HAL
archives-ouvertes.fr

De l'importance de la normalisation en bibliométrie

Emmanuel Nauer

▶ **To cite this version:**

Emmanuel Nauer. De l'importance de la normalisation en bibliométrie. Société Française de Bibliométrie Appliquée. Les systèmes d'information élaborée, 1999, Ile Rousse, France, 1999. <inria-00110354>

HAL Id: inria-00110354
https://hal.inria.fr/inria-00110354

Submitted on 27 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

De l'importance de la normalisation en bibliométrie

Emmanuel Nauer (LORIA / CNRS)

UMR LORIA 7503 - Bâtiment LORIA - 615, rue du Jardin Botanique - B.P. 239
54506 Vandoeuvre-lès-Nancy Cedex

Emmanuel.Nauer@loria.fr

Mots-clés

fédération de sources multiples, normalisation, lemmatisation, dédoublonnage, SGML, analyse statistique de l'information

Introduction

L'interrogation de plusieurs banques de données est une nécessité dans la constitution d'une bibliographie exhaustive sur un sujet pluridisciplinaire [Gehanno 1998]. Or, si la consultation de multiples sources d'information garantit une meilleure couverture du sujet, elle nécessite -en parallèle- de prendre en compte l'hétérogénéité des données. En effet, comme chaque producteur d'information possède son propre modèle de document, on obtient des représentations différentes d'une même information. Cette hétérogénéité occasionne deux problèmes majeurs pour la bibliométrie qui faussent statistiques et analyses. Ces difficultés concernent :

- la présence de variations pour représenter une même information (auteurs, descripteurs, affiliation, etc.) ;
- la présence de doublons (c'est-à-dire plusieurs représentations de la même référence bibliographique).

Dans cet article, nous proposons une approche visant à pallier ces problèmes. Nous discutons tout d'abord les choix retenus en terme de normalisation et de dédoublonnage.

Nous présentons ensuite la chaîne de traitement mise en place dans le but de fusionner de multiples sources. Cette chaîne repose sur une structure pivot de représentation des documents en SGML[1] et inclut la normalisation des structures et des contenus. Nous détaillons ici plus particulièrement nos travaux sur la normalisation des auteurs et des descripteurs, ainsi que sur le dédoublonnage des notices. Pour cela, nous décrivons une méthode générale pour établir des équivalences (d'auteurs, de descripteurs, de notices, etc.) et présentons l'algorithme que nous avons utilisé pour la normalisation des auteurs, des descripteurs et pour le dédoublonnage.

Nous insisterons, au fil de cet article, sur le caractère essentiel de la normalisation et du dédoublonnage dans le domaine de la bibliométrie, à travers une étude menée en collaboration avec l'INRS[2]. Cette étude porte sur un corpus documentaire concernant le stress professionnel et contenant 26251 références bibliographiques provenant de l'interrogation de 8 sources différentes (Medline, Embase, Biosis, Psyclit, NIOSHTIC, Cisilo, INRS-B et Pascal). Nous

1 Standard Generalized Markup Language [ISO 8879:1986]

2 Institut National de Recherche et Sécurité

montrons également comment nous avons utilisé le thésaurus de l'UMLS[3] pour normaliser les descripteurs.

1. Normalisation des données

L'approche que nous avons adoptée consiste à fédérer les multiples sources d'information avec un souci d'homogénéisation, afin de prendre en compte les diversités de représentation des informations, et afin d'aboutir à un modèle cohérent de représentation capable d'unifier les différents contenus.

Pour cela, nous avons défini un modèle de données vers lequel nous avons fait converger les différents modèles initiaux. La chaîne de traitement mise en place opère tout d'abord une normalisation de la structure des notices, puis une normalisation du contenu de chacun des champs, comme le préconise [Dou 91].

Nous décrivons les principales étapes nécessaires à la normalisation et renvoyons à [Jolibois 1999] pour une revue détaillée des normes.

Normalisation de la structure

Chaque base dispose d'une structure spécifique, constituée d'un nombre variable de champs. Si certains champs sont communs à toutes les bases (Auteurs, Titre, Date de publication, etc.), d'autres sont spécifiques à une ou deux bases seulement (les tags[4], par exemple, sont uniquement proposés par Medline, Embase et Psyclit). D'autres enfin apparaissent sous des formes différentes selon qu'ils sont fusionnés ou non. A titre d'exemple, le Titre du périodique, le Volume, le Fascicule et la Pagination apparaissent dans Medline PubMed dans un unique champ Source.

La normalisation de la structure peut être possible en utilisant des normes préexistantes. Celles-ci concernent la structure logique (de quelles informations une référence est-elle constituée ?) et la structure physique (quel format informatique utiliser pour gérer les références ?).

La normalisation de la structure logique peut reposer sur l'utilisation de différentes normes :

- des normes de catalogage, comme l'[ISBD 1992] (General International Standard Bibliographic Description) , l'[AACR2:1988] (Anglo-American Cataloguing Rules 2nd edition) ; et/ou
- des normes éditoriales, comme l'[ISO 690:1987], le style de Vancouver [ICMJE 1997], qui fait autorité en médecine, etc.

La normalisation de la structure physique, quant à elle, se base directement sur les formats d'échange de données, comme MARC (MAchine Readable Cataloguing) [LC 1999], SGML [ISO 8879:1986], TEI (Text Encoding Initiative) [ETC 1994], etc.

Après avoir pris en compte les différentes normes citées ci-dessus, nous avons défini un modèle de données homogène comprenant un ensemble bien défini de champs (structure logique), et avons opté pour l'utilisation de la norme SGML (structure physique) pour bénéficier des nombreux outils de manipulation de données SGML disponibles gratuitement.

Pour uniformiser les données provenant de différentes bases, il a donc été nécessaire :

- de supprimer des champs jugés non pertinents par rapport aux besoins de la bibliométrie (numéros de contrôle, date d'entrée et de mise à jour dans la base originale,

3 Unified Medical Language System

4 Les tags sont des descripteurs génériques qui renseignent sur le type de population (mâle, femelle, animal, etc.) ou la tranche d'âge (enfant, adulte, etc.), le type de publication (étude de cas, etc.), l'aire géographique, etc.

CODEN[5], cote, etc.) ;

- d'éclater certains champs afin de ventiler leur contenu dans des champs plus précis, dans le but de les exploiter directement. Par exemple le champ Source de Medline a été éclaté en Titre de périodique, Volume, Fascicule, Pagination, Lieu de publication, Editeur, Date de publication ;
- de créer de nouveaux champs, absents de certaines bases. Par exemple, dans le NIOSHTIC, le Type de publication n'est (initialement) pas un champ, mais l'information apparaît parmi les descripteurs ;
- de normaliser les intitulés des champs retenus (Auteurs, Titre, Descripteurs, etc.).

Normalisation des contenus

Une fois les structures de notices uniformisées, il est nécessaire de normaliser les contenus des champs les plus importants pour l'exploitation bibliométrique du corpus. Cela concerne les auteurs, l'affiliation de l'auteur principal, le pays d'affiliation de l'auteur principal, le type de publication, la date de publication, la langue de publication, le titre de périodique et les descripteurs.

Là aussi, on peut se référer à différentes normes déjà définies. On peut citer l'[ISO 3166:1997] pour les codes de pays, l'[ISO 639:1988] pour les codes de langues, l'[ISO 8601:1998] pour les dates, etc. Nous avons choisi de détailler, ci-après, nos travaux sur les auteurs et les descripteurs, deux champs fondamentaux pour les analyses bibliométriques, plutôt que de décrire l'approche mise en place pour la normalisation de l'ensemble des champs.

La description des auteurs subit de nombreuses variations [Degez 1998]. Le nom précède généralement le prénom, mais parfois c'est l'inverse. Le prénom est complet ou abrégé, ou bien seul le premier prénom apparaît en entier et le second est abrégé. Les noms peuvent être composés et présenter des particules (de, von, van, etc.). A ces variations s'en ajoutent d'autres, comme celles liées à des séparateurs (tiret, apostrophe, virgule) ou à la casse. Par exemple, l'auteur le plus référencé dans notre corpus documentaire sur le stress professionnel, Cary L Cooper apparaît représenté sous 10 formes différentes[6] : "Cooper,-Cary-L."(54), "Cooper-CL" (42), "Cooper CL" (34), "C. L. Cooper" (8), "Cooper C.L." (7), "COOPER CL" (5), "Cooper, C. L." (3), "COOPER-C-L" (2), "Cooper-C-L" (2), "Cooper,-C.-L" (1).

Les règles de catalogage (AACR2, [AFNOR NF Z 44-001:1995]) préconisent le rejet des particules selon les pratiques du pays auquel appartient l'auteur. Ainsi, un auteur français, espagnol ou portugais verra sa particule "de" rejetée (ex. : Roux, Jean de) tandis qu'un américain ou un italien la conservera en tête (De Sicca, Giovanni). Les noms composés sans espace sont classés au premier élément lorsqu'il s'agit de français, d'allemands, d'espagnols (ex. : Garcia Lorca, Federico), au dernier élément lorsqu'il s'agit d'américains ou de portugais (ex. : Mill, John Stuart). Ces règles sont difficilement automatisables. Aussi avons-nous opté pour un format plus simple qui ne procède à aucun rejet : le nom complet de l'auteur apparaît suivi des initiales de ses prénoms, sans point ; les noms composés sont classés au premier élément et les éléments sont systématiquement séparés par un trait d'union. Il est alors possible de mettre en place - pour chaque base - des règles de transformation qui établissent la correspondance entre les auteurs tels qu'ils sont représentés dans le format de la base, et le format cible.

Considérer les différentes formes d'un même concept comme étant des concepts différents est lourd de conséquences en statistique. Sur l'exemple précédent, le fait de normaliser les différentes formes de l'auteur, en une forme commune, a pour conséquence de quasiment tripler

5 Le CODEN est un code alphanumérique d'identification des périodiques, qui tend à disparaître au profit de l'ISSN.

6 le nombre entre parenthèses représente la fréquence d'occurence de chacune des formes

le nombre de travaux relatifs à cet auteur, en passant de 54 (fréquence de sa forme la plus occurente) à 158 (la somme des occurences des différentes formes). De même, l'extraction des collèges invisibles (collaboration d'auteurs) est fortement biaisée (rendue principalement incomplète) par cette même hétérogénéité des dénominations. Il est donc clairement primordial de normaliser les différentes formes de représentation pour une même information. Voyons maintenant en détail comment gérer cette hétérogénéité lorsqu'elle concerne les descripteurs.

2. Les descripteurs

Les différentes bases interrogées représentent les documents de différentes manières. La liberté de chaque base en matière d'indexation engendre de nombreux problèmes pour la fusion de données. Les variations peuvent concerner :

1. le vocabulaire utilisé : pour la représentation d'un même concept, le vocabulaire employé varie d'une base à l'autre (exemple : *Antidepressive Agents* et *Antidepressants* représentent le concept d'*antidépresseurs*) ;
2. l'étendue de l'indexation : chaque source représente l'information qui l'intéresse et utilise du vocabulaire spécialisé dépendant du domaine d'intérêt (Psyclit représente les documents prioritairement selon leur aspect psychologique ; Pascal a une vocation plus générale) ;
3. la complexité de l'indexation : la majorité des bases utilise, en matière de description, de simples groupes nominaux. Cependant, certaines bases utilisent une indexation à facettes, c'est-à-dire qu'un descripteur ne représentera pas simplement un concept, mais il associera un certain point de vue à ce concept (exemple : *Stress*, *Psychological_Epidemiology* indiquera qu'il est question du *stress psychologique*, abordé ici d'un point de vue *épidémiologique*).
4. la répartition de l'information dans différents champs : la plupart des bases fournissent pour chaque document un champ ***mots-clés*** dans lequel le contenu du document est représenté à l'aide de descripteurs. Certaines bases utilisent toutefois plusieurs champs pour représenter des aspects complémentaires à la description du document (type de document, population d'expérimentation, nom de pays, titre de périodique, divers mots-outils, etc.). D'autres bases fusionnent ces différentes informations dans un seul champ (on peut également noter ici, la difficulté de segmentation de l'information dans différents champs tant les limites ne sont pas évidentes entre les différentes parties d'information).

Nous exposons ci-après nos solutions qui prennent en compte ces difficultés.

Les descripteurs initiaux

L'ensemble des descripteurs initiaux sur le stress comprend 49597 formes très diverses :

Exemples de descripteurs	Source	Commentaire
Human, Animal	Pascal	
HUMAN, FATIGUE-	Biosis,Psyclit	
CAS 7439-92-1	Cisilo	Données propres à Cisilo
*stress-, sleep-disorder-etiology, human- (888), male- (41)	Embase	Descripteurs à facette, mais pas de marqueur spécifique pour la facette (impossibilité de les détecter automatiquement) ; codes de classification en addition aux descripteurs
*Epinephrine_Pharmacology--PD, Corticosterone_Blood--BL, Stress, Psychological:CO	Medline	Descripteurs à facette avec facettes détectables automatiquement en toutes lettres et/ou en abrégé

Humans, Mental stress, NIOSH Contract, NIOSH Publication, WOSTEH, 137586, 59461	Nioshtic	Mots courants et données propres au Nioshtic : - descripteurs comprenant "NIOSH" ; - codes de journaux comme WOSTEH qui correspondant à la revue "Work and Stress" ; - codes de classification.

Par ailleurs, il est important de noter que :

- le seul dédoublonnage des notices (suppresion des notices apparaissant plusieurs fois alors qu'elles représentent un même document, cf 3.) a éliminé 3762 formes initiales (!?) ; les descripteurs différents ne sont donc plus qu'au nombre de 45835 ;
- l'indexation d'un même document, dans différentes bases, varie fortement [Nauer 1999]. 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 | Hemodynamics/physiology | Job Satisfaction | Workload|Human | Program Evaluation | Support, Non-U.S. Gov't | Occupational Diseases/blood/prevention & control/physiopathology | Prolactin/blood | Prospective Studies | Questionnaires |Social Support | Stress, Psychological/blood/prevention & control/physiopathology

Une étude sera menée ultérieurement pour tirer profit de ces différences d'indexation : pour traiter des équivalences entre descripteurs, pour consolider l'indexation des références pour lesquels il existe des doublons, etc.

Homogénéisation des descripteurs

Un concept peut se voir associer un grand nombre de représentations terminologiques. C'est ici le cas pour le *stress psychologique* pour lequel on peut trouver diverses formes typographiques, ainsi que différents qualifiants associés. On comprend les biais statistiques que ces variations peuvent induire.

362	Psychological stress
350	PSYCHOLOGICAL-STRESS
347	Stress, Psychological:CO
321	Stress, Psychological:
300	Stress, Psychological:ET
294	Stress, Psychological:PX
235	Stress, Psychological
36	Stress, Psychological/etiology
…	
1	Stress, Psychological_Genetics--GE

Dès lors, il convient de diminuer le nombre de variations, en appliquant les opérations suivantes :

1. la normalisation typographique : pour aboutir à des descripteurs en minuscules avec leur première lettre en majuscule ;
2. l'élimination des caractères spéciaux (*, -, ...), des codes de classement - (888), etc ;
3. la transformation de certaines expressions (& devient and ; adoption du même séparateur de facette pour toutes les bases) ;
4. la suppression des descripteurs inutiles (codes de classifications, données propres à certaines bases).

Traitement des descripteurs à facettes

Pour des facilités d'accès à l'information, nous avons utilisé --- dans les bases contenant des descripteurs à facettes (Medline et Embase) --- une sorte d'autopostage. Ainsi, pour chaque

descripteur à facette composé d'un descripteur principal et d'un ou plusieurs qualifiants, nous avons ajouté dans la notice le descripteur principal seul (i.e. sans qualifiant) ainsi que le(s) qualifiant(s) seul(s). Ainsi, il n'y a aucune perte d'information par rapport aux données initiales. L'utilisateur peut alors effectuer des recherches sur le descripteur seul, sur le descripteur avec qualifiant, ou sur le(s) qualifiant(s) seul(s) selon qu'il désire accéder à tels ou tels points de vue. Par exemple :

*Stress, Psychological_Metabolism--ME sera tout d'abord épuré en

Stress, Psychological_Metabolism, concept pour lequel on auto-postera

Stress, Psychological et Metabolism

Dans les données issues du Nioshtic, on trouve une indexation exprimant parfois des liens entre concepts, sans pour autant être une indexation à facettes. Dans ce cas, nous avons éclaté ces descripteurs composés comme s'il s'agissait de deux descripteurs distincts (les recherches associées pouvant se faire par le biais d'une requête booléenne composée d'un "ET" entre les 2 termes). Par exemple, "Group dynamics / Industrial health" donnera naissance à deux descripeurs : "Group dynamics", et "Industrial health".

Résultat intermédiaire

L'homogénéisation des descripteurs et le traitement des descripteurs à facettes permettent l'élimination de plus de 11000 formes, ramenant ainsi à 34559 le nombre de descripteurs différents. Ce nombre, toujours aussi important, de descripteurs différents est fortement lié à l'utilisation d'une description à facettes dans laquelle on combine des descripteurs principaux avec des qualifiants. Dans notre corpus, les descripteurs à facettes différents sont au nombre de 13501, résultants de la combinaison de 5311 descripteurs principaux avec 146 qualifiants différents. En réalité, il n'y a donc que 21058 formes différentes, puisque 13501 formes sont des formes composées (descripteur + facette).

Pour réduire davantage les variations de concepts dans le fonds fédéré, nous avons mis en place les deux approches supplémentaires suivantes :

1. une convergence des descripteurs par rapport à une forme "appauvrie", incluant une normalisation des pluriels/singuliers ;
2. une convergence via l'UMLS qui contient le vocabulaire biomédical (et pour chaque concept, un ensemble de variations terminologiques, qui vont permettre d'effectuer le matching).

Forme "appauvrie" et pluriels/singuliers

La notion de forme appauvrie permet de prendre en compte les normalisations typographiques (majuscules/minuscules, séparateurs de mots (espace, tiret, etc.). Elle correspond à la transformation de la forme initiale d'un descripteur en sa forme en minuscule, n'incluant que les caractères alphanumériques. De plus, nous avons mis en place une méthode simple d'identification des singuliers / pluriels dans le vocabulaire présent dans le fonds ; ce qui permet d'améliorer les résultats. Ainsi, les quatre descripteurs *Adrenal Gland*, *Adrenal Glands*, *Adrenal glands*, *Adrenal-Glands*, auront la même forme appauvrie *adrenalgland*, et pourront être regroupés sous une forme commune, choisie parmi l'ensemble des variantes.

Pour l'ensemble du corpus, cette étape permet de ramener à 31697 les formes différentes. Parmi elles, on dénombre 13169 formes correspondant à des descripteurs à facettes (résultats de combinaison de 5058 formes principales et 132 facettes). Les formes principales différentes ne sont en fait plus que 18528.

Le mise en correspondance via l'UMLS

Comme l'UMLS contient un méta-thésaurus du vocabulaire biomédical, il permet de normaliser le vocabulaire en mettant en relation les différentes dénominations d'un même concept. Par exemple, le concept de Corticotropin aura plusieurs variantes, telles : ACTH, Acth Hormone, Adrenocorticotrophic Hormone, Adrenocorticotropic Hormone, que l'on retrouve dans le vocabulaire d'indexation des différentes bases.

Cette homogénéisation permet de supprimer 1007 formes pour aboutir à 30543 formes finales, constituées de 4862 formes principales combinées avec 131 qualifiants, soit 13022 descripteurs à facette. Au total, il ne reste que 17521 formes principales et qualifiants différents.

Parmi ces formes, 9133 sont dans l'UMLS

Evaluation de notre approche

Les apports de notre méthode ne sont pas directement quantifiables. En effet, même si il est possible de fournir des indicateurs bruts (comme la proportion de réduction du vocabulaire, par exemple), le véritable objectif de nos travaux est d'améliorer qualitativement les analyses produites (par l'utilisation de données normalisées plutot que brutes). Aussi, nous proposons simplement, pour illustrer les apports de notre méthode, de donner des exemples d'amélioration sur les listes des principaux auteurs et descripteurs, avant et après normalisation :

	Avant normalisation		Après normalisation	
	Fréquence	Forme	Fréquence	Forme
Auteurs	54	Cooper,-Cary-L.	158	Cooper CL
	50	Kvetnansky R	65	Theorell T
	42	Cooper-CL	56	Kvetnansky R
	38	Levine S	56	Burke RJ
	37	Kopin IJ	48	Chrousos GP
	35	Conforti N	45	McEwen BS
	34	Feldman S	43	Levine S
	34	Cooper CL	39	Murphy LR
	30	McCarty R	39	Kopin IJ
	28	Theorell T	36	Conforti N
Descripteurs	2536	Human	14275	Human
	2405	OCCUPATIONAL-STRESS	9547	Stress
	2302	Rats	6731	Adult
	2223	human-	5346	Male
	2063	ADULTHOOD-	3747	Stress, Psychological
	1847	Human:	2820	Rats
	1691	Stress	2439	Physiology <1>
	1335	Adult	2425	Occupational Stress
	1332	HUMAN	2303	Blood <1>
	1248	*stress-	2299	Pharmacology <1>
	1190	adult-	2287	Causality <1>
	1162	article-	2261	Psychology <1>
	972	priority-journal	2231	physiopathology
	943	Job stress	2056	Journal Article
	930	male-	2035	Metabolism <1>

Conclusion

La normalisation du vocabulaire utilisé dans l'indexation est capitale pour les statistiques concernant l'apparition de descripteurs, la co-occurence de descripteurs et la classification. Il est important de traiter les descripteurs à facettes pour homogénéiser l'accès avec celui des bases qui n'utilisent pas ce type de descripteurs. Il est cependant conseillé de ne pas les éliminer, de façon

à bénéficier de leur expressivité plus forte et à obtenir des analyses (partielles) plus fines. Enfin, la prise en compte du thésaurus est également essentielle, car même si le thésaurus ne permet pas une réduction importante en nombre de l'ensemble des descripteurs, il a l'avantage de fournir une réduction qualitative non-négligeable. Ce résultat s'explique également par le fait que cette étape a lieu en fin de traitement.

Les solutions proposées visent un objectif double : 1) réduire les cas les plus occurents ; 2) mettre en place des solutions simples. C'est pourquoi nous traitons uniquement les variations typographiques et variations simples, et non pas tous types de variations linguistiques (flexionnelle, syntaxique et morpho-dérivationnelle [Polanco 95]) qui nécessitent la mise en place de solutions plus couteuses.

3. Le dédoublonnage des notices

Les doublons désignent toutes les notices, au sein d'une ou de plusieurs BDD, qui font référence à la même publication logique : même(s) auteur(s), titre et support de publication. L'élimination des doublons est une opération préalable à toute analyse bibliométrique d'un corpus. Dans notre application, la présence de références en double au sein de la base sur le stress faussait les résultats statistiques et ne permettait pas l'édition de listes bibliographiques rigoureuses. Cependant, il est également possible de tirer profit des doublons. Par exemple, on pourra reconstruire des notices plus riches en sélectionnant judicieusement les différentes parties dans les sources les plus appropriées, comme le préconise [Grivel 99]. De même, il est possible de construire une *super-notice* par concaténation des informations présentes dans les doublons.

Quelle que soit la solution informatique retenue, le dédoublonnage s'opère en trois étapes [Desrichard 1997] :

- construction d'une ou de plusieurs clés de dédoublonnage ;
- identification des doublons, par comparaison et mise en correspondance des clés de dédoublonnage ;
- élimination des doublons du corpus.

La sélection des champs qui vont servir à la construction de la clé de dédoublonnage s'avère cruciale. Il convient de sélectionner les champs qui sont présents dans toutes les notices. Il est également essentiel que ces champs possèdent un format homogène, qu'ils soient significatifs et qu'ils identifient de façon univoque les notices. C'est pourquoi la normalisation est une étape critique dans la construction de cette clé.

La clé de dédoublonnage nécessite l'extraction d'une nouvelle information, à partir du traitement et de la concaténation de plusieurs données, sous une forme normalisée. Nous nous sommes inspirés du code Meyer-Uhlenried [Laisiepen 1980]. Cette clé alphanumérique de 13 caractères comprend :

- les 4 premières lettres du nom de l'auteur ;
- les 2 premières initiales du prénom de l'auteur ;
- les 2 dernières lettres de l'année ;
- la première lettre des 5 premiers mots du titre.

Nous avons apporté quelques aménagements à ce code :

- l'indication de la première page de l'article a été ajoutée, ce qui évite tout risque de confusion entre deux articles de périodique (cas des articles en plusieurs parties, par exemple) ;
- lorsque le titre comporte moins de 5 mots, nous avons retenu les initiales des mots présents que nous avons complétées par ajout des lettres du dernier mot du titre afin d'obtenir un code de 5 caractères ;
- nous avons constitué deux clés de dédoublonnage, la première (Clé 1) fonctionnant d'après le titre original, la seconde (Clé 2) d'après le titre traduit, ceci pour multiplier les chances de retrouver des doublons (certaines bases telles Biosis ou NIOSHTIC ne proposant pas le titre dans sa langue originale)
- pour le codage du titre, seules les caractères alphanumériques ont été conservés. Pour

l'ensemble de la clé, tous les caractères non imprimables de même que les tirets, points, apostrophes, etc., ont été supprimés ; chaque élément du code est séparé du suivant par une étoile et toutes les lettres ont été converties en majuscules ;

- dans le cas des non-périodiques, nous avons retenu le nombre total de pages du document (à la place de la première page) qui est facilement identifiable grâce à la mention "p." (ex. : 230 p.) ; lorsqu'il s'agit d'une contribution à une monographie, c'est l'indication de la première page de la contribution qui est sélectionnée et non la pagination totale.

Clés de dédoublonnage d'une notice type

Le calcul de la clé sur le document d'Arnetz BB donné précédemment en exemple donne :

- Clé 1 : *ARNE*BB*1996*TAPPS*53*
- Clé 2 : *ARNE*BB*1996*TUEPP*53*

L'utilisation de ces clés de dédoublonnage a permis l'élimination de 5849 références (23,28 % des données initiales) pour arriver à un total de 20402 références.

Pour la bibliométrie, cette étape est également essentielle, car il est inconcevable d'établir des statistiques sur un corpus contenant de l'information redondante. Par exemple, les auteurs d'une publication présente dans les 8 bases ne doivent pas comptabiliser 8 publications, mais une seule.

4. Méthodologie

D'un point de vue méthodologique, nous présentons un algorithme simple basé sur un tri préférentiel. Cet algorithme est utilisable dans de nombreuses situations (il est utilisé ici pour la normalisation des auteurs, des descripteurs et pour le dédoublonnage) et permet d'intégrer simplement des ressources annexes (thésaurus, règles de lemmatisation, etc.).

Algorithme pour la génération d'équivalence

Le principe de l'algorithme est élémentaire et consiste à :

1. générer des triplets < clé de regroupement, poids de classement, forme initiale >
2. trier l'ensemble des triplets ;
3. effectuer les regroupement des formes ayant la même clé de regroupement.

Le calcul de la clé de regroupement correspond à une fonction de transformation de forme initiale à classer. L'affectation d'un poids permet de classer entre-elles les formes initiales ayant la même clé de regroupement.

Cet approche très générale permet en fait de traiter un grand nombre de problèmes. Voyons à partir de l'extrait de descripteurs ci-contre, comment seront traités les variations morphologiques, flexionnelles et lexicales des descripteurs présents.

*antidepressant-agent	Antidepressant agent
ANTIDEPRESSANT AGENT	Antidepressants
ANTIDEPRESSANT-DRUG	Antidepressive Agents
ANTIDEPRESSANT-DRUGS	antidepressant-agent

Les variations morphologiques dans les descripteurs

Dans le cas des variations de casse (MAJUSCULES/minuscules), de séparateur de termes (tiret "-", underscore "_", etc.), de caractères parasites (astérisque *, etc.), la fonction de transformation consiste à réduire le descripteur à un ensemble de lettres minuscules, ignorant tous les signes ne faisant pas partie de caractères alphanumériques ([A-Z], [a-z], [0-9]).

Soit a_i les i lettres composant la forme A à transformer, la fonction f est définie comme :

$$\begin{cases} f(a_i) = a_i \text{ si } a_i \subset [a - z, 0 - 9\ ]; \\ f(A) = a;\ f(B) = b;\ \ldots\ ;\ f(Z) = z; \\ f(a_i) = \varnothing \text{ sinon} \\ f(A) = f(a_1a_2a_3....a_n) = f(a_1)f(a_2)f(a_3)...f(a_n) \end{cases}$$

Pour choisir une forme privilégiée (qui correspondra à la forme normalisée) plusieurs stratégies sont possibles, et dépendent uniquement de choix pris pour le calcul du poids. Si on veut privilégier la forme la plus esthétique, on peut par exemple affecter un poids en fonction du nombre de caractères parasites, du nombre de séparateurs autre que l'espace, du nombre de MAJUSCULE, etc.

Par exemple :

$$\begin{cases} p(a_i) = 0 \text{ si } a_i \in [a - z, 0 - 9, \text{espace}] \\ p(ai) = 1 \text{ sinon} \\ p(A) = p(a_1a_2a_3....a_n) = p(a_1) + p(a_2) + p(a3) + \ldots + p(an). \end{cases}$$

L'application des fonctions de transformation f et de poids p, suivi d'un tri donne comme résultats :

	Clé de regroupement f(A)	poids p(A)	forme intiale A
Clés identiques	antidepressantagent	1	Antidepressant agent
	antidepressantagent	1	antidepressant-agent
	antidepressantagent	2	*antidepressant-agent
	antidepressantagent	19	ANTIDEPRESSANT AGENT
	antidepressantdrug	19	ANTIDEPRESSANT-DRUG
	antidepressantdrugs	20	ANTIDEPRESSANT-DRUGS
	antidepressants	1	Antidepressants
	antidepressiveagents	2	Antidepressive Agents

Les 4 premières lignes ayant la même clé de regroupement, on considère les formes initiales comme équivalentes et on leur attribue la forme préférée ayant le plus petit poids (i.e. la forme initiale de la première ligne du regroupement).

Les variations flexionnelles

Le traitement des variations flexionnelles est également possible en appliquant le même principe. Chaque variation flexionnelle peut en effet être décrite selon une règle. Pour le traitement des pluriels par exemple, un ensemble de règles sera à prendre en compte : le "s" terminal (ex : birth -> births, day -> days, etc.), "an -> en" (ex : man -> men, woman -> women, etc.), "y -> ies "(ex : baby -> babies, etc.), "fe -> ves" (ex : wife -> wives, life ->lives, etc.), etc.

La fonction de transformation aura ici pour but - en plus de la normalisation morphologique - de supprimer les marques de pluriel et de singulier. Ainsi, si la suppression des marques du singulier et du pluriel conduisent deux formes initiales différentes, à une même clé de regroupement, alors l'une de ces formes est le singulier et l'autre est le pluriel. Il convient alors d'attribuer un poids en fonction de la transformation effectuée. Par exemple : p(A)=1 si suppression d'une marque de pluriel, p(A)=2 si suppression d'une marque de singulier. Sur l'exemple précédent, on obtiendrait notamment :

	Clé de regroupement f(A)	poids p(A)	forme intiale A
	antidepressantagent	1	Antidepressant agent
	antidepressantagent	1	antidepressant-agent
	antidepressantagent	1	*antidepressant-agent
	antidepressantagent	1	ANTIDEPRESSANT AGENT
Clés identiques	antidepressantdrug	1	ANTIDEPRESSANT-DRUGS
	antidepressantdrug	2	ANTIDEPRESSANT-DRUG
	antidepressants	2	Antidepressants
	antidepressiveagents	2	Antidepressive Agents

Par regroupement des clés identiques, avec un poids différent, on obtient les équivalences singuliers / pluriels.

Dans le cadre de notre étude, 2529 équivalences ont ainsi pu être établies.

Les variations lexicales

Les variations lexicales peuvent être prises en compte par l'utilisation d'un thésaurus. Voici par exemple, les différentes dénominations associées au concept d'antidépresseurs dans l'UMLS.

Concept :	Antidépresseurs
Forme préférée :	Antidepressive Agents
Variante morphologique :	Agents, Antidepressive
Terme synonyme 1 :	
Forme préférée :	Antidepressant agents
Variante flexionnelle :	Antidepressant agent
Variante morphologique :	Agents, Antidepressant
Terme synonyme 2:	
Forme préférée :	Antidepressants
Variante flexionnelle :	antidepressant
Terme synonyme 3:	
Forme préférée :	Antidepressant Drugs
Variante flexionnelle :	Antidepressant drug
Variante morphologique :	Drugs, Antidepressant

Pour attribuer aux différentes descripteurs la forme préférée du thésaurus, il convient de construire pour chaque variante V d'un concept un triplet de la forme < f(V), p(V), forme préférée du concept > ; f correspondant à la fonction de transformation de la page précédente et p(V)=0 par exemple permet de distinguer le vocabulaire issu du thésaurus de celui provenant des BDD). On obtiendrait pour l'exemple ci- dessus, la liste de triplets suivante, signifiant que toutes les formes appauvries du corpus trouvée dans la première colonne correspondent au concept "*Antidepressive Agents*" :

antidepressiveagents	0	Antidepressive Agents
agentsantidepressive	0	Antidepressive Agents
antidepressantagents	0	Antidepressive Agents
antidepressantagent	0	Antidepressive Agents
agentsantidepressant	0	Antidepressive Agents
antidepressants	0	Antidepressive Agents
antidepressant	0	Antidepressive Agents
antidepressantdrugs	0	Antidepressive Agents
antidepressantdrug	0	Antidepressive Agents
drugsantidepressant	0	Antidepressive Agents

Ensuite, en effectuant un tri avec les formes appauvries du corpus, puis en regroupant les clés de regroupement identiques, il est possible de générer des tables de réindexation pour l'ensemble du vocabulaire du corpus :

Clé de regroupement f(A)	poids p(A)	forme intiale A
agentsantidepressant	0	Antidepressive Agents
agentsantidepressive	0	Antidepressive Agents
antidepressant	0	Antidepressive Agents
antidepressantagent	0	Antidepressive Agents
antidepressantagent	1	*antidepressant-agent
antidepressantagent	1	Antidepressant agent
antidepressantagent	1	antidepressant-agent
antidepressantagent	1	ANTIDEPRESSANT AGENT
antidepressantagents	0	Antidepressive Agents
antidepressantdrug	0	Antidepressive Agents
antidepressantdrug	1	ANTIDEPRESSANT-DRUG
antidepressantdrugs	0	Antidepressive Agents
antidepressantdrugs	1	ANTIDEPRESSANT-DRUGS
antidepressants	0	Antidepressive Agents
antidepressants	1	Antidepressants
antidepressiveagents	0	Antidepressive Agents
antidepressiveagents	1	Antidepressive Agents
drugsantidepressant	0	Antidepressive Agents

Ici toutes les formes du corpus (identifiables par un poids de 1) ayant "*antidepressantagent*" comme forme appauvrie, devront être transformées en la forme initiale ayant "antidepressantagent" comme forme appauvrie, mais dont le poids à 0 montre qu'il s'agit de la forme préférée du concept dans le thésaurus.

Le dédoublonnage

La détection de doublons suit le même principe. Il convient tout d'abord de calculer la clé ; les références bibliographiques ayant la même clé peuvent alors être regroupées. On peut ainsi facilement éliminer les doublons en ne gardant qu'un seul exemplaire de notice pour chaque clé. De plus, en affectant à chaque référence un poids différent selon sa provenance, on peut privilégier un ordre de préférence dépendant de la base d'origine. Ainsi, nous avons choisi de retenir prioritairement les notices provenant de Medline, puis celles d'Embase, de Biosis, de PsycLIT, de Pascal, de NIOSHTIC, de Cisilo, et enfin celle d'INRS-B.

Voici une illustration des clés triées :

	Clé de regroupement f(A)	poids p(A)	base source/numéro enregistrement
	*ARNE*BB*1996*MAAHL*1108*	2	Embase/001351
	*ARNE*BB*1996*NDMED*1108*	8	INRS-B/000015
Clés identiques	*ARNE*BB*1996*TAPPS*53*	1	Medline/001021
	*ARNE*BB*1996*TAPPS*53*	3	Biosis/000612
	*ARNE*BB*1996*TAPPS*53*	6	NIOSHTIC/000014
	*ARNE*BB*1996*TAPPS*53*	6	NIOSHTIC/000121
	*ARNE*BB*1996*TAPPS*53*	7	Cisilo/000072
	*ARNE*BB*1996*TAPPS*53*	8	INRS-B/000059
	*ARNE*BB*1996*TUEPP*53*	7	Cisilo/000072
	*ARNE*BB*1996*TUEPP*53*	8	INRS-B/000059
	*ARNE*BB*1997*MSAPS*63*	1	Medline/001221

Toutes les références ayant une même clé de regroupement sont des doublons. On peut constater que notre algorithme résoud aussi bien les doublons intra-bases (cas des notices 000014 et 000121 du NIOSHTIC) que les cas inter-bases. L'obtention du corpus final se fait par élimination des références de poids les plus forts parmi les références ayant une clé de regroupement similaire[7]. Dans notre exemple, nous éliminerons les notices :

- Biosis/000612, NIOSHTIC/000014, NIOSHTIC/000121, Cisilo/000072, INRS-B/000059, pour la clé de regroupement *ARNE*BB*1996*TAPPS*53*
- INRS-B/000059 pour la clé de regroupement *ARNE*BB*1996*TUEPP*53*

Conclusion

Le dédoublonnage des notices et la normalisation des données sont des étapes indispensables à la constitution d'un corpus bibliographique multi-bases. Pour la bibliométrie, ces étapes permettent de résoudre les biais statistiques liés aux données brutes.

S'il est relativement simple de traiter l'homogénéisation des auteurs, il n'en est pas de même pour des champs qui varient plus fortement comme les descripteurs ou encore les affiliations. Pour les descripteurs, les travaux que nous avons mené traitent partiellement le problème. On peut toutefois s'interroger sur la qualité globale de l'indexation d'un tel corpus, tant les indexations sont originellement différentes dans les bases interrogées.

Bibliographie

[AACR2:1988] American Library Association. Anglo-American Cataloguing Rules, 2nd edition, 1988.

[AFNOR NF Z 44-001:1995] Technologies de l'information - Classement alphabétique des dénominations. AFNOR, 1995.

[Desrichard 1997] Y. Desrichard. Le dédoublonnage des banques de données bibliographiques : un état de l'art. Documentaliste - Sciences de l'information, 34(2):82-89, 1997.

[Dou 1991] H. Dou, Marie-Christiane Dionne, Clément Paoli, Albert La Tela. Les formats bibliométriques, choix et limites. In Les systèmes d'information élaborées, 122-127, SFBA, Ile Rousse, juin 1997.

[Gehanno 1998] J.F. Gehanno, C. Paris, B Thirion et al. Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology. Occup Environ Med, 55(562-566), 1998.

7 Ce qui ne revient pas au même que de garder pour chaque clé de regroupement les références de poids les moins forts. Ceci, à cause des doubles clés issues des titres originaux et traduits.

[ICJME 1997] International Commitee of Medical Journals Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. JAMA, 277:927-934, 1997.

[ISBD 1988] ISBD(G). General International Standard Bibliographic Description. Annotated text, 1988.

[ISO 639:1988] International Organization for Standardization. ISO 639. Code for the representation of names of languages, 1988.

[ISO 690:1987] International Organization for Standardization. ISO 690. Information and Documentation - Bibliographic references - Content, form and structure, 1987.

[ISO 3166:1997] International Organization for Standardization. ISO 3166. (revision of 1988) Code for the representation of names of countries, 1997.

[ISO 8601:1988] International Organization for Standardization. ISO 8601. Data elements and interchange formats -- Information interchange -- Representation od dates and times, 1988.

[ISO 8879:1986] International Organization for Standardization. ISO 8879. Information processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML), 1986.

[Grivel 99] Luc Grivel, Hélène Fagherazzi, Philippe Fourneret, Alain Zerouki. La conception de bases de données infométriques hybrides : analyse de la pratique de trois observatoires européens. In Les systèmes d'information élaborées, SFBA, Ile Rousse, septembre 99.

[Jolibois 1999] Samuel Jolibois, Emmanuel Nauer, Dominique Chouanière, Jacques Ducloy, Françoise Grandjean, and Marc Mouzé-Amady. Adaptation des normes et formats documentaires à la gestion informatisée de corpus bibliographiques. Bulletin des Bibliothèques de France, 1999. A paraître.

[Laisiepen 1980] K. Laisiepen, E. Lutterbeck, and K.H. Meyer-Uhlenried. Grundlagen der praktischen. Information und Dokumentation. Saur, 1980.

[MARC 1999] Library of Congress (LC). MARC Standards, 1999. Available from http://lcweb.loc.gov/marc/marc.html.

[Nauer 1999] Emmanuel Nauer. *Les problèmes de variations terminologiques dans l'indexation de références bibliographiques*. In Journées Internationales de Linguistique Appliquée (JILA'99). LILLA - Université de Nice, juin 1999.

[Polanco 95] Xavier Polanco, Jean Royauté, Luc Grivel et Alain Courgey. Une approche linguistico-infométrique au service de la veille scientifique et technologique. In Les systèmes d'information élaborées, SFBA, Ile Rousse, juin 1995.

[TEI 1994] Electronic Text Center (ETC). TEI Guidelines for Electronic Text Encoding, 1994. Available from http ://etext.virginia.edu/TEI.html.