Sélection de la bibliographie des maladies rares par la technique du Vocabulaire Commun Minimum

Michel Kerbaol, Jean-Yves Bansard

Département de Santé Publique INSERM Université de Rennes1

Abstract

Nowadays, searchers seeking for relevant data have to face more difficulties due to the dramatical inrease of the number of scientific publications, which moreover are indexed in a variable way according to data bases. An alternative to document selection by key words is proposed. This suggests to create a minimum common vocabulary on the bases of a sample of documents considered reliable. This vocabulary will be used for document research

Résumé

L'explosion du nombre des publications scientifiques rend la recherche des données pertinentes pour un chercheur d'autant plus difficile que ces publications sont indexées de manière variable selon les bases de données. Nous présentons une alternative à la sélection des documents par les mots clés, en construisant à partir d'un échantillon de documents, jugés pertinents, un vocabulaire commun minimum qui servira de sélecteur pour la recherche des documents.

Mots-clés : analyse factorielle des correspondances, bibliographie, dimension, vocabulaire commun minimum

1. Introduction

L'une des tâches les plus importantes, qui se présente aux chercheurs lors du démarrage d'une nouvelle recherche, est de réaliser un état des lieux. C'est ce que l'on désigne sous le nom de recherche bibliographique, celle-ci doit apporter une information de qualité et la plus exhaustive possible. Ces recherches reposent le plus souvent sur l'interrogation de multiples bases de données qui référencent les publications des chercheurs. Il n'existe pas de technique universelle pour l'interrogation de ces bases. Dans le détail, leurs structures diffèrent : certaines sont indexées par des thésaurus, d'autres par des procédures statistiques. Certaines possèdent un résumé d'auteur, d'autres un résumé construit par les indexeurs et au pire on n'accède qu'au titre, à la liste des auteurs et au nom de la publication.

De ce fait la recherche des documents pertinents pour une recherche est dépendante de facteurs non contrôlables par l'utilisateur final ; en particulier la politique scientifique des fournisseurs. Ainsi on a montré (Kerbaol 1996) que EMBASE et MEDLINE, deux des grandes bases mondiales en biomédecine avaient des lectures différentes d'un même document. Dans le cadre d'une étude sur le cancer du sein, l'analyse des mots clés a montré que l'une s'intéressait surtout à l'aspect clinique et l'autre à la chimiothérapie. De ce fait, un même article référencé dans les deux bases n'est pas accessible par la même interrogation.

Pour pallier ces problèmes, on a pris le parti de sélectionner les documents à partir d'analyses statistiques des mots des titres et des résumés. Le titre est important car il représente normalement la quintessence du travail décrit et le résumé de l'auteur contient les éléments dominants des résultats acquis.

Ce parti pris ne permettra pas la récupération de tous les articles concernant un sujet donné : en effet les auteurs s'expriment avec leurs mots, et *a priori*, ils ont à leur disposition un vocabulaire tel qu'ils pourraient écrire leur résumé sans jamais répéter un seul d'entre eux, ce qui rendrait leur résumé inaccessible à la statistique. Heureusement que dans la pratique, on constate que la plupart des auteurs usent de la répétition des mots pour asseoir leur discours permettant ainsi le recour aux statistiques.

2. Hypothèse

Un sujet de recherche, par exemple une maladie rare, peut être abordé de multiples façons. Des chercheurs vont l'étudier par une approche clinique, d'autre développeront une approche génétique, certains privilégieront une approche épidémiologique ou statistique ; bref les voies d'approches sont diverses et variées. Cette dispersion des approches ne neutralise pas pour autant le sujet dont il faut bien parler. Il doit donc exister un vocabulaire commun minimum à tous ces spécialistes, et c' est précisément ce vocabulaire commun que l'on cherche à cerner pour construire un sélecteur – i.e. un ensemble de mots qui permettra le repérage d'un maximum de documents utiles à la recherche bibliographique en cours.

3. Stratégie

La stratégie que nous proposons repose sur l' existence *a priori* d'un échantillon de textes en liaison directe avec le sujet de préoccupation du chercheur. Cet échantillon sera le noyau autour duquel reposent les différents calculs qui seront effectués. On procède à une analyse factorielle des correspondances du tableau croisant les mots et les documents du noyau. Cette analyse va nous permettre de repérer un ensemble de mots susceptible de contenir le vocabulaire commun minimum (VCM) recherché en se basant sur la notion de dimension d'un mot.

La figure 1 illustre de manière ensembliste notre proposition. Ainsi, parlant de la mucoviscidose, le chirurgien aura ses propres termes, le généticien n'a aucune raison d'abandonner son jargon, mais tous les deux, parlant, travaillant sur la même maladie, auront un vocabulaire commun minimum. Cette configuration de mots se représente facilement en analyse factorielle des correspondances (figure 2), chaque spécialité sera explicative d'un facteur, le vocabulaire commun minimum sera explicatif sur les deux facteurs. Nous appellerons dimension d'un mot, le nombre de facteurs où il est explicatif.

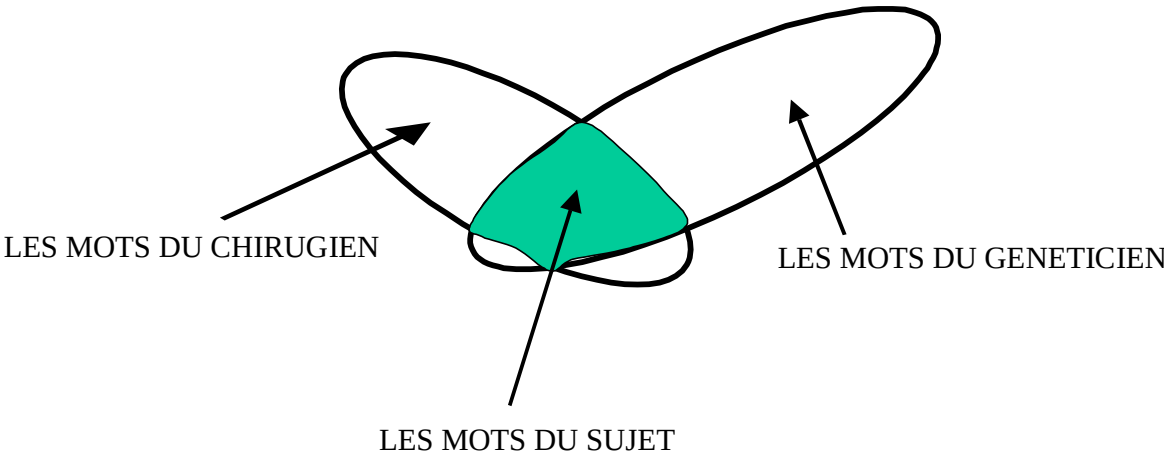
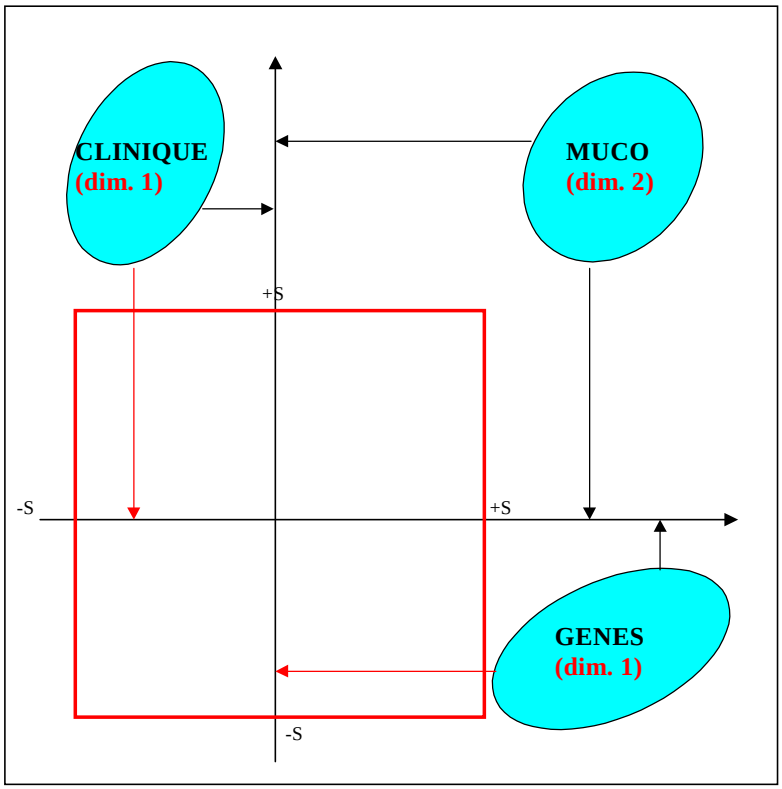


Figure 1 : *Représentation ensembliste de la notion de dimension d'un mot*



LA NOTION DE DIMENSION

Figure 2 : *Représentation factorielle et notion de la dimension d'un mot*

4. Rappel sur les mots explicatifs

Si tous les mots (i) analysés avaient le même pouvoir explicatif, chacun d'eux aurait une contribution (ctr(i)) à la valeur propre (λ_i) d'un facteur (i) égale à cette valeur propre divisée par le nombre N de mots analysés. (ctr(i) $= \lambda_i / N$)

Nous dirons donc qu'un mot est explicatif si sa contribution à la valeur propre est supérieure à un seuil S qui est souvent fixé de manière arbitraire à 1 ou 2 fois la contribution moyenne. On induit ainsi une hiérarchie sur la capacité explicative des mots.

5. La dimension des mots

Cette notion est relative à une analyse donnée. Elle est totalement dépendante du choix de l'unité textuelle choisie : plus cette unité textuelle est grande, plus la dimension du mot sera élevée. Ceci est dû au fait que plus l'unité textuelle est grande, plus le mot a de chance pour s'associer à d'autres.

Nous avons divisé l'étendue des dimensions des mots en 3 classes arbitraires :
1) l'une est censée repérer les mots associés à des thèmes : la dimension est inférieure à la moyenne moins l'écart type (quand la répartition des dimensions est symétrique), sinon elle est inférieure au premier quartile
2) la seconde est supérieure à la dimension moyenne plus un écart-type, ou au troisième quartile et contient les mots en relation avec le sujet
3) la classe intermédiaire reste un sujet de préoccupation…

6. Création d'un Vocabulaire Commun Minimum (VCM)

MATERIEL UTILISE POUR LA PREMIERE RECHERCHE :

Une base noyau de 612 documents écrits par des chercheurs connus pour leur participation à des recherches sur une maladie rare, la mucoviscidose, nous a été fournie par les ingénieurs en information de l' INSERM. Une première analyse factorielle de correspondances de cette base nous permet d'extraire 81 mots ayant une dimension supérieure à 10 (tableau 1).

ACID	**ADENOSINE**	**ADENOVIRUS**	**ADHESION**
AERUGINOSA	**ALPHA**	**ALVEOLAR**	**AMILORIDE**
ANTIGENS	**ANTITRYPSIN**	**ASPERGILLOSIS**	**ATP**
AUREUS	**BRONCHIAL**	**CAMP**	**CASES**
CELL	**CELLS**	**CFTR**	**CHANNEL**
CHANNELS	**CHILDREN**	**CHROMOSOME**	**CIRRHOSIS**
CONCENTRATIONS	**CYSTIC**	**DIAGNOSIS**	**DOUBLE**
DRUG	**ELASTASE**	**ELASTIN**	**EMPHYSEMA**
EPITHELIUM	**EXPRESSION**	**FETAL**	**FIBROSIS**
FLOW	**FLUID**	**HLA**	**INHIBITOR**
LEFT	**LIVER**	**LUNG**	**MARKERS**
MUCIN	**MUCINS**	**MUCUS**	**MUTATIONS**
NASAL	**NEONATAL**	**NEUTROPHILS**	**PATCHES**
PEPTIDE	**PERFORMANCE**	**PLASMA**	**PNEUMONIA**
PRENATAL	**PROPERTIES**	**PROTEASE**	**PROTEIN**
PROTEINASE	**PSEUDOMONAS**	**RAT**	**RATS**
RECEPTOR	**RECEPTORS**	**REJECTION**	**RIGHT**
SCREENING	**SECRETION**	**SECRETIONS**	**SPUTUM**
STRAINS	**THERAPY**	**TRANSFER**	**TRANSPLANTATIO**
TRANSPORT	**TRYPSIN**	**VENTRICULAR**	**VIVO**
WATER			

Tableau 1 : *Les 81 mots de dimension 10 et plus*

Selon notre hypothèse, le VCM recherché est inclus dans l' ensemble des mots de dimension élevée. Nous dirons qu'un e dimension élevée est une condition nécessaire pour qu'un mot puisse appartenir à un VCM. Cependant il ne s' agit pas d'un e condition suffisante car les mots candidats au statut de VCM peuvent se séparer en au moins deux types : les mots communs au sujet et ceux communs à la discipline (ex : la biologie, la médecine..). Pour séparer les mots du VCM, des mots communs à la discipline, il faut introduire dans notre stratégie des bases de textes provenant d' autres thèmes de la recherche biomédicale. On a ainsi sélectionné 4 bases témoins se rapportant à la leucémie aiguë lymphoblastique, aux polyamines, au cancer du sein et au tumor infiltrating leucocyte. Le tableau 2 indique pour chacune des bases, le nombre de documents ayant au moins un des 81 mots. Il montre que la seule sélection par la dimension n' est qu'un e condition nécessaire.

Type de sujet		Taux de récupération	Total de la base
Base 0 MUCO	mucoviscidose	612 (100%)	612
Base 1 LAL	leucémie aiguë lymphoblastiqe	1990 (96.5%)	2063
Base 2 POLY	polyamines	11912 (93.6%)	12726
Base 3 KC	cancer du sein	8728 (88.4%)	9871
Base 4 TIL	Tumor Infiltrating leucocyte	546 (93.2%)	586

Tableau 2 : *Taux de récupération des 5 bases avec les mots candidats*

7. Analyse des bases

On a construit un tableau mettant en correspondance les 5 bases et les 81 mots, ce tableau est soumis à une AFC

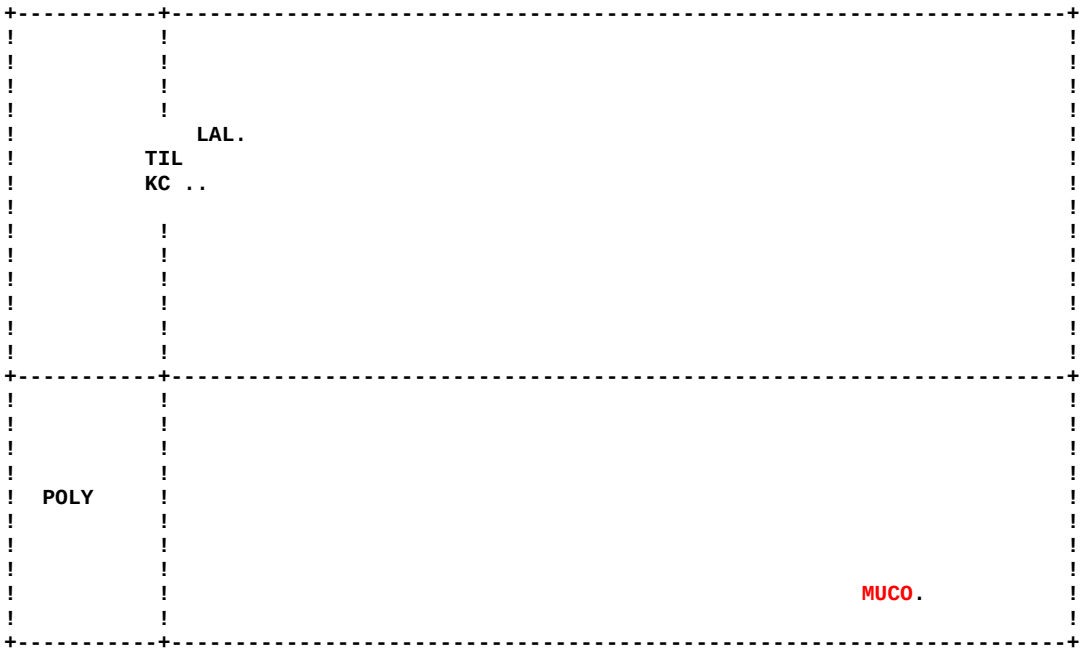


Figure 3 : *Analyse des 5 bases*

Sur ce graphique, la base MUCO se sépare des 4 autres bases. Dans le quadrant inférieur droit du graphe, 20 mots sont particulièrement attirés par cette zone. La classification ascendante hiérarchique (CAH) réalisée sur les facteurs de l' analyse confirme cette séparation des 81 mots en deux sous-ensembles.

ADENOVIRUS	AERUGINOSA	AMILORIDE	ANTITRYPSIN
AUREUS	BRONCHIAL	CFTR	CIRRHOSIS
CYSTIC	ELASTASE	ELASTIN	EMPHYSEMA
FIBROSIS	MUCUS	NASAL	PATCHES
PROTEINASE	PSEUDOMONAS	SECRETIONS	SPUTUM

Tableau 3 : *Liste alphabétique des 20 mots du vocabulaire commun minimum*

Notre hypothèse sera vérifiée si ces 20 mots codent un maximum de documents de la base MUCO et peu de documents dans les autres bases. Le tableau (4) donne la répartition par base des taux d' extractions des documents.

Type de sujet		Taux de récup	total
Base 0 MUCO	mucovicidose	550 (89.9%)	612
Base 1 LAL	leucémie aiguë lymphoblastiqe	38 (1.8%)	2063
Base 2 POLY	polyamines	341 (2.7%)	12726
Base 3 KC	cancer du sein	202 (2.1%)	9871
Base 4 TIL	Tumor Infiltrating leucocyte	9 (1.5%)	586

Tableau 4 : *Taux de récupération par le VCM de 20 mots*

COMMENTAIRE

Les 62 articles éliminés ont été lus par des spécialistes de la maladie rare. Il s'est avéré que ces fiches étaient incomplètes. Parmi les documents récupérés dans les bases témoins, la moitié de ceux-ci étaient récupérés à bon escient.

8. Confirmation

Pour vérifier si les résultats obtenus n'étaient pas le fruit d'un hasard, nous avons appliqué cette stratégie à une autre recherche bibliographique. Dans ce travail, il s'agit de trouver le VCM d'un domaine de recherche, à savoir les neurosciences. L'obligation de procéder à une recherche multibase, nous amène à mettre en œuvre une stratégie semblable à celle utilisée pour les maladies rares. La sélection de la base noyau a été réalisée par l'extraction dans une base générale de toutes les fiches bibliographiques publiées dans les journaux qui possédaient la chaîne de caractères NEURO dans leur titre. On a ainsi récupéré 2882 notices publiées au cours de l'année 1998. La sélection des bases témoins s'est effectuée selon une technique analogue : nous avons sélectionné les fiches publiées dans les revues dont le titre possédait les mots : GASTRO, HEPATO, DERMATO, URO, ENDOCRI, RESPIR, CARDIO.

9. Analyse de la base noyau

L' analyse des correspondances des 2882 notices décrites par les 998 mots permet de sélectionner 251 mots ayant une dimension supérieure à 10 (valeur du 3 ième quartile). Le tableau (5) donne les taux de récupération par base. Ces taux, comme dans le cas des maladies rares, sont très élevés.

Type de sujet		Taux de récupération	Nombre de notices
Base 0 NEURO	neurologie	2879 (99.9)	2882
Base 1 DERMATO	dermatologie	334 (98.8)	338
Base 2 GASTRO	gastro-entérologie	535 (98.2)	545
Base 3 HEPATO	hépatologie	776 (99.6)	779
Base 4 ENDOCRI	endocrinologie	1057 (99.6)	1061
base 5 URO	urologie	661 (97.9)	675
base 6 CARDIO	cardiologie	462 (99.8)	463
Base 7 RESPIR	maladies respiratoires	361 (99.4)	363

Tableau 5 : *Récupération avec les 251 mots candidats*

Pour repérer les mots spécifiques de la base noyau, on a construit le tableau de correspondance entre les 8 bases et les mots retenus dans cette première analyse. La figure 4 donne la représentation des bases lors de cette nouvelle analyse.

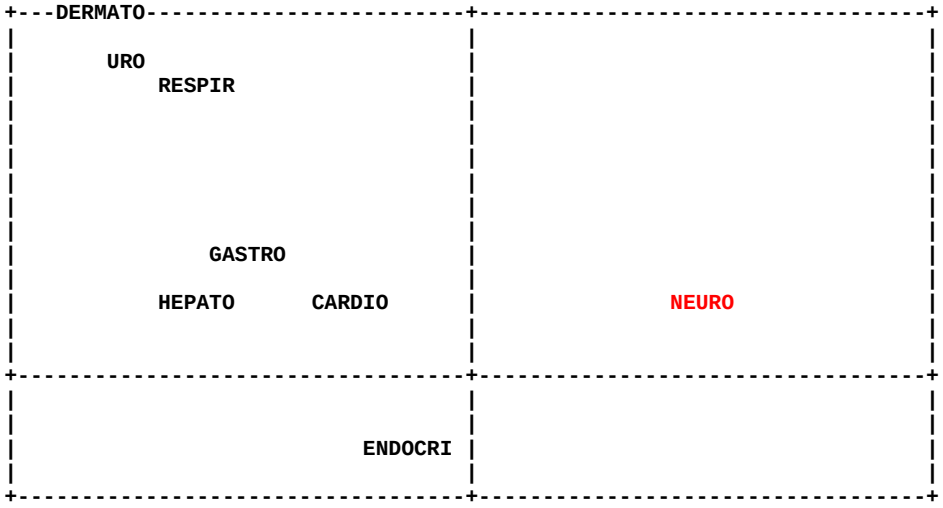


Figure 4 : *Graphe des bases témoins et de la base neuro*

La base NEURO se sépare des autres bases sur le premier facteur. Les mots voisins de ce point sont les mots qui sont les plus caractéristiques du noyau (base neuro). Les mots sont sélectionnés selon 2 critères : avoir une coordonnée sur le facteur 1 positive et avoir une contribution supérieure à la valeur moyenne de la contribution.

NEURONS	**BRAIN**	**NUCLEUS**
CORTEX	**GLUTAMATE**	**CEREBRAL**
SPINAL	**SYNAPTIC**	**DORSAL**
GABA	**MOTOR**	**DOPAMINE**

Tableau 6 : *Liste des mots du vocabulaire commun minimum du sujet neuro*

Le tableau ci-dessous fourni les taux de récupération dans les différentes bases :

Type de sujet	Récupération Avec 12 mots	Nombre de notices
Base 0 NEURO	**2186 (75.8)**	**2882**
Base 1 DERMATO	**4 (1.2)**	**338**
Base 2 GASTRO	**18 (3.3)**	**545**
Base 3 HEPATO	**20 (2.4)**	**779**
Base 4 ENDOCRI	**154 (14.5)**	**1061**
Base 5 URO	**18 (2.7)**	**675**
Base 6 CARDIO	**19 (4.1)**	**463**
Base 7 RESPIR	**21 (5.8)**	**363**

Tableau 7 : *Taux de récupération sur les 8 bases avec le VCM de la neuro*

COMMENTAIRE

La critique majeure que l'on p eut formuler sur cet exercice, porte principalement sur la méthode de sélection des différentes bases. Le fait d' avoir "NEURO" dans le titre d'un e revue n' implique pas une sélection d' articles spécifiques à la recherche en neurosciences : on trouve dans ces revues de plus en plus d' articles dédiés à la biologie moléculaire appliquée aux fonctions cérébrales. Le mauvais score obtenu sur la base "ENDOCRI", montre au contraire la qualité de la stratégie car nous accédons dans cette base de l'endocrinologie aux documents

produits par une branche de cette science nommée neuroendocrinologie et donc pertinents dans notre interrogation.

10. Conclusions

La dimension des mots, n' est que le reflet de propriétés liées à la structure des données analysées. Dans le cas des données bibliographiques bio-médicales, au vu de nos expériences, le vocabulaire des auteurs semble se scinder en quatre grandes classes.
1) Les mots techniques à dimension faible lors de l' analyse d'un corpus.
2) Les mots liés au sujet traité (le Vocabulaire Minimum Commun).
3) Les mots liés à la discipline.
4) Les mots non analysés, ceux du vocabulaire ordinaire (articles, prépositions…).

La création d'un VMC sur un sujet donné va permettre l' alimentation des équations logiques des moteurs de recherches, qui pourront alors agir avec plus d' efficacité sur les mots du texte.

Références

Benzécri J.P et coll (1973). *L'analyse des données*. Editeur Dunod Paris
Lebart L, Salem A, (1994). *L'analyse des données textuelles*. Editeur Dunod Paris.
Kerbaol M., Bansard J.Y (1999).*Pratique de l'analyse des données textuelles en bibliographie.* Ecole Modulad SFdS, INRIA, Bases de données et statistique, Editeur sous presse
Kerbaol M., Bansard J.Y, Favier L. (1997). *Sélection de la bibliographie de "Maladies rares": une approche expérimentale*. 4ème séminaire d' Obernai DICDOC INSERM gestion informatisée en réseaux de l' information biologique, médicale et en santé. Editeur INSERM Paris
Kerbaol M (1996). *Bibliographie cancer du sein Medline vs Embase*. 1er séminaire d' Obernai DICDOC INSERM. Editeur INSERM Paris.