

UNIVERSITE DE DROIT, D'ECONOMIE ET DES SCIENCES AIX-MARSEILLE III
FACULTE DES SCIENCES ET TECHNIQUES DE SAINT JEROME
CENTRE DE RECHERCHE RETROSPECTIVE DE MARSEILLE

THESE

présentée le 13 octobre 1997 par

Pascal FAUCOMPRE

pour obtenir le grade de Docteur en Sciences

spécialité : **SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION**

**LA MISE EN CORRESPONDANCE AUTOMATIQUE DE BANQUES
DE DONNEES BIBLIOGRAPHIQUES SCIENTIFIQUES ET TECHNIQUES
A L'AIDE DE LA CLASSIFICATION INTERNATIONALE DES BREVETS**

CONTRIBUTION AU RAPPROCHEMENT DE LA SCIENCE ET DE LA TECHNOLOGIE

Membres du Jury

M. Richard BOUCHE	École Nationale Supérieure des Sciences de l'Information et des Bibliothèques, Professeur
M. Henri DOU	Université Aix-Marseille 3, Professeur
M. Daniel DUFRESNE	Université Aix-Marseille 2, UNIMECA, Professeur
M. Bernard MARX	Institut National de la Propriété Industrielle
M. Xavier POLANCO	Institut de l'Information Scientifique et Technique
M. Luc QUONIAM	Université Aix-Marseille 3, Maître de Conférence

SOMMAIRE

INTRODUCTION	7
1 CONTEXTE GÉNÉRAL, PROBLÉMATIQUE ET OBJECTIFS	11
1.1 LE CONTEXTE : PERFORMANCES TECHNOLOGIQUES ET COMPÉTITIVITÉ ÉCONOMIQUE	11
1.1.1 <i>L'innovation comme processus : l'appel du marché</i>	13
1.1.2 <i>Les PME/PMI et l'objet technique</i>	15
1.1.3 <i>Les voies d'accès à l'innovation et aux technologies</i>	17
1.1.4 <i>Les facteurs d'assimilation et d'incertitude</i>	20
1.1.5 <i>Réseaux de collaboration et lieux d'appropriation</i>	23
1.1.6 <i>La nécessaire lisibilité de l'information</i>	25
1.2 LE DOMMAGEABLE CLOISONNEMENT DES INFORMATIONS	27
1.2.1 <i>Veille technologique et organisation de l'entreprise</i>	27
1.2.2 <i>Rendre significantes des informations éparées</i>	30
1.2.3 <i>Détecter des liens et favoriser des rapprochements</i>	32
1.2.4 <i>Une relation plus explicite entre science et technologie</i>	33
1.2.5 <i>Un lien entre informations formalisées</i>	35
1.2.6 <i>Une correspondance entre banques de données</i>	36
1.3 CONCLUSION DE L'EXPOSÉ DU CONTEXTE ÉCONOMIQUE	37
2 LA MISE EN RELATION DES LANGAGES DOCUMENTAIRES	39
2.1 LE CONTEXTE : UN ENVIRONNEMENT NATUREL MONOBASE	39
2.1.1 <i>De l'utilité des langages documentaires</i>	39
2.1.2 <i>Les langages classificatoires</i>	45
2.1.3 <i>Les langages à structure combinatoire</i>	50
2.1.4 <i>Quel type de langage utiliser ?</i>	54
2.1.5 <i>Conséquences du choix des langages pour l'utilisateur</i>	57
2.2 DE LA NÉCESSITÉ DE RECHERCHES MULTIBASES	60
2.2.1 <i>Une typologie des niveaux de mise en correspondance</i>	64
2.2.2 <i>Le niveau des traitements</i>	64
2.2.3 <i>Le niveau des requêtes</i>	66
2.2.4 <i>Le niveau des données</i>	70
2.2.5 <i>Conséquences du niveau de mise en relation</i>	71
2.3 UNE TYPOLOGIE DES MÉTHODES DE MISE EN RELATION	72
2.3.1 <i>Un langage unique</i>	74
2.3.2 <i>Les systèmes intermédiaires ou systèmes de correspondance</i>	75
2.3.3 <i>Intégration et fusion des langages</i>	83
2.3.4 <i>Conséquences des méthodes de mise en relation</i>	88
2.4 CONCLUSION DE L'EXPOSÉ DU CONTEXTE DOCUMENTAIRE	91

3	CHOIX PRÉALABLES À UNE MISE EN CORRESPONDANCE	99
3.1	LE CONTEXTE : LES TYPES DE GISEMENT À METTRE EN RELATION	99
3.1.1	<i>L'information scientifique : la base multidisciplinaire PASCAL</i>	99
3.1.2	<i>L'information technologique : la base spécialisée du CETIM</i>	108
3.1.3	<i>L'information d'entreprise : les nomenclatures NAF et CPF</i>	110
3.1.4	<i>L'information technique : les brevets d'invention</i>	112
3.2	UNE CLASSIFICATION TECHNIQUE COMME LANGAGE COMMUN	124
3.2.1	<i>La Classification Internationale des Brevets</i>	124
3.2.2	<i>La structure hiérarchique de la CIB</i>	129
3.2.3	<i>Le fonctionnement de la CIB : classer et rechercher</i>	134
3.2.4	<i>Les difficultés liées à la CIB : appliquer et utiliser</i>	140
3.3	LES CATCHWORDS COMME SYSTÈME PIVOT	145
3.3.1	<i>Le rôle d'orientation des catchwords</i>	145
3.3.2	<i>La structure hiérarchique des catchwords</i>	147
3.3.3	<i>La CIB et les catchwords français, anglais et allemands</i>	152
3.3.4	<i>Une représentation comparée de la CIB à travers ses index</i>	163
3.4	CONCLUSION : LE CHOIX D'UN LANGAGE INTERMÉDIAIRE	169
4	MISES EN CORRESPONDANCE EXPÉRIMENTALES	171
4.1	LE CONTEXTE : UN MODÈLE DE MISE EN CORRESPONDANCE	171
4.1.1	<i>Le principe d'un système intermédiaire</i>	171
4.1.2	<i>La sélection automatique des liens</i>	173
4.1.3	<i>Le choix d'une technique de pivot relationnel</i>	176
4.2	LA MISE EN CORRESPONDANCE DES VOCABULAIRES	179
4.2.1	<i>Les classifications NAF et CPF</i>	179
4.2.2	<i>Le thésaurus du CETIM</i>	181
4.2.3	<i>Les vocabulaires PASCAL</i>	182
4.3	LA RÉINDEXATION À L'AIDE DES CATCHWORDS	189
4.3.1	<i>Les références bibliographiques du CETIM</i>	189
4.3.2	<i>Les notices de la section Métallurgie PASCAL</i>	192
4.3.3	<i>Conséquences de la réindexation</i>	195
4.4	LA RÉINJECTION DIRECTE DES CODES CIB : LA BASE PACA-INIST	199
4.4.1	<i>Les résultats de la réindexation</i>	200
4.4.2	<i>Bruit et pertinence documentaire</i>	203
4.4.3	<i>Conséquence de la réinjection des codes</i>	208
4.5	CONCLUSION DE LA MISE CORRESPONDANCE	209
4.5.1	<i>La première solution : le traçage du lien</i>	209
4.5.2	<i>La seconde solution : sans traçage du lien</i>	213
4.5.3	<i>De l'ergonomie au conceptuel</i>	213

5	ÉVALUATION, PROBLÈMES ET PERSPECTIVES DU SYSTÈME RELATIONNEL	217
5.1	LE CONTEXTE : UNE VALIDATION MULTIEXPERTISES	217
5.1.1	<i>Une évaluation documentaire est-elle suffisante ?</i>	218
5.1.2	<i>Présomption de liens et indicateurs documentaires</i>	219
5.1.3	<i>Une place centrale pour l'utilisateur final</i>	223
5.2	UNE PREMIÈRE ÉVALUATION TECHNIQUE	225
5.2.1	<i>Types et méthodes d'évaluation</i>	225
5.2.2	<i>Une micro-évaluation : comparer indexation et classement</i>	226
5.2.3	<i>Une macro-évaluation : identifier des thématiques globales</i>	230
5.2.4	<i>Conclusion de l'évaluation : le veilleur comme cible</i>	231
5.3	DE L'HORIZONTAL AU VERTICAL : UN EXEMPLE POUR LE VEILLEUR	232
5.3.1	<i>Considérer l'objet en soi ou ses utilisations</i>	234
5.3.2	<i>Rechercher l'implicite, protéger l'explicite</i>	236
5.3.3	<i>Indexer les parties, classer le tout</i>	237
5.3.4	<i>D'un lien logique et structurel</i>	239
5.3.5	<i>A une relation science/technologie ouverte</i>	241
5.4	SOLUTIONS, DIFFICULTÉS ET PERSPECTIVES DU SYSTÈME PIVOT	247
5.4.1	<i>Les apports de la seconde solution</i>	247
5.4.2	<i>Les problèmes à résoudre</i>	249
5.4.3	<i>Les perspectives du système de mise en correspondance</i>	253
5.5	CONCLUSION : VERS UNE ÉVOLUTION DES CATCHWORDS DE LA CIB ?	256
	CONCLUSION GÉNÉRALE	259
	RÉFÉRENCES BIBLIOGRAPHIQUES	263

TABLE DES FIGURES

<i>Figure 1 De la science au client : un simple flux linéaire et irréversible</i>	14
<i>Figure 2 Les flux d'information et de coopération dans le modèle de la chaîne liée</i>	14
<i>Figure 3 Impact des innovations techniques sur le chiffre d'affaires des entreprises innovantes</i>	17
<i>Figure 4 Modes d'accès aux technologies : l'opportunité d'une double circulation ?</i>	18
<i>Figure 5 Les multiples rétroactions de la recherche scientifique</i>	20
<i>Figure 6 Aider un dialogue entre des logiques et des intérêts différents</i>	25
<i>Figure 7 L'environnement comme espace d'informations prometteuses ou menaçantes</i>	28
<i>Figure 8 Une mission de la veille : redonner une unité à des informations éparses</i>	31
<i>Figure 9 Lier des informations documentaires pour relier des domaines hétérogènes</i>	34
<i>Figure 10 En amont de la veille, mettre en correspondance des informations formelles</i>	35
<i>Figure 11 Un détour par la représentation documentaire des banques de données</i>	38
<i>Figure 12 L'indexation comme point de rencontre entre analyse et réponse documentaire</i>	40
<i>Figure 13 Le double codage de la médiation documentaire</i>	43
<i>Figure 14 Un apprentissage préalable ou a posteriori selon la coordination et le contrôle des outils</i>	58
<i>Figure 15 Stabilité introduite par les outils contrôlés (classification ou indexation)</i>	59
<i>Figure 16 Évolution du nombre de banques de données dans le monde</i>	61
<i>Figure 17 Degré d'engagement de l'utilisateur et niveau de correspondance</i>	71
<i>Figure 18 Les niveaux de correspondance et d'investissement des acteurs</i>	72
<i>Figure 19 La technique de mise en concordance dépend des investissements consentis</i>	89
<i>Figure 20 La CIB, passage obligé entre références scientifiques et références techniques ?</i>	92
<i>Figure 21 Quelques méthodes pour lier les outils documentaires issus de sources multiples</i>	93
<i>Figure 22 La relation science/technologie repose d'abord sur les compétences de l'utilisateur</i>	93
<i>Figure 23 Élargir l'espace des recherches par les banques de données elles-mêmes</i>	94
<i>Figure 24 Diversification ou unification de l'accès à l'information : deux évolutions opposées</i>	95
<i>Figure 25 Quelle passerelle utiliser entre des langages de type différent ?</i>	96
<i>Figure 26 La nature des outils et des ratios dépend des objectifs de la recherche</i>	96
<i>Figure 27 De multiples états quantifient l'intensité des correspondances</i>	97
<i>Figure 28 Principaux domaines couverts par la base PASCAL (1973-1995)</i>	100
<i>Figure 29 Un exemple de notice de la base PASCAL</i>	101
<i>Figure 30 Les 3 types de vocabulaire de la base PASCAL</i>	106
<i>Figure 31 Les 6 systèmes de correspondance produits-activités</i>	111
<i>Figure 32 Un exemple de notice de brevet dans les Chemical Abstracts</i>	118
<i>Figure 33 JOPAL : évolution du nombre de références (1981-1995)</i>	120
<i>Figure 34 JOPAL : classement CIB des références bibliographiques</i>	121
<i>Figure 35 Structuration des catchwords français en format d'enregistrement de base de données</i>	148
<i>Figure 36 Exemple de catchword français structuré sur plusieurs niveaux hiérarchiques</i>	149
<i>Figure 37 Catchwords français : traduction des niveaux hiérarchiques en champs de données</i>	150
<i>Figure 38 Le chrome : un exemple de sélection opérée par les catchwords</i>	154
<i>Figure 39 Catchwords anglais : structuration en champs des niveaux hiérarchiques</i>	155
<i>Figure 40 Les catchwords diversifient l'accès à un même objet technique</i>	156
<i>Figure 41 Catchwords allemands : les 3 niveaux hiérarchiques</i>	159
<i>Figure 42 Catchwords allemands : permutation des entrées principales et secondaires</i>	160
<i>Figure 43 Les premières définitions associées au stichword asbestos (amiante)</i>	162
<i>Figure 44 Un outil proche de la langue naturelle comme pont entre classification et indexation</i>	170
<i>Figure 45 Le principe d'économie d'un système pivot</i>	172
<i>Figure 46 Les étapes de la mise en correspondance automatique des vocabulaires</i>	175
<i>Figure 47 Les 3 techniques du pivot relationnel</i>	176
<i>Figure 48 Le maïs : exemple d'un lien code d'activité - codes CIB</i>	179
<i>Figure 49 D'un concept large aux activités spécialisées</i>	181
<i>Figure 50 Homogénéiser les structures et les écritures</i>	182
<i>Figure 51 Correspondance d'un concept technique et de deux procédés industriels</i>	183
<i>Figure 52 Une mise en perspective industrielle d'un principe de thermoélectricité</i>	183
<i>Figure 53 Quelques fonctions et applications techniques associées à un mot clé PASCAL</i>	184
<i>Figure 54 Un lien multicatchwords à un descripteur PASCAL</i>	185
<i>Figure 55 Les 3 causes de rejet de liens CIB/PASCAL</i>	188
<i>Figure 56 Méthode de sélection d'un code principal</i>	190
<i>Figure 57 Références du CETIM liées aux catchwords de la CIB</i>	192
<i>Figure 58 Notice PASCAL-Métallurgie liée aux catchwords CIB</i>	193
<i>Figure 59 Croiser les objectifs des acteurs à travers l'information</i>	195

Figure 60	La CIB comme passerelle entre deux banques de données non techniques	197
Figure 61	Un lien entre plusieurs banques de données de types différents	198
Figure 62	La réindexation de la base de l'Observatoire Provence-Alpes-Côte-d'Azur	200
Figure 63	Un lien entre information scientifique et information technique	202
Figure 64	Le passage d'une référence fondamentale à une référence de brevet	203
Figure 65	Référence théorique non associée à l'information technique	204
Figure 66	La notion d'exploration comme concept technique multivoque	204
Figure 67	Le processus de réindexation globale d'une référence bibliographique	215
Figure 68	Comment concilier un ensemble d'expertises individuelles ?	219
Figure 69	Les 3 étapes de la validation : l'utilisateur final doit valider les évaluations techniques	223
Figure 70	Un cycle itératif de validations descendantes puis ascendantes	224
Figure 71	Les trois dimensions complémentaires du projet de mise en correspondance	225
Figure 72	Requête documentaire et veille : une même information pour deux attentes différentes ?	231
Figure 73	Le passage fonction/application par un langage d'indexation	236
Figure 74	Quelques principes généraux de l'indexation par mots clés et de l'attribution de codes CIB	239
Figure 75	Les chemins entre fonctions et applications	240
Figure 76	Une nouvelle complémentarité entre une indexation et une classification	241
Figure 77	Un double passage fonction / application et science / technique	242
Figure 78	Les déposants de brevets divisés entre fonction et application	244
Figure 79	Une relation indirecte élargit un lien documentaire direct	245
Figure 80	Hydrocarbures et micro-organismes (Marseille 1993-1995)	246
Figure 81	Une retranscription automatique complémentaire du discours de l'expert	247
Figure 82	La règle des 80/20, descriptive des traitements de l'information	250
Figure 83	Toute mise à jour des outils implique une nouvelle mise en correspondance	252

LISTE DES TABLEAUX

Tableau 1	Influence de la méthode d'indexation sur le rappel et la précision (selon Johnston)	56
Tableau 2	Normalisation des 5 degrés d'équivalence entre 2 langages	80
Tableau 3	Classement de la proximité des bases dans le VSS	82
Tableau 4	Types de couverture de PASCAL et des catchwords	104
Tableau 5	Les vocabulaires PASCAL sélectionnés pour la mise en correspondance	106
Tableau 6	PASCAL : profil des descripteurs contrôlés français et anglais	107
Tableau 7	PASCAL : profil des candidats descripteurs français et anglais	107
Tableau 8	PASCAL : profil des mots clés libres français et anglais	107
Tableau 9	PASCAL : le profil commun des descripteurs	108
Tableau 10	Thésaurus CETIM : profil des entrées	109
Tableau 11	Relation entre Classification des produits et Nomenclature des activités (France)	110
Tableau 12	Liste générale des activités et des emboîtements CITI-NACE-NAF	111
Tableau 13	Banques de données produites ou coproduites par l'INPI (serveur Questel-Orbit)	113
Tableau 14	Nombre de banques de données par domaine (Questel-Orbit)	114
Tableau 15	JOPAL : répartition des sous-classes les plus fréquentes	121
Tableau 16	JOPAL : répartition sectorielle des codes CIB	122
Tableau 17	JOPAL : distribution des codes CIB par section	122
Tableau 18	Nationalité des brevets (million de documents)	127
Tableau 19	Nombre de codes d'indexation dans la CIB (4 ^e , 5 ^e et 6 ^e éd.)	137
Tableau 20	Catchwords français : fréquence des niveaux hiérarchiques	150
Tableau 21	Catchwords français : profils des 4 niveaux hiérarchiques	150
Tableau 22	Catchwords français : nombre moyen de mots par niveau hiérarchique	151
Tableau 23	Catchwords français : profil des définitions	152
Tableau 24	CIB : distribution des codes	152
Tableau 25	Catchwords français - distribution des codes	153
Tableau 26	Catchwords anglais : nombre d'entrées par niveau hiérarchique	156
Tableau 27	Catchwords anglais : profil des entrées et distribution des définitions complètes	157
Tableau 28	Catchwords allemands : profils des définitions	161
Tableau 29	Catchwords allemands : profils des entrées et des définitions complètes	161
Tableau 30	Fréquence de quelques thèmes dans les catchwords français, anglais et allemands	162
Tableau 31	Les codes CIB pointés par les catchwords	163
Tableau 32	Répartition des codes par niveau hiérarchique	164
Tableau 33	Proportion de codes par rapport à la CIB	165

<i>Tableau 34 Taux de subdivision logique de la CIB</i>	<i>166</i>
<i>Tableau 35 Les taux de subdivision à travers les catchwords</i>	<i>167</i>
<i>Tableau 36 Le rapport des taux de subdivision entre catchwords et CIB</i>	<i>168</i>
<i>Tableau 37 Nombre de catchwords par section</i>	<i>169</i>
<i>Tableau 38 Définition des contraintes en fonction du type de relations</i>	<i>173</i>
<i>Tableau 39 Sélection des correspondances en fonction de contraintes a priori</i>	<i>173</i>
<i>Tableau 40 Nombre de correspondances entre catchwords et section 001 PASCAL</i>	<i>186</i>
<i>Tableau 41 Nombre de liens entre les vocabulaires PASCAL et les catchwords</i>	<i>187</i>
<i>Tableau 42 Codes dédoublonnés liés aux descripteurs de la base PASCAL</i>	<i>187</i>
<i>Tableau 43 PASCAL-métallurgie, CETIM & CPF : nombre de liens avec les catchwords</i>	<i>195</i>
<i>Tableau 44 PASCAL-métallurgie, CETIM & CPF : nombre moyen de catchwords liés</i>	<i>196</i>
<i>Tableau 45 Profil des entrées des 4 vocabulaires liés aux catchwords</i>	<i>196</i>
<i>Tableau 46 Antidictionnaires descripteurs PASCAL/catchwords</i>	<i>205</i>
<i>Tableau 47 Taux d'erreur en fonction des antidictionnaires PASCAL</i>	<i>205</i>
<i>Tableau 48 Constitution des corpus d'évaluation de l'INIST</i>	<i>226</i>
<i>Tableau 49 Identification des thématiques présentes dans les corpus PASCAL-PACA et EPAT</i>	<i>230</i>
<i>Tableau 50 Catchwords français : nombre d'expressions désignant la fonction ou l'application</i>	<i>240</i>