et dans une tentative de relier intellectuellement leur conceptualisation. Cette idée ne devrait pas se confondre avec celle de la construction de meta-thesauri. En établissant des concordances, il n'y a pas de tentative de normaliser des mots traduisant des concepts existants. La concordance comprend seulement une union partielle de systèmes terminologiques existants, incluant l'aspect incontournable de la problématique de la concordance de termes. De telles concordances proposent la plupart du temps des alignements terminologiques (voir Tableau 1 et 2) de relations de synonymie ou de hiérarchie / similitude mais aussi de relations induisant une règle déductive.
- L'approche quantitative et statistique : le problème du transfert peut être de façon générale représenté comme un problème de flou entre deux langages de description de contenu. Différentes opérations automatiques ont été tentées (procédures fondées sur la probabilité, approches floues et réseaux neuronaux) pouvant être utilisées dans le cas de transferts problématiques (Hellweg et al. 2001) afin de répondre au problème de l'imprécision entre des termes équivalents dans le cadre de la recherche d'information, aussi bien dans la requête de l'utilisateur que dans les collections de données. Le document isolé peut être indexé à l'intérieur d'autres documents dans deux schémas conceptuels, deux différents documents indexés différemment pouvant être mis en relation l'un avec l'autre. Des procédures de ce type doivent être expérimentées. Pour la recherche d'information multilingue, le même texte doit exister dans deux langues.

Quand le traitement de l'hétérogénéité sémantique (concordances) est implémenté dans un scénario de recherche d'information distribuée, on peut rechercher différentes informations par le schéma de recherche organisé en métadonnées sujet avec lequel on est familier. Les alignements terminologiques peuvent étayer une recherche d'information abordée de différentes façons. D'abord et avant tout, ils doivent permettre une recherche transparente

entre divers systèmes de métadonnées sujet. De plus, ils peuvent servir d'outil dans le cadre de vocabulaire élargi dès lors qu'ils présentent un ensemble de termes organisés avec des relations de type équivalence, générique, d'association ou spécifique (voir les exemples de termes dans la Table 1 et 2).Troisièmement, cette organisation de termes peut aussi être utilisée dans le cadre de requêtes élargies ou de reformulations.

La requête n'est pas seulement formulée dans des formulations précises de recherche mais le service de cartographie sémantique traduit automatiquement la requête dans les autres terminologies présentes dans le signalement des ressources de la bibliothèque électronique. Un chercheur peut passer de façon transparente d'une ressource à une autre car la traduction sémantique entre les différents termes se fait automatiquement.

En ce qui concerne des systèmes d'information interdisciplinaires, l'intégration sémantique accroît les chances de succès dans des recherches impliquant différents schémas de métadonnées sujet mais elle procure aussi au chercheur une ouverture vers un cadre disciplinaire différent et, dans le cas où les vocabulaires alignés sont disponibles, dans un langage spécifique (voir Tableau 1).

Les alignements sémantiques jouent aussi un rôle important en apportant une méthodologie pour le transfert de données entre des bases de données dans des langues étrangères. Dans la mesure où une cartographie de termes peut être établie entre des vocabulaires contrôlés provenant de différentes bases données ou de différentes disciplines, cela peut contribuer à effectuer une traduction au sens traditionnel du terme : voir par exemple dans le tableau 1 d'une terminologie allemande à une terminologie anglaise.
Le tableau 1 présente deux termes de recherche (colonne de gauche) dans le thesaurus allemand en sciences sociales (TheSoz) (« Weiterbildung », traduction anglaise : « further education » et « Meinungsforchung » traduction anglaise « opinion research ») et des termes associés dans des vocabulaires en alignement sémantique. La typologie des relations entre les termes est expliquée dans le tableau 2.

Start term TheSoz	**Relation**	**End term**	**End vocabulary**
Weiterbildung engl: "further education"	=	Weiterbildung	Psyndex, STW, Infodata, SWD, BISp, DZI
Weiterbildung engl: "further education"	^	Berufsfortbildung	FES
Weiterbildung engl: "further education"	=	Further education	CSA-ASSIA
Weiterbildung engl: "further education"	=	Continuing education	CSA-PEI
Weiterbildung engl: "further education"	=	Adult Education	CSA-SA
Weiterbildung engl: "further education"	<	Education	CSA-WPSA
Weiterbildung engl: "further education"	=	Erwachsenenbildung	IBLK

Meinungsforschung engl: "opinion research"	0		Psyndex
Meinungsforschung engl: "opinion research"	^	Einstellungsforschung	IAB
Meinungsforschung engl: "opinion research"	=	Opinion Polls	CSA-ASSIA
Meinungsforschung engl: "opinion research"	=	Opinions + Research	CSA-SA
Meinungsforschung engl: "opinion research"	<	Research	CSA-PEI
Meinungsforschung engl: "opinion research"	=	Public Opinion Research	CSA-WPSA
Meinungsforschung engl: "opinion research"	=	Public Opinion Polls	ELSST
Meinungsforschung engl: "opinion research"	=	Meinungsumfrage/Meinungs -forschung	IBLK

Table 1. Start or seed terms in the TheSoz vocabulary and a selection of end terms (semantic mappings).

Récemment, différentes organisations ont développé des initiatives afin de permettre une intégration sémantique dans les systèmes d'information. Aux États-Unis, OCLC a initié le projet Teminology Services[5] (Vizine-Goetz, 2004, 2006) afin d'offrir des services sur le web avec des alignements terminologiques entre différents vocabulaires contrôlés comme la CDD, LCC, LCSH ou le MeSH. En Europe, le programme Delos 2 du Network of excellence in Digital Libraries a consacré un ensemble de travaux (WP5) au problème de l'extraction du savoir et de l'opérabilité sémantique (Patel et al. 2005). Un autre rapport mandaté par JISC dresse un panorama des services relatifs à la terminologie avec un accent particulier sur les réalisations du Royaume-Uni (Tudhope, Koch et al., 2006).
Le projet CRISSCROSS[6] mené conjointement par la Bibliothèque nationale allemande et l'Université des sciences appliquées de Cologne a établi un vocabulaire multilingue à structure de thesaurus réunissant les fichiers d'autorité matières (SWD) et les notations de la classification de Dewey (CDD) (voir Panzer, 2008). Le département Agricultural Information Management Standards de la FAO[7] participe à de nombreuses initiatives d'alignements sémantiques (Liang & Sini, 2006). Le projet HILT[8] (High-Level Thesaurus Project) de l'Université de Strathclyde est encore un exemple de projet mené sur le long terme dans le domaine des technologies relatives à l'alignement sémantique (Macgregor et al., 2007).

3. L'approche de la cartographie terminologique au GESIS-IZ

L'interopérabilité sémantique peut être obtenue de différentes façons. Pour une vue d'ensemble des différentes méthodologies et des projets d'alignements terminologiques, voir Zeng & Chan (2004, 2006a, 2006b), Doerr (2001, 2004), et Hellweg et al. (2001).

Le projet KoMoHe s'est centré sur les concordances. Nous définissons les concordances comme des passerelles créées intellectuellement (manuellement) déterminant des relations

[5] http://oclc.org/research/projects/termservices/
[6] http://www.d-nb.de/eng/wir/projekte/crisscross.htm
[7] http://www.fao.org/aims/
[8] http://hilt.cdlr.strath.ac.uk/

d'équivalence, de hiérarchie et d'association entre des termes de deux vocabulaires contrôlés.

Typiquement, les vocabulaires sont mis en relation bilatéralement, c'est-à-dire qu'on établit une concordance reliant des termes du vocabulaire A au vocabulaire B et aussi des termes du vocabulaire B au vocabulaire A. Les relations bilatérales ne sont pas nécessairement symétriques. Par exemple, le terme « ordinateur » dans le système A est aligné avec le terme « système d'information » dans le système B mais ce même terme « système d'information » est aligné avec un autre terme, « base de données » dans le système A.
Notre approche permet les relations 1 : 1 ou 1: n

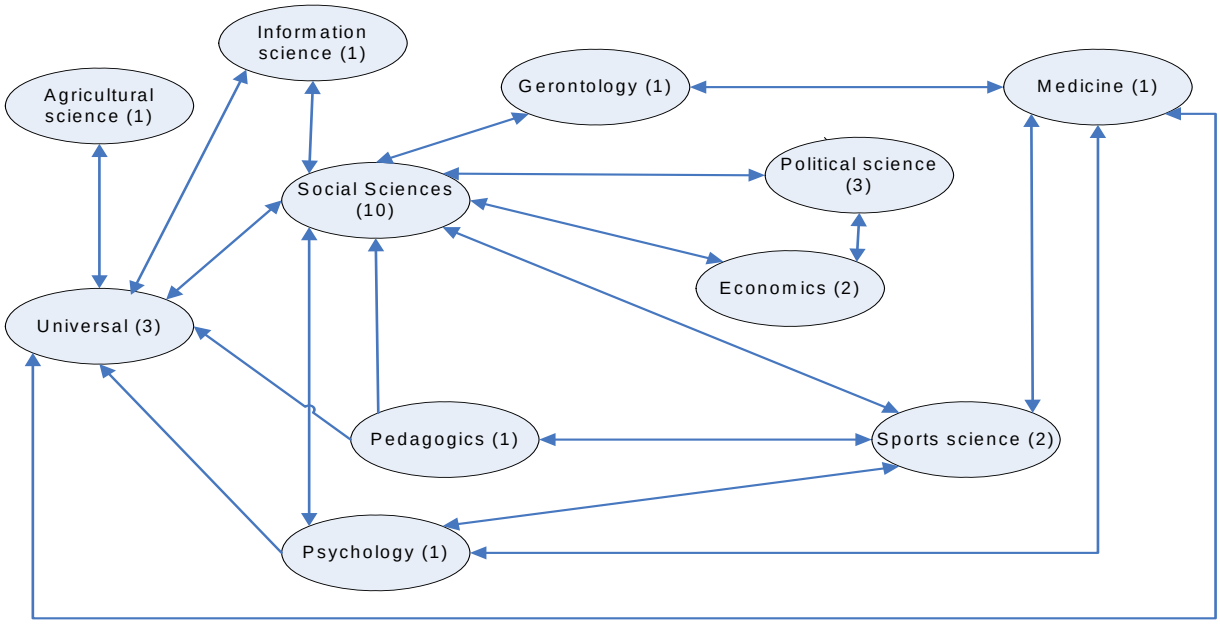
- Equivalence (=) signifie identité, synonymie ou quasi-synonymie
- Hiérarchie (termes génériques < ; termes spécifiques >)
- Association (^) pour des termes en relation
- Une exception est la relation nulle, ce qui signifie qu'un terme ne peut pas être aligné avec un autre (voir tableau 2, n°4).

De plus, chaque relation doit être affectée d'un indice de pertinence (fort, moyen et faible). L'indice de pertinence est un instrument secondaire et peu efficace pour améliorer la qualité des relations. Il n'est pas utilisé dans nos applications courantes.
Le tableau 2 présente des concordances classiques unidirectionnelles entre deux vocabulaires A et B

N°	Vocabulaire A	Relation	Vocabulaire B	Description
1	hacker	=	hacking	Relation d'équivalence
2	hacker	^+	computer + crime	2 relations d'association (^) à des combinaisons de termes (+)
3	hacker	^+	internet + security	
4	isdn device	0		Relation nulle. Le concept ne peut être aligné, terme trop spécifique
5	isdn	<	telecommunications	Relation de termes spécifiques
6	documentation system	>	abstracting services	Relation de termes génériques

Table 2. Cross-concordance examples (unidirectional).

Les alignements sémantiques dans le projet KoMoHe concernent la totalité ou la majeure partie des vocabulaires. Les vocabulaires sont analysés du point de vue des chevauchements de sens et syntaxiques avant que l'alignement ne soit réalisé. Tous les alignements terminologiques sont créés par des chercheurs ou des experts en terminologie. La compréhension du sens et de la sémantique des termes ainsi que les relations internes des vocabulaires concernés est essentielle pour construire des alignements pertinents. Ceci inclut des contrôles syntaxiques de radicaux de termes mais aussi des connaissances sémantiques afin de relever des synonymes et autres termes qui seraient en relation.
Le processus d'alignement sémantique est établi à partir d'un ensemble de guides et de règles pratiques (voir Patel et al., 2005). Pendant le processus d'alignement sémantique, toutes les relations intra-thesaurus (y compris les notes d'application) sont consultées. La précision et le taux des rappel des relations effectuées doivent être vérifiés dans des bases de données associées. Cela est particulièrement important pour des combinaisons de termes (1:n relations). Les relations un-à-un (1:1) sont privilégiées. Les groupements de mots et les ajustements de pertinence doivent être consistants.



Enfin, les termes sémantiques des alignements sont revus par des experts et des échantillons sont testés empiriquement pour le taux de rappel et la précision.
Toutes choses considérées, c'est un service de qualité mais qui requiert un effort conséquent en termes de coûts et de temps pour produire un réseau reposant uniquement sur des concordances.

3.1 Les résultats de l'initiative

A ce jour, 25 vocabulaires contrôlés trilingues (allemand, anglais et russe) provenant de 11 disciplines ont été connectés entre eux avec des vocabulaires comportant entre 1000 et 17000 termes alignés (voir le tableau 2 pour une vue détaillée des relations). Plus de 513 000 relations se trouvent dans 64 passerelles (30 concordances bilatérales[9] et 4 unilatérales). La figure 1 retrace le réseau établi de concordances par disciplines.

Figure 1. Network of terminology mappings in the KoMoHe project. The numbers in brackets contain the number of mapped controlled vocabularies in a discipline.

Le projet a généré des concordances entre les vocabulaires suivants (thesauri, listes de descripteurs, classifications et vedettes matière) qui ont tous une place dans les collections spécifiques par sujet de Vascoda. Plusieurs concordances provenant des projets antérieurs CARMEN [10] et infoconnex[11] ont été incorporées.

Les vocabulaires présents dans le projet KoMoHe sont surtout allemands, anglais (N= 8), russes (N=1) ou multilingues (par exemple AGROVOC, IBLK, DDC). Quelques vocabulaires comportent des traductions de termes allemands ou anglais (par exemple THESOZ, PSYNDEX, MESH, INION, STW).

[9] Une concordance bilatérale est comptée comme deux intersections
[10] http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en
[11] http://www.infoonnex.de/

Thesauri alignés (N=16) :

- AGROVOC Thesaurus (AGROVOC): A vocabulary in the *agricultural* domain which contains round 39,000 terms. Mapping to: SWD.
- CSA Thesaurus Applied Social Sciences Index and Abstracts (CSA-ASSIA): A vocabulary in the *social science* domain which contains round 17,000 terms. Mapping to: THESOZ.
- CSA Thesaurus PAIS International Subject Headings (CSA-PAIS): A vocabulary in the *political science* domain which contains round 7,000 terms. Mapping to: IBLK.
- CSA Thesaurus Physical Education Index (CSA-PEI): A vocabulary in the *sports science* domain which contains round 1,800 terms. Mapping to: THESOZ.
- CSA Thesaurus of Political Science Indexing Terms (CSA-WPSA): A vocabulary in the *social and political science* domain which contains round 3,100 terms. Mapping to: THESOZ.
- European Language Social Science Thesaurus (ELSST): A vocabulary in the *social science* domain which contains round 3,200 terms. Mapping to: THESOZ.
- INFODATA Thesaurus (INFODATA): A vocabulary in the *information science* domain which contains round 1,000 terms. Mapping to: THESOZ and SWD.
- Psyndex Terms (PSYNDEX): A vocabulary in the *psychological* domain which contains round 5,400 terms. Mapping to: THESOZ, SWD, BISP, MESH and BILDUNG.
- Standard Thesaurus Wirtschaft (STW): A vocabulary in the *economics* domain which contains round 5,700 terms. Mapping to: THESOZ, SWD, IAB and IBLK.
- Thesaurus Bildung (BILDUNG): A vocabulary in the *pedagogic* domain which contains round 50,000 terms. Mapping to: THESOZ, SWD, PSYNDEX and BISP.
- Thesaurus Internationale Beziehungen und Länderkunde (IBLK): A vocabulary in the *political science* domain which contains round 8,400 terms. Mapping to: THESOZ, STW, TWSE and CSA-PAIS.
- Thesaurus Sozialwissenschaften (THESOZ): A vocabulary in the *social science* domain which contains round 7,700 terms. Mapping to: GEROLIT, DZI, FES, CSA-WPSA, CSA-ASSIA, CSA-SA, CSA-PEI, ELSST, IAB, IBLK, STW, SWD, BILDUNG, PSYNDEX, INFODATA and BISP.
- Thesaurus für wirtschaftliche und soziale Entwicklung (TWSE): A vocabulary in the *political science* domain which contains round 2,800 terms. Mapping to: IBLK.
- Thesaurus of Sociological Indexing Terms (CSA-SA): A vocabulary in the *social science* domain which contains round 4,300 terms. Mapping to THESOZ.
- Thesaurus of the Deutschen Instituts für soziale Fragen (DZI): A vocabulary in the *social science* domain which contains round 1,900 terms. Mapping to THESOZ.
- Thesaurus of the Deutschen Zentrums für Altersfragen (GEROLIT): A vocabulary in the *gerontology* domain which contains round 1,900 terms. Mapping to THESOZ and MESH.

Listes de descripteurs alignés (N=4):

- Descriptors of the Bundesinstitut für Sportwissenschaft (BISP): A vocabulary in the *sports science* domain which contains round 7,400 terms. Mapping to THESOZ, MESH and BILDUNG.
- Descriptors of the Friedrich-Ebert Stiftung (FES): A vocabulary in the *social science* domain which contains round 4,000 terms. Mapping to THESOZ.
- Descriptors of the Institut für Arbeitsmarkt- und Berufsforschung (IAB): A vocabulary in the *social science* domain which contains round 6,800 terms. Mapping to THESOZ and STW.
- Descriptors of the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences (INION): A vocabulary in the *social science* domain which contains round 7,000 terms. Mapping to THESOZ.

Classifications alignées (N=3):

- Dewey Decimal Classification (DDC): An *univeral* vocabulary which contains thousands of notations. Mapping to RVK.
- Journal of Economic Literature Classification System (JEL): A vocabulary in the *economics* domain which contains round 1,000 notations. Mapping to STW.
- Regensburger Verbundklassifikation (RVK): An *univeral* vocabulary which contains thousands of notations. Mapping to DDC.

Listes de vedettes matière alignées (N=2):

- Medical Subject Headings (MESH): A vocabulary in the *medicine* domain which contains round 23,000 terms. Mapping to PSYNDEX, GEROLIT, BISP and SWD.

- Schlagwortnormdatei (SWD): An *universal* vocabulary which contains round 650,000 terms. Mapping to THESOZ, MESH, STW, AGROVOC and INFODATA.

La figure 2 donne un aperçu de la totalité des 64 passerelles. Le Thesaurus Socialwissenchaften (THESOZ) est le vocabulaire qui comprend le plus d'ajustements terminologiques en entrée et en sortie et en raison de sa position centrale, il se déploie au milieu du réseau. D'autres vocabulaires comme SWD ou PSINDEX jouent un rôle central dans la connexion à d'autres domaines. L'alignement terminologique CDD-RVK est la seule concordance qui n'est pas connectée. Le travail effectué à l'occasion du projet CRISSCROSS alignant SWD et CDD pourrait être utilisé à cette fin. L'alignement terminologique JEL-STW est un exemple d'une concordance uni-directionnelle de JEL à STW.

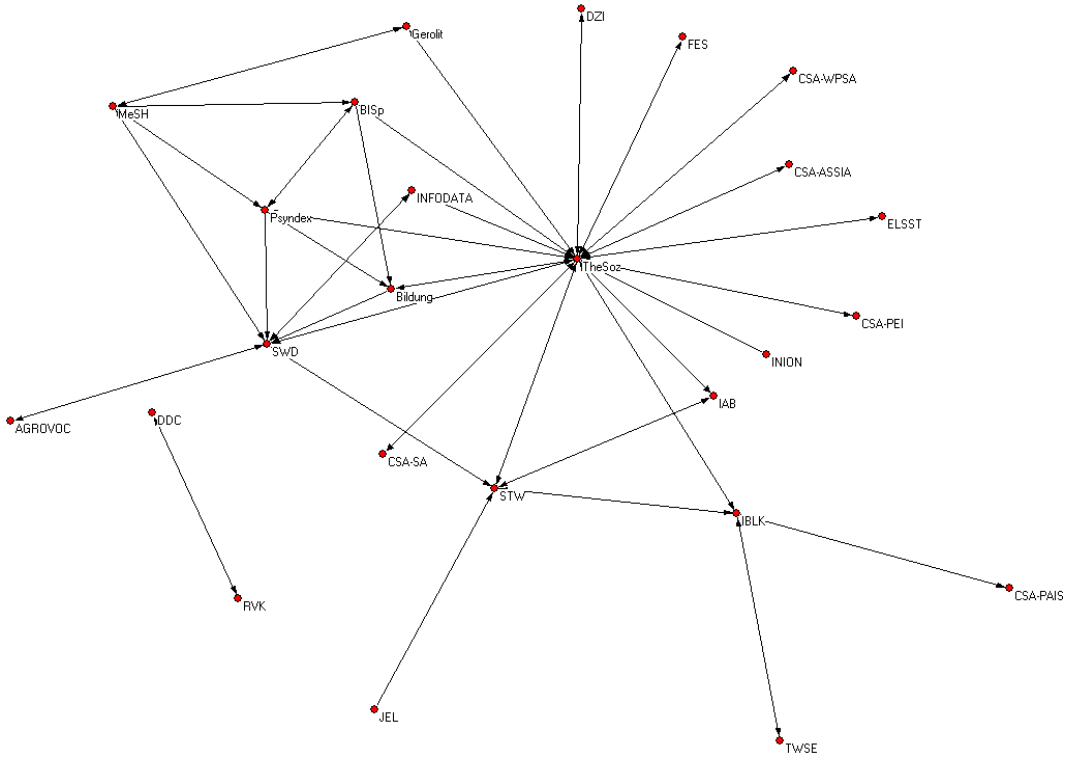


Figure 2. Net of mapped vocabularies in the KoMoHe project.

Les 513 000 relations disponibles dans notre base de concordances concernent plus de 181000 concepts uniques (descripteurs contrôlés uniques ou combinaisons de descripteurs ou de notations) et plus de 270 000 termes de départ (descripteurs contrôlés uniques). En moyenne (par concordance), 6 500 termes du vocabulaire initial sont alignés à 3 600 termes dans le vocabulaire final (1.2 relations par terme).

La figure 3 montre la répartition des types de relations dans le projet (à comparer avec le tableau 2). La relation d'équivalence (autour de 45%) est le type de relation rencontrée le plus fréquemment entre des termes. Seulement 12% de toutes les relations sont « nulles » (pas d'alignement de termes possible).

Un panorama plus détaillé ainsi qu'une analyse quantitative de la répartition sera le sujet d'un article prochain.

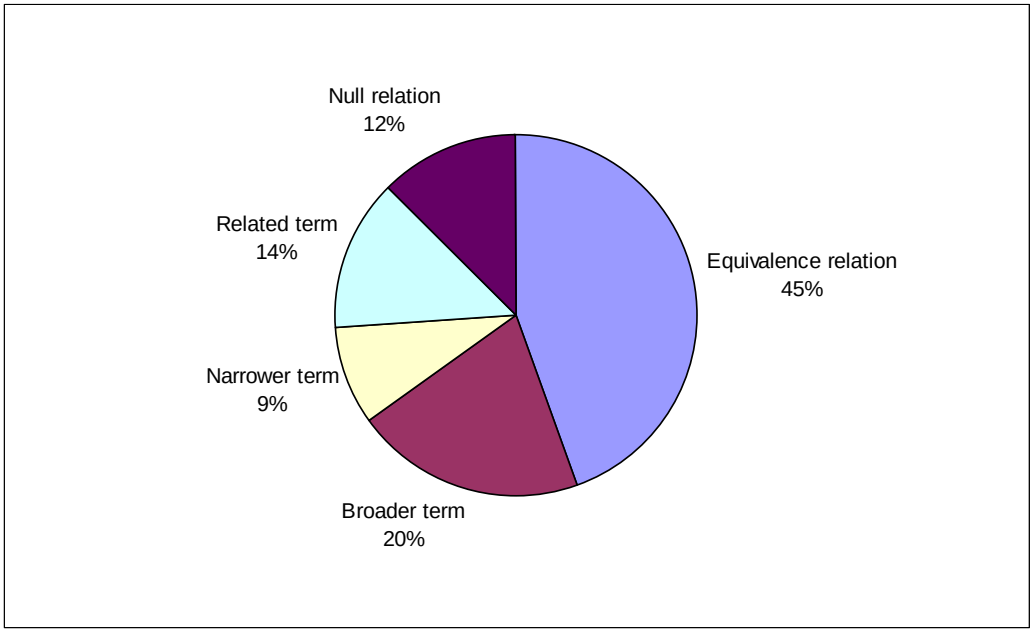


Figure 3. Distribution of relationship types across all cross-concordances.

3.2 Implémentation des concordances

Une base de données relationnelle a été créée afin de conserver les concordances pour une utilisation prochaine. Il a été mis en évidence que la structure relationnelle est capable de reprendre de façon appropriée le nombre de différents vocabulaires contrôlés, termes et combinaisons de termes, et relations. Les vocabulaires et termes sont représentés sous forme de listes, indépendantes les unes des autres sans prendre en compte particulièrement la structure syndétique (de coordination) des vocabulaires concernés.
L'orthographe et l'exploitation des termes issus du vocabulaire contrôlé sont normalisées.
Les combinaisons de termes (c'est à dire computer + crime en tant que combinaison reliée au terme hacker) sont également stockées en tant que concepts distincts.
Afin de rechercher et d'extraire les données terminologiques de la base, un service web (appelé service de l'hétérogénéité ou HTS dans le tableau 4, voir Mayr & Walter, 2008) a été créé afin d'accompagner les recherches de concordances concernant des termes isolés lors de la recherche initiale, des termes alignés, des vocabulaires initiaux ou finaux ainsi que les différents types de relation. Une implémentation qui utilise les relations d'équivalence, recherche les termes dans la liste des termes issus du vocabulaire contrôlé et ajoute automatiquement à la requête tous les termes équivalents de tous les vocabulaires disponibles. Si les vocabulaires sont dans des langues différentes, le service de l'hétérogénéité traduit également le terme original dans l'autre langue. Si la requête originelle contient un opérateur booléen, il subsiste après le déploiement de la requête (chaque mot de la requête est développé séparément). En raison des indices de performance, le développement de la requête portant sur les concordances ne fait pas la différence entre les différentes bases de données et leurs termes issus de vocabulaires contrôlés préférentiels mais inclut tous les termes équivalents à la requête. En principe, cette utilisation d'un réseau terminologique permet une requête étendue aux synonymes ou aux quasi synonymes des termes originaux.

4. Évaluation des concordances

4.1 Questions d'ordre général

Bien que le besoin d'alignement terminologique soit généralement reconnu et que de nombreux projets d'alignements soient menés, l'efficacité et l'utilité réelles de ces projets sont rarement évaluées avec rigueur. De nombreuses questions peuvent être soulevées au sujet de ces réseaux terminologiques engendrés par les alignements, telles que :

- Combien d'expressions peut-on trouver pour un même concept ?
- Quels sont les concepts liés ?
- Les vocabulaires ont-ils une portée générique ou spécifique ?
- Quelles terminologies sont très semblables ?
- Quelles disciplines/quels sujets sont adjacents ou éloignés ?
- Quel chevauchement y-a-t-il entre les différentes bases de données ou les vocabulaires contrôlés sur un sujet donné ?

La question la plus importante, et celle pour laquelle sont établis le plus d'alignements, est celle de l'utilité et de l'efficacité de ces derniers pour la recherche. Dans un portail avec nombre de bases de données différentes, la question de savoir si les concordances permettent une recherche distribuée devient cruciale. Peuvent-elles combler les différences de langages pour faciliter une recherche transparente à l'aide de la même requête dans les différentes bases de données ?

Quand on évalue les alignements terminologiques, le point de départ analytique est très important. Que contrôle-t-on : la qualité des alignements eux-mêmes ou la qualité des recherches qu'ils permettent ? La qualité des alignements est un pré-requis pour une meilleure qualité de recherche. Dans les cas des concordances du projet KoMoHe, des experts des institutions partenaires ont vérifié chaque alignement en fonction de leur domaine de spécialité. La création manuelle et les corrections soigneuses offrent la certitude que les alignements sont correctement faits, de bonne qualité et adaptés.

Les caractéristiques intrinsèques des concordances (et leur impact sur la recherche) peuvent différer en fonction des vocabulaires contrôlés qui ont été alignés et de facteurs externes dans le processus de création de concordances. Par exemple, la date de création d'une concordance peut avoir des conséquences sur le nombre de relations à partir d'un terme. Plus en amont dans le projet, on a formé moins de relations. Des concordances d'un projet antérieur (CARMEN) ont été discutées dans un groupe d'experts, elles sont plus sélectives. Des modifications dans les vocabulaires contrôlés ou les pratiques d'indexation peuvent aussi avoir un impact sur la qualité des concordances. On observe d'autres différences :

- Taille du langage source/du langage cible
- Différences entre vocabulaires pré-coordonné et post-coordonné
- Nombre de relations
- Nombre de termes alignés (couverture, chevauchement)
- Différents types de relations (équivalence, terme générique, terme spécifique, terme associé, absence de relation)
- Différents types de pertinence (haute, moyenne, basse)
- Alignement de termes identiques
- Disparité dans la spécificité (c'est-à-dire vocabulaires dont la portée est soit très spécifique soit très générique)

- Combinaisons de termes alignés (l'alignement se fait vers plus d'un terme)

Une analyse quantitative offre une vision des caractéristiques essentielles d'une concordance mais ne peut déterminer les améliorations qualitatives obtenues en utilisant des alignements spécifiques dans la recherche. Nous avons conçu un test de recherche d'information qui a pour but d'évaluer l'application de concordances dans un réel scénario de recherche.

4.2 Conception d'un test de recherche d'information

Dans la recherche, de nombreux facteurs entrent en jeu quand il s'agit d'évaluer la qualité des alignements terminologiques : les concordances elles-mêmes, mais aussi le contenu des bases de données concernées, leur couverture ou les chevauchements de leurs contenus, l'interface de recherche ou le classement des résultats de la recherche. Le but étant d'évaluer l'impact des concordances, les conditions effectives de recherche (interface, classement des réponses, etc.) ont été maintenues aussi stables que possible.

L'idée basique pour utiliser les concordances est de traduire les termes de la recherche dans d'autres terminologies afin de faciliter la recherche dans différentes bases de données et terminologies. Se lancer dans les concordances devrait agrandir l'espace de recherche, corriger les ambiguïtés et les imprécisions dans la formulation de la recherche et donc trouver plus de documents pertinents pour une requête donnée.

L'application des concordances dans la recherche présente aussi un caveat : elles pourraient avoir un impact sur la vitesse et la facilité d'emploi du processus de recherche lui-même. Un des principes pour l'implémentation technique des alignements terminologiques devrait être que le chercheur puisse les utiliser sans même s'en apercevoir. Les alignements devraient améliorer l'expérience de la recherche sans augmenter l'effort de l'usager. En faisant usage d'une approche strictement automatique pour injecter des concordances au cours de l'évaluation, aucune intervention manuelle sous la forme d'une reformulation de la requête par un humain ne fut nécessaire.

Deux tests de recherche d'information furent conçus pour évaluer la qualité des concordances pour la recherche :

Test 1 : *L'application d'alignements de termes apporte-t-elle une amélioration à la recherche par rapport à une recherche sujet non modifiée (c'est-à-dire en vocabulaire contrôlé) ?*
Dans le test 1, une requête a été traduite dans les termes du vocabulaire contrôlé (A) et recherchée dans les champs des termes contrôlés de la base de données bibliographique avec un vocabulaire contrôlé différent (B). La recherche fut répétée à l'aide de la concordance A → B, traduisant les termes de la recherche du vocabulaire contrôlé initial en termes du vocabulaire contrôlé de la base de donnée cible. La figure 4 est une représentation graphique de ce processus. Les résultats de la recherche furent comparés.

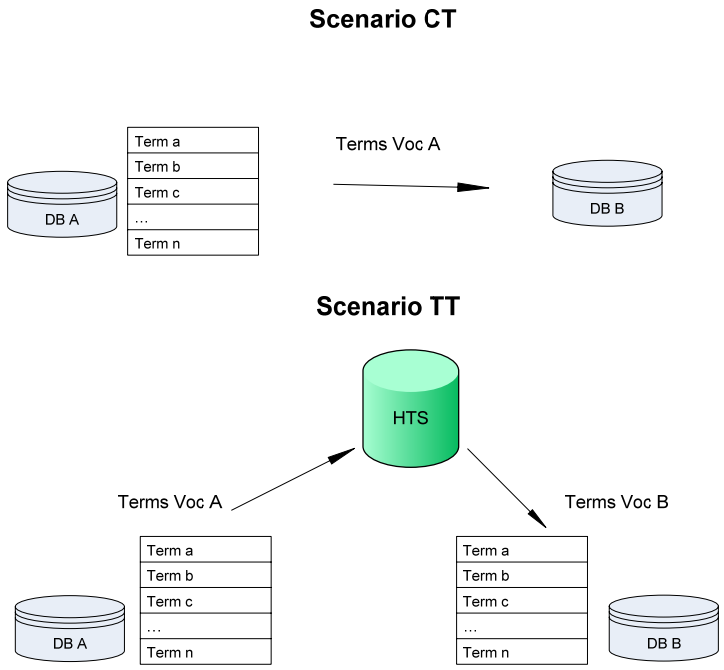


Figure 4. Cross-concordance information retrieval evaluation set-up. Scenario CT = controlled term search. Scenario TT = term-transformed search. HTS = heterogeneity service.

Si les résultats des deux recherches sont identiques, c'est que l'application de concordances pour la transformation des termes est sans effet. Si les résultats sont moins intéressants, c'est que la concordance a un effet négatif, et si les résultats sont meilleurs, ce que l'on espère, c'est que la concordance a un effet positif sur la recherche.

Test 2 : *L'application d'alignements terminologiques améliore-t-elle une recherche plein texte ?*
Dans le test 2, la requête initiale était utilisée dans un scénario de recherche plein texte (la plupart des chercheurs n'utilisent pas les vocabulaires contrôlés appropriés). Une recherche plein texte cherche les termes de la requête non seulement dans les champs en vocabulaire contrôlé, mais aussi dans les champs titre et résumé. Dans l'expérience, la requête initiale fut d'abord lancée dans les champs en plein texte de la base de données. En second lieu, les termes de la requête ont été recherchés dans la concordance (si toutefois il existait un alignement). Les nouveaux termes issus de la concordance ont été ajoutés à la requête initiale.

Un exemple illustre les différences entre les tests 1 et 2 : une requête en langage naturel cherche des documents sur les « family relations ». « Family relations » est un terme contrôlé en vocabulaire A et n'a de ce fait pas besoin d'être traduit en terme contrôlé, il est donc utilisé pour la première recherche dans les champs en vocabulaire contrôlé de la base de données B : *Test 1 CT : Family relations*
La concordance A → B aligne la locution « family relations » qui vient du vocabulaire A avec la combinaison de termes « family » ET « social relations » du vocabulaire B. La seconde recherche dans la base de données B est donc : *Test 1 TT : Family ET social relations*

Pour le test 2, la requête originale est lancée dans les champs en plein texte de la base de données B (titre, résumés et termes contrôlés) : *Test 2 FT : Family relations.*

Comme le terme de la requête apparaît dans le vocabulaire A, on peut trouver un terme aligné pour la requête dans la concordance. On ajoute les termes à la requête initiale et on les cherche dans les champs en plein texte de la base de données B : *Test 2 FT+TT : Family relations OR (Family AND social relations)*

Le test 1 ne cherche que dans les champs de termes contrôlés, tandis que le test 2 cherche aussi dans d'autres champs où le terme de la requête pourrait apparaître par hasard (titre et résumé). Le test 2 est en général un test moins fiable car les termes alignés sont ajoutés à la requête et ne remplacent pas la requête originale comme dans le test 1. Comme le test 2 recherche aussi dans les champs dans lesquels le test 1 cherche, on pourrait estimer que le test 2 englobe le test 1. Dans le test 2 néanmoins il y a moins d'ajouts de termes car tous les termes de la requête n'apparaissent pas dans le vocabulaire contrôlé initial et on trouve moins de termes alignés.

Pour la création de requêtes, l'aide des producteurs ou des hébergeurs des bases de données furent sollicités afin de s'assurer que des requêtes réalistes étaient lancées. On leur a demandé de fournir entre 3 et 10 requêtes (en moyenne : 6-7) basées sur leur expérience quotidienne, qui étaient traduites dans les vocabulaires contrôlés des bases de données en test. Les requêtes en langage naturel lancées dans le plein texte comprenaient entre 1 et 3 mots par requête, alors que les requêtes booléennes de termes contrôlés comprenaient entre 2 et 6 mots. Pour tous les alignements dans les expériences de recherche d'information, seules les relations d'équivalence furent utilisées. Tous les documents énumérés par le système d'information furent retrouvés avec une limite de 1000 documents classés. Enfin, les ensembles de documents trouvés dans chaque expérience ont été évalués par pertinence.

Pour évaluer l'effet des concordances, les mesures classiques de recherche d'information, le taux de rappel et la précision, basés sur l'évaluation de pertinence des documents trouvés ont été utilisées.
Les mesures suivantes ont été analysées :

- Trouvés (retrieved) : nombre moyen de documents trouvés (tous types de recherches)
- Pertinents (relevant) : nombre moyen de documents pertinents trouvés (tous types de recherches)
- Rel_ret : nombre moyen de documents pertinents trouvés en fonction de types de recherches
- Taux de rappel (recall) : proportion de documents pertinents trouvés par rapport au nombre total de documents pertinents (en moyenne sur toutes les requêtes dans un type de recherche)
- Précision : proportion de documents pertinents trouvés par rapport à l'ensemble des documents trouvés (en moyenne sur toutes les requêtes dans un type de recherche)
- P10 : Précision à 10 = précision après 10 documents trouvés
- P20 : Précision à 20 = précision après 20 documents trouvés

P10 et P20 ont été calculés afin de représenter un scénario de recherche réaliste : en effet les usagers ne regardent en général pas plus loin que la première ou la deuxième page de résultats. Pour les systèmes de recherche qui ne classent pas mais énumèrent les résultats par année ou par auteur, P10 et P20 ne sont pas significatifs.

4.3 Concordances et bases de données en test

Dans ces deux expériences, les concordances étaient divisées par disciplines (intra- ou interdisciplinaires) et par langues (mono- ou bilingue). Les concordances intradisciplinaires recouvrent des vocabulaires en sciences sociales puisque la plupart des concordances créées dans ce projet se situent dans cette discipline. Les concordances interdisciplinaires sont des vocabulaires alignés dans les domaines de l'économie, de la médecine, des sciences politiques, de la psychologie et des sciences sociales. Les concordances monolingues comportaient des vocabulaires en allemand ; les concordances bilingues comportaient un vocabulaire allemand et un anglais. Le tableau 3 montre le nombre de concordances en test par expérience :

Test 1: Controlled term search	
Intradisciplinary cross-concordances	5 (1 bilingual)
Interdisciplinary cross-concordances	8
Test 2: Free-text search	
Intradisciplinary cross-concordances	6 (1 bilingual)
Interdisciplinary cross-concordances	2

Table 3. Number of cross-concordances tested

Derrière la séparation de ces concordances de cette manière se trouve l'hypothèse que les concordances entre des vocabulaires dans la même discipline (intradisciplinaires) vont se chevaucher et contiennent plus de termes identiques et ont donc un effet moindre sur les résultats de la recherche que les concordances interdisciplinaires. Les alignements terminologiques entre des vocabulaires dans un langage naturel différent (par exemple anglais → allemand) ou entre des systèmes différents de notations (par exemple CDD → LCC) auront aussi plus d'impact parce que les occurrences de termes identiques ou les chevauchements sont moins probables.

Pour les expériences, on a inclus les bases de données bibliographiques qui contiennent entre 70 000 et 16 millions de documents, produites et hébergées pour la plupart en Allemagne. Parmi elles se trouvaient des bases de résumés et d'indexation, mais aussi des catalogues de bibliothèques. Le tableau 4 montre les bases de données en test et les vocabulaires associés :

Vocabulary	**Discipline**	**Database**	**Documents in DB**
TheSoz – Thesaurus Sozialwissenschaften (GESIS-IZ)	Social Sciences	SOLIS	345,086
DZI – Thesaurus des Deutschen Instituts für soziale Fragen	Social Sciences	SoLit	151,925
SWD – Schlagwortnormdatei	General (Social Sciences Excerpt)	USB Köln Sowi OPAC	72,729
CSA – Thesaurus of Sociological Indexing Terms (Cambridge Scientific Abstracts)	Social Sciences	CSA Sociological Abstracts	294,875
Psyndex - Psyndex Terms	Psychology	Psyndex (ZPID)	Ca. 200,000
STW – Standard Thesaurus Wirtschaft	Economics	Econis (ZBW Kiel)	Ca. 3,000,000
IBLK - Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)	Political Science	World Affairs Online WAO (SWP Berlin)	643,420
Mesh – Medical Subject Headings	Medicine	Medline (Dimdi)	Ca. 16,800,000

Table 4. Vocabularies and databases in the KoMoHe IR test

De nombreuses concordances ainsi que les bases de données qui leur sont associées ont pu être mises en test sur place en indexant les documents grâce au système libre de recherche de l'information Solr[12] qui utilise les mêmes modules de traitement et de classement pour toutes les bases de données. Pour les bases de données non disponibles sur place, les hébergeurs étaient priés de fournir des listes de résultats ordonnés pour des requêtes pré-déterminées.

Pour la plupart des bases de données, les alignements terminologiques furent testés dans les deux directions, allant aussi bien du vocabulaire A vers le vocabulaire B (A → B) que du vocabulaire B vers A (B → A). Comme ils forment des recherches différentes (les différentes requêtes étant dépendantes du vocabulaire de départ) et des bases de données différentes, ils sont indépendants.

5. Résultats de l'évaluation

5.1 Test 1 : Recherche en terme contrôlé

Le test 1 avait pour objet d'évaluer si le remplacement d'une requête avec les termes du vocabulaire A (CT) par des termes contrôlés du vocabulaire B (transformation par alignement terminologique) (TT) améliorait la recherche dans la base de données B. Si l'alignement est imprécis ou ambigu ou si les vocabulaires se chevauchent, alors la traduction de la requête initiale dans la requête alignée pourrait amener du bruit dans la formulation de la requête, ce qui peut donc faire obstacle à la qualité de la recherche..

Le tableau 5 montre les résultats moyens de 13 concordances mises en test. La dernière ligne montre la différence en pourcentage entre les types de recherche :

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	156.5	144.8	42.0	0.3152	0.2214	0.1987	0.1748
TT	325.4	144.8	88.2	0.6047	0.3391	0.3052	0.2848
				91.8%	**53.2%**	**53.6%**	**62.9%**

Table 5. Test 1 evaluation results for all cross-concordances (N=13)

La recherche qui utilise les transformations de termes multiplie par deux le nombre de documents trouvés, on obtient davantage de documents comportant les termes de la requête. Le taux de rappel augmente de presque 100%, tandis que la précision augmente de près de 45%. L'utilisation d'une concordance dans cette recherche particulière trouve non seulement davantage de documents pertinents (taux de rappel) mais est aussi plus précise (précision) qu'une recherche sans transformation de termes.

Néanmoins cette nette amélioration est due en partie à la traduction entre l'anglais et l'allemand dans la concordance bilingue. Tandis que des alignements terminologiques monolingues pourraient être sans effet car les termes alignés sont identiques, ce n'est pas le cas dans un alignement traduit. Le tableau 6 montre les résultats de la recherche quand la concordance bilingue est ôtée de l'ensemble mis en test :

[12] http://lucene.apache.org/solr/

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	169.6	141.2	45.5	0.3415	0.2399	0.2153	0.1894
TT	320.5	141.2	87.6	0.6113	0.3431	0.3126	0.2877
				79.0%	**43.1%**	**45.2%**	**51.9%**

Table 6. Test 1 evaluation results for all monolingual cross-concordances (N=12)

En raison du chevauchement de termes, les résultats de la recherche devraient être différents selon que la concordance couvre deux disciplines (interdisciplinaire) ou une seule (intradisciplinaire). Si l'on sépare les résultats du test en fonction des disciplines, on relève des changements significatifs dans les résultats de la recherche. Dans le cas des concordances intradisciplinaires, taux de rappel et précision augmentent mais pas autant. Un changement moindre ou négatif dans la précision serait en fait attendu puisque, de façon générale dans la recherche d'information, précision et taux de rappel sont dans une situation inversée l'un par rapport à l'autre (si le taux de rappel augmente, la précision diminue).

Le tableau 7 montre le taux de rappel moyen et les mesures de précision pour toutes les concordances et pour les concordances monolingues intradisciplinaires. Pour les concordances monolingues intradisciplinaires, précision et taux de rappel augmentent mais bien moins que pour toutes les concordances.

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	126.6	101.3	36.2	0.3726	0.2491	0.2002	0.1637
TT	238.9	101.3	59.9	0.5189	0.3335	0.2784	0.2352
				39.3%	**33.9%**	**39.1%**	**43.7%**
Monolingual							
CT	158.2	79.7	45.3	0.4657	0.3113	0.2503	0.2046
TT	202.9	79.7	51.0	0.5174	0.3441	0.2939	0.2315
				11.1%	**10.5%**	**17.4%**	**13.2%**

Table 7. Test 1 evaluation results for intradisciplinary cross-concordances (N=5)

Une amélioration extraordinaire est visible pour le taux de rappel et la précision quand on applique une concordance interdisciplinaire. Taux de rappel et précision augmentent bien plus que pour une concordance moyenne (cf. Tableau 8) :

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	175.2	171.9	45.6	0.2794	0.2041	0.1978	0.1817
TT	379.4	171.9	105.9	0.6583	0.3426	0.3220	0.3157
				135.6%	**67.8%**	**62.8%**	**73.7%**

Table 8. Test 1 evaluation results for interdisciplinary cross-concordances (N=8)

Utiliser les concordances a plus qu'un effet positif sur la recherche de terme contrôlé. L'ensemble de résultats est non seulement plus grand mais aussi plus précis. On observe l'impact le plus important pour des concordances qui couvrent plus d'une discipline.

5.2 Test 2 : Recherche plein texte

Le test 2 avait pour objet d'évaluer si l'ajout de termes de vocabulaire contrôlé obtenus en alignant des termes de requête en langage naturel d'une base de données (FT-CK) à une requête plein texte (FT) améliorerait les résultats. Pour certaines des requêtes individuelles dans les tests, aucun changement n'était fait dans les requêtes car aucun terme contrôlé correspondant ne pouvait être trouvé. Le tableau 9 montre les résultats pour les 8 concordances mises en test :

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	155.3	106.4	56.2	0.6026	0.4551	0.4101	0.3682
FT-CK	266.8	106.4	72.8	0.7273	0.3934	0.3203	0.3083
				20.7	**-13.6**	**-21.9%**	**-16.3%**

Table 9. Test 2 evaluation results for all cross-concordances (N=8)

Les résultats montrent que non seulement davantage de documents mais aussi des documents plus pertinents sont trouvés. Le taux de rappel moyen augmente encore de 20%. En général, les termes contrôlés ajoutés à une requête peuvent améliorer les résultats. Néanmoins on observe une chute dans la précision, qui n'est néanmoins pas aussi importante que l'augmentation du taux de rappel.

Le tableau 10 montre les résultats pour les termes alignés dans la même discipline, tandis que le tableau 11 montre les résultats pour deux concordances interdisciplinaires.

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	163.8	115.8	60.2	0.5934	0.5025	0.4635	0.4090
FT-CK	244.9	115.8	77.1	0.7096	0.4449	0.3826	0.3681
				19.6	**-11.5%**	**-17.5%**	**-10.0%**

Table 10. Test 2 evaluation results for intradisciplinary cross-concordances (N=6)

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	129.9	78.2	44.3	0.6303	0.3129	0.2500	0.2459
FT-CK	332.4	78.2	59.8	0.7805	0.2388	0.1333	0.1292
				23.8%	**-23.7%**	**-46.7%**	**-47.5%**

Table 11. Test 2 evaluation results for interdisciplinary cross-concordances (N=2)

Contrairement à l'analyse faite pour le test 1, les différences entre les concordances intradisciplinaires et interdisciplinaires ne sont pas aussi grandes. Pour les deux ensembles, le taux de rappel augmente par rapport à une simple recherche plein texte tandis que la précision chute. Le taux de rappel augmente légèrement pour les concordances interdisciplinaires tandis que la précision chute beaucoup plus. Cela pourrait être dû au fait que seules deux concordances interdisciplinaires ont été évaluées, ce qui pourrait biaiser les résultats et ne pas montrer de tendance claire.

Les résultats des l'expérience de la recherche plein texte pourraient sans doute être bien améliorés si les termes contrôlés et les termes en langage naturel étaient mieux intégrés dans la requête plutôt que d'être juste ajoutés les uns aux autres. Une possibilité serait de traduire (de manière automatique) les termes en langage naturel en termes contrôlés a priori de telle sorte qu'il y ait plus de chances pour que l'alignement se fasse vers différents vocabulaires contrôlés et bases de données (cf. Mayr et al., 2008).

En conclusion, ces expériences de recherche d'information montrent les effets positifs des concordances pour la recherche dans des bases de données hétérogènes. Les résultats s'améliorent pour toutes les concordances, néanmoins, les concordances interdisciplinaires ont un impact plus important (positif) sur les résultats de la recherche. Pour toutes les concordances testées dans ces scénarios, davantage de documents pertinents sont trouvés par rapport aux types de requêtes sans l'utilisation de concordances ; dans des cas particuliers l'ensemble des réponses était en outre plus précis (augmentation de la précision également).

6. Conclusion et perspectives

Après avoir montré que se lancer dans les concordances pour la recherche peut avoir un effet positif sur les résultats, nous envisageons de les implémenter sur le portail vascoda. Nous avons déjà employé de nombreuses concordances pour la recherche dans le portail allemand d'information en sciences sociales sowiport[13] qui propose des informations bibliographiques et autres (dont 15 bases de données avec 10 vocabulaires différents et environ 2.5 millions de références bibliographiques).

L'implémentation dans sowiport, qui, comme les expériences, n'utilise que des relations d'équivalence, cherche les termes dans le vocabulaire contrôlé et ajoute ensuite automatiquement à la requête tous les termes équivalents de tous les vocabulaires alignés disponibles. C'est donc similaire à la méthodologie appliquée dans l'expérience 2 (recherche plein texte). Les opérateurs booléens fonctionnent comme des séparateurs entre les locutions de recherche : ils demeurent intacts après le développement de la requête (c'est-à-dire que chaque partie de la requête est développée séparément). L'alignement terminologique est automatique et invisible pour l'usager, une petite icône symbolise la transformation (cliquer sur l'icône permet de lister les termes ajoutés à la requête)

Pour effectuer plus de recherches et avoir plus de mémoire, une base de données de relation fut créée pour stocker les concordances. Pour rechercher et obtenir des données de la base de données, un service web (appelé service hétérogénéité, cf Mayr & Walter, 2008) fut conçu pour les recherches de concordance à partir d'un terme, de termes alignés, de vocabulaires de départ et d'arrivée ainsi que différents types de relations. Néanmoins, il est aussi possible de faire des recherches dans la base elle-même. Les données de l'alignement terminologique ainsi que le service web sont rendus disponibles dans un but de recherche. Quelques alignements sont déjà en usage pour des disciplines spécifiques à la conférence de CLEF (Cross-Language Evaluation Forum) (Petras, Baerisch & Stempfhuber, 2007). D'autres caractéristiques et applications des concordances comme la permutation, l'interaction ou la manipulation feront l'objet de recherches ultérieures.

Une autre opportunité de conserver et faire des requêtes est le langage SKOS (Simple Knowledge Organization System)[14]. Le but de ce langage, qui est à l'état de version préliminaire au W3C, est la représentation formalisée de vocabulaires contrôlés, dans le cadre du Web sémantique. L'ébauche comporte une section sur l'alignement terminologique. Une fois que le langage SKOS sera stabilisé, nous rendrons nos alignements disponibles dans ce format.

[13] http://www.sowiport.de/
[14] http://www.w3.org/2004/02/skos/

Un domaine intéressant est la création d'alignements à partir d'un vocabulaire pivot pour des ressources où aucun moyen ne permet d'alignement direct. Si le vocabulaire A est aligné avec le vocabulaire B et le vocabulaire B avec le vocabulaire C, il devrait être possible de créer un alignement de A → C en utilisant le vocabulaire B comme pivot. Nous espérons qu'avec le développement d'un langage pour la présentation et l'échange, davantage d'alignements et de vocabulaires vont être disponibles pour des recherches ultérieures.

Acknowledgements

The project was funded by BMBF, grant no. 01C5953. We wish to thank all our project partners for their collaboration. We especially appreciate the help of our colleagues Anne-Kathrin Walter and Stefan Baerisch, who implemented most of the technical infrastructure and helped with the evaluation.

References

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level. D-Lib Magazine, 12(6).

Depping, R. (2007). vascoda.de and the system of the German virtual subject libraries. In A. R. D. Prasad & D. P. Madalli (Eds.), International Conference on Semantic Web & Digital Libraries (ICSD 2007) (pp. 304-314). Bangalore, India: Documentation Research & Training Centre, Indian Statistical Institute.

Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1(8).

Doerr, M. (2004). Semantic interoperability: Theoretical Considerations.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., et al. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften.

Krause, J. (2003). Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. Paper presented at the IFLA 2003, World Library and Information Congress: 69th IFLA General Conference and Council, Berlin.

Liang, A. C., & Sini, M. (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review in Hypermedia and Multimedia, 12(1), 51-62.

Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. Bangalore, India.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed Digital Libraries: treatment of term vagueness and document re-ranking. Library Review, 57(3), 213-224. URL: http://www.emeraldinsight.com/10.1108/00242530810865484

Mayr, P., & Walter, A.-K. (2008). Mapping Knowledge Organization Systems. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte der Wissenorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 80-95). Würzburg: Ergon.

Panzer, M. (2008). Semantische Integration heterogener und unterschiedlichsprachiger Wissensorganisationssysteme: CrissCross und jenseits. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte in der Wissensorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 61-69). Würzburg: Ergon.

Patel, M., Koch, T., Doerr, M., & Tsinaraki, C. (2005). Semantic Interoperability in Digital Library Systems.

Petras, V., Baerisch, S., & Stempfhuber, M. (2007). The Domain-Specific Track at CLEF 2007, Cross Language Evaluation Forum Workshop (CLEF) 2007. Budapest.

Tudhope, D., Koch, T., & Heery, R. (2006). Terminology Services and Technology: JISC state of the art review.

Vizine-Goetz, D., Hickey, C., Houghton, A., & Thompsen, R. (2004). Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4).

Vizine-Goetz, D., Houghton, A., & Childress, E. (2006). Web Services for Controlled Vocabularies. ASIS&T Bulletin, 2006 (June/July).

Zeng, M. L., & Chan, L. M. (2004). Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. Journal of the American Society for Information Science and Technology, 55(3), 377-395.

Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. D-Lib Magazine, 12(6).