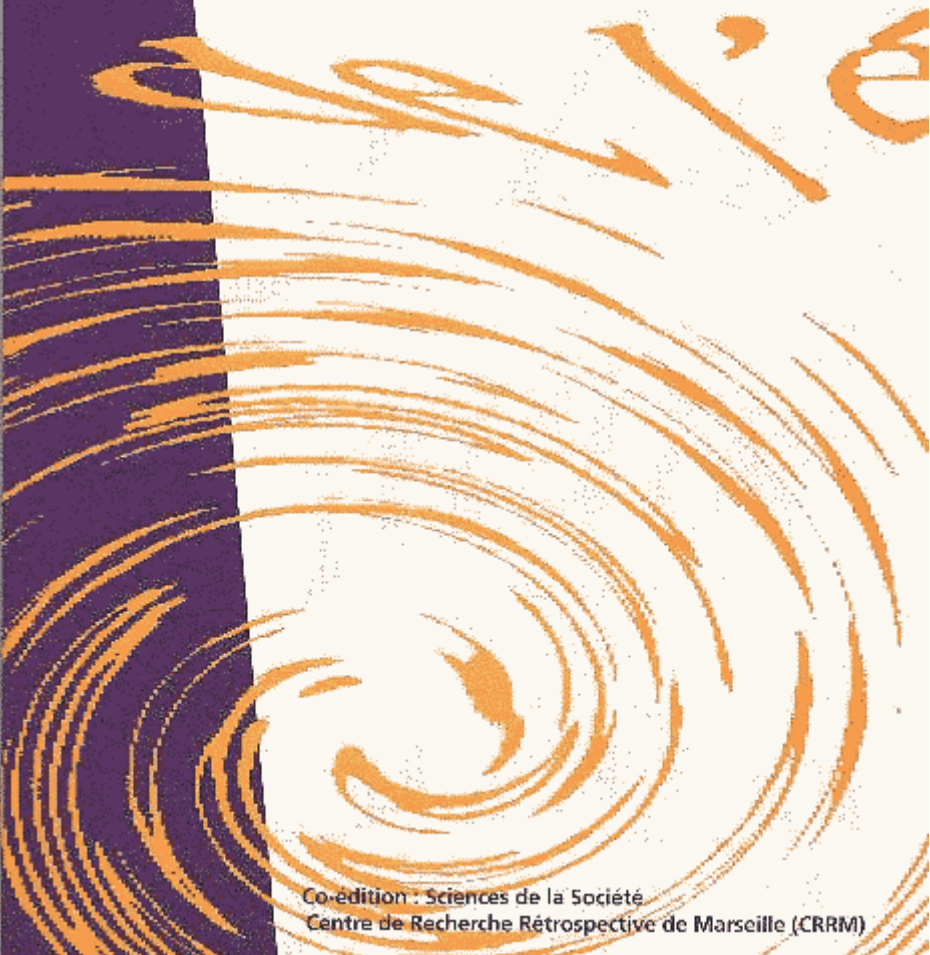
Sciences de la Société
Collection "Outils et méthodes"

La bibliométrie et ses techniques

Hervé Rostaing



Co-édition : Sciences de la Société
Centre de Recherche Rétrospective de Marseille (CRRM)

Ce document est la version électronique d'un ouvrage publié sous le même titre en co-édition *Sciences de la société* - *CRRM* dans la collection "Outils et méthodes", Supplément au n°38 -1996, ISSN : 1168 - 1446.

115, route de Narbonne
31077 Toulouse Cedex
Tél : (33) 5.62.25.82.80 Fax : (33) 5.62.25.80.01

Centre Scientifique de Saint Jérôme
Avenue Escadrille Normandie-Niemen
13397 Marseille Cedex 20
Tél : (33) 4.91.28.87.46 Fax : (33) 4.91.28.87.12

TABLE DES MATIERES

⌘ ⌘ ⌘

PRÉFACE

La Veille Technologique a pour objectif d'utiliser rationnellement l'ensemble des informations disponibles pour permettre de mieux connaître l'environnement de l'entreprise. Cette définition lapidaire met en évidence la nécessité de s'appuyer sur une méthodologie et sur des outils qui permettent un accès, une analyse et une gestion cohérente des données.

Parmi les types d'informations qui sont les plus facilement accessibles, les références bibliographiques plus ou moins codifiées et/ou richement indexées constituent un matériau de choix pour le veilleur. Que les informations soient internes ou externes à l'entreprise, elles fourniront, pour un coût somme toute assez faible, une matière première de qualité, prenant en compte des productions mondiales, souvent spécialisées dans divers domaines.

C'est en partant de ce constat qu'Hervé Rostaing a développé au cours de sa thèse un outil d'analyse puissant, DATAVIEW, en s'appuyant sur les travaux antérieurs réalisés au CRRM par Albert La Tela, puis par Luc Quoniam.

Ce travail de thèse, qui a reçu en 1994 le prix scientifique du SGDN (Secrétariat Général de la Défense Nationale), l'a conduit à se pencher sur l'analyse systématique de corpus structurés. Une meilleure connaissance de la répartition des informations obtenues, ainsi que sur les lois de répartition de celles-ci s'imposait. C'est dans cet esprit qu'il a entrepris une étude systématique qui a donné naissance à l'ouvrage présenté ici.

La connaissance des travaux antérieurs, de l'historique, des résultats de la bibliométrie et de sa modernité permet de replacer l'analyse des informations structurées dans le contexte de la Veille Technologique. Ces analyses, en

s'appuyant sur les données validées de la littérature, permettent de situer globalement le sujet, et souvent de mieux cerner l'environnement des facteurs critiques de succès. Très utile dans l'analyse des brevets, les résultats obtenus doivent être complétés, et il ne faut pas l'oublier, par des informations informelles plus « fraîches », ainsi que par des informations économiques.

Le mérite de cet ouvrage est de permettre rapidement à un non-spécialiste de la bibliométrie de situer le contexte de cette discipline, ainsi que les types de résultats attendus. Cette approche permettra ensuite d'intégrer plus rationnellement ces données nouvelles dans l'ensemble des informations constituant le système de Veille Technologique. Un tel ouvrage en langue française, facile à lire, devrait constituer un outil de choix pour tous les spécialistes de la Veille et de la Documentation Scientifique, Technique et Économique.

Henri Dou
Professeur à l'Université Aix-Marseille III
Directeur du CRRM

INTRODUCTION

Les prémisses du concept de bibliométrie remonteraient au début du XIX^{ème} siècle. Depuis ce temps, les fondements, les techniques, les applications des méthodes bibliométriques ont grandement évolué grâce à de nombreuses expérimentations. Cette abondance de travaux et surtout la variété des objectifs recherchés ont rendu confuse l'idée que l'on peut se faire de la bibliométrie. À un tel point que les spécialistes eux-mêmes ont des difficultés à se mettre d'accord sur des notions aussi cruciales que la terminologie employée (bibliométrie, scientométrie, infométrie, technométrie), sur les limites de sa mise en application (bibliothéconométrie, évaluation de la recherche, sociologie de la science, évaluation macro-économique, sociologie de l'innovation, veille technologique, traitement du langage naturel) et, plus important, sur l'évolution et l'avenir de ses praticiens[1].

Or ces dernières années, la bibliométrie est l'objet d'un sursaut d'intérêt parmi la communauté scientifique ainsi que dans les milieux industriels. La rareté des ouvrages[2] exposant la richesse et la variété des techniques conduit au maintien des confusions en ce qui concerne le concept de bibliométrie. Cet ouvrage espère aider à mieux faire connaître les techniques bibliométriques. Nous essayerons de faire un tour d'horizon de ces techniques sans entrer dans la complexité du débat concernant les facteurs sociologiques mis en jeu. Cet exercice respecte une approche très pragmatique de ce domaine. Le résultat n'est

[1] Se reporter aux débats soulevés par une intervention lors d'un colloque international de bibliométrie-scientométrie et rapportés dans un numéro spécial de la revue *Scientometrics* (*Scientometrics*, Vol 30, N°2-3, 176 p., 1994).

[2] Ouvrages de synthèse sur le sujet : VAN RAAN, 1988 ; WHITE et McCAIN, 1989 ; COURTIAL, 1990 ; DUTHEUIL, 1991 ; DESVALS et DOU, 1992 ; CALLON et alii, 1993.

certainement pas parfait. Il n'est pas imaginable de présenter de façon exhaustive toutes les approches existantes. Mais nous espérons avoir abordé les principales écoles de pensées. Les descriptions des méthodes soulignent plus particulièrement les principes des traitements qu'elles impliquent. Les raisons théoriques qui ont permis de les étayer et les critiques de leur validité ne seront pas toujours abordées.

L'une des principales raisons du regain d'intérêt pour les méthodes bibliométriques est probablement leur récente mise en application dans le monde industriel. L'effervescent attrait des entreprises pour les activités d'Intelligence Economique, de Veille Stratégique, ou plus simplement, de Veille Technologique a fait de la bibliométrie un outil parfaitement adapté à l'analyse de la concurrence. Ce récent domaine d'application donne une nouvelle jeunesse à des méthodes dont la mise en application a souvent été décriée par le passé (aide à l'évaluation des chercheurs, ou aide à la restructuration des organismes de recherche publics). L'introduction de la bibliométrie comme outil de surveillance de la concurrence ou de l'environnement scientifique d'une entreprise donne une nouvelle envergure stratégique et économique aux recherches menées dans ce domaine[1].

Cet ouvrage retrace brièvement l'évolution de la bibliométrie jusqu'à son exploitation dans le contexte industriel. Un premier chapitre évoque son émergence, ses précurseurs et ses fondements. Puis un rapide historique présente les progrès de la bibliométrie et les facteurs qui ont stimulé ces progrès. Ce chapitre s'achève par une proposition de définition de la bibliométrie et un inventaire, probablement imparfait, des différents acteurs de la bibliométrie en France.

Les chapitres suivants ont été découpés selon les différentes méthodes et techniques rencontrées en bibliométrie. La bibliométrie appliquée à l'évaluation de l'information scientifique est répartie en trois moments. Tandis que la récente application à l'information brevet est regroupée dans un dernier chapitre. Les trois chapitres consacrés à la quantification de la science classent les techniques bibliométriques en trois catégories : les distributions statistiques, les indicateurs univariés, les cartes relationnelles. Le chapitre consacré à la quantification de la technique présente les indicateurs univariés et les cartes relationnelles appliquées à l'information brevet.

[1] Les principes de l'activité de surveillance de la concurrence en milieu industriel ne seront pas mentionnés, aussi les lecteurs pourront-ils se reporter avantageusement aux nombreux et souvent très récents ouvrages qui traitent du sujet : LESCA, 1988 ; MORIN, 1988 ; HUNT et ZARTARIAN, 1990 ; VILLAIN, 1990 ; JAKOBIAK, 1991 ; LAINE, 1991 ; MARTINET, 1991 ; DESVALS et DOU, 1992.

CHAPITRE I
HISTOIRE, FONDEMENTS ET CONCEPTS

Pour bien comprendre ce qui a très rapidement poussé certains chercheurs à concevoir les techniques d'analyse bibliométrique, il est indispensable de prendre connaissance de certains faits.

Le monde de la science et de la technique se transforme rapidement et profondément, allant vers toujours plus de complexité. Les spécialités se multiplient vite. Les frontières qui les délimitent sont mouvantes et laissent apparaître bien souvent des zones floues au moment de l'émergence de nouveaux concepts. La prise de conscience de l'existence de ces domaines et la compréhension de leurs activités sont de plus en plus difficiles à maîtriser. Très rapidement, l'homme de science a senti qu'il ne pourrait plus maîtriser l'ensemble des connaissances. Rester au fait de toutes les dernières découvertes ne fut bientôt plus possible car la quantité des écrits dépassa les capacités de lecture et d'entendement humain. L'homme de science subissait de plein fouet l'explosion de la croissance des connaissances qui sera bien plus tard modélisée par de Solla Price comme étant régie selon une courbe à tendance exponentielle (voir p. 54). Dans ce décor changeant, l'exploitation de méthodes et d'outils devient indispensable pour mieux appréhender cette complexité.

Bien évidemment, les premiers concepteurs de ces méthodes ont recherché la maîtrise des écrits scientifiques puisque c'était (et cela reste toujours) le vecteur privilégié de la communication des connaissances. Comme cette maîtrise n'était plus possible par la simple lecture de ces écrits, ils ont alors appliqué les techniques statistiques à ce type de données que constituent les écrits scientifiques. Ce phénomène explique la parenté des techniques

bibliométriques avec celles des traitements statistiques linguistiques. Cette parenté est de plus en plus marquée, puisque de nombreux centres de recherche en bibliométrie ont pour nouvel objectif le traitement automatique des textes rédigés en langage naturel. Le fossé qui reste persistant entre ces deux domaines scientifiques est tout simplement la nature des écrits traités. La bibliométrie cherche à analyser des écrits scientifiques, tandis que les traitements statistiques linguistiques ont été développés initialement pour l'analyse des textes littéraires. Or les vocabulaires, et donc les lexiques et les dictionnaires, sont bien trop différents pour permettre une totale adaptation des outils linguistiques aux études bibliométriques.

Les statistiques, quelles que soient leurs mises en application, sont basées sur le principe du dénombrement. Les techniques bibliométriques reposent donc toutes sur ce même principe. Ce qui différencie chaque usage des statistiques est l'entité dénombrée. Ainsi, les statistiques en démographie cherchent-elles à comptabiliser des individus selon diverses caractéristiques de classement. L'emploi des statistiques en linguistique comptabilise les mots présents dans les textes étudiés. Là encore, la bibliométrie se distingue des statistiques linguistiques en considérant comme entités à dénombrer les "signalements" des écrits scientifiques (références bibliographiques) et non le contenu même de ces écrits.

LES PRECURSEURS

La première étude, considérée comme pouvant respecter les conditions bibliométriques, est attribuée à Cole et Eales (COLE et EALES, 1917). Cette étude avait pour objet de répertorier toute la littérature, publiée entre 1850 et 1860, concernant l'anatomie. L'étude statistique de cette littérature a permis de montrer les fluctuations d'intérêts scientifiques pendant cette période.

Dix ans plus tard, Gross et Gross (GROSS et GROSS, 1927) furent les premiers à comptabiliser non plus les documents scientifiques mais les citations que les chercheurs faisaient, dans leurs propres documents, des travaux précédemment publiés. Ils effectuèrent ces comptages pour les journaux cités qui touchaient tous les domaines de la chimie, puis rangèrent ces journaux par ordre décroissant du nombre de citations reçues. Ils venaient d'établir la liste des journaux qu'ils considéraient comme indispensables à consulter dans le domaine de la chimie. Nous pouvons remarquer, pour cette étude, que le dénombrement reste toujours basé sur le comptage des articles, ceux-ci étant regroupés par catégories : l'appartenance aux mêmes journaux.

En 1934, l'anglais Bradford soumit à ses pairs une théorie (BRADFORD, 1934) qui sera par la suite dénommée *loi de Bradford*. Cette première publication de son travail ne fit pas beaucoup d'émules. Par contre, lorsqu'il le présenta de nouveau dans son livre en 1948, ce travail fut rapidement repris par d'autres chercheurs et fit l'objet d'un intérêt remarquable (BRADFORD, 1948). Depuis, c'est certainement l'article qui a fait couler le plus d'encre dans la

discipline. Cette théorie cherchait à modéliser la répartition des journaux selon leur aptitude à représenter un domaine scientifique, cette aptitude étant évaluée en fonction du nombre d'articles concernés par ce domaine pour chaque journal (voir p. 36).

L'étude de Bradford intéressa tout particulièrement les gestionnaires de bibliothèques, puisque la mise en pratique de ce modèle leur permettait d'optimiser le nombre de leurs abonnements aux revues selon les centres d'intérêts de leur clientèle. La loi de Bradford devait permettre une amélioration de la gestion économique des bibliothèques, et c'est pour cette raison que cette partie de la bibliométrie est reconnue comme appartenant au domaine de la bibliothéconométrie.

Cet intérêt pour la loi de Bradford explique pourquoi la plupart des travaux bibliométriques qui suivirent étaient consacrés à la recherche de formules mathématiques qui s'ajustent le mieux possible aux données bibliographiques expérimentales. Par la suite, certains auteurs se sont aussi penchés sur la détermination des relations mathématiques qui permettraient de relier la loi de Bradford à d'autres lois, celles de Lotka et de Zipf. L'objectif ambitieux était de découvrir une formulation mathématique unifiant toutes ces lois sous une même loi universelle (voir p. 44). L'école de pensée bibliométrique anglo-saxonne reste encore maintenant très fortement attachée aux statistiques distributionnelles et à la mesure de la circulation de l'information. Les chercheurs anglais comptent parmi les meilleurs spécialistes de cette approche de la bibliométrie.

LA "SCIENCE DE LA SCIENCE"

Aux États-Unis, sous l'impulsion de sociologues, d'autres courants de pensée se sont développés au cours des années 1960 et 1970. De Solla Price a été l'un des plus fervents artisans de l'explosion de la discipline pendant cette période. Il se dissocia du mouvement anglo-saxon de l'époque. Il renonça à l'emploi de l'outil statistique selon la rigueur mathématique exigée mais le mit en oeuvre au service de l'idée selon laquelle l'activité scientifique est régie selon des règles sociologiques (PRICE, 1963). Ses investigations sociologiques de la science lui permirent de proposer plusieurs lois qui sont encore reconnues de nos jours dans la discipline. On peut citer parmi toutes celles qu'il avança les trois suivantes : la prolifération de la connaissance scientifique suivrait une courbe en S (voir p. 54), le phénomène de collaboration entre chercheurs serait dépendant d'un ensemble de règles sociologiques dont la principale est la création de "*collèges invisibles*", la plupart des phénomènes de reconnaissance en science, que ce soit pour les travaux, les individus, les paradigmes, les terminologies, respecteraient une même règle qu'il nomma "*l'avantage du cumul*" (voir p. 45). L'importance de sa contribution à la discipline est reconnue de tous, à tel point que *Scientometrics*, ayant décidé de délivrer chaque année un

prix au chercheur de la discipline le plus méritant, a choisi de le nommer *Prix Derek John de Solla Price.*

Cette approche novatrice de l'exploitation de l'outil statistique à l'information scientifique créa un engouement considérable en faveur de nouveaux axes de recherche. Ainsi, Price fut accompagné dans ses découvertes par de nombreux pionniers de la discipline. Cet essor de la bibliométrie se déploya de part et d'autre du rideau de fer.

C'est à cette époque que certains chercheurs reconnurent ne plus pouvoir regrouper leurs méthodes sous l'appellation "bibliométrie", puisque l'emploi de l'outil statistique n'avait plus du tout la même finalité. Leur approche étant plus générale, ils préférèrent faire connaître leur activité sous le nom "*science de la science*" dans le sens d'utilisation de techniques scientifiques pour analyser l'histoire sociologique de la science. Ils décidèrent donc d'appeler les techniques qu'ils employaient pour leurs analyses "scientométrie". Cette appellation étant tout simplement la traduction du terme russe *"nauko-vometrica"* attribué par Doborov et Korennoi aux techniques statistiques donnant accès à la mesure de la science (DOBOROV et KORENNOI, 1969). Cette modification de l'appellation des méthodes n'est qu'artificielle, puisque les techniques sont les mêmes, seuls les objectifs de leur mise en oeuvre sont différents. En effet, les mesures de la science sont établies à partir du dénombrement des textes scientifiques ; qu'ils soient regroupés par noms d'auteurs, dates, domaines ou journaux ne modifie aucunement le fondement même du principe de comptage.

DE L'ETUDE DE LA SCIENCE A L'EVALUATION DE LA RECHERCHE

La création, à Philadelphie, au début des années 1960, par E. Garfield, de l'*Institute for Scientific Information* (ISI) a permis à la discipline d'étayer la partie instrumentale des méthodes et concepts mis en avant par Price. Une nouvelle technique d'évaluation de l'activité scientifique, fondée sur l'étude des citations que se distribuent les auteurs d'articles, se greffe autour de cet institut.

Garfield a eu l'idée de constituer un répertoire ayant une couverture interdisciplinaire et qui regrouperait uniquement les articles publiés par les principaux périodiques scientifiques. Ce "coeur" des revues est déterminé par le taux de citations dont elles font l'objet (GARFIELD, 1979). L'activité principale de l'ISI est donc de collecter les articles publiés dans les revues les plus prestigieuses à travers le monde dans toutes les branches de la science. Il avait déjà imaginé, en 1955, le principe de ce *Science Citation Index (SCI).* La première édition papier du *SCI* est parue en 1963 et couvrait la littérature scientifique de 1961. Elle s'étendait à 613 revues et contenait 1,4 millions de citations en 5 volumes. Actuellement, le SCI couvre à peu près 4 200 périodiques. Depuis, deux nouveaux répertoires, concernant les autres branches de la science, sont venus se joindre au *SCI* : le *SSCI* (*Science Social Citation*

Index), publié depuis 1973, suit 1400 autres périodiques, et depuis 1978, l'*A&HCI* (*Arts & Humanities Citation Index*) est venu le compléter.

Au début le SCI comprenait 3 répertoires :

- le *Citation Index* (répertoire des citations par noms d'auteurs)
- le *Source Index* (répertoire des publications par noms d'auteurs)
- le *Permuterm Subject Index* (répertoire des mots du titre des publications)

Depuis 1976, un quatrième répertoire est venu s'y joindre, le *Journal Citation Reports (JCR)* qui contient d'importantes informations au sujet des périodiques scientifiques par le reflet de leurs citations. Il fournit entre autres :

- le nombre de citations reçues par une revue (c'est-à-dire les citations des articles publiés dans la revue) pendant une année
- le facteur d'impact (I_F) de chaque revue (voir formule p. 55)
- le classement des revues par thèmes (128) selon leurs valeurs de I_F
- les 15 revues les plus citées pour chaque revue
- les 15 revues qui citent le plus chaque revue

Ces journaux de l'ISI ont été et restent à la base de multiples études bibliométriques. Ils ont été les premiers instruments de travail. Mis à la disposition de tous sous la forme de périodiques, ils donnent l'opportunité aux chercheurs de toutes les disciplines d'utiliser des outils bibliométriques dans l'évaluation de leur propre domaine. Ceci explique l'attrait que le monde scientifique, et plus particulièrement la communauté américaine, a eu pour les documents fournis par l'ISI (sans oublier les *Current Contents*[1]).

Alors que ces outils de travail ont été initialement imaginés comme une aide à la compréhension de l'évolution de la communauté scientifique et de ses paradigmes, les instances dirigeantes ont vu là l'occasion de disposer d'un système d'évaluation de la recherche. Les buts ne sont bien évidemment plus d'étudier l'aspect sociologique de la science, mais de créer des indicateurs qui permettent d'évaluer l'activité de la recherche, c'est-à-dire la productivité et la position stratégique des différents acteurs participant à cette recherche. Ces indicateurs font partie intégrante de l'ensemble des facteurs qui influencent les prises de décision dans les politiques de recherche. Les répertoires de l'ISI n'ont pas été créés pour mesurer les "performances" des chercheurs, des équipes ou des laboratoires, mais plutôt pour établir des relations pouvant exister entre les divers travaux de recherche menés n'importe où dans le monde. Or très rapidement, on a fait dévier le but originel de ces données pour évaluer, plus ou moins légitimement, les acteurs de la recherche scientifique.

Nous pouvons facilement imaginer le danger qu'il peut y avoir à utiliser des données, conçues au départ pour des analyses globales et qualitatives, à des fins

[1] Revues hebdomadaires de l'ISI regroupant par disciplines les tables des matières des principaux périodiques nouvellement édités. Ces revues traitent des sciences exactes, des sciences sociales, des sciences humaines et des arts.

de prises de décision tout à fait rationnelles et donc fortement basées sur des évaluations quantitatives. Ceci est d'autant plus dangereux que tous les centres nationaux d'évaluation[1] au service des politiques de recherche ne se servaient principalement que de la source d'information produite par l'ISI. Or la base de la construction de ces données, c'est-à-dire le "coeur" des revues les plus prestigieuses de la science, ne peut absolument pas représenter de façon équitable à la fois toutes les disciplines, tous les pays et toutes les périodes. Ce sujet réapparaît régulièrement dans la discipline au centre des débats car bien que l'informatisation des outils bibliométriques ait permis aux instituts nationaux d'évaluation de la recherche d'élaborer leurs propres indicateurs, les indicateurs de l'ISI restent fortement employés. En effet, cet institut est le seul producteur d'information qui prend en charge le phénomène de la citation scientifique.

C'est dans ce contexte, qu'au début des années 1970, la *National Science Foundation* américaine demanda à une petite firme de Philadelphie, la *Computer Horizon Incorporated* (*CHI*), dirigée par Francis Narin, de mettre les répertoires de l'ISI sous une forme les rendant utilisables pour la production d'indicateurs. Ce "nettoyage", fondé sur un comptage plus rigoureux des publications et des citations, a nécessité un travail considérable : normalisation des noms des pays, vérification minutieuse des orthographes des auteurs, traitement des articles publiés par plusieurs auteurs, sélection des types de documents retenus, classements de ces documents par spécialité, discipline ou domaine. En ce qui concerne la difficile question de la sélection des revues, CHI a décidé de maintenir constant l'ensemble des périodiques constituant le "noyau", soit un peu plus de deux mille cent revues choisies parmi celles dépouillées par l'ISI dès 1973. Il en résulte bien évidemment qu'aucune des revues lancées depuis cette date ne figure dans la base et que, parmi celles qui s'y trouvent encore, certaines sont moins représentatives de la science en cours. C'est en biologie et en mathématiques que ce décalage est le plus marqué. La représentation de la science par ce noyau est donc conservatrice et statique. Rigidité des classifications, couverture incomplète de certains secteurs, définitions parfois contestables des domaines et des sous-domaines par une liste intangible de revues, telles sont les limites les plus évidentes de la banque du CHI. Plus on s'intéresse à des disciplines étroitement définies et récentes, plus ces défauts deviennent rédhibitoires (rapport de l'*Advisory Board for the Research Councils*, 1986). Outre la restructuration et l'exploitation de la banque SCI de l'ISI, le CHI a constitué sa propre banque bibliométrique de brevets. Depuis 1975, les brevets déposés au *US Patent Office* sont saisis informatiquement par le CHI. Cette *Technological Activity and Impact Indicators Database* est une source courante d'études bibliométriques

[1] Exemples d'instituts nationaux : National Science Foundation (NSF), États Unis ; Science Policy Research Unit (SPRU), Royaume Uni ; Information Science and Scientometrics Research Unit (ISSRU), Hongrie ; Centre for Science and Technology Studies (CWTS), Pays-Bas ; Center for Science Studies (CSS), Allemagne ; Observatoire des Sciences et Technologies (OST), France.

américaines concernant l'aspect technologique (voir p. 93). Ces données bibliométriques qui présentent une certaine qualité au niveau de l'exploitation informatique ne sont pas accessibles au public. Elles ne sont exploitées qu'à titre privé pour renseigner la NSF, ou dans le cadre de services de courtage pour des études à la demande.

L'INFORMATISATION DE LA BIBLIOMETRIE

Le dernier facteur d'évolution de la bibliométrie est l'amélioration des technologies de la gestion de l'information. Ces technologies ont permis tout d'abord un stockage et une ré-exploitation rapide de l'information grâce à l'informatique, puis une diffusion internationale et immédiate de ces informations par le réseau de télécommunications. Dernièrement la technologie des disques compacts (CD-ROM) offre un nouveau moyen de diffusion de l'information qui n'a pas la même capacité de stockage et la même rapidité de diffusion que les précédentes mais qui, par contre, a un coût bien plus économique.

En fait, les grands producteurs de répertoires scientifiques, soumis à l'accroissement de la production mondiale de documents et de la demande, ne pouvaient plus se contenter des versions papier des répertoires. Dans les années 1960, l'informatique est venue leur apporter des solutions satisfaisantes, les ordinateurs se prêtant bien au stockage et à la recherche rapide d'information. Depuis les années 1970, avec la mise au point de la communication des données par télématique, les versions informatisées des grands bulletins signalétiques sont devenues accessibles à tous sous la forme de banques de données. Plus récemment, avec les progrès des technologies de l'information et de la communication, deux autres modes de diffusion des répertoires informatisés sont venus compléter les deux précédents : disque compact et vidéotex (conception purement française).

Ainsi, les plus grands répertoires bibliographiques en science sont maintenant disponibles sous leur version informatique : en chimie, *Chemical Abstract*, produit par *Chemical Abstract Services,* en physique-électricité-électronique, *Inspec*, produit par *IEEE* et *IEE*, en sciences médicales, *Medline*, produit par la *National Library of Medicine* et l'*INSERM*, en sciences de la vie, *Biosis*, produit par *Biosis*, en sciences (pluridisciplinaire), *Pascal*, produit par l'*INIST*, en sciences humaines et sociales, *Francis*, produit par l'*INIST*, en sciences de l'éducation, *Eric*, produit par *ERIC* et *US Depart. of Education*, etc. On retrouve aussi une partie des répertoires produits par l'ISI en version informatique : le *SCI*, le *SSCI*, et le A*&HCI*. Le *JCR*[1] est la seule exception. Avant l'apparition de ces banques de données, les chercheurs qui développaient de nouvelles théories bibliométriques se trouvaient devant d'énormes problèmes pour valider leurs théories par des données expérimentales. Ils se voyaient

[1] *Journal of Citation Reports*

contraints d'effectuer manuellement de lourdes opérations de collecte et de comptage de documents. Maintenant tous ces répertoires sont consultables à partir d'un micro-ordinateur connecté par les réseaux de télécommunication à des distributeurs internationaux, appelés plus communément serveurs (de banques de données). Les principaux serveurs de banques de données scientifiques et techniques sont *Questel-Orbit* filiale de *France Télécom*, *Cedocar* de la *Direction Générale de l'Armement*, *Dialog* et *Data-Star* du groupe américain *Thomson & Tomson*, et *STN*, propriété conjointe de *Chemical Abstratcs Services* et de l'institut allemand *FIZ Karlsruhe*.

Le fait que l'information à analyser soit sur un support informatique donne l'opportunité de développer de nouvelles méthodes bibliométriques. Ces méthodes peuvent bénéficier de l'existence de volumes de données sans commune mesure avec le passé et surtout permettent d'exploiter des techniques statistiques gourmandes de calculs mathématiques. C'est ainsi que les premières recherches sur les réseaux de citations se sont développées.

C'est tout d'abord Kessler, en 1963, qui eut le premier l'idée de se servir du phénomène de la citation (voir p. 66) comme critère de mise en relation de documents scientifiques. Cette idée a ensuite été reprise par l'école de pensée qu'avaient fondée Price et Garfield. Small, un collaborateur de Garfield, avec l'aide de l'Université de Drexel, mit en application cette idée sur la banque de donnée SCI de l'ISI, dès 1973. La méthode est connue sous le nom d'analyse des co-citations. Le principe de dénombrement consiste à comptabiliser le nombre de co-apparitions de couples de références cités, les co-citations (voir p. 67). En 1978, cette méthode a été appliquée sur l'ensemble de la banque de donnée de l'ISI pour construire des agrégats de références bibliographiques selon la similarité des citations énumérées à la fin de chacune de ces références. Le résultat de ces agrégats, après avoir été reproduit dans un espace plan selon les liens entretenus entre tous les agrégats, fut exposé comme la première cartographie de la science. Cette méthode bibliométrique a immédiatement séduit la discipline, car le résultat statistique ne se réduit pas à un simple classement d'éléments ou à un simple modèle de type distribution, mais permet de disposer les documents selon leurs "similarités" dans un espace optimal pour l'esprit humain, c'est-à-dire la représentation plane.

Malgré cet attrait, il fallut attendre quelques années pour que tous puissent mettre en oeuvre une telle technique bibliométrique. Les contraintes principales étaient les méthodes d'analyse statistique employées : classification automatique et analyse d'inertie. Ces techniques mathématiques, déjà fortement utilisées dans d'autres disciplines, exigeaient d'avoir à sa disposition une puissance de calcul considérable pour l'époque, c'est-à-dire de pouvoir accéder à l'équivalent d'un calculateur d'un centre universitaire. Il fallait ensuite transférer de grands volumes de données bibliographiques provenant des banques de données internationales, puis développer des programmes de comptage afin de construire un tableau exploitable par une méthode mathématique conventionnelle. Seuls des organismes comme l'ISI pouvaient disposer à la fois de tels moyens

matériels et de personnels en informatique. Aussi, il fallut attendre l'arrivée de la micro-informatique pour que de telles techniques bibliométriques se démocratisent.

L'apparition de la micro-informatique donne naissance à une profusion de nouvelles approches bibliométriques. Les ordinateurs personnels offrent à chaque chercheur la possibilité de consulter les banques de données à distance avec un modem, de rapatrier les références bibliographiques choisies grâce à son propre logiciel de télécommunication, d'homogénéiser ces références par des outils commerciaux, d'effectuer les traitements bibliométriques par des programmes personnels et d'injecter les résultats dans des logiciels d'analyse statistique commercialisés. De telles facilités informatiques sont devenues courantes dans les centres de recherche de tous les pays industrialisés. Aussi, ces nouvelles approches bibliométriques ont-elles principalement développées en Europe occidentale avec la ferme volonté de se marginaliser du mouvement hégémonique américain existant jusqu'à présent dans la discipline. C'est alors que des méthodes bibliométriques comme l'*analyse des mots associés*, l'*analyse des co-classific*ations, les *analyses des co-auteur*s, l'*analyses des co-opération*s ou les *analyses de tableaux de contingence* (voir p. 73-85) émergent, apportant chacune d'elles un regard nouveau et différent sur la compréhension de l'activité des travaux scientifiques. Chaque méthode fait appel à de nouveaux dénombrements de co-apparitions d'éléments bibliographiques respectivement, co-occurrence de mots, co-occurrence de codes de classification documentaire, co-occurrence d'auteurs, co-occurrence des pays collaborant, et finalement co-occurrence de deux éléments bibliographiques distincts (Pays-Codes, Pays-Mots, Villes-Codes, Pays-Années...).

La conjonction entre ces atouts informatiques opérationnels et le besoin grandissant d'une information scientifique et technique élaborée de la part des sphères décisionnelles des entreprises, favorise l'adaptation des techniques bibliométriques au monde industriel (voir p. 93). Spontanément, les techniques bibliométriques se révèlent des outils parfaitement adaptés à l'évaluation de l'activité scientifique ou de l'activité en propriété industrielle. C'est ainsi que la bibliométrie se fait connaître comme un outil d'aide à l'activité de veille industrielle ou veille technologique. Cette pratique a débuté dans les années 1980 aux Etats-Unis avec les travaux de Narin sur la base des brevets américains de CHI (voir p. 93). Depuis ces dernières années, l'intérêt se porte vers l'application des techniques bibliométriques sur les données de propriété industrielle. L'adaptation de la méthodologie bibliométrique à ces données est facilitée par l'existence de banques de données qui répertorient les dépôts de brevets nationaux ou internationaux sous forme de références bibliographiques. Les principales banques de données brevets sont *World Patent Index* et *US Patents*, produites par *Derwent*, *INPADOC*, de *International Patent Documentation Center in Vienna*, *FPAT EPAT et EDOC*, bases de l'*Institut National de la Propriété Intellectuelle*, *CLAIMS*, de *IFI-Plenum Data Corp*, *JAPIO* et *PATOLIS*, créées par *Japanese Patent Information Office*. Ces

banques de données brevets sont présentes sur les mêmes serveurs commerciaux que les banques de données scientifiques.

Les pays de l'Europe de l'Est ont gardé une forte activité tout au long de ces dernières années, mais ne disposant pas toujours d'autant de commodités informatiques que ceux de l'Ouest, ils ont conservé une approche plus théorique et n'ont que très peu souvent appliqué la bibliométrie aux brevets.

QUELLE DEFINITION POUR LA BIBLIOMETRIE ?

Avant de donner une définition de la bibliométrie, il est bon de rappeler les deux postulats de travail implicites à toute méthode d'analyse bibliométrique.

Premier postulat : un écrit scientifique est le produit objectif de l'activité d'une pensée. Dans un contexte scientifique, une publication est une représentation de l'activité de recherche de son auteur. Le plus grand effort de cet auteur est de persuader les autres scientifiques que ses découvertes, ses méthodes et techniques sont particulièrement pertinentes. Le mode de communication écrit fournira donc tous les éléments techniques, conceptuels, sociaux et économiques que l'auteur cherche à affirmer tout au long de son argumentation.

Second postulat : l'activité de publication scientifique est une perpétuelle confrontation entre les propres réflexions de l'auteur et les connaissances qu'il a acquises par la lecture des travaux émanant d'autres auteurs. La publication devient par conséquent le fruit d'une communion de pensées individuelles et de pensées collectives. Ainsi, les chercheurs, pour consolider leur argumentation, font souvent référence à des travaux d'autres chercheurs qui font l'objet d'un certain consensus dans la communauté scientifique. Par conséquent, il existe une relation entre tous les travaux scientifiques publiés, que cette relation soit directe ou indirecte, reconnue ou dissimulée, consciente ou inconsciente, en accord ou en désaccord.

Par l'acceptation de ces deux postulats, l'étude des publications scientifiques permettrait d'appréhender les connaissances et leurs structures suivant les écoles de pensée et leurs évolutions. Ces postulats, qui ont été définis initialement pour la recherche scientifique, ont ensuite été admis pour les publications rassemblant les connaissances techniques ou technologiques, c'est-à-dire les publications des dépôts de brevets. S'appuyant sur ces deux postulats, le principe de la bibliométrie est d'analyser l'activité scientifique ou technique par des études quantitatives des publications. Les données quantitatives sont calculées à partir de comptages statistiques de publications ou d'éléments extraits de ces publications. La bibliométrie est donc un terme générique qui rassemble une série de techniques statistiques cherchant à quantifier les processus de la communication écrite.

Les auteurs anglo-saxons attribuent l'invention du terme bibliométrie à Pritchard (PRITCHARD, 1969), tandis que certains auteurs français l'attribuent à Otlet (ESTIVALS, 1969). En fait, en introduisant le terme bibliométrie,

Pritchard suggérait de le substituer à l'expression bibliographie statistique qui était employée depuis 1923, date à laquelle Hulme avait pour la première fois présenté son travail (HULME, 1923). Il estimait que bibliographie statistique pouvait prêter à confusion, et être interprétée comme une bibliographie sur la statistique. De plus, le terme bibliométrie se rapprochait du même coup de termes établis comme biométrie ou économétrie. Il en profita pour donner sa propre définition de la bibliométrie comme étant *"... l'application de méthodes mathématiques et statistiques aux livres et aux autres médias de communication"*.

Cette définition de Pritchard ne donne aucune indication sur la finalité de la bibliométrie. A l'époque, son application s'insérait dans le domaine de la gestion des bibliothèques comme le montre la définition donnée par Raising en 1962, alors que ce genre de méthodes est toujours connu sous le nom de bibliographie statistique : *"l'assemblage et l'interprétation de statistiques relatives aux livres et aux périodiques... pour démontrer des mouvements historiques, pour déterminer l'utilisation par la recherche nationale et universelle des livres et des journaux, et pour s'assurer dans de nombreuses situations locales de l'utilisation générale des livres et des journaux"* (RAISING, 1962). Or, la bibliométrie a depuis fortement repoussé son application au-delà des frontières de la bibliothéconométrie. Hawkins plus récemment a défini la bibliométrie comme *"des analyses quantitatives des caractéristiques bibliographiques d'un corps de littérature"* (HAWKINS, 1977). Mais cette définition est trop restrictive car elle ne prend pas en compte une des activités de la bibliométrie : l'étude de la circulation des publications.

C'est pour distinguer ces deux types d'applications qu'un autre terme est apparu, celui de scientométrie. Dans une conférence, Brookes a précisé de nouveau cette distinction. "*Alors que la bibliométrie aurait pour objet d'étudier les livres ou les revues scientifiques et pour objectif de comprendre les activités de communication de l'information, la scientométrie aurait pour objet l'étude des aspects quantitatifs de la création, diffusion et utilisation de l'information scientifique et technique et pour objectif la compréhension des mécanismes de la recherche comme activité sociale*" (BROOKES, 1987). Donc, la bibliométrie regrouperait l'ensemble des méthodes aidant à la gestion des bibliothèques et la scientométrie rechercherait les lois qui régissent la science, d'où son appellation *"science de la science"* par Price.

Pour notre part, nous aurons une approche plus pragmatique de la bibliométrie:

> La bibliométrie est l'application de méthodes statistiques ou mathématiques sur des ensembles de références bibliographiques.

Cette définition rejoint celle de White et McCain dans leur article faisant le point sur la bibliométrie (WHITE et McCAIN, 1989). Elle permet d'intégrer l'ensemble des traitements cités par les précédentes définitions. La scientométrie

serait alors une conception englobant la bibliométrie comme un outil parmi d'autres pour dresser des bilans de santé d'un système de recherche. Les études scientométriques prennent en compte, dans leurs analyses, d'autres facteurs que le simple acte de publier, telles les ressources et la façon dont ces ressources sont transformées en connaissances et savoir-faire (TURNER, 1990). Pourtant, comme le fait remarquer Callon, "*jusqu'à une date récente elle s'est presque exclusivement intéressée à l'analyse des documents rédigés par les chercheurs et technologues*" (CALLON et alii, 1993). C'est-à-dire que la scientométrie ne fait appel qu'à des techniques bibliométriques, mettant de côté tous les autres facteurs à analyser. Cela explique la confusion qui s'est établie entre ces deux termes tout au long de l'histoire de la "science de la science".

TYPOLOGIE DES METHODES ET DES APPLICATIONS

Avant d'entrer dans le détail de chacune des méthodes bibliométriques, nous avons essayé de les classer par grandes catégories. Dans le même souci de clarté, il nous parait utile de présenter également la diversité des applications obtenues par de telles méthodes.

Pour introduire ces classements, il est nécessaire de s'attarder un instant sur la notion suivante : dans le terme bibliométrie, le suffixe "métrie" renvoie aussi bien à la mesure qu'à la métrique. Dans un rapport qu'il a rédigé pour le compte du SGDN[1], Dutheuil précise ces deux sens : "*La métrique s'applique à un ensemble pour lequel on accepte une convention permettant de définir les "distances" entre les éléments, ce qui conduit à les classer par ressemblance ou dissemblance. La mesure est l'évaluation d'une grandeur faite d'après son rapport à une grandeur de même espèce prise pour unité et comme comparaison (étalon)*" (DUTHEUIL, 1991).

La bibliométrie s'inscrit tout à fait dans ces deux significations. Le concept de mesure est bien représenté par les études bibliométriques utilisant des indicateurs univariés où chaque élément à étudier est soumis à une mesure selon une dimension choisie (voir p. 51). Le classement et la comparaison des éléments les uns par rapport aux autres, selon cette dimension, sont alors possibles. C'est le cas des indicateurs fournis dans les répertoires de l'ISI. Par contre, le concept de métrique est plus spécifique aux indicateurs relationnels (voir p. 65). Dans ce cas, les comparaisons entre les éléments ne se font plus sur un référentiel à une seule dimension mais à partir d'un ensemble de facteurs influents. Les méthodes employées chercheront à disposer les éléments selon des calculs de "distances" qui devront estimer les degrés de ressemblance ou de dissemblance entre les éléments. C'est le cas de la méthode d'analyse des co-citations.

La bibliométrie est donc un outil de "mesure" basé sur l'emploi de techniques statistiques, qui a pour objet d'aider à la comparaison et à la

[1] Secrétariat Général de la Défense Nationale

compréhension d'un ensemble d'éléments bibliographiques. Comme White et McCain l'ont si bien relevé : "*la bibliométrie est aux publications ce que la démographie est aux populations*" (WHITE et McCAIN, 1989). Le bibliomètre exploite statistiquement des signalements bibliographiques comme le démographe étudie les populations : il n'est pas sensé avoir lu les publications qu'il catégorise et comptabilise, comme le démographe n'est pas sensé connaître les individus qu'il étudie. Heureusement pour le technicien des méthodes bibliométriques, puisqu'il ne pourrait, de toute évidence, lire et synthétiser dans des temps raisonnables les ensembles de documents qu'il analyse !

Les techniques

Il est possible de classer sommairement les techniques selon les méthodes employées. Ces méthodes peuvent être découpées selon la liste suivante :

- *la modélisation des distributions* des éléments bibliométriques : répartition de type coeur/dispersion, loi de Bradford, loi de Lotka, loi de Zipf et unification en une loi universelle (voir p. 29)
- *l'élaboration d'indicateurs univariés*, c'est-à-dire de mesures purement quantitatives basées sur du simple dénombrement ou des calculs de ratio à partir des différents éléments bibliographiques : la date de publication, les revues, les auteurs, les organismes, les pays, les thèmes (voir p. 51)
- *l'élaboration d'indicateurs relationnels*, c'est-à-dire l'exploitation des méthodes d'analyse des données statistiques pour décrire les relations entretenues entre différents éléments bibliographiques : analyses des co-citations, des mots associés, des co-classifications, des co-publications, des co-opérations, des tableaux de contingences (voir p. 65)
- *la modélisation de la diffusion des connaissances* : lois sur la circulation des ouvrages et théories de la communication.

Les domaines d'application

Les trois premières techniques bibliométriques mentionnées ci-dessus sont abordées dans ce travail. La dernière ne sera pas évoquée car ces méthodes statistiques sont plus particulièrement développées à des fins biblio-théconométriques ou pour des problématiques propres aux sciences de la communication.

En ce qui concerne l'application des trois premières techniques bibliométriques, plusieurs domaines sont concernés :

- La sociologie et l'histoire des sciences et des techniques
- L'évaluation de la recherche et des techniques
- La veille technologique et concurrentielle.

Les travaux menés dans le premier domaine sont dans la directe continuité des recherches amorcées par des théoriciens comme Price et Moravcsik. Il est à noter que cette branche d'activité de la scientométrie est en plein déclin de nos

jours. Ce constat s'est vérifié au cours d'une récente conférence internationale[1] : le nombre d'intervenants dans ce domaine était très faible.

Par opposition, le nombre de travaux présentés dans le domaine d'évaluation de la recherche et des techniques a été prédominant au cours de cette conférence. L'élaboration de macro-indicateurs comme aide à la politique des programmes scientifiques et des techniques reste un domaine d'application prépondérant, malgré les débats de fond sur les risques encourus lors de l'interprétation un peu trop hâtive de certains indicateurs. Ces indicateurs sont employés pour avoir des visions macroscopiques[2] de l'activité scientifique et technique. La mise en pratique de ces indicateurs s'amplifie grâce à leur utilisation massive par de nouveaux pays. Cette conférence a marqué l'intérêt grandissant pour ces évaluations institutionnelles de pays comme l'Inde, l'Espagne, le Mexique, les pays d'Amérique Latine, les jeunes nations de l'Europe de l'Est ou certains pays du continent africain.

Les travaux concernant le domaine d'application strictement industrielle (comme l'information brevet) ont été proportionnellement peu nombreux. De plus, les analyses des banques de données de brevets sont effectuées dans un contexte d'évaluation "macroscopique". L'objet de tels travaux répond, comme précédemment, à des demandes provenant d'instituts nationaux d'évaluation qui cherchent à positionner des pays les uns par rapport aux autres selon leurs niveaux en développement d'innovations industrielles. Les expériences de mise en application dans un système d'évaluation concurrentiel à l'échelle d'une entreprise sont pratiquement inexistantes, alors qu'il parait incontournable que de telles études soient réalisées avec régularité et minutie dans les industries de dimension internationale. Cette absence de présentation en public s'explique parfaitement par les contraintes de confidentialité qu'impose ce genre d'étude stratégique pour les industries. C'est pour cette même raison que la partie de cet ouvrage consacrée à ce domaine d'application est relativement réduite et très peu agrémentée par des exemples concrets.

LES ACTEURS FRANÇAIS

En bibliométrie, la France ne détient qu'une place assez modeste à l'échelle internationale. Cette faiblesse est due à l'absence de grands centres de recherche ou d'un grand institut national comme il en existe dans les autres pays. Il faut tout de même noter la création depuis peu d'un Groupement d'Intérêt Public (l'Observatoire des Sciences et Techniques) qui, par sa taille, peut difficilement égaler l'activité d'instituts comme la NSF aux États-Unis ou le CWTS aux Pays Bas. Néanmoins, il ne faut pas minimiser l'action des groupes de travail français

URL

URL

[1] Fourth International Conference on Bibliometrics, Informetrics and Scientometrics, Berlin, 11-15 Septembre, 1993

[2] Les unités de dénombrement et de comparaison dans ces évaluations sont des entités du type continent, communauté économique, nation, région, grande université ou grand institut de recherche.

dans leur contexte national et même leur visibilité internationale pour certains d'entre eux. Les principaux centres français qui mènent des activités de recherche en bibliométrie ou scientométrie sont les suivants :

- CEDOCAR *(CEntre de DOCumentation de l'ARmement)*, 26 bd Victor, 00460 ARMEES (Paris 15ème)
  Paoli C, Longevialle C

Outre l'activité de serveur et de producteur de banques de données spécialisées dans les questions militaires, le CEDOCAR est le promoteur du développement d'une plate-forme de travail bibliométrique. La mise en place de cet instrument de travail a été commanditée par le SGDN[1] auprès du CEDOCAR (PAOLI et alii, 1993). Cette plate-forme, nommée ATLAS[2], intègre les différents outils d'analyse de l'information scientifique et technique élaborés à l'échelle nationale. On y retrouve donc les méthodes développées par le CEMAP, le CERESI, le CRRM, le CSI, l'INIST et l'IRIT[3] de l'Université de Toulouse 3 (DOUSSET et alii 1991, 1993). De ce fait, le CEDOCAR n'est pas lui-même concepteur de techniques bibliométriques mais il a beaucoup contribué à l'aide au développement de certaines méthodes grâce au soutien du SGDN.

- CEMAP *(Centre Européen de Mathématiques APpliquées)*, 68-76 quai de la Rapée, 75592 Paris Cedex 12 ***URL***
  Huot C, Bédécarrax C, Coupet P

Le CEMAP, centre de recherche scientifique d'IBM, met en application l'Analyse Relationnelle dans des études bibliométriques d'analyse de brevets (voir p. 110). Ces études sont proposées sous forme de prestations de service auprès des industriels. Actuellement, l'équipe du CEMAP, en collaboration avec le Centre de Traitement de la Langue Française d'IBM, cherche à mettre au point des analyses statistiques sur des textes en langage naturel. Ces centres d'IBM espèrent pouvoir l'appliquer aux résumés des signalements bibliographiques, et pourquoi pas, aux textes scientifiques dans leur intégralité (WARMESSON et alii, 1993).

- CERESI *(Centre d'Etude et de REcherche en Sciences de l'Information)*, 1 pl. Aristide Briand, 92195 Meudon
  Turner W, Lelu L, Georgel A, de Saint Léger M

Laboratoire du CNRS, le CERESI développe des logiciels d'IST dont les applications vont de l'aide à la documentation jusqu'à la bibliométrie (CARDINE et alii, 1993 ; GEORGEL et alii, 1993). Turner a contribué au lancement de la technique des mots associés et de l'école de pensée française qui en est issue (voir p. 73), ainsi que tout récemment au développement d'une nouvelle méthode d'analyse des données mise au point pour des données

[1] Secrétariat Général de la Défense Nationale, Services du Premier Ministre

[2] Analyses et Traitements Automatiques de la Littérature Scientifique

[3] Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex

bibliographiques (avec le soutien de Lelu) : l'analyse en composantes locales angulaires (GEORGEL, 1992).

- CETIM *(Centre d'Etudes Techniques en Industrie Mécanique)*, 52 av Félix-Louat, BP 67, 60304 Senlis Cedex **URL**
  Devalan P, Belot J-M, Dumas S

Le CETIM, comme tous les CTI[1] , a pour mission d'améliorer les progrès techniques dans son domaine de compétences et de veiller à la parfaite adaptation de ces techniques aux industries adhérentes. Le CETIM contribue à cette mission par de nombreuses actions. L'une de celles-ci est d'assurer une veille technologique adaptée aux besoins de l'industrie mécanique. Les techniques d'analyse statistique et bibliométrique les aident dans cette lourde tâche (voir p. 76).

- Collaboration quadripartite LEPI (Miquel J F) / CERCOA (Gilbert J) / Laboratoire d'information chimique et biologique du Muséum National d'Histoire Naturelle (Doré J C) / CNIC (Dutheuil C)

Ces quatre chercheurs mènent en collaboration ou isolément des travaux en bibliométrie depuis 1979. Spécialistes en chimie, ils appliquent les méthodes bibliométriques pour la chimie et la pharmacologie, que ce soit sur des données bibliographiques (DORÉ et alii, 1987) ou des données factuelles en chimie. Dutheuil a dernièrement intégré la société Synthélabo pour continuer ses activités de recherche en analyse de l'information scientifique en chimie. Miquel, du LEPI[2], a montré un intérêt tout particulier pour l'étude des collaborations internationales dans ses derniers articles (voir p. 88)

- CRRM *(Centre de Recherche Rétrospective de Marseille)*, Université de Aix-Marseille III, 13397 Marseille Cedex 20 : **URL**
  Dou H, Hassanaly P, Quoniam L, La Tela A, Giraud E, Rostaing H

Le CRRM, centre de recherche et de développement de méthodes et de logiciels bibliométriques, est l'un des premiers en France à avoir montré l'utilité et l'opérationnalité des techniques bibliométriques comme outil de veille technologique (DOU et alii, 1990c), (DESVALS et DOU, 1992). Laboratoire universitaire, il assure aussi des formations où ces techniques sont enseignées (DEA Veille Technologique, DESS Gestion d'Information Scientifique et Technique, formation professionnelle). Quelques travaux de cette équipe sont rapportés dans cet ouvrage (voir p 34, 48, 56, 78, 82, 89, 99, 108, 111, 112).

- CSI *(Centre de Sociologie de l'Innovation)*, École des Mines de Paris, 62 bd Saint-Michel, 75006 Paris **URL**
  Callon M, Courtial J P, Penan H, Sigogneau A

Le CSI axe ses recherches sur l'aspect socio-cognitif de la science et des techniques. Callon et Courtial, en collaboration avec Turner sont les "pères

[1] Centres Techniques Industriels

[2] Laboratoire d'Evaluation et de Prospective Internationales, CNRS

fondateurs" de l'école de pensée tournant autour de l'utilisation de l'analyse des mots associés comme méthode scientométrique (voir p. 73). Ils mettent aussi au point d'autres outils méthodologiques pour des études sociologiques du développement de la science (LAREDO et alii, 1993) et ont intégré à leurs outils de travail l'analyse des co-citations grâce à l'arrivée de Penan dans leur équipe (CALLON et alii, 1993).

- ENSSIB *(École Nationale Supérieure des Sciences de l'Information et des Bibliothèques)*, 17/21 bd du 11 Novembre 1918, 69623 Villeurbanne Cedex URL URL
  Lafouge T, Boucher R

Centre de formation des conservateurs de bibliothèques, l'ENSSIB, et plus particulièrement le laboratoire CERSI dont une équipe est spécialisée dans les systèmes statistiques pour la circulation des ouvrages (LAFOUGE, 1991, 1993) et la modélisation de la communication de l'information (voir p. 48).

- INIST *(Institut National d'Information Scientifique et Technique)*, 2 allée du Parc de Brabois, 54514 Vandoeuvre-lès-Nancy Cedex URL
  Polanco X, Grivel L

Ancien centre de documentation du CNRS et maintenant institut à part entière, l'INIST a pour principale vocation la production d'information scientifique (producteur des banques de données *Pascal*, *Francis* et autres services). En matière d'analyse bibliométrique, l'INIST a lancé un programme de recherche (POLANCO et alii, 1993) pour la conception d'une "boîte à outils" de techniques d'exploitation des données provenant de ses propres banques de données. L'INIST offre ses compétences d'analyse sous forme de prestations de services auprès de ses clients. Cette "boîte à outils" est principalement composée de deux méthodes d'analyse bibliométrique : l'analyse des mots associés (renommée SDOC, voir p. 73) et l'analyse en composantes locales angulaires (renommée NEURODOC)

- LERECO INRA-ESR, La Géraudière, BP 527, 44026 Nantes Cedex 03 URL
  Zitt M, Bassecoulard-Zitt E

M et E Zitt ont mis au point une chaîne d'analyse scientométrique (nommée SINBAD) mettant en oeuvre des techniques inspirées à la fois de l'analyse des co-citations et de l'analyse des mots associés, et donc par là même laissant une certaine liberté méthodologique dans l'analyse des références bibliographiques. Ils appliquent bien évidemment cet outil dans l'élaboration d'études pour le compte de l'INRA (BASSECOULARD-ZITT et ZITT, 1993 ; ZITT et BASSECOULARD, 1994 ; voir aussi p. 104).

- OST *(Observatoire des Sciences et des Technologies)*, 93 rue de Vaugirard, 75006 Paris URL
  Barré R, Laville F

L'OST élabore des indicateurs macro-bibliométriques pour décrire les activités scientifiques et techniques régionales et nationales françaises et la position de la France à l'échelle européenne ou internationale (voir p. 86 et 104) ainsi que

(BARRÉ et LAVILLE, 1993). Ces indicateurs sont basés sur des données scientifiques, techniques, technologiques, économiques et sociologiques et sont régulièrement publiés[1] (indiquons la participation de Zitt M pour l'élaboration de la cohérence de ces indicateurs).

Pour compléter cette liste, il convient d'y adjoindre les nombreux acteurs qui mettent en pratique ces techniques bibliométriques comme aide à l'évaluation de leurs spécialités : CNRS (Bauin de l'UNIPS[2], Vergnes & Mossetti de l'IN2P3[3]), INRA (De Looze de SERD[4]), IFP[5] (Moureau & Girard, voir p. 98), LERASS[6] (Jeannin, Devillard et Suraud, voir p. 82)...

Il est à noter l'absence d'entreprise privée dans cette liste. Pour les raisons de confidentialité, déjà évoquées un peu plus haut, les entreprises préfèrent ne pas informer leurs concurrents directs de l'utilisation de ces techniques en espérant garder un avantage. Comme toute citation de société privée employant ces méthodes ne pourrait pas être justifiée par une communication écrite, nous préférons conserver cet anonymat délibéré, mis à part les quelques sociétés qui l'ont affirmé dans des communications publiques : l'Air Liquide (CARDINE et alii, 1993), Atochem (JAKOBIAK, 1994 ; voir p. 98), L'Oréal (NIVOL, 1993 ; voir p. 107 et 108), Elf Aquitaine (BERNAT, 1993), Synthélabo (DUTHEUIL et alii, 1993), Total (DIMO, 1990 ; voir p. 102)

Bien que l'importance de l'activité des techniques bibliométriques en France soit toute relative par rapport à d'autres pays (États-Unis, Pays-Bas, Hongrie, Inde...), il faut remarquer que les groupes de travail énumérés ci-dessus sont bien représentatifs de l'ensemble des axes de recherche en bibliométrie. Ainsi l'aspect bibliothéconométrie est-il assuré par les recherches de l'ENSSIB, celui de la sociologie de la science et des techniques par le CERESI, le CSI, l'INIST et le LERECO, tandis que les indicateurs nationaux d'évaluation de la recherche et des techniques sont élaborés par l'OST et, plus récemment le LEPI, et que l'intégration des techniques à la problématique industrielle est couverte par le CEMAP, le CETIM et le CRRM.

[1] OST, *Science et technologie, indicateurs 1992*, Editions Economica, 1992

[2] Unité d'Indicateurs de Politique Scientifique, Paris ***URL***

[3] Institut National de Physique Nucléaire et de Physique des Particules, Orsay ***URL***

[4] Sociologie et Économie de la Recherche Développement, Université Pierre Mendès France, Grenoble

[5] Institut Français du Pétrole, Rueil-Malmaison ***URL***

[6] Laboratoire d'Etudes et de Recherches Appliquées en Sciences Sociales, Université Paul Sabatier, Toulouse ***URL***

CHAPITRE II
LES DISTRIBUTIONS BIBLIOMÉTRIQUES

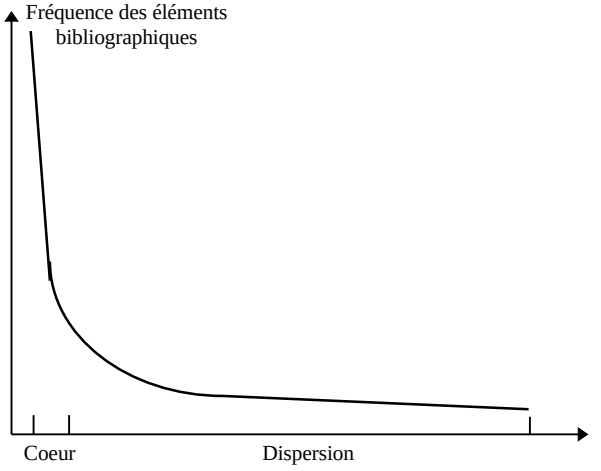
Nous commencerons ce chapitre par les plus anciens travaux de recherche en bibliométrie. Ils font tous appel à la notion de regroupement et de présentation sous forme de distributions des résultats obtenus à la suite de dénombrements bibliographiques. Les chercheurs, après avoir remarqué que le classement de ces valeurs numériques suivaient certaines régularités, ont souhaité découvrir les modèles et les lois reproductibles qui engendrent ces valeurs. Ces approches modélisées permettent de mieux faire connaître ces distributions, et donc de mieux maîtriser leur évolution et leur signification.

LE "COEUR" ET LA "DISPERSION"

La bibliométrie est – rappelons le – une technique basée sur le recensement des travaux scientifiques ayant les mêmes particularités. Ainsi, on peut vouloir connaître la quantité de travaux réalisés sur un sujet donné, publiés à une date précise, rédigés par un auteur ou par un organisme, diffusés par une revue scientifique... Toutes ces propriétés, qui caractérisent les documents comptabilisés, sont présentes dans les références bibliographiques de ces documents. Le plus souvent, les études bibliométriques cherchent plutôt à connaître la variété d'une propriété lorsqu'une autre est déjà connue. Par exemple, on peut vouloir connaître tous les auteurs travaillant sur un sujet, les revues concernées par un thème, les auteurs pris comme référence sur un sujet, les auteurs publiant dans une revue... Toutes les combinaisons de croisements entre ces caractéristiques bibliographiques peuvent faire l'objet d'une étude dès lors qu'elles ont un sens et un intérêt. Lorsque les précurseurs de la bibliométrie

ont obtenu les premiers résultats sur les variétés des éléments que procure la combinaison de deux caractéristiques bibliographiques, ils ont immédiatement cherché à les classer. En classant ces éléments par ordre de fréquence d'apparition décroissante, ils ont construit les premières distributions statistiques en bibliométrie. Ces dernières suivent pratiquement toutes le même mode de répartition et ont le même aspect graphique (seules les distributions qui font intervenir comme unité de comptage le temps, par exemple la date de publication, ne suivent pas la même loi de distribution, voir p. 54). Ces distributions sont souvent présentées comme comportant une partie nommée le "coeur" (*core)* et une autre la "dispersion" *(scatter*). Cette appellation est suggérée par la forme de représentation graphique des distributions (figure 1).

Fig. 1 - Coeur et dispersion d'une distribution bibliométrique



Éléments bibliographiques triés par ordre de fréquence décroissante

- le **coeur** représente le groupe d'éléments qui apparaissent le plus fréquemment dans l'ensemble des références bibliographiques étudiées. Par exemple, dans le cas de la loi de Lotka, ce coeur symbolise les auteurs les plus prolifiques dans un domaine donné (voir p. 40).
- la **dispersion** représente les nombreux autres éléments à basse fréquence dans l'ensemble des références bibliographiques étudiées. Dans le cas de la loi de Lotka, la dispersion correspond à la grande diversité des auteurs qui ont très peu publié sur ce même domaine (voir p. 40).

Bien qu'elle n'ait pas toujours été sujette à des travaux de modélisation, une bonne partie des éléments bibliographiques suit ce type de concentration en forme de coeur et de dispersion. Dans le cas des lois de Bradford, de Lotka et de Zipf, la répartition de la variété d'une catégorie d'éléments bibliographiques (réciproquement les périodiques, les auteurs, les mots) est étudiée pour un sujet précis. En définitive, ces lois mettent en évidence la distribution des co-

occurrences entre les éléments et le sujet. Mais d'autres études, relatées par White et McCain (WHITE et McCAIN, 1989), présentent des systèmes de concentrations équivalents alors qu'elles ne sont pas forcément établies à partir du choix d'un sujet (voir encadré ci-dessous). Elles montrent que les distributions présentant un coeur et une dispersion sont des phénomènes caractéristiques de la plupart des données bibliographiques.

De façon générale, le coeur des termes présents dans les titres des articles est induit par celui qui les génère : Pour un auteur ce coeur représente ses thèmes de prédilection, pour un périodique les sujets principaux traités par ses articles, pour un sujet d'étude les termes du thème ou les termes des thèmes connexes.

- *Les revues concentrent les termes des titres :*

Paisley (PAISLEY, 1986) a analysé 300 titres d'articles venant de 6 revues, les 5 premiers termes pour deux de ces périodiques sont :

Journal of American society for information science

information	126
science	59
system	49
retrieval	41
searching	36

Public opinion quarterly

polling	31
opinion	20
public	19
survey	18
media	17

La redondance entre les mots du nom de la revue et les mots des titres est remarquable.

- *Les descripteurs concentrent les termes du titre :*

Etude de Lawson et al. sur le thème de l'analyse énergétique (LAWSON et alii, 1980). Parmi 349 publications, les termes des titres les plus présents sont :

energy	97
energy analysis	50
energy cost(s)	28
energy equipment	20
energy use	17

Les résultats de cette étude se rapprochent de ceux de l'étude de Zipf (voir p. 42 et suivantes).

- *Les auteurs cités concentrent les revues :*

White (WHITE, 1981a), cherchant sur le sujet *"Prehistoric great basin ecology"*, trouve 21 articles citant conjointement J. Steward et D.H. Thomas, deux auteurs en archéologie spécialistes de la région Utah-Nevada. Les principaux journaux qui les avaient publiés étaient :

American antiquity	8
Annual review of anthropology	3
American anthropology	2
Journal of anthropological research	2

Ceci montre qu'une recherche d'article par co-citation concentre les journaux où sont publiés les articles de ces auteurs.

- *Les auteurs concentrent les auteurs cités :*

Lors d'une étude de citations, Lawani a trouvé que les citations des auteurs étaient le plus souvent des auto-citations, des citations de collaborateurs, des citations cachant une relation étudiant-enseignant et des citations des co-auteurs de l'article (LAWANI, 1982). Toutes ces citations font intervenir des liens sociaux entre les auteurs. Lorsque ces relations entre auteurs, qui n'appartiennent pas forcément aux mêmes organismes de recherche, forment une certaine cohésion, ce noyau d'auteurs constitue ce que l'on nomme un "collège invisible" (*invisible college*).

- *Les revues concentrent les revues citées :*

Garfield (GARFIELD, 1979) a remarqué que les auteurs qui publient dans une revue ont tendance à citer des articles du même périodique. Il a aussi remarqué que les journaux cités venant ensuite ont des thèmes proches de ceux abordés par les articles du journal. Il donne l'exemple des articles du *Journal of experimental medicine* qui citent, en second après lui, le *Journal of immunology*. Ce qui révèle que les articles du *Journal of experimental medicine* abordent plutôt les thèmes de l'immunologie que ceux de la médecine clinique.

- *Les références citées concentrent les descripteurs :*

White et Griffith ont comptabilisé les descripteurs des articles co-cités pour 18 concepts en science de la connaissance médicinale (WHITE et GRIFFITH, 1987). Leur bibliographie étant tirée de la banque de donnée *Medline*, pour les 5 plus grands articles co-cités concernant la culpabilité des étudiants au sujet du sexe, les principaux descripteurs étaient:

guilt	5
sex behaviour	4
analysis of variance	3
arousal	3
personality	3

- *Les références citées concentrent les termes des titres des articles citants :*

C'est le principe même de la prospection des "fronts de recherche" (*research fronts*) utilisée par l'Institute for Scientific Information (voir p 67). Small et Griffith donnent une illustration de cette concentration par l'étude des termes des titres pour les documents qui citent un livre de Lederer sur la physique du nucléaire et des particules (SMALL et GRIFFITH, 1974). Les principaux termes sont:

decay	12
reactions	7
isotopes	5
neutron activation	5

- *Les descripteurs concentrent les descripteurs :*

Pour un descripteur qui apparaît au moins dans trois articles dans une banque de données, les descripteurs qui co-apparaissent avec lui forment une distribution produisant un coeur et une dispersion. L'utilisation des termes de ce coeur pour retrouver d'autres documents concernés par le même sujet que ceux trouvés avec le descripteur initial, est une technique très connue pour améliorer les recherches en ligne. Martin a décrit une recherche sur la banque *Inspec* des articles sur le sujet *growth of crystals under weightlessness* en commençant uniquement sa requête par les descripteurs *gravity* et *crystal(s)* (MARTIN, 1983). Pour les 103 documents retenus, l'emploi de la commande statistique en ligne lui a permis de connaître les autres principaux descripteurs utilisés dans ces documents. En ré-interrogeant avec ces nouveaux descripteurs, il a obtenu ces 4 principaux termes lors de sa nouvelle commande statistique:

zero gravity experiments	60
solidification	20
crystal growth	16
crystallisation	15

Ce qui lui a permis de savoir de quelle manière les indexeurs ont exprimé le concept dans la base, sans employer de thesaurus.

Récapitulatif :

White et McCain ont ensuite récapitulé les relations de concentration par la table suivante et ils ont précisé qu'elle n'était probablement pas exhaustive.

Les auteurs prolifiques concentrent :

- leurs propres termes de titres
- leurs propres auteurs cités

- les revues auxquelles ils ont contribué
- les descripteurs affectés à leurs propres travaux

Les termes de titre concentrent :

- les noms d'auteurs
- d'autres termes de titres
- les citations des références
- les revues
- les descripteurs

Les références fortement citées concentrent :

- les auteurs des citations
- les titres des citations
- les références co-citées
- les descripteurs

Les revues concentrent :

- les noms d'auteurs
- les termes de titres
- les références citées
- les descripteurs

Les descripteurs concentrent :

- les noms d'auteurs
- les termes de titres
- les références citées
- les revues
- d'autres descripteurs

Quel sens donner aux distributions qui présentent ce phénomène de coeur et de dispersion ? Les exemples évoqués ci-dessus illustrent bien les significations que les auteurs ont données à ces deux zones : *le coeur entretient l'identité, la redondance, tandis que la dispersion contient l'individualisation, la variété.* Dans le premier exemple, les termes du titre des articles d'une revue contiennent principalement les termes du nom du périodique. Ceci indique bien la volonté d'entretenir l'identité du périodique à travers les titres des articles publiés que ce soit par les auteurs ou par l'éditeur. Un autre exemple caractéristique est celui des auteurs des références citées par un auteur. Un auteur reconnu cite principalement ses propres travaux, ce qui vérifie parfaitement que le coeur entretient l'identité. De plus, les autres auteurs fréquemment cités ont tous un lien social avec l'auteur. Ils entretiennent là encore une identité sociale. Moed et Van Raan ont même précisé (MOED et VAN RAAN, 1986) que les auteurs cités faisant partie de la dispersion sont souvent des chercheurs que l'auteur connaît uniquement intellectuellement.

Tous les exemples peuvent être facilement expliqués par cette règle d'identité. Mais pour quelles raisons les divers acteurs de la communication de

la connaissance scientifique auraient-ils tendance à la suivre ? De nombreux auteurs se sont penchés sur cette question, et notamment Nelson et Tague suivis de Price et de Weinberg. Ils ont donné deux explications opposées :

1) l'avantage du cumul (*cumulative advantage*) : plus un élément bibliographique est à forte fréquence, plus il sera réutilisé souvent (voir p. 45)

2) la spécialisation : plus un élément bibliographique est fréquent, moins il contient d'information, d'où une plus grande probabilité pour qu'il soit rejeté en faveur d'un mot plus spécifique ou nouveau, plus adapté.

Initialement, ces explications ont été données pour les mots qu'un auteur emploie dans ses publications et peut très bien s'appliquer à tous les autres types d'éléments bibliographiques précédemment récapitulés. Ils agissent, comme le concept du Yin et du Yang, de façon antagoniste : l'avantage du cumul tend vers l'effet de coeur et la spécialisation vers l'effet de dispersion. En fait, ce phénomène n'est pas encore bien compris.

Les collèges invisibles (voir p. 81) expriment le phénomène de coeur et de dispersion dans un exemple de perversion de la science. Deux modes de comportement peuvent se présenter chez les chercheurs :

1) la création d'un effet de coeur en ne cherchant, lisant et parlant qu'avec le même groupe de travail, ce qui aboutit au collège invisible, au monopole d'une revue ou d'une certaine terminologie.

2) la création de l'effet de dispersion en balayant continuellement les nouveautés et les différentes personnes et idées, ce qui provoque une extension continuelle du réseau de travail pour, en fait, éviter toute redondance.

La conduite idéale est celle où le chercheur garde en équilibre ces deux modes de comportement.

La plupart des auteurs proposent un découpage naturel des distributions en deux zones, le coeur et la dispersion. Un découpage en trois ou quatre zones a été proposé récemment par Quoniam (QUONIAM, 1992). Ce chercheur estime qu'il faut considérer deux grands groupes d'éléments bibliographiques : ceux qui sont issus de champs bibliographiques ayant un vocabulaire contrôlé, et ceux qui sont issus de champs bibliographiques ayant un vocabulaire libre. Selon que les éléments bibliographiques appartiennent à l'un ou l'autre de ces deux groupes, leurs distributions sont découpées soit en trois zones (figure 2) soit en quatre zones (figure 3).

Il propose de considérer comme champs comportant un vocabulaire contrôlé les champs suivants : auteurs, affiliations, sources, citations, codes de classification documentaire et les descripteurs d'indexation. Les champs restants sont donc considérés comme comportant un vocabulaire libre : les titres et les résumés. Les champs à vocabulaire libre sont ceux qui sont constitués d'un texte en langage naturel (ou approchant). L'auteur donne la signification d'un tel découpage :

"*1) L'information triviale* (zone I) *est celle qui définit les thèmes centraux du corpus...*

2) Le bruit (zone III) *est caractéristique, soit de concepts non encore émergents (il est impossible de dire s'ils sont porteurs ou non), soit d'erreurs...*

3) Entre les deux, se situe l'information intéressante (zone II) *qui montre soit des thèmes périphériques oubliés, soit de l'information potentiellement innovante. C'est là que les transferts de technologie, des thèmes nouveaux sont éventuellement envisageables...*

4) Dans le cas d'un vocabulaire libre, à plus forte fréquence, apparaît une zone supplémentaire (zone A) *contenant les mots vides de sens...*".

Fig. 2 - Distribution d'un vocabulaire contrôlé

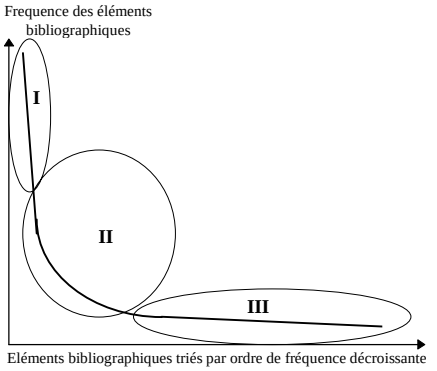
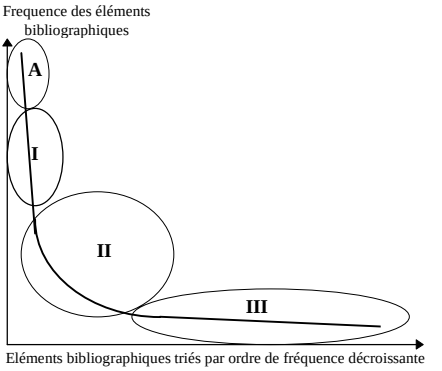


Fig. 3 - Distribution d'un vocabulaire libre



Ce découpage des distributions a été inspiré de l'approche linguistique de l'analyse des mots contenus dans un texte. Cette approche appliquée aux données bibliométriques impose de différencier les situations selon le mode de création du champ à traiter. L'auteur avoue qu'un tel découpage est encore purement intellectuel. Il ne dispose pas de "*critères objectifs pour le calcul des frontières entre les différentes zones*", mais il n'élimine pas l'éventualité de posséder un jour des techniques qui le permettent. Il espère pouvoir appliquer des caractéristiques de la théorie de l'information de Shannon et Weaver.

Les chercheurs, après avoir découvert que certains types de comportements bibliographiques suivaient des règles, ont alors cherché à connaître ces règles pour les exprimer sous la forme de lois et de modèles. En bibliométrie, ces lois s'intéressent principalement aux relations qui existent entre une quantité de sources et une productivité. Elles ne sont en aucun cas analogues à celles de la physique car elles n'expliquent pas le phénomène bibliographique. Elles ne font que le représenter. Certains, comme Brookes, ont cherché à les exprimer de manière à ce que l'on puisse les utiliser dans des situations pratiques, tandis que d'autres ont étudié leur formulation et leur similitude avec les distributions statistiques standards. Les bibliomètres ont développé de nombreuses techniques pour traiter ces distributions. Celles-ci font très souvent appel à un regroupement des éléments étudiés sous forme de rangs : les statisticiens parlent

alors de statistiques non paramétriques. Elles s'écartent en cela des statistiques classiques.

Nous allons donc évoquer dans les paragraphes suivants ces travaux, mais les descriptions des concepts seront présentées sans la rigueur mathématique et statistique auxquelles nous devrions faire appel. Ces lois seront essentiellement discutées en termes intuitifs.

LA LOI DE BRADFORD

Les travaux de Bradford

Bradford considérait que l'activité de gestionnaire de bibliothèque est soumise au "*chaos documentaire*" de la littérature (BRADFORD, 1948). L'un des problèmes qui se posait à Bradford était le suivant : s'abonner à tous les périodiques concernant un domaine reviendrait trop cher, aussi a-t-il pensé à sélectionner parmi tous les périodiques ceux qui seraient les "meilleurs" représentants du domaine. Un article ne parle pas uniquement d'un thème mais, bien souvent, il touche plusieurs domaines en même temps. En se basant sur ce fait, Bradford admet que des périodiques puissent contenir des articles ne concernant pas uniquement le sujet de prédilection du périodique. Les périodiques ont bien souvent des articles qui peuvent intéresser plusieurs spécialités.

Bradford voulait donc pouvoir connaître le "noyau" (*nucleus*) des périodiques cernant le mieux un sujet. Il voulait pouvoir ranger les périodiques en "zones" dégressives de productivité, en fonction de leurs proportions d'articles traitant du sujet donné. Il peut paraître normal que le nombre de périodiques dans chaque zone augmente alors que la productivité diminue. Il formula cette réflexion sous forme mathématique (voir encadré ci-contre) et en déduisit une conclusion qu'il formula ainsi : "*Si les périodiques scientifiques sont rangés par ordre décroissant de productivité sur un sujet donné, ils peuvent être divisés en un noyau de périodiques plus particulièrement reliés au sujet et en plusieurs groupes contenant le même nombre d'articles que le noyau, quand les nombres de périodiques dans le noyau et dans les zones successives suivent la série : 1 ; n ; n² ...*".

Les zones des périodiques jouent le même rôle qu'une famille avec ses générations successives ayant des liens de parenté de plus en plus faibles. Chaque génération est plus importante en nombre que la précédente, et chaque élément d'une génération a un lien de parenté avec le "noyau" inversement proportionnel à son degré de parenté.

Bradford proposa que cette distribution puisse se formuler mathématiquement ainsi :

soit m = nombre d'articles dans le noyau
m_1 = nombre ..... la deuxième zone...
m_x = nombre ..... la $x^{ième}$ zone

et p = nombre de périodiques dans le noyau
p_1 = nombre ..... la deuxième zone...
p_x = nombre ..... la $x^{ième}$ zone

et r = nombre moyen d'articles par périodique dans le noyau
r_1 = nombre ..... la deuxième zone...
r_x = nombre ..... la $x^{ième}$ zone

on a $r = m/p$, $r_1 = m_1/p_1$, ..., $r_n = m_x/p_n$

comme $p < p_1 < ... < p_x$ et $r > r_1 > ... > r_n$

on peut imaginer de choisir les zones de façon à obtenir
$p.r = p_1.r_1 = ... = p_x r_x$ ($1^{ère}$ hypothèse)

Par conséquent
$p_1/p = r/r_1 = n_1$
$p_2/p_1 = r_1/r_2 = n_2$...
où n_1, n_2, ... sont des constantes

Alors on peut écrire
$p_1 = n_1.p$
$p_2 = n_2.p_1 = n_1.n_2.p$
...

On peut encore choisir d'établir les zones de façon à avoir
$n_1 = n_2 = ... = n_x = n$ (2^{nde} hypothèse)

d'où $p_1 = n .p$
$p_2 = n^2.p$...
$p_x = n^x.p$

Bradford montra la vraisemblance de cette loi à partir de données expérimentales et d'une représentation graphique de ces données. Il constitua deux séries de données expérimentales qui sont le résultat de deux bibliographies réalisées en consultant la collection des résumés de périodiques de *Science Library*. L'une concernait le sujet *applied geophysics* pour la période 1928-31; l'autre sujet, *lubrification*, couvrait la période 1931-juin 1933. Selon sa loi les zones ont, pour un nombre d'articles constant, un nombre de publications croissant avec l'exponentielle d'une constante *n* (nommée coefficient multiplicateur de Bradford). Cette relation suggère que la somme cumulée des articles est proportionnelle au logarithme du nombre de publications correspondantes. Il constitua pour les deux séries une table qui comportait 5 colonnes :

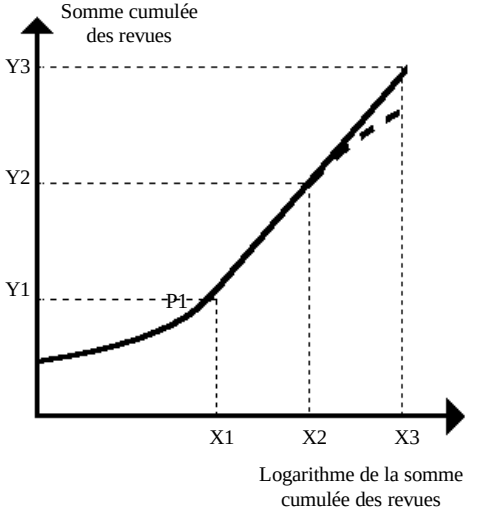
1) la colonne A contient le nombre de revues produisant le même nombre d'articles
2) la colonne B, le nombre d'articles correspondant
3) la colonne C, la somme cumulée du nombre de journaux en colonne A
4) la colonne D, la somme cumulée du produit du nombre d'articles en deuxième colonne, multipliée par le nombre de revues en colonne A
5) la colonne E, le logarithme de la valeur présente dans la colonne C.

Tabl. 1 - Table des données expérimentales collectées par Bradford

A	B	C	D	E
n_1	f_1	$c_1=n_1$	$d_1=f_1*n_1$	$\log(c_1)$
n_2	f_2	$c_2=c_1+n_2$	$d_2=d_1+f_2*n_2$	$\log(c_2)$

La première ligne du tableau comporte la revue (ou les revues) qui a (ou ont à égalité) la plus forte productivité d'articles dans le domaine étudié, la seconde ligne la ou les revues étant en seconde position etc... On a bien un regroupement des revues par rang de nombre équivalent d'articles. A partir des deux tableaux construits selon ce principe, Bradford à construit la courbe logarithmique en distribuant les points selon les valeurs de la colonne E en abscisses et les valeurs de la colonne D en ordonnées (figure 4). Lorsque la courbe est tracée, les points sont presque alignés sur une droite et donc elle vérifierait la progression régulière formulée mathématiquement. Seule la première partie des courbes ne suivait pas cette loi exponentielle. Bradford conclut que le cumul des articles pour un sujet donné est proportionnel au logarithme du nombre de producteurs concernés, quand ceux-ci sont classés par ordre de productivité. Il précisa que ceci se trouvait vérifié sur tous les articles mis à part ceux produits par la première zone de production.

Fig. 4 - Courbe construite lors de l'étude de Bradford



Il a construit trois zones pour ces données séparées par deux frontières arbitraires. La première frontière est déterminée par le point P1 de la courbe où commence la partie linéaire. Puis la seconde frontière a été fixée en reportant la distance Y1 qu'il avait entre l'origine et la première frontière. Il obtient donc trois zones où l'accumulation d'articles entre elles est constante, la première zone définissant le noyau des "meilleurs" producteurs concernant le sujet. Mais

la question de la définition de ces zones pour un sujet quelconque a été posée par Bradford et est restée sans réponse.

Vérification de la loi

Des études bibliographiques dans de nombreux domaines ont confirmé que la dispersion des articles d'un ensemble de périodiques est conforme à la distribution statistique proposée par Bradford : LAWANI, 1973 ; ALABI, 1979 ; AIYEPEKU, 1977... L'allure générale de la loi est correcte, mais certains détails animent encore, de nos jours, des débats passionnés :

- *Le nombre de zones et le coefficient multiplicateur n* : des auteurs ont noté la difficulté de distinguer le noyau des revues de la dispersion dans une distribution de Bradford. Bradford découpa arbitrairement la courbe en trois zones équivalentes et trouva le coefficient multiplicateur égal à 2. Mais McCreery et Pao divisèrent la littérature en ethnomusicologie en 14 zones en ayant comme coefficient moyen n = 1,63 (McCREERY et PAO, 1984). Wallace découpa la littérature "dessalement" en 10 et trouva un coefficient de 1,8 (WALLACE, 1986). Pontigo et Lancaster firent le découpage de la littérature sur la bactérie méthagonique en 4 parties avec n = 3,68 (PONTIGO et LANCASTER, 1986). De nombreuses méthodes moins empiriques pour déterminer ces zones ont été discutées par Egghe, Brookes et Rousseau mais toutes nécessitent forcément l'introduction d'une donnée arbitraire. Seule la réflexion de Bradford "*qu'une relative petite proportion de revues peut satisfaire la requête d'une grande proportion d'articles sur un sujet*", paraît pertinente.
- *La formulation mathématique* : la volonté de trouver la formulation mathématique qui s'ajuste le mieux aux données expérimentales a été la source d'un farouche débat pendant de longues années. L'expression mathématique qu'avait donnée Bradford a été maintes fois controversée. Dans un premier temps, Leimkuhler reprit la loi établie par Bradford pour en proposer une expression générale (LEIMKUHLER, 1967). L'année suivante, Brookes montra son désaccord sur la formulation donnée par Leimkuhler et livra une formulation plus simple dressée à partir des représentations graphiques données par Bradford (BROOKES, 1968). En 1972, Wilkinson montra que leur discorde venait du fait qu'ils étaient partis des deux formulations données par Bradford (WILKINSON, 1972). Or Bradford avait fait une erreur dans son analyse algébrique et sa formulation verbale était incorrecte. En 1948, Vickery avait déjà noté (VICKERY, 1948) qu'il n'y avait pas concordance entre sa représentation graphique et sa formulation verbale. Wilkinson trouva que la formulation graphique était plus proche des données empiriques que de l'expression verbale. Plus récemment, en 1984, Maia et Maia ont su donner finalement une formulation mathématique de la loi de Bradford satisfaisante bien que relativement complexe (MAIA et MAIA, 1984). Elle est proche de

celle de Brookes, à la différence qu'elle ne décrit pas seulement la partie rectiligne de la courbe (R(n) = k Log n), mais la totalité de la distribution.

• *L'affaissement de Groos (Groos droop)* : une autre remarque a été faite en 1967 par Groos en ce qui concerne la fin de la distribution telle que l'avait présentée Bradford (GROOS, 1967). Il nota que la courbe a alors une tendance à s'affaisser (partie en pointillé sur la figure 4). Cette nouvelle partie est appelée l'affaissement de Groos (*Groos droop*) en l'honneur de son inventeur. Plusieurs interprétations ont été données, mais celle qui est la plus communément acceptée estime que cet affaissement reflète le caractère d'une bibliographie réalisée de façon incomplète pour le domaine étudié.

Bien que la formulation mathématique et graphique soit encore un sujet de discussions, la proposition de Bradford a été retenue depuis longtemps. Cette loi présente parfaitement la tentative de modélisation d'une répartition de type coeur/dispersion basée sur la fréquence d'articles abordant un domaine scientifique parmi les revues.

LA LOI DE LOTKA

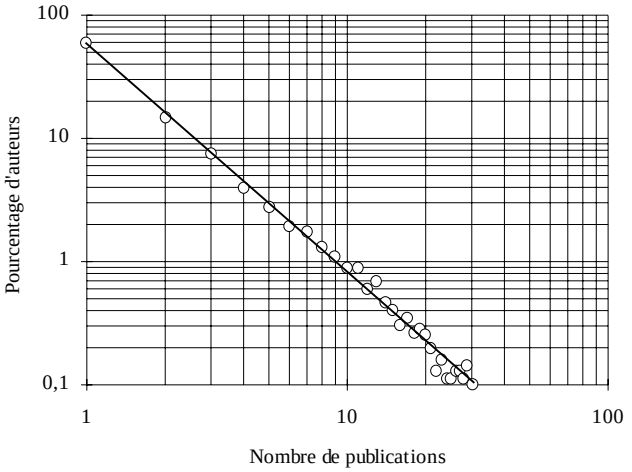
Les travaux de Lotka

Il a semblé intéressant à Lotka de déterminer la part de contribution de chaque chercheur au progrès de la science. Il a proposé cette réflexion pour la première fois dans un article publié en 1926 (LOTKA, 1926). Il a mis en application son idée dans le domaine de la chimie. Pour cela, il a comptabilisé le nombre d'entrées de l'index du *Chemical Abstracts 1907-1916* pour tous les auteurs dont le nom commence par les lettres A et B. Puis, il a cumulé tous les auteurs qui n'avaient qu'une entrée (c'est-à-dire les auteurs présents uniquement dans un article collecté par le *Chemical Abstracts Services*), puis ceux qui en avait 2, puis 3, etc. La même manipulation a été aussi faite avec l'index des auteurs du journal *Auerbach's Geschichtstafeln der Physik 1910.* Ceci correspond à un regroupement des auteurs par rang de fréquence de publications.

Lotka a d'abord présenté ses résultats sous la forme d'un histogramme construit en distribuant le pourcentage d'auteurs (en ordonnée) en fonction du nombre d'articles publiés. Il a obtenu pour les deux séries de données expérimentales des distributions pratiquement similaires dont la forme générale est celle de la courbe de type coeur/dispersion après inversion entre les deux axes (voir figure 1 p. 30). C'est-à-dire, très peu d'auteurs ont publié un grand nombre d'articles (le coeur) et une grande variété d'auteurs ont très peu publié. Puis les mêmes résultats ont été présentés selon des échelles logarithmiques pour les deux axes (figure 5). Les points étant pratiquement alignés selon une droite, il était possible d'estimer que ces résultats suivaient la relation $y = x^n . C$, où x est le nombre de publications, y le nombre (ou pourcentage) d'auteurs ayant x publications, n la pente de la droite et C une constante (dont la valeur est

égale à l'intersection de la droite avec l'axe des *y*, c'est-à-dire le nombre d'auteurs n'ayant publié qu'une fois). La pente de la droite étant pour ces deux séries de valeurs proches de la valeur -2 (-2,021 ± 0,017 et -1,888 ± 0,007), il a obtenu la relation empirique : $y = C / x^2$. Cette relation peut s'exprimer par la locution suivante : *par rapport au nombre d'auteurs qui publient 1 article (y = C), 4 fois moins d'auteurs en publient deux (y=C/4) , 9 fois moins en publient trois (y=C/9) , ... n² fois moins en publient n (y=C/n²)*. On a donc une productivité scientifique qui diminue selon une loi de type carré inverse. Lotka n'a pas pu tester son modèle à partir d'autres périodiques car ceux-ci n'avaient pas d'index des auteurs.

Fig. 5 - Distribution de Lotka sous sa forme logarithmique



Controverse sur le modèle Lotka

Price a précisé (PRICE, 1963) que la formule de Lotka tendait à surestimer le nombre d'auteurs à forte productivité. En fait, les données que Price a collectées ont montré que le nombre de personnes à plus forte productivité a une diminution plus proche de l'inverse du cube que de l'inverse du carré. Plusieurs autres études sur la validité de la loi de Lotka ont produit des résultats négatifs (RADHAKRISHNAN et KERNIZAN, 1979 ; VOOHS, 1974). Mais Subramanyam attribue ces résultats à la mauvaise sélection de l'échantillon ou à une négligence relative aux articles à auteurs multiples (SUBRAMANYAM, 1979). Murphy a même montré que la littérature de l'humanité en général suivait une loi de type Lotka (MURPHY, 1973). Cette étude a été, elle aussi, contestée par Coile qui a testé statistiquement les résultats de Murphy (COILE, 1977). Sur des données expérimentales provenant de références brevets, la répartition des inventeurs déposant des brevets dans un domaine technique précis suit aussi parfaitement la formule générale $y = C / x^n$. Par contre, la valeur de l'exponentiation varie selon les domaines techniques étudiés. Par exemple, les

régressions linéaires d'une telle courbe appliquée aux inventeurs issus de deux ensembles de brevets ont donné une pente de -2,71 pour un domaine de la pharmacologie et une pente de -3,04 pour un domaine de l'électroménager (ROSTAING, 1993).

En conclusion, l'universalité de la formule de Lotka est peut-être à remettre en cause. Mais cette étude est reconnue dans le monde de la bibliométrie sous le nom de loi de Lotka. La présentation générale et la régularité de la distribution selon une répartition hyperbolique ne sont plus remises en cause. Seule l'exactitude de la formulation mathématique reste un sujet de discussion.

LA LOI DE ZIPF

Les travaux de Zipf

Zipf a repris une idée qui avait déjà été exposée en 1919 par Estoup dans *Gammes sténographiques* dont il a étendu grandement la portée. Parmi tous les sujets traités (ZIPF, 1949), il a cherché à répondre à la question suivante : *à quelle fréquence les mots apparaissent-ils dans un texte littéraire ?* Zipf a comptabilisé les occurrences des 29 899 mots différents trouvés dans *Ulysses* de Joyce. Il les a classées par ordre décroissant de fréquence et il a affecté à chaque mot un rang, de 1 pour le mot le plus fréquemment apparu jusqu'à 29 899 pour le mot le moins fréquemment apparu (voir tableau 2). Il a créé une troisième colonne en multipliant la valeur de chaque rang *r* par la valeur de la fréquence correspondante *f* et il a obtenu un produit *C* proche d'une constante pour l'ensemble de la liste de mots. D'où la formule : $f \bullet r = C$. Zipf a nommé cette loi *"le principe du moindre effort"* car l'analyse d'un tel classement suggérait la réflexion que l'être humain choisit et utilise plus facilement des mots familiers que des mots insolites par pure paresse. Donc la probabilité d'occurrence d'un mot familier est bien plus élevée que celle des autres mots dans les discours de tous les jours comme dans les discours plus soignés.

Tabl. 2 - Echantillon des valeurs étudiées par Zipf

Rang (I) (r)	Fréquence (II) (f)	Produit de I et II (r * f = C)
10	2.653	26.530
30	926	27.780
50	556	27.800
100	265	26.500
300	84	25.200
500	50	25.000
1.000	26	26.000
3.000	8	24.000
5.000	5	25.000
10.000	2	20.000
20.000	1	20.000
29.899	1	29.899

Il n'a pas proposé de représentation graphique de sa loi, mais en restant dans le même ordre d'idée que les précédentes lois, il est facile d'imaginer que la loi de Zipf s'applique parfaitement au modèle de distribution de type coeur/dispersion (voir figure 1 p. 30). Selon la loi de Zipf, cette courbe suivrait la forme *1 / x*, à ceci près que la courbe est continue là où l'histogramme est discontinu. Cette précision est valable pour toutes les lois bibliométriques auxquelles les auteurs appliquent abusivement des formules mathématiques continues alors que les données sont discrètes.

Formulation mathématique

Là encore cette loi a donné lieu à la formulation de divers modèles mathématiques par de nombreux auteurs. Chacun relève que la loi de Zipf ne s'ajuste pas correctement pour les fréquences faibles comme pour les fréquences élevées. Les nouvelles formulations ont toutes pour objectif d'améliorer la représentation des données empiriques.

Une première approche mathématique encore assez simple a été donnée par Fairthorne (FAIRTHORNE, 1969). Sa description de la distribution de Zipf est la suivante : "*1/2 du nombre total des mots sont à fréquence d'occurrence de 1 ; 1/6 sont à fréquence d'occurrence de 2 ; 1/12 sont à fréquence d'occurrence de 3 ; 1/20 sont à fréquence d'occurrence de 4 et ainsi de suite...*" Donc le ratio $1/n \bullet (n+1)$ donne la fraction du nombre total de mots différents apparaissant à la fréquence n.

Bien plus tard, Fedorowicz a présenté, puis appliqué une formulation de la loi de Zipf faite par Booth et inspirée de l'approche de Mandelbrot (FEDOROWICZ, 1982). Il a proposé une formulation qui améliore celle de Zipf pour les mots à très faibles fréquences d'occurrences. En raison de l'abondance des mots qui apparaissent rarement, de nombreux termes devraient avoir le même rang. La formule générale de Booth divise la distribution en groupes de fréquences (ou zones). Chaque groupe G_m est égal au nombre d'occurrences des mots compris dans la gamme de fréquences 2^{m-1} à 2^m-1, m étant un entier positif, selon la formule $G_m = (kT)^{1/ß} [1/(2^{m-1})^{1/ß} - 1/(2^m)^{1/ß}]$, *T* étant la longueur du texte et *ß* une constante > 0. Fedorowicz a testé cette formulation sur la banque de données *Medline* pour les fichiers d'index des champs dont le contenu est en langage libre (titre et résumé). Cette étude lui a permis de vérifier que la relation entre le fichier index (mots et adresses dans le fichier *postings* associé) et le fichier *postings* (liste des références correspondant à chaque entrée c'est-à-dire à chaque mot) est bien du type Zipf. Il a voulu, par cette modélisation, aider les serveurs à mieux gérer leur stockage d'information par des prédictions de quantité d'information pour chaque champ en fonction du nombre de journaux examinés, de la période de temps etc. Ceci pour améliorer le compromis entre la taille de la base et le temps de réponse à la recherche.

Une dernière formulation de la loi de Zipf, inspirée de la théorie de l'information, est principalement le fruit du travail de Mandelbrot. Celui-ci a

proposé une formule plus générale et ajustée (MANDELBROT, 1953) selon la relation : $f(r) = k.(r+c)^{-\mu}$, $f(r)$ étant la fréquence d'un mot et r le rang de ce mot. La constante c améliore l'ajustement pour les mots communs dont les rangs sont peu élevés. L'exposant $-\mu$ améliore l'ajustement pour les rangs très élevés qui correspondent aux mots rares. Pour la plupart des langages naturels, μ est généralement plus grand que 1. Les langages ayant des contraintes de vocabulaire ou utilisant des règles d'usage ont un μ inférieur à l'unité. Pour $\mu = 0$ ceci indiquerait que tous les termes sont employés en moyenne aussi souvent les uns que les autres.

La loi de Zipf n'est pas fondée sur des principes bibliométriques mais sur des principes linguistiques. Malgré cela, elle est devenue incontournable en bibliométrie puisque de nos jours de nombreux groupes de recherche s'orientent vers l'exploitation de texte en langage naturel présents dans les banques de données. Néanmoins, il reste beaucoup à faire dans cette nouvelle perspective, car la formulation mathématique ne permet pas d'affecter systématiquement les mots selon leur rang à des ensembles différents donnant une idée sur l'apport statistique plus ou moins grand de chacun des mots. Or, c'est le problème que rencontre tout bibliomètre lors de l'étude d'un texte en langage naturel. Les outils linguistiques permettant de faire des regroupements sémantiques n'étant pas parfaitement adaptés au type de vocabulaire rencontré dans les textes scientifiques ou techniques, ces textes génèrent une trop grande variété de termes. Ce qui diminue l'impact statistique de chacun des termes. Reste à sélectionner les mots qui sont les plus représentatifs des textes initiaux. Il faut donc réaliser un découpage en zones et donner un sens statistique à chacune d'elles. Ceci nous rapproche du problème du sens de ces distributions et de leur exploitation en statistiques (voir p. 33-35).

UNIFICATION DES LOIS

Les lois de Bradford, Zipf, Lotka ont été formulées indépendamment pour expliquer des phénomènes disparates, mais leur ressemblance laisse penser qu'elles sont régies par un même principe. Par conséquent, les auteurs ont souvent cherché à mettre en évidence les relations entre ces trois lois. Certains ont même essayé de formuler des principes permettant de les unifier sous une seule et même loi.

Ressemblance des lois Zipf-Bradford

Quand il a publié sa dérivation simplifiée de la distribution de Bradford, Brookes découvrit qu'elle avait une forte ressemblance avec la loi de Zipf (BROOKES, 1968). Celle-ci ne décrit pas exactement la distribution de Bradford. Il est vrai que multiplier le rang d'un périodique par son nombre d'articles contribuant à un thème aboutira à un nombre voisin d’une constante. Mais ceci ne s'applique qu'à la portion rectiligne de la courbe de Bradford. Le

"noyau" ne confirme pas la relation de Zipf $f \cdot r = C$. Bien avant Brookes, en 1960, Kendall avait montré que "Zipf" et "Bradford" étaient deux lois structurellement similaires (KENDALL, 1960). Plus récemment, Egghe a encore travaillé sur la concordance entre la loi de Zipf et les deux formulations de la loi de Bradford, graphique et verbale (EGGHE, 1991).

Ressemblance des lois Lotka-Pareto-Zipf

Parker-Rhodes et Joyce (PARKER et JOYCE, 1956) ont présenté la distribution des mots d'un texte comme Lotka l'avait fait pour les auteurs dans le passé : $n(u) = k.u^{-2}$, $n(u)$ étant le nombre de mots qui apparaissent, et u la fréquence. Price a aussi indiqué que la formule de la loi de Lotka ressemblait à une autre loi économique, celle de Pareto, utilisée pour représenter la répartition des revenus dans la société.

Ressemblance des lois Bradford-Lotka

De la même façon que Parker-Rhodes et Joyce l'avaient fait avec la répartition des mots par la loi de Lotka, récemment Chung a choisi de représenter la répartition des auteurs selon le modèle proposé par Bradford (CHUNG, 1994). Les auteurs étudiés provenaient d'une bibliographie dans le domaine des systèmes de classification documentaire. Les données expérimentales que Chung a traitées présentaient bien une répartition selon la forme de Bradford mis à part qu'au lieu d'avoir l'affaissement de Groos en fin de courbe, les données s'écartaient de la droite en sens opposé. Dans le même ordre d'idée, une expérience menée sur des données issues de références de brevets montre que les sociétés déposantes d'un domaine technique précis peuvent être représentées selon une répartition de type Bradford (ROSTAING, 1993). Les données expérimentales ont même permis d'établir un découpage de la courbe en quatre zones avec un coefficient multiplicateur très proche de trois et donc d'isoler le "noyau dur" dans les sociétés déposantes dans ce domaine selon le critère de Bradford.

L'unification

L'existence d'un rapprochement mathématique entre ces trois lois a été démontrée par Hubert (HUBERT, 1978) et Chen (CHEN, 1985), mais une formule unificatrice reste encore à trouver ! Pourtant de nombreuses recherches se sont dirigées vers la découverte de cette loi unique.

Le premier auteur à avoir ressenti le besoin de réunir ces trois lois sous un même principe a été Price. Price a proposé une théorie unifiée pour toutes les lois statistiques bibliométriques. Celle-ci y intègre aussi la règle d'accumulation des citations d'articles. Il voulait par sa "*théorie du processus de l'avantage du cumul*" (*theory of cumulative advantage process*), formuler le phénomène connu sous le nom d'*"effet Saint-Mathieu"*. Cette théorie retranscrit l'idée que le

succès est récompensé alors que l'échec n'a aucune conséquence. Il explique sa formule ainsi : supposons une population d'individus essayant d'atteindre un but : un nombre de publications (Lotka), une acquisition d'articles pertinents (Bradford)... Si un élément (le scientifique, le périodique...) a du succès, sa probabilité de succès augmentera pour une tentative ultérieure, alors qu'un échec ne réduit pas la probabilité de succès pour la prochaine tentative. La distribution de l'avantage cumulatif est présentée par Price selon une formulation de densité. Cette formulation ne contient donc qu'un seul paramètre en la présence de la variable *m*.

Formule de l'avantage du cumul selon Price:

$$f(n) = (m+1).\mathrm{Beta}(n, m+2)$$

avec
- n = nombre de succès
- $f(n)$ = fraction d'individus avec n succès
- m = constante pour les individus d'une population pour tous les n
- $\mathrm{Beta}(n,m+2)$ = fonction "Beta" (*Beta Function*) dont la valeur pour les deux arguments entre parenthèse peut être lue dans une table. En fait une valeur de cette fonction est approximativement égale à $\mathrm{Beta}(a,b) = (b-1)!\ a^{-b}$.

Mais cette loi unifiée n'est pas reconnue comme telle par tous. Egghe et Rousseau la trouvent trop approximative pour satisfaire les lois de Zipf et Mandelbrot (EGGHE et ROUSSEAU, 1986). Concernant l'application de la théorie de Price pour modéliser la fréquence de citation d'une publication, Budd et Hurt dans un récent article l'ont expérimentée, puis comparée avec des données réelles recueillies dans les bases de données de l'ISI (BUDD et HURT, 1991). Les résultats montrent que l'amorce des citations d'une publication suit une pente bien plus escarpée que celle obtenue par le modèle de Price. Les auteurs n'ont pas pu déduire si le modèle de Price était déficient ou si les cas qu'ils ont considérés pour leur exemple sont de mauvaises représentations du phénomène en général.

Jusqu'à présent, aucune formule statistique permettant de décrire toutes les caractéristiques bibliométriques n'a été reconnue. Haitun a récapitulé les principales lois hyperboliques concernant de près ou de loin les distributions bibliométriques. Il a argumenté sur le fait qu'il y avait deux types de distributions : Gaussienne et Zipfienne. Alors que les distributions Gaussiennes sont les bases des sciences naturelles, Haitun considère les distributions Zipfiennes comme les bases de la vie sociale, la loi quantitative fondamentale de l'activité humaine (HAITUN, 1982). Même si sa formule exacte n'est pas encore bien définie, on peut retenir que la forme de la loi de Zipf est la distribution de base en bibliométrie.

MESURES SYNTHETIQUES DES DISTRIBUTIONS

La simple connaissance de la formule mathématique d'une distribution n'apporte en fait que peu d'intérêt. Mais c'est cette formulation qui permet

d'établir l'appartenance de la distribution à une catégorie de distributions. A chaque catégorie, est associée la définition de paramètres qui mesurent les caractéristiques d'une distribution dans sa catégorie. Les deux paramètres synthétiques les plus connus sont ceux établis pour les distributions appartenant à la catégorie des lois normales : la moyenne et l'écart-type. Ces paramètres synthétiques sont calculés à partir de la théorie des moments. La moyenne permet d'étudier la tendance centrale de la distribution tandis que l'écart-type informe sur la dispersion des données autour de cette tendance centrale. La simple connaissance de ces deux mesures définit suffisamment bien la distribution pour permettre sa comparaison avec d'autres. Haitun a précisé que les distributions, contrairement aux lois des sciences naturelles, ne sont pas Gaussiennes mais hyperboliques, et les dénomma sous le nom de Zipfiennes (HAITUN, 1982). Contrairement aux Gaussiennes, les distributions hyperboliques n'ont pas des moments stables mais des moments que Haitun a caractérisés comme infinis. Bien sûr, les échantillons finis ont des moments finis mais ceux-ci sont principalement dépendants de la taille de l'échantillon. Donc ils ne paraissent apporter que peu d'indications sur les caractéristiques des distributions. Ainsi, les techniques statistiques Gaussiennes ne peuvent pas s'appliquer aux distributions Zipfiennes. Ce qui signifie que les calculs de moyenne et d'écart-type appliqués aux distributions bibliométriques ne permettent pas de mesurer leurs caractéristiques.

Des recherches ont été menées pour trouver des indicateurs de mesure synthétique des distributions bibliométriques. Une première approche a tenté de mesurer des concentrations indiquant un trait de caractère propre aux distributions. Une seconde fait appel à une théorie beaucoup plus généraliste qui a déjà été appliquée à de nombreux domaines scientifiques : la théorie de la communication.

Mesures de concentration

Conscient de cette absence de mesures synthétiques, Pratt a proposé des mesures de concentration (dispersion) des distributions bibliométriques et de leurs éléments (PRATT, 1977). Une première mesure de concentration servait de valeur caractéristique des distributions offrant ainsi un repère pour les comparer. Par une seconde mesure, une concentration relative, Pratt a voulu pouvoir estimer la concentration de chaque élément pour la distribution. Appliqué à la distribution de Bradford, cet indice devait mesurer le degré de concentration des articles sur un sujet dans une collection de revues. Juste après la publication de Pratt, Carpenter a fait remarquer la ressemblance entre l'indice de Pratt et l'indice de Gini (CARPENTER, 1979). L'indice de Gini a été formulé par Corrado Gini en 1908 pour mesurer la concentration totale de la courbe de Lorenz qui représente la répartition de la richesse (revenus) chez les citoyens. On retrouve là encore la ressemblance entre les lois Lotka-Pareto-Zipf (voir p. 45). L'interprétation bibliométrique de l'indice de Pratt n'est pas reconnue par

tous. Drott a observé que l'application de la formule de Pratt dépendait plus de la taille de l'échantillon que des concentrations intrinsèques à la littérature. Un procès similaire a été fait à cet indice par Hustopecky et Vlachy dans (HUSTOPECKY et VLACHY, 1978). Il y a peu, Egghe a modifié l'indice de Pratt pour l'appliquer dans des cas particuliers aux distributions bibliométriques (EGGHE, 1987, 1988)[1].

Cette première approche n'a pas apporté les fruits attendus. La mesure de concentration reste une indication plus dépendante de la taille de l'échantillon que des caractéristiques de la distribution. Donc, elle ne peut être satisfaisante.

L'entropie de Shannon

De nombreux auteurs émettent l'idée que la communication scientifique suit des processus de transmission que l'on trouve par ailleurs dans la nature. Par exemple, Goffman et Newill ont proposé de rapprocher le phénomène de propagation des idées scientifiques du modèle "épidémiologique" (GOFFMAN et NEWIL, 1964). Ils estimaient que la communication des idées par les publications est régie par un processus formellement équivalent à la transmission d'une maladie par un organisme ou à la communication d'un signal dans une machine. La recherche d'information est assimilable alors au processus de croissance de la propagation de l'infection en favorisant les contacts entre les systèmes infectés et ceux susceptibles de l'être.

La théorie de l'information (renommée par la suite théorie de la communication) a été formulée par Shannon en 1948 (WEAVER et SHANNON, 1975). Elle a souvent été employée pour modéliser des processus naturels, alors qu'à l'origine Shannon l'avait établie pour l'étude de la transmission des signaux par voie téléphonique. L'outil essentiel de cette théorie est la mesure de la variété moyenne ou complication des signaux par une équation mathématique. Cette théorie a été largement utilisée par les écologistes (LEGENDRE et LEGENDRE, 1984). En écologie, le concept de diversité des espèces est prépondérant pour évaluer la richesse d'un milieu. La richesse est caractéristique de la maturité et de la stabilité de ce milieu. Cette théorie permet de définir des indicateurs pour mesurer la diversité d'une distribution en écologie.

Comme en biologie, cette théorie mathématique de la communication a donné lieu dans le domaine des sciences de l'information à de nombreuses applications. Nous n'aborderons ici que l'application qui peut en être faite dans le cadre de l'aide à l'élaboration de mesures synthétiques d'une distribution bibliométrique. Lafouge et Quoniam (LAFOUGE et QUONIAM, 1992) ont rappelé comment cette mesure de diversité peut s'appliquer aux distributions bibliométriques. Cette notion de diversité représente une toute autre information que celle induite par la théorie des moments. L'applicabilité et la validité de

[1] Pour des discussions générales sur les mesures de concentrations, le lecteur se doit de consulter EGGHE et ROUSSEAU, 1990 ; BURREL, 1991.

cette théorie pour des données bibliométriques expérimentales n'ont pas encore été confirmées. Mais dans le cas où cette théorie se montrerait adaptée aux données expérimentales, la théorie de la communication offrirait des mesures synthétiques pour caractériser des distributions bibliométriques. Selon les conclusions, non définitives, des auteurs, les deux premières mesures caractéristiques seraient la taille du lexique étudié et l'entropie de Shannon.

En considérant une distribution bibliométrique où f_i est l'occurrence du $i^{ème}$ terme et n le nombre de termes dans l'échantillon étudié :
la probabilité d'apparition du $i^{ème}$ terme, p_i, est calculée selon la relation

$$p_i = \frac{f_i}{\sum_{i=1}^{n} f_i}$$

l'entropie généralisée à l'ordre *a* formulée par Rényi et utilisée en biologie est transposable en science de l'information :

$$H_a = \frac{1}{1-a} \times \log \sum_{i=1}^{n} p_i^a$$

L'approche logarithmique n'étant pas forcément parlante, on utilise le concept de diversité généralisée d'ordre a développé par Hill:

$$N_a = \text{Exp } H_a$$

En développant l'entropie généralisée on obtient

à l'ordre 0 $H_0 = \text{Log } n$
$N_0 = n$ diversité = nombre de formes

à l'ordre 1 $\lim_{a \to 1} H_a = H_1 = -\sum_{i=1}^{n} p_i \text{ Log } p_i = H_{shannon}$
$N_1 = \text{Exp } H_{Shannon}$ diversité spécifique

Donc, le nombre de termes de la distribution revient à donner la diversité à l'ordre 0 de Hill.
La diversité spécifique d'ordre 1 fournit un renseignement supplémentaire : elle procure la quantité d'information (ou de diversité) de chacun des termes. Cette dernière se trouve être la formulation de Shannon.
On a $H_1 = 0$ si le nombre de termes est réduit à 1
et $H_1 = \text{Log } N$ (maximum) si les occurrences pour chaque terme sont identiques, $p_i = 1/n$

Remarque : une augmentation de l'entropie en thermodynamique correspond à un accroissement du désordre, ce qui entraîne une diminution de l'information. De façon stricte, l'information est une entropie négative, une néguentropie. Ce n'est que pour des raisons de simplicité qu'on la qualifie d'entropie.

On peut donc dire qu'en l'état actuel des recherches, il n'existe aucune solution mathématique totalement validée par des données expérimentales pour disposer de mesures synthétiques caractéristiques des distributions bibliométriques. La solution par l'approche de la mesure de l'entropie paraît prometteuse bien qu'une partie des premiers résultats rejoigne les conclusions de l'approche par la mesure de concentration (la mesure à l'ordre 0 serait la taille du lexique analysé).

CONCLUSION

Les lois bibliométriques s'intéressent principalement aux relations qui existent entre une quantité de sources et une productivité. La loi de Bradford répartit les revues scientifiques selon leur aptitude à représenter un domaine scientifique par la mesure du nombre d'articles le traitant. La loi de Lotka répartit les auteurs selon l'aptitude à représenter un domaine scientifique par la mesure de la quantité de leurs travaux. Et la loi de Zipf répartit les mots selon leur aptitude à être représentatifs du texte. Ces lois n'expliquent aucunement les phénomènes bibliographiques mais ne font que les représenter. Il est regrettable de constater que la majorité des travaux scientifiques concernant cette branche de la bibliométrie se soit principalement axée sur la capacité des modèles mathématiques à s'ajuster aux données expérimentales. Dès lors, nous n'avons que très peu d'explications sur le sens profond de ces distributions, sur les potentialités de découpage des éléments distribués en groupes significatifs (au sens statistique ou même sémantique). L'absence de mesures synthétiques en est l'une des causes.

Malgré ces déficiences, la connaissance de ces lois reste indispensable à tous ceux qui ont l'intention de mener des analyses bibliométriques. La caractéristique hyperbolique de ces distributions est une notion fondamentale à toujours conserver à l'esprit. Toutes les méthodes statistiques ne sont pas bonnes à employer car elles sont bien souvent construites sur le principe d'une répartition normale. L'emploi de la moyenne en est un très bon exemple. La valeur moyenne de la fréquence des mots dans un texte n'a aucun sens. De même, la valeur moyenne de la fréquence de publications des auteurs dans un domaine scientifique n'a aucune raison d'être mentionnée. Il est également sans objet de comparer deux moyennes de fréquences d'auteurs pour deux domaines scientifiques. Actuellement la seule comparaison possible est celle qui oblige à évaluer la différence entre les deux distributions complètes. Ces règles doivent être respectées lors de la construction en bibliométrie d'indicateurs univariés ou relationnels. C'est pourquoi les méthodes d'analyses des données appliquées en bibliométrie sont des techniques purement descriptives et non des techniques qui cherchent découvrir des modèles régis par des lois de distributions Gaussiennes.

CHAPITRE III
LES INDICATEURS UNIVARIÉS

Le principe de la bibliométrie univariée est de constituer des indicateurs qui permettent de comparer entre eux les éléments d'un ensemble de références bibliographiques. La difficulté dans l'élaboration de ces indicateurs se situe dans le choix du système de mesure pour comparer de façon équitable les éléments. Ces indicateurs univariés sont généralement considérés comme livrant des informations purement quantitatives.

Le système de mesure le plus élémentaire est tout simplement le résultat brut du dénombrement de l'élément bibliographique étudié. On différencie deux types de dénombrements.

- *Le comptage des références* où est présent l'élément bibliographique : ce simple comptage est généralement considéré comme la mesure même de la productivité de cet élément. Dans l'absolu, un tel nombre ne veut pas dire grand chose. Selon les périodes considérées, selon les spécialités et les disciplines, selon les pays, les volumes de publications peuvent être très variables. Ainsi les chercheurs ont voulu élaborer des mesures adaptées à chaque objectif d'évaluation. Ces nouvelles mesures sont construites de façon à relativiser le taux de publication en fonction de certains critères. Certaines de ces mesures seront présentées dans les lignes qui vont suivre. Mais quelle que soit la mesure employée, l'évaluation et l'interprétation qui en est faite restent contestables. Les évolutions temporelles de ces mesures sont toujours beaucoup plus significatives. Connaître la vitesse d'un objet est intéressant pour le classer parmi un ensemble d'objets qui évoluent, mais connaître son accélération donne une meilleure idée sur ses capacités de mobilité et sur son prochain classement

parmi les autres. Les évolutions temporelles pour les indicateurs univariés jouent le même rôle. Elles seront toujours source de plus grands renseignements.

• *Le comptage des citations reçues* par l'élément bibliographique (nombre d'articles qui citent l'élément) : juger de la productivité par une mesure de quantité (nombre de publications) est bien, mais pas satisfaisant aux yeux des bibliomètres. Ils ont immédiatement cherché à connaître quel indicateur leur permettrait d'évaluer la qualité d'une publication. Plusieurs facteurs rentrent en jeu dans l'évaluation de la qualité d'une publication. En éliminant l'évaluation du contenu du document lui-même qui imposerait sa lecture et une parfaite connaissance du domaine scientifique traité, les facteurs considérés comme importants sont :

1) le type de publication (livre, périodique, rapport...)
2) la collaboration : la publication est-elle le résultat d'une collaboration entre différents groupes de recherche (la collaboration internationale étant la plus valorisante) ?
3) la nature du contenu de la publication (fondamental, méthodologique, expérimental, synthèse,...)
4) la renommée du périodique concerné
5) le nombre de fois où la publication est citée par la communauté scientifique.

Le facteur longtemps considéré comme donnant une indication de qualité est le taux de citation. Certains spécialistes en bibliométrie ont estimé qu'un article fortement cité par d'autres auteurs avait un contenu reconnu dans son domaine et certainement utile à la communauté scientifique, puisque d'autres en tenaient compte dans leurs propres recherches. Mais les raisons qui poussent les auteurs à citer d'autres publications sont fort complexes. Ainsi, de nombreuses critiques sont venues s'opposer à cette hypothèse :

1) les citations erronées qui renvoient à des sources secondaires plutôt qu'à l'auteur principal de la découverte (McROBERTS et McROBERTS, 1989)
2) l'auto-citation, évaluée de 10 à 30% par article (McROBERTS et McROBERTS, 1989)
3) la différence de nature entre les citations : certaines sont faites dans un contexte de critique (NADEL, 1983), d'autres non
4) l'inertie de la citation : laps de temps important entre la publication et la citation (McCAIN et TURNER, 1989)
5) l'influence de la revue où est paru l'article. A un point tel que Van Raan a montré que l'impact d'un article pouvait s'estimer par le facteur d'impact de la revue (VAN RAAN, 1988)
6) la variation de la pratique de la citation dans chaque discipline (MOED et alii, 1985 ; GARFIELD, 1982)
7) le taux de citation dépend du type de document (article ou synthèse). Par exemple, les articles de méthodologie sont plus cités car les auteurs, pour

éviter de décrire la méthode qu'ils utilisent, citent les documents où elle est expliquée (PERITZ, 1983)

8) un auteur cite plus facilement ses compatriotes (STEVENS et NARIN, 1989)

9) le taux de citation varie selon la nationalité des auteurs (McROBERTS et McROBERTS, 1989)

10) l' "effet S^{t} Mathieu" qui veut que l'on prête plus facilement aux riches. Les auteurs d'articles cherchent à faire référence à des articles des chercheurs renommés afin de mieux convaincre de la solidité de leur argumentation (COURTIAL, 1990).

Pour ajouter à ces critiques de fond quelques critiques plus techniques, il faut rappeler que les données concernant le taux de citation ne sont accessibles que sur la base de données de l'ISI. Il faut donc, en plus, prendre en considération les limites propres à la source ISI : les mauvaises couvertures thématiques, géographiques, temporelles (CARPENTER et NARIN, 1981) et la variation des saisies des citations dans les références (retranscrites telles qu'elles sont mentionnées dans l'article d'origine).

A la suite de ces critiques, certains auteurs ont réévalué le rôle à donner au dénombrement des citations. Il y a actuellement un consensus pour dire que la citation ne mesure pas la qualité de la recherche mais plus exactement ce qu'on pourrait appeler l'impact des publications, que cet impact soit dû au contenu de l'article ou aux autres facteurs influençant la pratique de citation. Certains auteurs restent même très pessimistes en ce qui concerne l'information fournie par cet indicateur : *"En conclusion, le taux de citation d'un document est un indicateur grossier de son impact au sens où il permet au moins d'opposer deux types de publications : celles qui passent "inaperçues" (et, a fortiori, dans la communauté scientifique internationale telle qu'elle est définie par la base ISI si on utilise le Science Citation Index) et celles qui sont réutilisées par les autres chercheurs, sans que leur réutilisation ait une valeur précise"* (COURTIAL, 1990).

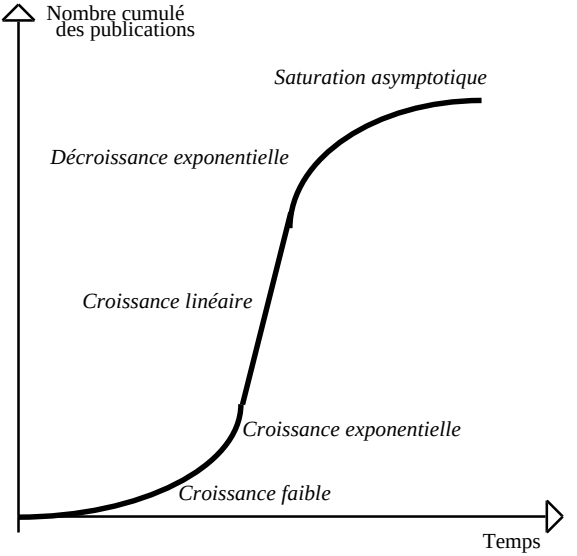
Les deux catégories de dénombrement qui viennent d'être énumérées sont appliquées à la plupart des éléments présents dans une référence bibliographique : auteurs, revues, affiliations, pays, domaines, dates. Tous ces comptages peuvent être combinés pour élaborer des indicateurs plus complexes et pour pondérer certains effets sous-jacents à chacun de ces comptages. Quelques exemples seront présentés dans les paragraphes suivants.

L'EVOLUTION DE L'ACTIVITE DE RECHERCHE

Il est légitime de commençer par des études qui traitent de la notion du temps (l'élément bibliographique "date") et, tout particulièrement, par celle entreprise par Price dans les années 1960. Il a suggéré que la courbe d'évolution des connaissances en science au cours du temps doit suivre une loi. Il a étudié l'accumulation des écrits scientifiques et essayé de montrer que la courbe de

cette accumulation se développe selon une courbe logistique. Cette courbe, encore appelée courbe en S, est employée comme modèle dans divers domaines : le suivi du tonnage obtenu par l'extraction d'un minerai ou l'évolution de la production industrielle d'un produit résultant d'une innovation. Price a voulu l'appliquer à la production des travaux scientifiques en imaginant que la création de connaissances scientifiques nouvelles viendrait à s'épuiser avec le temps (figure 6). Son étude avait prévu pour 1950 le point d'inflexion de la courbe vers le déclin. Les travaux de Tague en 1981 ont montré que nous étions bien dans une phase de croissance linéaire du nombre cumulé des publications scientifiques avec le temps, toutes sciences confondues (TAGUE, 1981). Mais rien ne nous a prouvé jusqu'à présent la validité d'une telle thèse au niveau de la connaissance mondiale.

Fig. 6 - Courbe logistique de croissance de la science



Par contre, cette courbe est souvent établie pour estimer l'évolution des recherches sur un domaine spécifique. Un très bon exemple d'étude de l'évolution de la recherche dans le domaine du laser a été réalisé par des chercheurs indiens (ASHOK et GARG, 1992). Ils ont modélisé la courbe des publications dans le domaine du laser à l'échelle mondiale et celle de l'Inde (voir encadré ci-dessous). La formule mathématique s'est très bien ajustée aux données expérimentales.

Les chercheurs indiens Ashok et Garg sont partis de la formule suivante donnée par Sternman :

$$dw/dt = E(t)\ (W-w(t))$$

où

- E(t) est la fonction qui représente l'effort de la communauté contribuant à l'évolution des connaissances dans un domaine.
- W le nombre maximum de publications que cette communauté peut produire avant que la connaissance du domaine soit épuisée
- w(t) le nombre de publications déjà parues

donc W - w(t) est le nombre de publications qu'il reste à éditer au temps t et E(t) = p + q w(t)

où
- p représente les ressources déployées pour résoudre les énigmes posées
- q représente le nombre de chercheurs qui se rallient à cette tâche

donc $$dw/dt = (p + q\, w(t))\,(W - w(t))$$

Cette formulation est très similaire à celle qu'avait donnée Bass pour le modèle de la diffusion de la technologie. Dans ce modèle W symbolisait le niveau de saturation: le nombre de personnes qui vont adopter la technologie et p et q étaient des coefficients d'innovation et d'imitation. Pareillement, le modèle, décrit par Sternman, exprime la diffusion ou l'adoption du paradigme. Les auteurs aboutissent à la forme discrète de la formule suivante :

$$w(t+1) - w(t) = (p + q\, w(t))\,(W - w(t))$$

L'EVALUATION DES REVUES

Beaucoup d'investigations en bibliométrie s'intéressent à la revue scientifique en tant qu'unité d'analyse. Ceci n'est pas étonnant puisqu'elle joue un rôle essentiel dans la communication des résultats de recherches. Quand on sait qu'une grande part des budgets d'un centre de documentation est consacrée aux revues spécialisées, la création de moyens d'évaluation de ces derniers est fondamentale pour bien gérer les abonnements. Mais ce n'est pas la seule raison de l'utilisation des revues comme élément d'analyse bibliométrique. L'étude des revues scientifiques spécialisées peut, dans une première approximation, donner des indications sur l'état de ces domaines de spécialité. Lorsque l'on sait que l'acte de publication est le principal vecteur de reconnaissance scientifique, l'importance de ces revues dans la communauté scientifique et l'évolution de cette importance au cours du temps sont des signes marquants de l'état de santé d'un domaine scientifique.

La citation

La création du *Journal citation reports* (*JCR*) par l'ISI (voir p 14) a permis d'offrir de nouvelles données pour juger de l'importance des revues par l'intermédiaire des citations. L'ISI y recueille le nombre de citations dont font l'objet 4200 revues scientifiques. Ce comptage signifie en fait, pour chaque année, le cumul du nombre de citations dont font l'objet les articles parus dans une revue. Il est évident que plus une revue a d'articles et plus elle a de chance d'être citée. Pareillement, plus la reconnaissance de cette revue est importante et plus son taux de citation augmente. Pour nuancer ces facteurs, de nombreux indices ont été proposés. Les plus connus sont les deux indices introduits par Garfield en 1969 dans le *JCR* :

- Le facteur d'impact : $I_F = \dfrac{c_x}{p_{x-1} + p_{x-2}}$

c'est-à-dire le nombre de citations reçues l'année *x* pour les articles publiés par une revue pendant les deux années précédentes, divisé par le nombre d'articles publiés par cette même revue pour ces deux années précédentes.

• l'indice d'immédiateté : $I_I = \frac{c_y}{p_y}$

c'est-à-dire le nombre de citations, reçues l'année *y* pour les articles publiés la même année par une revue, divisé par le nombre d'articles publiés cette année là. Garfield voulait, par cet indice, donner une indication sur la rapidité d'utilisation des articles.

Des algorithmes plus élaborés pour déterminer l'importance d'une revue ont été ensuite développés par Bennion et Karschamroon (BENNION et KARS-CHAMROON, 1984), ainsi que par He et Pao (HE et PAO, 1986). Pour donner une idée de l'emploi du *JCR* et des données accessibles en ligne pour l'étude des revues, un article de Buffeteau présente comment les données de l'ISI permettent d'estimer l'impact de la revue de l'Institut Français du Pétrole (BUFFETEAU, 1991) dans son domaine.

Mais il faut constamment se rappeler que les données fournies par le *JCR* sont à prendre comme des indices. En plus de tous les biais introduits par la pratique de la citation (voir p. 52 et 47), des critiques propres au *JCR* entrent en jeu. Tous les chercheurs utilisant le *JCR* ont remarqué des inadéquations ou des problèmes (VLACHY, 1985 ; CARPENTER et NARIN, 1981 ; RICE, 1989). Certains ont rapporté la possibilité d'avoir jusqu'à 25% d'erreur de mesure en utilisant le *JCR* du SSCI 1977-85, la plupart des erreurs étant dues à des comptages d'abréviations aberrantes. Mais il faut reconnaître que le *JCR* est la seule source des données des citations pour un important nombre de revues à travers une longue période.

La typologie

La comparaison des revues peut être menée par des approches bibliométriques plus classiques ne faisant pas intervenir la pratique des citations. L'article de Dou et al., par les solutions graphiques qui y sont proposées, est un bon exemple de ce genre d'étude (DOU et alii, 1990a). Les auteurs établissent des cartes typologiques de thèmes par revues permettant de mieux connaître les thèmes de prédilection et les spécialités qui les distinguent. Dans ce cas, la comparaison se fait sur une estimation du contenu scientifique des revues. L'article présente une étude pour six revues en chimie. Toutes les références des articles des six revues ont été collectées de janvier 1982 à juillet 1987 par consultation du *Chemical Abstracts Services* (CAS) accessible en ligne. Ces données ont été traitées automatiquement par des logiciels spécifiques pour produire des cartographies exprimant l'ampleur et la fréquence des thèmes abordés dans les articles par l'intermédiaire des codes de CAS. De telles présentations graphiques mettent en évidence la typologie des articles publiés dans chacune des revues et, par conséquent, des similarités et complémentarités des contenus de ces revues. Un découpage des références d'une revue par année fournit aussi l'évolution thématique des articles par une succession de cartes typologiques.

L'EVALUATION DES CHERCHEURS

Il est bien connu que le nombre d'articles est encore le principal moyen d'évaluation des chercheurs pour les instituts de recherches. Ce simple chiffre est l'élément de référence pour juger de la productivité du chercheur, pour apprécier son mérite et, probablement, pour décider de sa promotion (DEMAZURE, 1992). Est-ce que cet indicateur est suffisant ? Bien évidemment, la réponse est négative, mais peu de solutions alternatives sont proposées. L'objet de la question est trop sensible pour être traité à la légère.

Le nombre de publications - la loi de Lotka

Le premier traitement bibliométrique qui a considéré l'auteur scientifique comme unité de travail a été réalisé par Lotka. Cette étude, exposée précédemment, n'avait pas pour objectif l'évaluation de la productivité des auteurs. Toute trace nominative des auteurs est perdue par leurs regroupements par rangs de fréquence égale. Si les auteurs considérés par ces regroupements sont les individus qui contribuent au développement d'un domaine, la loi de Lotka représentera le profil des fréquences de publications caractéristiques de ce domaine. Dans l'absolu, le nombre de publications d'un chercheur n'a de sens que s'il est replacé par rapport à la pratique de publication dans son domaine. Il faudrait donc toujours accompagner la valeur d'un nombre de publications par la distribution caractéristique associée. Comment juger d'une valeur sans référentiel ? La suite logique à la prescription de cette contrainte est de savoir choisir le bon référentiel. Tout le problème de l'évaluation est dans ce choix ! Quel ensemble de références bibliographiques peut jouer le rôle d'étalon de la mesure ?

La citation

Comme pour les périodiques de nombreuses études bibliométriques présentent la citation dont un auteur fait l'objet comme un indicateur de l'impact de cet auteur dans la communauté. Les données, comme pour toutes celles concernant la citation, sont uniquement produites par l'ISI, soit sous forme papier, soit par consultation en ligne (GARFIELD, 1981). Ces données sont là encore fortement critiquées, non seulement parce que le calcul des citations n'est réalisé qu'à partir de 4200 revues, mais parce que viennent s'ajouter des biais spécifiques aux auteurs (McROBERTS, 1989) :

1) seul le premier auteur de l'article cité est considéré
2) risque d'homonymie
3) risque d'accentuation de l'auto-citation
4) erreurs d'orthographe.

Les co-signatures

Pour être totalement irréprochable lors du comptage de l'élément auteur, il faudrait prendre en considération le phénomène de la co-signature. Une étude sur l'implication des différentes procédures de comptabilisation de la productivité des auteurs a été menée par Pravdic et Oluic-Vukovic (PRAVDIC et OLUIC-VUKOVIC, 1991). Quatre procédures de comptabilisation ont été menées simultanément sur le même ensemble d'auteurs :

1) Comptage normal (*normal count*) : il donne un crédit équivalent à tous les auteurs d'une même publication ; il y a donc comptabilisation pour un auteur de tous les articles dont il est signataire.

2) Paternité fractionnée (*authorship fractional*) : la contribution de l'auteur est pondérée par le nombre d'auteurs de l'article. La productivité de l'auteur est alors la somme de toutes ses participations aux publications.

3) Comptage direct (*straight count*) : seul le premier auteur reçoit la paternité de la publication.

4) Comptage direct modifié (*modified straight count*) : chaque publication est attribuée à un seul auteur, celui qui a la plus forte productivité.

Les trois dernières procédures ont en commun de conserver le nombre total de publications, tandis que la première perd cette valeur en introduisant un effet multiplicateur. Les troisièmes et quatrièmes procédures réduisent les corpus des auteurs.

Les auteurs concluent l'article en affirmant que chaque procédure apporte une spécificité et que lors d'une étude de distribution de productivité d'auteurs et encore plus pour l'appréciation d'une contribution d'un seul auteur, toutes ces procédures devraient être utilisées.

Mis à part le problème du comptage, la co-signature peut être une source importante d'information sur le niveau de collaboration qu'entretient l'auteur. Publie-t-il avec d'autres laboratoires ? Maintient-il ses relations de collaboration dans de longs programmes? Quelles sont les nationalités de ces collaborations ? Autant de questions qui peuvent aider à mieux évaluer l'activité de recherche d'un auteur. Les méthodes bibliométriques à employer pour y répondre s'éloignent des indicateurs univariés pour s'approcher d'indicateurs mettant en jeu des mesures de relation (voir p. 81).

L'EVALUATION DES AFFILIATIONS

Les méthodes sont toujours les mêmes mais ici elles sont utilisées pour une unité d'analyse prenant en compte non plus l'individu mais les centres de recherche. On se trouve dans le même cas d'étude que lors de l'évaluation d'un chercheur, mis à part qu'il ne faut plus considérer un individu mais plusieurs appartenant aux mêmes organismes.

Il est pratiquement impossible pour le traitement bibliométrique de considérer comme unité d'analyse le champ affiliation des références

bibliographiques fournies par les banques de données. Les noms des affiliations y sont très mal représentés pour trois raisons :

1) la plupart des bases n'indiquent que l'affiliation du premier auteur de l'article
2) il n'existe aucune norme pour la saisie des noms et des adresses des centres
3) l'information sur l'affiliation est inexistante dans les citations des banques de l'ISI.

Bien que la tâche soit grande, certains auteurs ont tout de même réalisé des analyses d'organismes portant sur des milliers de références. Un cas remarquable de pugnacité est l'étude de Bauin et al. qui s'étaient fixés pour objectif de connaître la part de publications nationales pour les grands organismes de recherche français, en se focalisant plus particulièrement sur le CNRS, leur propre organisme (BAUIN et alii, 1993). Ils ont collecté à partir de la banque SCI 86-90 de l'ISI toutes les références de publications comportant une affiliation française. Puis ils ont normalisé toutes les affiliations, c'est-à-dire 280 000 occurrences pour 150 000 références françaises. Au cours de cette étude, ils ont eu recours à un comptage fractionnel des affiliations pour relativiser le poids selon le nombre d'affiliations de co-signataires (comme dans le cas des comptages d'auteurs).

Une autre solution pour l'étude des affiliations est de comptabiliser les publications de chaque individu appartenant au même organisme et de les compiler. Le traitement par les noms pose deux difficultés. Il faut d'abord connaître l'ensemble des chercheurs pour chaque affiliation. Des techniques de regroupement automatique des auteurs par chaînage donnent une première approche. Une fois que les noms des chercheurs sont connus, il reste toujours quelques ambiguïtés dues à l'homonymie lors du comptage de leurs publications mais cet artefact peut être facilement maîtrisé.

L'EVALUATION DES DOMAINES D'ACTIVITE DES PAYS

C'est peut-être l'unité d'analyse la plus employée en bibliométrie. La comparaison des pays en fonction de leurs contributions scientifiques a toujours été importante dans les études bibliométriques car elle s'intègre bien dans des investigations scientométriques à l'échelle nationale. Connaître la situation du pays dans les divers domaines scientifiques par rapport à celle des autres pays est une indication précieuse pour budgétiser les programmes de recherche nationaux. Les centres d'évaluation nationaux excellent dans l'élaboration de ce type d'indicateur pour leurs études d'évaluation macroscopique de la recherche internationale.

À cette échelle de mesure, la part des contributions scientifiques d'un pays est fortement dépendante de son poids démographique et économique. Aussi, dans un souci d'évaluation plus équitable, ces études emploient de multiples finesses mathématiques pour pondérer l'importance de chaque pays. Elles ne mesurent plus les productivités scientifiques fournies par les données brutes des comptages, mais préfèrent, par des jeux de pondération, contrebalancer la

productivité brute par la valeur que l'on aurait pu espérer avoir de l'ensemble des données.

Indice d'avantage

Pour évaluer la contribution d'un pays à la science, les études macro-bibliométriques aiment mesurer deux facteurs : les domaines scientifiques et l'évolution dans le temps. Il faut donc tout d'abord diviser l'activité scientifique en disciplines homogènes. Le choix des mesures pour sonder un domaine est toujours aussi restreint : soit la mesure de la productivité par le nombre de publications dans le domaine, soit la mesure de l'impact par le nombre de citations reçues pour des articles du domaine. Ensuite, une fois les mesures connues, les auteurs ont pris l'habitude de les pondérer selon la part réelle qu'elles représentent par rapport aux autres domaines et aux autres pays. Le premier a avoir introduit ce calcul est Price (PRICE, 1981). Il voulait montrer comment rendre un tableau de données plus accessible et plus rapidement interprétable. Dans le cas de l'évaluation d'un pays, le calcul qu'il a mis en place revient à pondérer la mesure de la contribution du pays dans un domaine par la valeur qu'on aurait pu espérer avoir en fonction des parts que représentent les contributions totales de ce pays par rapport aux autres pays et les contributions totales du domaine par rapport à l'ensemble des domaines. Sous forme mathématique, ceci correspond à la formule:

$$M'_{pd} = \frac{M_{pd}}{\frac{M_p \cdot M_d}{M}}$$

avec M = nombre total de publications tous pays et tous domaines confondus
M_p = nombre de publications du pays *p* étudié
M_d = nombre de publications du domaine *d* étudié
M_{pd} = nombre de publications réelles pour le pays *p* dans ce domaine *d*
M'_{pd} = nouvelle valeur pondérée

où le dénominateur symbolise la valeur escomptée

Cette pondération est connue sous le nom d'indice d'avantage.

Indice d'activité

D'autres interprétations de cet indice peuvent être données si l'on présente la formule ainsi :

$$M'_{pd} = \frac{M_{pd} / M_d}{M_p / M}$$

Ici le numérateur peut représenter le poids (ou la performance) du pays *p* dans le domaine *d* tandis que le dénominateur peut représenter le poids (ou la performance) du pays *p* tous domaines confondus. Le poids du pays dans un domaine est pondéré par son poids au niveau international. Dans les deux approches, si la valeur de l'indice est inférieure à un, cela signifie que le pays à une contribution faible dans le domaine par rapport à la contribution qu'il a en général. Et inversement, si l'indice est supérieur à un. Les auteurs ont pris

l'habitude de nommer celui-ci indice d'activité *(activity index)* lorsqu'il est appliqué à l'analyse de la contribution d'un pays dans un domaine scientifique.

Barré utilise cette technique dans le cadre des évaluations conduites par l'OST pour comparer les publications de 11 pays au travers de la base *Pascal* dans (BARRÉ, 1991) (voir aussi p 86). Callon et Leydesdorff se sont servis de cet indice pour estimer l'état de santé de la recherche française (CALLON et LEYDESDORFF, 1987) en l'appliquant à la mesure de productivité (nombre de publications) ainsi qu'à la mesure de l'impact (nombre de citations). Ils ont fait remarquer que la citation comme indicateur n'est pas une valeur sûre puisque l'ISI privilégie plus particulièrement la couverture de certains pays par rapport à d'autres.

Certaines études exploitent cet indice sous forme graphique en disposant les pays par domaine selon deux axes. L'un des axes porte le numérateur et le second le dénominateur (pour la seconde formule). La diagonale symbolise l'état d'équilibre entre le poids du pays dans le domaine et son poids au niveau international, c'est-à-dire un indice égal à un. Ensuite les pays positionnés de part et d'autre de cette diagonale ont soit une contribution trop importante dans le domaine, soit trop faible. Schubert et Braun exploitent la version graphique de cette balance pour la mesure de la citation dans des études du ISSRU[1]. Ils ont, par exemple, évalué 25 pays pour la recherche en chimie (SCHUBERT et BRAUN, 1986). Nous verrons que cet indice et sa présentation graphique sont aussi appliqués pour les études des dépôts de brevets par nation (voir p. 93).

La collaboration internationale comme élément d'évaluation

Cette technique de pondération est aussi appliquée dans de nombreuses études pour analyser la part d'activité de chaque pays dans les collaborations internationales. Au lieu de travailler sur le nombre de publications de chaque pays pour chaque domaine, l'indice est calculé sur le nombre de publications co-signées avec un organisme étranger de chaque pays pour chaque domaine. Donc, dans les libellés des éléments composant la formule de l'indice d'avantage (voir p 60), il faut remplacer les termes "nombre de publications" par "nombre de publications faisant l'objet d'une collaboration internationale". Parmi les plus récents des nombreux travaux traitant ce type de données et multipliant les indices de mesure, nous pouvons citer NAGPUL et LALITA SHARMA (1994), MIQUEL et OKUBO (1994), (voir aussi p. 88) sans oublier les résultats calculés régulièrement par l'ISSRU et publiés dans la revue *Scientometrics* à partir de leur propre banque de données *(Scientometric indicators datafile)* fondée sur des données distribuées par l'ISI (BRAUN et GLANZEL, 1993).

[1] Information Science and Scientometrics Research Unit, Hongrie

Problème de la description des domaines d'activité

Ces indicateurs imposent de pouvoir affecter chaque publication à un domaine scientifique et à un seul, parmi un ensemble de domaines déterminés. Pour faciliter l'analyse, il est pratique que cet ensemble soit constitué d'un nombre relativement restreint de domaines, par exemple chaque domaine pourrait représenter une grande discipline de la science exacte. Malheureusement, ce type de classement simpliste n'est généralement pas présent dans les grandes banques de données, puisque à l'origine celles-ci ont été conçues pour pourvoir retrouver les références bibliographiques de manière aisée et rapide. Ainsi, les éléments bibliographiques qui décrivent les domaines abordés dans les travaux scientifiques appartiennent à un découpage, très fin de la science, mettant en évidence la spécificité du travail scientifique. De plus, une publication n'est pas décrite par un seul élément de ce découpage mais par plusieurs, indiquant la multiplicité des spécificités scientifiques abordées. Ces éléments bibliographiques ne peuvent donc pas servir au découpage par grands domaines utile à ces indicateurs. Seule la banque de données *SCI* du CHI (voir p. 16) comporte une classification par domaine spécialement conçue pour répondre à ce genre d'indicateurs "macroscopiques". Ceci explique pourquoi pratiquement toutes ces études sont obtenues par l'exploitation des données du CHI, et donc avec toutes les imperfections que cette base comporte.

CLASSEMENT DES INDICATEURS BIBLIOMETRIQUES UNIVARIES

Vinkler a recueilli un ensemble d'indicateurs bibliométriques univariés qu'il a classés selon plusieurs critères (VINKLER, 1988). Ce classement trie les indicateurs selon trois catégories :

- *Nature du comptage :* les indicateurs peuvent se diviser en deux groupes qui dépendent de la donnée de départ de la mesure : indicateurs de publications ou indicateurs de citations.
- *Nature du calcul :* ces deux types d'indicateurs peuvent eux-mêmes se partager en fonction des différents types de mesure qu'ils mettent en valeur :

1) mesures à caractéristique simple (un comptage simple comme le nombre d'articles, le nombre de citations reçues)

2) mesures à caractéristique spécifique (productivité en fonction d'un autre facteur comme le nombre d'articles par an en fonction du nombre de chercheurs ou du budget)

3) mesures à caractéristique de balance (comparaison entre une entrée et une sortie tel le nombre de citations données comparé au nombre de citations reçues)

4) mesures à caractéristique de distribution (mesure d'une donnée sous forme de part comme le nombre d'articles non cités par rapport au total d'articles publiés)

5) mesures à caractéristique relative (mesure par rapport à une valeur étalon tel le nombre de citations par article par rapport à la moyenne du nombre de citations par article dans la discipline).

• *Nature de la mesure finale :* ces indicateurs sont des mesures concernant l'impact scientifique et/ou la quantité de publications scientifiques :
a) quantité
b) impact
c) quantité/impact

Il a aussi précisé que les indicateurs pouvaient être mis en oeuvre pour différents niveaux d'évaluation (tableau 3).

Tabl. 3 - Niveaux d'évaluation des indicateurs bibliométriques

	Niveau d'évaluation		
Type d'évaluation	Micro	Méso	Macro
organisation	personne, équipe	institut, département	instituts, groupes de pays, monde
thématique	projet	sous-domaine de recherche	Discipline scientifique, nature de la science
publication	un article	ensemble de publications	toutes les publications

CONCLUSION

Ces indicateurs sont utilisés comme des mesures permettant des comparaisons entre pays, organisations, thèmes, types de publications... La difficulté majeure de leur emploi est de savoir à quel niveau il faut se situer et quel type d'évaluation il faut élaborer pour être sûr que la mesure soit totalement adaptée à l'unité bibliographique étudiée.

Bien souvent, pour évaluer une discipline ou une spécialité de la science, toute une batterie d'indicateurs est appliquée aux références bibliographiques sélectionnées pour représenter l'activité scientifique de cette discipline ou de cette spécialité. Ces indicateurs sont chargés de clarifier les principales tendances en quelques chiffres clés. Ils mettent en évidence, sous forme quantifiée, l'évolution des publications dans le temps, la répartition des publications par pays, l'évolution des citations dans le temps, les auteurs ou les organismes importants, les revues les plus spécifiques... Ces indicateurs sont encore plus appréciés lors d'études d'évaluation internationale. Là où des milliers de documents scientifiques sont concernés, ces indicateurs macro-

bibliométriques prennent tout leur intérêt. Les tendances lourdes de chaque pays sont identifiées et comparées.

L'efficacité et l'attrait de ces indicateurs univariés relève fatalement d'une simplification poussée à l'extrême des éléments à mesurer. Ces techniques bibliométriques vont quantifier la contribution de chaque élément de façon à pouvoir les classer entre eux mais elles ne donneront aucune indication sur les dépendances et les interactions pouvant exister entre ces éléments. Quels sont les liens entretenus par les acteurs de la science (auteurs, organismes, pays) ? Quelles sont les structures qui dessinent les dépendances et les frontières entre chaque discipline ? Autant de questions que les méthodes relationnelles vont permettre d'élucider grâce à la construction de cartographies qui retranscrivent sous forme de graphiques la complexité de ces relations.

CHAPITRE IV
LES CARTES RELATIONNELLES

Les bibliomètres ont très vite voulu présenter sous forme imagée le coeur et la dispersion du contenu des indicateurs. Les représentations graphiques des distributions ne permettent de disposer les éléments étudiés que selon un ordonnancement unique. Ils ont cherché à disposer les éléments sur des cartes en deux dimensions plus adaptées à résumer les phénomènes de coeur et de dispersion des littératures et permettant de découvrir les relations entretenues entre les éléments de ces deux catégories. Cette idée de carte implique de pouvoir positionner les éléments les uns par rapport aux autres grâce à une métrique ou une distance. Cette notion de relation entre éléments ne se fait plus par comparaison binaire de mesures comme pour les indicateurs univariés, mais de façon à prendre en compte l'ensemble des mesures. Ces distances sont relatives à l'ensemble des relations qu'entretient chaque élément avec tous les autres. Elles décrivent un degré de ressemblance ou de dissemblance entre les éléments et donc la mesure d'une intensité de relation entre ces éléments.

Les méthodes mathématiques qui offrent de telles caractéristiques font appel à la statistique descriptive. Celles-ci, très gourmandes en calcul, n'ont été praticables qu'après le développement des ordinateurs alors que leur création était déjà ancienne. Ces méthodes descriptives ont donc été introduites tardivement en bibliométrie lorsque les instruments mathématiques ont été suffisamment maîtrisés pour permettre leur vulgarisation.

Une fois l'adaptation de ces méthodes aux études bibliométriques vérifiée, de nombreux auteurs ont vu en elles des possibilités bien plus variées que la

simple représentation du coeur et de la dispersion de la science. Leur exploitation s'est alors diversifiée pour fournir des informations de plus en plus précieuses. Bien que les méthodes mathématiques employées soient construites sur la définition d'une métrique, on peut estimer que les informations fournies par ces cartes ont plutôt un aspect qualitatif que quantitatif. La lecture des résultats ne se fixe pas sur l'interprétation de valeurs numériques, mais plutôt sur la répartition des éléments dans l'espace, des agrégats formés, des éléments isolés... Les auteurs ont alors tendance à parler d'indices "qualitatifs" en comparaison avec les indices quantitatifs élaborés par les mesures univariées.

Ce chapitre présente les principales méthodes de construction de cartes relationnelles employées en bibliométrie. Dans la mesure du possible, la présentation de la méthode est accompagnée d'un exemple d'application trouvé dans la littérature. Les procédés mathématiques ne seront pas exposés lorsqu'ils font appel à des principes d'analyse des données classiques et reconnus. En général, les bibliomètres exploitent pour leurs traitements des modules d'analyse des données inclus dans des logiciels statistiques commercialisés[1].

Les méthodes abordées ont été classées en trois parties. Les méthodes qui font appel aux principes de la co-citation (les plus anciennes). Celles qui sont basées sur le principe de la co-occurrence des mots (méthodes inspirées par l'école de pensée française). Celles qui mettent en valeur d'autres types de relations bibliographiques (relations entre les codes de classification documentaire, relations entre auteurs, relations entre pays, relations entre deux entités bibliographiques différentes).

LES METHODES DES CO-CITATIONS

Les premières constructions de cartes ont été fondées à partir de réflexions se servant de la pratique de la citation comme élément relationnel entre les documents. Ces méthodes ont été développées par l'école de pensée américaine initiée par Garfield et Price. Le développement des instruments en eux-mêmes est le fruit de collaborations entre l'ISI (Garfield, Small) et l'Université de Drexel (White, Griffith, McCain). Les données traitées sont, par la nature même du principe de ces méthodes, collectées à partir des sources de l'ISI.

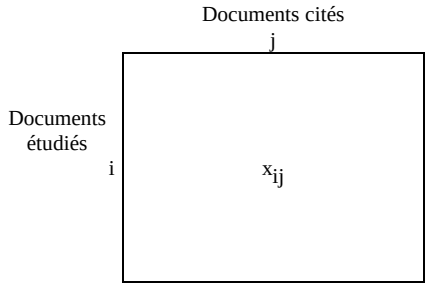
L'association bibliographique (*bibliographic coupling*)

C'est Kessler qui, le premier, a imaginé d'enrichir la méthode statistique des citations par l'apport de techniques mathématiques servant à formaliser et à mesurer les liens d'interaction entre des groupes d'auteurs. Inspiré par les travaux de Fano, il a élaboré la méthode d'analyse bibliométrique par

[1] Les logiciels traditionnellement employés sont SAS et SPSS. Très certainement, les logiciels les plus complets et les plus performants mais qui obligent à un investissement considérable à la fois en argent et en temps d'apprentissage. Mis à part les grands instituts qui disposent de ces outils installés sur leur centre de calcul, les autres centres de recherche en bibliométrie se servent de logiciels statistiques beaucoup plus modestes : StatGraphics, Statistica, Statitcf, Clustan, Spad.N, Tétralogie...

association bibliographique (*bibliographic coupling*) (KESSLER, 1963). Kessler a postulé que des articles scientifiques entretiennent une relation significative entre eux quand ils ont une ou plusieurs citations identiques. Le nombre de ces citations communes détermine la force de l'association. Il a conçu un tableau en affectant à chaque ligne un document étudié et à chaque colonne l'ensemble des citations effectuées (tableau 4). Ce tableau traité par classification automatique permet de construire des agrégats de documents selon la ressemblance de la pratique de citation de leurs auteurs. Kessler, comparant les résultats de sa méthode à ceux d'une analyse sur les thèmes indexés, a conclu qu'il y avait une très forte corrélation des groupes formés par ces deux méthodes (KESSLER, 1965). Cette méthode est malheureusement tombée en désuétude, probablement parce qu'elle nécessitait une masse de données trop importante pour les systèmes informatiques de l'époque : elle impose la présence d'une entrée de tableau (ligne) par document à étudier.

Tabl. 4 - Relations dans l'association bibliographique



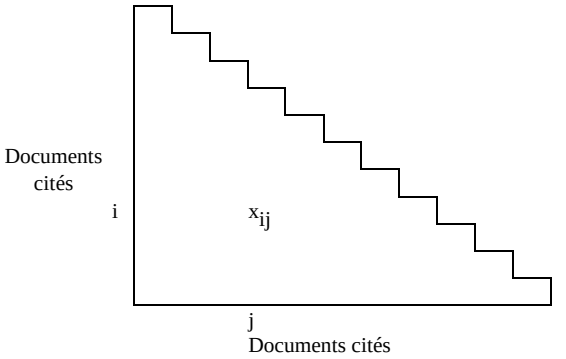
x_{ij} = 1 si le document *i* cite le document *j* sinon 0

L'analyse de la co-citation de documents (*document co-citation analysis*)

Small s'est inspiré des premiers résultats de Kessler pour développer une méthode de cartographie certainement la plus employée en bibliométrie : l'analyse de la co-citation. Pour pallier le problème du surplus de données à traiter, cette méthode ne conserve pas dans l'analyse statistique l'information concernant les documents "citants". Ainsi, la méthode statistique ne va pas servir à mesurer les ressemblances entre les documents citants mais les ressemblances entre les documents cités par ceux-ci. Pour estimer cette ressemblance entre les documents cités, la métrique calculée est basée sur la mesure de co-citation. La co-citation entre deux documents cités correspond au nombre de documents qui citent simultanément ces deux documents. Le tableau de relation, introduit en entrée de l'analyse statistique, est donc la matrice carrée mettant en regard les documents cités avec eux-mêmes (tableau 5). Ce tableau perd la trace des documents à l'origine des citations.

Il est d'abord traité pour normaliser la mesure de distance entre les documents cités, et ensuite introduit dans une analyse d'agrégation de type classification à liens simples. Les regroupements obtenus de ces citations sont traditionnellement dessinés sur un plan où les points, symbolisant ces citations, sont disposés par une méthode de cadrage multidimensionnelle non métrique.

Tabl. 5 - Relations dans l'analyse des co-citations de documents



x_{ij} = nombre de documents qui ont cité le document *i* et le document *j* en même temps

Une fois cette structure établie, on cherche à connaître les ensembles de documents ayant cité ces groupes de documents. Selon ce concept, les groupes de documents cités constituent les "souches de littérature-coeur" agrégeant un "front de recherche" (les documents "citants") autour du consensus scientifique qu'elles représentent. Ces souches correspondent généralement à des spécialisations intellectuelles pour un sujet. Les documents citants peuvent, par conséquent, être associés à plusieurs souches et donc être présents dans plusieurs fronts de recherche (certains documents peuvent faire référence à des travaux des différents domaines). Ces recouvrements permettent de calculer la force de lien entre les différentes spécialités que représentent les souches.

Cette méthodologie développée en 1973 (SMALL, 1973), a été utilisée pour de multiples études, mais la plus importante est celle que Small et Garfield ont dirigée pour découper la base SCI en fronts de recherche. En 1978, la première version de l'analyse a strictement suivi la procédure de la méthodologie exposée. Le nombre de documents cités à traiter étant colossal pour l'ensemble de la base SCI, les seuils de citations et de co-citations des documents étaient très élevés. Par conséquent, l'analyse défavorisait les disciplines dont la pratique de citation est faible. Des disciplines, pourtant importantes, comme les mathématiques et certaines sciences appliquées ne structuraient aucun front de recherche.

En 1985, un nouveau modèle est construit (SMALL et alii, 1985). La mesure des co-citations est calculée à partir d'une pondération. Chaque citation est divisée par le nombre de citations présentes dans le document citant. De plus, le modèle consiste à créer des agrégats de taille identique en utilisant des seuils de co-citations variables. Troisièmement, la carte finale sera "dégrossie"

par itération de la procédure pour créer des agrégations emboîtantes. Ainsi, pour les documents de l'année 1984, à partir de 13 931 paires de liens entre documents cités, on obtient 3 932 agrégats de 49 documents maximum. Pour ces agrégats, des forces d'association sont alors calculées en fonction des recouvrements des fronts de recherche. La procédure d'agrégation est ré-exécutée pour ces nouvelles relations, 502 agrégats sont obtenus en deuxième génération. Puis une troisième itération crée 57 agrégats finaux. Ce résultat nous suggère une décomposition de la science en 57 "secteurs" liés entre eux. Leurs liens sont représentés sur une carte que l'ISI aime nommer "Atlas de la science".

L'analyse de la co-citation d'auteurs (*author co-citation analysis*)

L'étude des co-citations d'auteurs déplace l'unité d'analyse du document individualisé au groupe de documents identifiables comme l'oeuvre d'un auteur. Ce changement entraîne une perte de finesse des structures des connaissances obtenues par l'analyse de co-citations de documents, mais il focalise l'attention sur une durable et intéressante unité à mi-chemin entre les documents et les revues : les auteurs eux-mêmes.

Les données sont donc obtenues par comptage du nombre de documents qui citent deux auteurs simultanément. La matrice d'entrée est donc un tableau de paires d'auteurs co-cités. Et les traitements statistiques sont les mêmes que ceux exposés pour l'analyse des co-citations de documents. Le résultat des "constellations" de points sur ces cartes représente des structures non plus construites par des consensus autour de travaux précis mais sur l'apport global d'un auteur à sa discipline. Les coeurs de littérature obtenus symbolisent en général les grandes écoles de pensée dans le domaine étudié.

La technique a été introduite par White et Griffith pour cartographier la science de l'information (WHITE et GRIFFITH, 1981b). Ils estiment que les auteurs sont proches sur la carte non seulement parce qu'ils ont en commun un thème de prédilection ou une méthodologie, mais ces rapprochements expriment aussi des liens de type collaboration. Une telle cartographie décrit à la fois le jeu social et la structure intellectuelle des relations.

Beaucoup de chercheurs l'ont appliquée pour diverses études. Nous mentionnerons notamment les travaux du français Penan qui propose d'ajouter en fin de procédure une analyse lexicale des titres des documents des fronts de recherche (simple comptage de mots). Ceci doit permettre de donner des libellés à chacun des fronts de recherche de façon pratiquement automatique alors que, jusqu'à présent, les cartes imposaient l'intervention d'un expert pour donner des noms aux groupes (PENAN, 1992).

Critique des méthodes de co-citations

De nombreuses critiques ont été exprimées concernant la conception théorique des analyses de co-citations et de la technique mathématique appliquée. Nous avons déjà énuméré les critiques portées sur la validité des

mesures basées sur la citation (voir p. 52 et 57). Nous n'y reviendrons pas. Nous donnerons ici uniquement les critiques se rapportant à la méthode des co-citations en elle-même.

En premier lieu, nous pouvons évoquer l'exemple caricatural de Sigogneau qui présente un agrégat-coeur construit par l'analyse de co-citation de la base SCI en 1978 (SIGOGNEAU et alii, 1990). Cette souche est formée de 7 articles qui se répartissent en fait en deux groupes distincts de collaborations. L'examen du front de recherche associé à ce coeur laisse apparaître que la quasi-totalité des documents citants ont été rédigés par des auteurs de ces deux groupes de travail. Cet exemple démontre que cette souche n'a pas été dégagée à partir d'un consensus de publications concernant les travaux antérieurs, mais uniquement sur un fort taux d'auto-citations ou, tout simplement, par une hégémonie de ces deux groupes dans la spécialité.

On peut aussi relever une étude approfondie de validation de l'analyse des co-citations (ainsi que de l'analyse des mots associés) commanditée par le *United Kingdom Advisory Board for Research Council* (HEALEY et alii, 1986). Les résultats livrés par le CSI ont été confrontés à des interviews de spécialistes scientifiques. Les principaux points de critiques formulés en conclusion sont :

1) la méthode des co-citations réduit les risques de mauvaises estimations en se basant sur l'évaluation des articles par le "plébiscite" qu'en font d'autres scientifiques du domaine en question. Ce conservatisme a pour effet de limiter la capacité à prendre en compte les travaux récents. Ceci couplé à l'inertie de la citation (période d'attente avant que le taux de citations soit notable), montre que cette méthode ne révèlera jamais l'émergence des nouvelles spécialités en temps réel.

2) ce qui intéresse le plus les décideurs ce ne sont pas les finesses de description des domaines, mais plutôt les interfaces entre les domaines car elles sont sources de renseignements précieux. Or le déséquilibre de la pratique de citation pour chaque discipline de la science déforme ces zones interdisciplinaires.

Des critiques ont été aussi formulées sur les méthodes mathématiques que ces analyses de co-citations appliquent. Leydesdorff a, par exemple, fait remarquer que la technique "standard" de classification à lien simple est totalement inadaptée pour les analyses de co-citations (LEYDESDORFF, 1987a). Comme pour l'évaluation britannique citée ci-dessus, l'*Advisory Council for Science Policy (RAWB)* hollandais a financé une étude sur le sérieux de la méthode. Le RAWB n'a pas fait reposer son jugement uniquement sur l'interprétation des résultats par des experts, mais aussi sur une estimation de la solidité statistique des techniques mathématiques employées. Comme il n'y a aucun moyen d'accéder à la même source d'information que celle utilisée par l'ISI pour construire cette étude, il a établi un projet de conception d'une simulation informatique par génération aléatoire, sous des contraintes, d'un tissu de co-citations dans un corpus bibliographique virtuel (répartition Zipfienne des

citations, nombre de citations par référence, valeur de la citation maximale...) (OBERSKI 1988). En conclusion, l'instabilité statistique des résultats des structures de groupes de l'analyse de co-citations ne semble pas avoir été correctement appréciée par l'ISI. Et l'étude révèle de sérieux problèmes. Ces problèmes suggèrent que les résultats de l'analyse ne peuvent être pris sérieusement comme une preuve pertinente de formulation de l'activité de recherche. Une ultime précision formule qu'il est inimaginable, dans une étude approfondie, de ne pas pouvoir accéder aux données dont l'ISI se sert pour leur analyse.

En conclusion, on peut dire que la co-citation reflète la science comme les scientifiques la perçoivent et plus sous un aspect historique et sociologique que sous la dimension d'aide à la prospective.

L'analyse des citations croisées de revues (*cross-citation analysis*)

Pour décrire un domaine de recherche plus étendu qu'un front de recherche et au moyen des citations, les études bibliométriques se fondent sur une unité d'analyse encore plus large que l'unité des auteurs : les revues. Elles tracent des cartes de citations reliant les revues du domaine étudié. Cette méthode part du postulat que les revues qui se citent mutuellement mettent en évidence des rattachements disciplinaires. Donc elle cherche à décrire, pour l'ensemble des périodiques, leur réseau de communication pour identifier les revues centrales et périphériques dans la spécialité, l'existence de sous-spécialités, et pour modéliser le flot d'information transitant entre les revues.

Différentes approches pour identifier les groupes de revues sont proposées, mais toutes construisent un tableau de relation des citations-croisées (*cross-citation*) comportant en ligne les périodiques citants et en colonnes les périodiques cités (tableau 6). Le choix de la période T_C est très important pour construire des mesures appropriées et ensuite les interpréter. Pour que ce genre de matrice soit interprétable il faut généralement qu'il y ait peu de décalage entre T_C et T_r pour représenter un sujet de manière "constante".

Tabl. 6 - Relations dans l'analyse des citations croisées

		Revues citées pour la période T_C j
revues citantes à la période T_r	i	x_{ij}

x_{ij} = nombre d'articles dans la revue *i* qui citent des articles du périodique *j*

Il est évident qu'une forte activité de publication pour une revue influencera son taux de citations. Pour réduire cette influence, les auteurs ont envisagé de nombreuses possibilités de normalisation. Dans un article Todorov et Glanzel

ont récapitulé les principaux calculs proposés dans les travaux antérieurs (TODOROV et GLANZEL, 1987). Ils ont traité aussi des propositions faites pour inhiber les fortes valeurs de la diagonale caractéristiques de ce type de matrices (les articles citent souvent des articles de la même revue).

Différentes méthodes mathématiques peuvent être ensuite appliquées à cette matrice normalisée. Leydesdorff (LEYDESDORFF, 1986) a appliqué des méthodes d'analyse factorielle ou de positionnement multidimensionnel qu'il a ensuite comparées avec des méthodes de classifications automatiques – simple lien et Ward – (LEYDESDORFF, 1987a). Doreian a mis au point une technique basée sur la méthode du *block-modelling* (DOREIAN, 1985). Narin et Carpenter (CARPENTER et NARIN, 1973), Niyamoto et Nakayama (NIYAMOTO et NAKAYAMA, 1983) ont eux exploité les méthodes de regroupement traditionnelles. Agirre et al. ont appliqué l'analyse des correspondances (AGIRRE, 1991).

La principale critique à l'encontre de cette méthode est que la seule source des données est le *Journal Citation Reports* de l'ISI qui, comme on l'a déjà indiqué (voir p 56), contient un pourcentage d'erreurs non négligeable et une couverture restreinte et hétérogène.

LES METHODES DES CO-OCCURRENCES DE MOTS

Cette méthode établit l'analyse des concepts introduits dans les publications scientifiques en considérant que certains mots-clés présents dans les références bibliographiques, mots du titre (affectés par les auteurs) ou descripteurs (affectés par les indexeurs), reflètent les étapes de l'argumentation scientifique des auteurs. Quand une paire de mots(-clés) est utilisée pour indexer un grand nombre d'articles, ces mots représentent une forte association entre les problèmes ou les concepts auxquels ils se réfèrent. Donc cette méthode est basée sur l'étude des co-occurrences de mots par des méthodes statistiques pour découvrir les agrégats de mots, symbolisant les thèmes scientifiques, et leurs situations les uns par rapport aux autres.

La méthode des co-occurrences de mots a été presque exclusivement conçue par une école de pensée française. Elle est, en fait, la conséquence d'une conjonction entre le principe de modélisation de la science (bibliométrie-scientométrie) et l'approche sociologique de la science (représentation sociale de la connaissance). La collaboration entre le CSI[1] et l'INIST[2] est à l'origine de la recherche et du développement de cette méthode principalement mise au point pour traiter les termes d'indexation de la banque scientifique multidisciplinaire *Pascal* produite par l'INIST. Elle a été ensuite reprise par divers auteurs, principalement européens (allemands, britanniques, hollandais, français).

[1] Centre de Sociologie de l'Innovation

[2] Institut National d'Information Scientifique et Technique

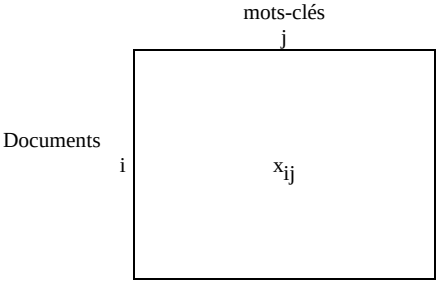
L'analyse des mots associés (*co-word analysis*)

La méthode d'analyse des mots associés est un projet de longue haleine, mené conjointement par le Centre de Sociologie de l'Innovation de l'École des Mines de Paris et l'Institut National d'Information Scientifique et Technique du CNRS. Sa dernière version a été mise au point par la thèse de Michelet et a abouti à la création d'un logiciel spécifique nommé *Leximappe* (MICHELET, 1988). C'est ce logiciel qui est exploité actuellement dans de nombreuses études (LAW et alii, 1988 ; CALLON et alii, 1991 ; DE LOOZE et JOLY, 1993) et qui sera exposé dans ces lignes. Il fait aussi l'objet de projets de développement en tant que moteur de nouveaux logiciels d'analyse bibliométriques. Ces projets conduits indépendamment au CERESI[1] (CARDINE et alii, 1992) et à l'INIST (POLANCO et GRIVEL, 1993) ont des objectifs similaires : favoriser la navigation entre la structure obtenue par l'agrégation statistique et les références bibliographiques analysées.

• *La méthode Leximappe*

La première étape de la procédure est la construction d'une matrice Documents *x* Mots-clés à partir de l'extraction des termes d'indexation (ou mots du titre) présents dans l'ensemble des références étudiées, jusqu'à concurrence de 1500 mots-clés (tableau 7). Puis, une matrice carrée d'association des mots-clés est calculée à partir du coefficient d'équivalence pour mesurer l'éloignement statistique des mots-clés (analogue à une matrice de co-citation normalisée).

Tabl. 7 - Relations dans l'analyse des mots associés



x_{ij} = nombre de fois que le mot *j* est présent dans le document *i*.

Ensuite, la méthode réalise des agrégats de mots-clés sur la base de mesure de "distance" entre mots-clés, mesures récapitulées dans la matrice carrée d'association (le principe de regroupement employé a été spécialement développé pour la méthode des mots associés). Contrairement aux autres méthodes bibliométriques, celle-ci n'applique pas une méthode statistique reconnue et entérinée par les mathématiciens. L'algorithme commence comme une classification à lien simple mais il impose au cours des regroupements une

[1] Centre d'Etude et de Recherche en Sciences de l'Information

taille maximale de 10 mots-clés par agrégat. Les classes résultantes sont donc très hétérogènes : la première sera constituée des mots-clés les plus fortement liés alors que la dernière sera constituée de tous les mots-clés "rebuts", puisque très faiblement liés à tous les autres.

Le regroupement en agrégats étant terminé, la méthode les dispose sur un plan selon un système à deux axes perpendiculaires. L'un des deux axes répartit les agrégats selon leur "densité" (mesure de la cohésion interne de l'agrégat), et le second selon les liens qu'ils entretiennent avec les autres agrégats, la "centralité" (mesure de l'intensité des liaisons avec des mots-clés d'autres agrégats). Chaque agrégat est symbolisé sur le graphe par un des mots-clés du groupe sélectionné automatiquement par un indice. Cette représentation graphique est désignée sous le nom de "*diagramme stratégique*" par ses concepteurs.

• *Les inconvénients de la méthode Leximappe*

Ce n'est pas le principe de l'analyse qui est remis en cause ; une analyse des mots présents dans les références, qu'ils soient donnés par l'auteur lui-même ou par un indexeur, est sans aucun doute source de grands renseignements. Le principal reproche, concernant la chaîne des traitements informatiques, est de ne pas avoir suffisamment permis de jouer sur les critères de ces traitements en fonction des données à étudier : seuil de troncature automatique, agrégat de mots de même taille, mesure des "distances" entre mots fixée une fois pour toutes et principe d'agrégation irrévocable. *Leximappe* automatise complètement l'analyse d'un corpus, de l'information brute jusqu'à sa présentation analytique, sous la forme d'un diagramme. Cette méthode fermée fait du logiciel un outil très facile à mettre en oeuvre (ce qui a permis à Michelet de la présenter dans sa thèse comme une *"boite noire"*). Mais en contrepartie, l'utilisateur du logiciel n'a pas la possibilité de nuancer son étude en fonction de la spécificité des données de départ.

Un récent débat s'est instauré entre Leydesdorff et Courtial sur les spécificités des résultats statistiques produits par *Leximappe*. Leydesdorff a avancé, sur la base d'une démonstration mettant en oeuvre des calculs d'entropie, l'existence d'une dépendance entre les deux axes stratégiques de *Leximappe* et la non-reproductibilité des résultats de *Leximappe* avec une quelconque autre méthode statistique (LEYDESDORFF, 1992 ; COURTIAL, 1992).

La méthode a été conçue pour mettre en évidence le réseau des relations entre les problématiques socio-techniques qui existent dans les divers secteurs de la recherche. Les descripteurs présents dans les notices bibliographiques sont considérés comme les représentants synthétiques des ces problématiques. Ainsi *Leximappe* a été mis au point pour traiter les descripteurs de la banque scientifique *Pascal*. Lors d'études ultérieures, cette méthode a été employée avec d'autres champs où les mots présents ont été jugés comme de bons représentants de ces problématiques : le champ "titre" pour les publications scientifiques ou le champ "titre normalisé" pour les références de brevets de la

base *Derwent*. Cette méthode est donc confinée à l'analyse du contenu d'un seul champ et pas n'importe lequel.

• *Analyse des mots associés modifiée par Law et Whittaker*

Law et Whittaker ont conduit une étude sur les mots-clés d'un échantillon de références collecté sur *Pascal* (LAW et WHITTAKER, 1992). Ils ont ensuite découpé l'ensemble des références en cinq périodes de temps pour les examiner à l'aide d'une analyse des mots associés basée sur les mêmes principes que *Leximappe*. L'apport méthodologique réside essentiellement en deux points. Le premier point correspond à une amélioration des graphes fournis pour chacune de ces périodes. Ils sont construits de façon :

1) à disposer les agrégats concernant les mêmes concepts à peu près aux mêmes positions sur les 5 graphes

2) à symboliser les intensités de relation entre et à l'intérieur des agrégats par des nuances d'épaisseur de traits

3) à rendre compte du nombre de documents contribuant à la création de ces agrégats par des surfaces de carrés différents.

La seconde amélioration est le calcul de deux nouveaux indices pour mesurer le chevauchement entre les thèmes de sujets similaires qui surviennent au cours des périodes successives. Ce calcul permet de construire des graphes qui retracent la "génération" des thèmes.

Les autres méthodes d'analyse des co-occurrences de mots

D'autres auteurs ont repris l'idée formulée par l'École française. Nous présenterons quelques uns de ces travaux en nous attachant particulièrement aux différences avec la méthode d'origine.

• *Analyse des mots du titre par Leydesdorff*

Pour éliminer un possible "effet d'indexation" pendant une étude expérimentale (analyse de seulement 57 documents d'un laboratoire), Leydesdorff a utilisé les mots du titre et du résumé des documents originaux (LEYDESDORFF, 1987b). Après élimination des mots en dessous d'un seuil de fréquence et des mots triviaux, il a construit la matrice carrée et symétrique du croisement des mots avec une diagonale vide qu'il transforme en matrice de corrélation de Pearson. Il a traité cette matrice par la méthode de Ward pour constituer les agrégats des mots. Il a expliqué qu'il a choisi cette procédure mathématique car la méthode des liens simples lui a semblé totalement inadaptée. Comme la matrice est presque "vide", elle génère un effet de chaînage sur le premier groupe. Il a remarqué que les mots des résumés sont moins spécifiques que ceux des titres et concordent moins bien avec les thèmes des documents. Mais la co-occurrence des mots semble très bien traduire la spécificité des différents axes de recherche auxquels le laboratoire se consacre.

• *Analyse de tableaux de contingence de mots-clés*

Dans les analyses présentées jusqu'ici, tous les mots-clés sont considérés dès lors que leurs fréquences sont supérieures à un seuil. La technique mathématique va donc traiter une matrice symétrique où chaque mot joue le même rôle. Or les mots-clés appartiennent à plusieurs catégories de sens selon qu'ils représentent un aspect technique, technologique, un composant chimique, une caractéristique physique, un traitement, une condition expérimentale, un matériel utilisé, un secteur d'activité... Il peut paraître intéressant de distribuer les mots-clés en deux catégories pour étudier l'influence de l'une sur l'autre (et vice-versa). Dans ce cas, le tableau construit ne contient plus tous les mots-clés mais que ceux intéressant l'étude. Ces tableaux croisant deux ensembles d'éléments distincts sont nommés par le vocabulaire statistique des "tableaux de contingences". Une technique statistique a été spécialement développée pour analyser ce type de tableau par le français Benzécri[1] : l'analyse de correspondance. Cette technique est certainement la plus adaptée pour l'analyse des tableaux de correspondances.

Les études menées par le CETIM sont de très bons exemples de ce genre d'approche. Dans une étude (DEVALAN et alii, 1990) menée pour l'entreprise Burton-Corbin (fabricant de compresseurs), le CETIM a, dans un premier temps, construit une matrice croisant les mots-clés correspondant à des composants avec les mots-clés caractérisant une technologie ou un type de sollicitation (pour 30 000 références provenant d'une quinzaine de bases). Le graphe factoriel, obtenu par une analyse des correspondances de cette matrice, a donné une image assez générale qui a incité Burton-Corbin à recentrer l'étude sur un des agrégats dégagés, le domaine des compresseurs volumétriques. Une seconde matrice, croisant les mêmes catégories de mots-clés mais ceux-ci étant choisis à un niveau plus fin (plus de 1 000 documents concernés), a fourni une seconde structure plus détaillée, après une analyse des correspondances et une classification automatique.

La même technique a servi pour d'autres études bibliométriques du CETIM : analyse de tableaux de contingences technologies *x* marchés dans le domaine de la productique, équipements *x* composants pour les références de trois années de la base du CETIM, technologies *x* marchés dans le domaine des revêtements de surfaces (DEVALAN et alii, 1989 ; DEVALAN et alii, 1991).

La construction au CETIM de ce type de matrice a beaucoup évolué au cours du temps. Les premiers tableaux ont été construits manuellement. Une seconde phase a été de les constituer par l'utilisation détournée de certaines fonctions évoluées des langages d'interrogation de banques de données. Maintenant, ces matrices sont élaborées de façon automatique à partir des références télédéchargées grâce à l'appui du logiciel bibliométrique *DATAVIEW* ***URL*** conçu par le CRRM (DUMAS, 1994).

[1] BENZÉCRI JP, *L'analyse des données*, Tome 1 : La taxinomie, Tome 2 : L'analyse des correspondances, Editions Dunod, Paris, 1973

LES AUTRES ANALYSES DE RELATIONS BIBLIOGRAPHIQUES

Plus récemment, des chercheurs ont élargi l'éventail des analyses bibliométriques pour connaître des dépendances et interactions entretenues par d'autres éléments bibliographiques que les citations et les mots. Les premières méthodes de cartes relationnelles étaient fondées sur des principes sociologiques. Les méthodes statistiques employées ne sont, dans ces conditions, que des outils pour essayer de représenter ces concepts par l'intermédiaire de la littérature scientifique.

Les nouvelles méthodes s'éloignent des premières par leur approche moins conceptuelle et plus pragmatique. Plus aucun phénomène sociologique n'est à mettre en évidence, mais simplement une volonté de mieux comprendre le contenu d'un ensemble de références bibliographiques sélectionnées pour leur pertinence. La statistique descriptive est exploitée exactement pour ce qu'elle est : un outil de description de données et non de modélisation de phénomènes sociologiques. Dans ce contexte, toute catégorie d'information est analysable par des méthodes statistiques dès lors qu'elle peut être représentée sous une forme quantifiable. Pour l'information contenue dans les références bibliographiques, tout élément bibliographique peut faire l'objet d'une étude. Des analyses sur des unités bibliographiques comme les codes documentaires, les pays, les villes, les organismes, les auteurs sont alors apparues. Nous en présenterons dans les lignes qui suivent quelques exemples.

Nous commencerons par des méthodes qui n'expriment que les relations entre des éléments d'une seule catégorie (relation intra éléments bibliographiques, suivant le même principe que les analyses de co-citation et de mots associés). Toutes ces méthodes ont eu l'attribution du préfixe "co-". Une dernière partie présentera la richesse des méthodes cherchant à structurer les relations qui peuvent s'établir entre deux entités bibliographiques différentes, par exemple expliquer les ressemblances entre les pays (ou villes ou organismes ou mêmes auteurs) selon leurs activités scientifiques (traduites par une des catégories de descripteurs, codes, mots-clés contrôlés ou libres). Ces dernières décrivent alors les relations inter éléments bibliographiques.

L'analyse des co-classifications documentaires

Un des systèmes d'attribution de descripteurs à des références dans les banques de données est l'affectation de codes de classification documentaire. La classification documentaire découpe la science en secteurs d'activité disjoints. Chaque secteur est identifié par un code (succession de caractères alphanumériques). Par exemple, la banque de données *Chemical Abstracts* découpe la chimie en 80 sections ; elle est donc identifiée par 80 codes différents. Un article en chimie aborde bien souvent plusieurs aspects de la chimie et concerne donc plusieurs sections du découpage. Autant de codes CAS seront attribués à la référence de cet article que de sections concernées. L'utilisation de ces codes s'est avérée propice à la réalisation d'études

bibliométriques. Un tel découpage permet l'analyse des thèmes de recherche de la même façon que les études portant sur les mots-clés. L'indexation par codes est parfois préférée à l'indexation par mots-clés dans les études bibliométriques pour les raisons suivantes :
1) condensation de concepts en un seul code donc sens plus synthétique
2) absence de synonymie donc aucune ambiguïté sémantique
3) pérennité dans le temps donc pas de déviances du langage
4) diversité des termes plus faible donc meilleure qualité statistique
5) traitement plus facile donc gain de temps.

Plusieurs travaux bibliométriques exploitent cette ressource documentaire en analysant la structure des relations des paires de codes, c'est-à-dire les co-occurrences de codes.

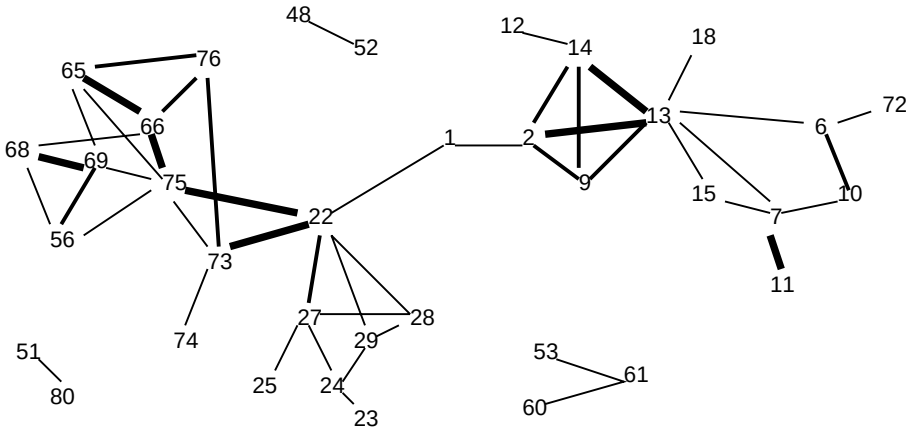
• *Réseau de paires de codes documentaires*

Dans les analyses bibliométriques relationnelles, le traitement le plus simple est la construction du réseau des co-occurrences. Cette construction est simple parce qu'elle ne fait appel à aucun calcul mathématique. Elle offre l'avantage de livrer des cartes dont l'interprétation est immédiate car elle exprime sous forme visuelle la notion de fréquence de co-apparition des éléments bibliographiques. En fait, on peut considérer que le principe est le même que celui d'une analyse statistique. La mesure de la "distance" entre deux éléments n'est pas calculée par une "normalisation" des données mais correspond au résultat brut du dénombrement des co-occurrences : la fréquence de co-apparitions. Cette méthode privilégie donc les relations entre les éléments à très fortes fréquences en négligeant les relations entre les éléments rares.

L'équipe du CRRM exploite depuis longtemps ce mode de représentation des relations entre des unités bibliographiques et tout particulièrement pour l'analyse des codes documentaires. Elle s'est particulièrement penchée sur le développement de logiciels spécifiques pour traiter les références bibliographiques collectées sur les serveurs de banques de données. Ces logiciels effectuent à la fois les comptages des occurrences des codes mais aussi des co-occurrences pour générer les graphes de réseaux de relations entre les codes (DOU et alii, 1989b ; DOU et alii, 1989c ; QUONIAM et alii, 1991). Le graphe d'un réseau de codes symbolise les codes sous la forme de points dans un plan et les relations entre ces codes sous la forme de traits (figure 7).

Les nuances d'épaisseur des traits indiquent la variation d'intensité du lien, c'est-à-dire la fréquence de co-apparition des deux codes. Les chercheurs ont, par exemple, collecté sur la banque de données des *Chemical Abstracts* tous les articles publiés en chimie par des laboratoires de Marseille pendant dix ans. L'analyse des cartographies de réseaux de codes CAS tracées année par année leur a permis de montrer l'évolution de la politique recherche en chimie à Marseille, ainsi que la dégénérescence de certaines structures thématiques au profit de nouvelles.

Fig. 7 - Cartographie d'un réseau de paires de codes documentaires CAS



• *Analyse des co-codes (co-heading analysis):*

Todorov et Winterhager ont imaginé, en 1990, un nouveau traitement bibliométrique pour analyser les codes de classification documentaire (TODOROV et WINTERHAGER, 1990). Le principe de la méthode est le suivant :

1) comptage des occurrences et des co-occurrences de codes, puis construction de la matrice triangulaire des relations entre codes (voir tableau 8) pour les codes apparaissant le plus souvent (un logiciel spécifique a été développé)

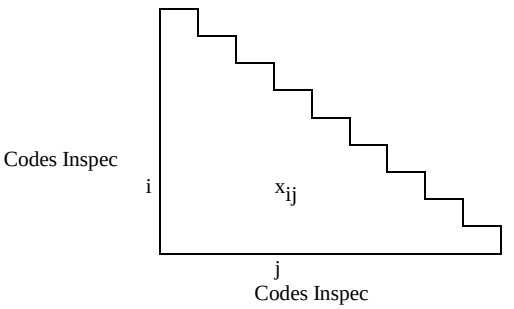
2) calcul de la matrice des "distances" des relations entre les codes (mesure par l'indice d'inclusion)

3) cadrage multidimensionnel de la matrice des relations pour positionner les points sur un graphe plan (programme ALSCAL)

4) définition des agrégats de codes par une classification hiérarchique ascendante à partir de la matrice des relations

5) dessin sur le graphe de la symbolisation des agrégats et des forces de relations les plus élevées par des traits reliant les points

Tabl. 8 - Relations dans l'analyse des *co-headings*



x_{ij} = nombre de publications indexées conjointement par le code *i* et le code *j*

Les auteurs ont appliqué leur méthode dans des études à partir de la classification documentaire présente sur la banque de données *Inspec* (TODOROV et WINTERHAGER, 1991 ; TODOROV, 1992). Pour eux, les codes sont considérés comme des mots-clés, mais à un niveau d'agrégation de sens plus élevé. Ils estiment que leur emploi offre des avantages considérables tels que garder un sens constant entre les sous-domaines, ne pas dépendre du langage de l'auteur, ne pas avoir de limites de couverture comme pour le SCI (limité aux articles des périodiques les plus cités), et être plus objectif que les mots employés par l'auteur car ils sont généralement basés sur un jugement extérieur opéré par un indexeur.

Dans leurs études, ils ont parallèlement étudié l'information contenue par les termes contrôlés d'*Inspec* présents dans le champ CT *(Controlled Terms)*. Les résultats sont identiques, mais les renseignements fournis pour les termes contrôlés sont plus détaillés.

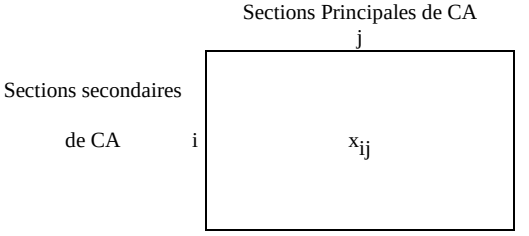
• *Analyse des co-sous-domaines (co-subfield analysis)*

Après avoir employé dans une étude précédente trois méthodes d'analyse bibliométrique pour caractériser la dynamique de l'ingénierie en chimie (une analyse de citations entre revues, une analyse des relations entre revues et thèmes par l'intermédiaire de codes de classification et une analyse des *co-words* sur les mots-clés), Van Raan et Peters ont présenté une méthode basée sur l'étude des relations entre les codes de classification pour le même domaine (VAN RAAN et PETERS, 1989). Le but est de décrire le transfert des connaissances entre différents thèmes et, si possible, de tracer des processus de synthèse des connaissances.

La méthode propose les étapes suivantes :

1) traitement des 80 codes de classifications des *Chemical Abstracts* pour construire le tableau des co-occurrences (tableau 9) entre les codes principaux du champ SC (*Main Sections*) et les codes secondaires du champ SX (*Cross Sections*), de façon à construire les structures en réseaux des thèmes pour des périodes de temps successives (1977-79, 1980-82, 1983-85). Pour cette étude, le dénombrement des croisements de codes est effectué par interrogation en ligne en réalisant toutes les combinaisons SC *x* SX *x* période.

Tabl. 9 - Relations dans l'analyse des *co-sous-domaines*



x_{ij} = nombre de publications, pendant une période donnée, indexées conjointement par le code principal *j* et le code secondaire *i*

2) application de la méthode d'analyse par quasi-correspondance *(quasi-correspondence analysis* QCA) qu'ils ont mise au point pour disposer les codes sur un espace à deux dimensions.

En comparant ces résultats à ceux obtenus par les précédentes études réalisées sur les données, les auteurs concluent sur deux observations majeures. Premièrement, l'analyse des co-occurrences de mots-clés produit des résultats plus fins et, contrairement à une classification focalisée sur le découpage d'une seule discipline (ici la chimie), elle décrit bien l'émergence d'applications de techniques appartenant à une autre discipline. Dans le cas étudié, l'analyse des co-occurrences de mots a montré le récent développement du thème modèles mathématiques, tandis que dans l'analyse des "co-thèmes" ce thème est noyé dans une section générale (section 48 : *unit operations & processes*). Par contre, l'analyse des co-occurrences de mots étant trop précise sur les aspects spécifiques, il n'est pas possible comme dans la méthode des co-thèmes d'obtenir un aperçu global des relations entre la discipline étudiée et les autres disciplines. Deuxièmement, les inconditionnels de l'analyse des co-citations affirment qu'il est important de présenter une discipline selon une structure établie sur ses anciens fondements. Mais ceci ne permet pas de donner un aperçu clair en terme de relations entre les sous-domaines à un niveau macroscopique.

Ils considèrent donc que cette technique est à prendre en considération au même titre que les autres, chacune apportant une vision différente.

L'analyse des co-signatures

Price et Beaver ont été les premiers à avoir utilisé les relations de co-auteurs pour "enquêter" sur les structures sociales et leurs influences en science, et spécialement les réseaux de la communication scientifique (PRICE et BEAVER, 1966). Au cours de leurs "manipulations expérimentales", ils ont découvert ce qu'ils ont nommé les "collèges invisibles" (*Invisible colleges*). Ces manipulations avaient pour finalité la reconstitution des groupes de collaboration autour d'un auteur. Ils ont commençé par rechercher tous les auteurs qui avaient travaillé avec l'auteur en question, puis les nouveaux auteurs qui avaient publié avec ces derniers et ainsi de suite...

• *L'analyse des co-auteurs (co-author analysis)*

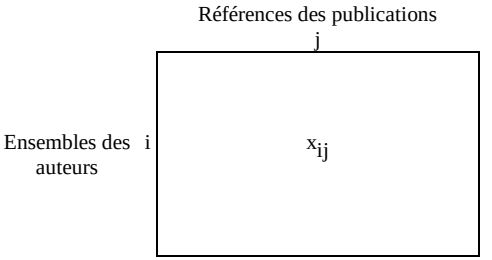
En 1991, Peters et Van Raan ont repris le concept de Price et Beaver et imaginé pouvoir appliquer une méthode de classification automatique pour agréger les auteurs en groupes (PETERS et VAN RAAN, 1991). La méthode passe évidemment en premier lieu par une étape d'homogénéisation des noms d'auteurs et de leurs affiliations (cette étape restant encore manuelle). Puis vient ensuite la succession d'étapes assez classiques :

1) construction d'une matrice de co-auteurs (voir tableau 10)

2) regroupement des auteurs par la méthode de classification à lien simple sur la mesure d'association dite du cosinus entre auteurs

3) représentation plane des groupes par construction manuelle ; c'est une présentation similaire à un réseau de paires de codes, où les nuances des traits entre les auteurs symbolisent des intervalles de valeurs des mesures d'associations du cosinus.

Tabl. 10 - Relations dans l'analyse des co-auteurs



x_{ij} = 1 si l'auteur i a signé ou co-signé pour l'article j

Ils finissent par conclure que l'analyse des co-auteurs permet de dégager plusieurs types de renseignements. Tout d'abord, elle révèle les liens intellectuels et/ou les cohésions sociales entre les individus mais ne permet pas de différencier les groupes liés pour une raison intellectuelle de ceux liés pour une raison sociale. Par des comparaisons dans le temps, on peut aussi mettre en évidence les évolutions des groupes. Mais de plus, cette structuration permet d'identifier les spécialités phares ou pivots ainsi que les *leaders* dans les spécialités.

• *Réseaux de compétences*

Comme pour les réseaux de paires de codes, l'équipe du CRRM a mis au point des programmes informatiques pour construire des réseaux de relations entre chercheurs (ou inventeurs). L'un de ces logiciels comporte un module de création automatique de groupes de compétences scientifiques (ou techniques) sur une structure de réseaux d'auteurs (ou d'inventeurs). Le principe de cette méthode a été décrit au colloque *IDT 93* à la suite d'une collaboration effectuée avec le CEDOCAR (HAON et alii, 1993). Le traitement informatique enchaîne automatiquement trois traitements successifs à partir d'un télé-déchargement de références d'une banque de données (dans l'étude la base JICST) :

1) Constitution des groupes de collaboration entre auteurs. Cette procédure itérative est la parfaite reproduction de ce qui avait été réalisé à l'époque par Price et Beaver, si ce n'est qu'elle est automatisée jusqu'à épuisement du chaînage des relations (aucun nouvel auteur raccroché au groupe constitué à l'étape n - 1)

2) Attribution des affiliations des auteurs constituant les groupes (*JICST* a la caractéristique de posséder les multi-affiliations avec détermination de leurs attributions aux auteurs, contrairement à la base *SCI* de l'ISI)

3) Attribution des activités spécifiques à chaque groupe (les codes de la classification documentaire utilisés par les auteurs du groupe sont rattachés à celui-ci)

Chaque groupe est finalement représenté sous la forme d'un graphe de réseaux d'auteurs surligné par les attributions d'affiliations et d'activités (figure 8).

Cette procédure présente de nombreux avantages. Premièrement, l'étude des collaborations entre organismes devient possible par l'intermédiaire de la structure des relations appartenant à ces organismes avec un minimum d'erreur (problème de la variante de la saisie des affiliations des organismes, voir p. 58). Deuxièmement, la méthode détecte les réseaux de compétences, c'est-à-dire l'identification des collaborations entre des spécialistes de différents domaines. Troisièmement, l'analyse des réseaux de compétences met en évidence les frontières entre les différentes spécialités du domaine étudié par l'intermédiaire des acteurs de ce domaine. De plus, le découpage par années des références bibliographiques, puis la construction des réseaux de compétences pour chaque période, a permis aux auteurs d'étudier la dynamique de ces collaborations, ainsi que l'évolution des centres de compétences.

Plus récemment, cette méthode a été mise en pratique au cours d'une collaboration avec le LERASS[1] pour la détermination des physiciens français ayant une activité de recherche dans le domaine du processus de fragmentation en physique fondamentale (SURAUD et alii, 1994, 1995). L'identification de ces spécialistes entrait dans le cadre d'un projet de mise en place de Groupe de Recherche du CNRS sur le thème de la fragmentation. Cette méthode a été particulièrement pertinente pour ce projet, puisqu'il n'existe pas encore de consensus autour de ce thème et que ses spécialistes ne sont pas encore connus.

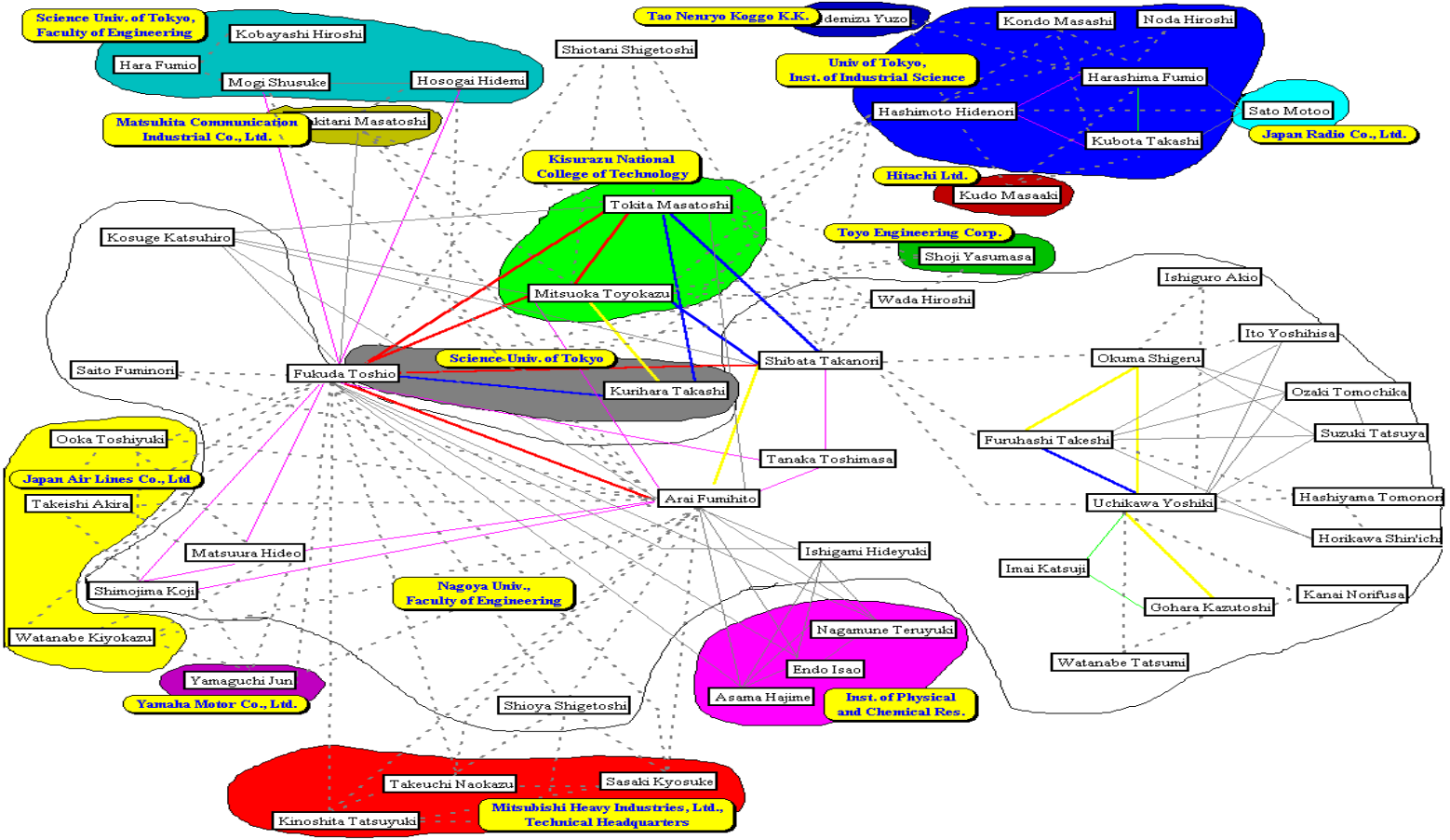
L'analyse des coopérations internationales

Pour cette méthode basée sur les relations intra entités bibliographiques représentant les pays, nous ne citerons d'une seule technique, celle employée par Moed et alii qui ont présenté sous forme cartographique les collaborations internationales des scientifiques des pays de la communauté européenne, c'est-à-dire représenté graphiquement toutes les publications fruits de l'aboutissement d'un travail entre deux équipes de pays différents, mais dont l'un des deux appartient à la communauté européenne (MOED et alii, 1991).

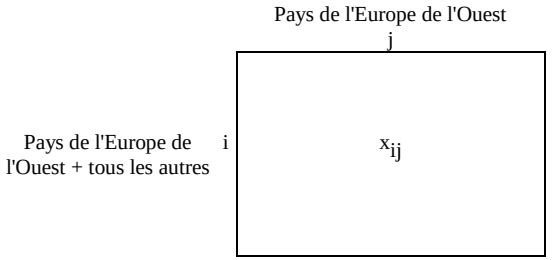
Il était donc indispensable d'utiliser une source qui contienne les données des différentes affiliations des auteurs. Les chercheurs ont collecté leurs données dans les banques de l'ISI (*SCI*, *SSCI* et *A&HCI*), puisque ce sont les seules banques multidisciplinaires qui saisissent toutes les affiliations des co-signataires. Deux ensembles de pays sont donc croisés pour constituer une matrice de fréquences de co-publications internationales (tableau 11).

[1] Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales, Université Paul Sabatier, Toulouse

Fig. 8 - Graphe d'un réseau de compétences pour un groupe de collaborations d'auteurs



Tabl. 11 - Relations dans l'analyse des co-opérations internationales



x_{ij} = nombre de publications co-signées par un organisme du pays *i* avec un organisme du pays *j*

Dans l'étude, la matrice traitée est composée de 9 colonnes comportant les pays de l'Europe de l'Ouest (Pays-Bas, Belgique, Danemark, RFA, Grande-Bretagne, France, Italie, Suisse et Suède) et 18 lignes comportant les 9 précédents pays plus : Afrique, Asie, Australie, Canada, Amérique centrale et Amérique du Sud, Europe de l'Est, Japon, reste de l'Europe de l'Ouest, États-Unis.

La carte est obtenue par injection de cette matrice "brute" dans une analyse des correspondances. On a donc bien évidemment les pays de l'Europe de l'Ouest représentés deux fois sur le graphe, une première fois en tant que pays qui entretiennent des liens avec l'ensemble des pays du monde et une seconde fois en tant que pays qui collaborent avec les pays de l'Europe de l'Ouest. Ainsi, la place du premier point est dépendante de ses relations mondiales et celle du second de ses relations intra-européennes. Les auteurs ont fait remarquer que les positions des pays sont étrangement similaires de leurs rapprochements géographiques. Ceci démontre que les chercheurs ont tendance à plus facilement entretenir des collaborations avec les pays qui partagent des frontières communes avec eux. Cette remarque a été confirmée par un travail de modélisation de la dépendance entre les collaborations intra-universités canadiennes et les distances géographiques qui séparent ces universités (KATZ, 1994).

Les analyses de tableaux de contingences bibliographiques

Nous venons de vérifier que certaines méthodes d'analyse bibliométrique ne sont pas basées sur la construction d'un tableau carré et symétrique. Toutes les techniques mathématiques n'imposent pas ce genre de tableau. Les bibliomètres ont su extraire de la multitude des techniques mathématiques de nouvelles possibilités. S'étant aperçus que certaines d'entre elles étaient bien adaptées aux traitements de matrices asymétriques, ils ont vu là de nouvelles ouvertures. Au lieu de combiner uniquement les données du même champ, pourquoi ne pas chercher des corrélations entre les données de champs différents ?

Dès lors que l'on croise des données de nature différente, la première analyse que l'on cherche à résoudre permet de répondre à l'interrogation du type

"qui fait quoi ?". Les travaux qui vont être décrits tentent tous de répondre à cette question mais pas toujours avec les unités bibliographiques à analyser.

• *Analyse de l'avantage national (relations pays x domaines)*

Ce type d'analyse est l'une des plus employées car il intéresse directement les centres d'évaluation nationaux qui élaborent des indicateurs pour définir la stratégie des politiques de programmation de la recherche. Dans ce contexte, le "qui" désigne les principaux pays industrialisés et le "quoi" les grands domaines de recherche.

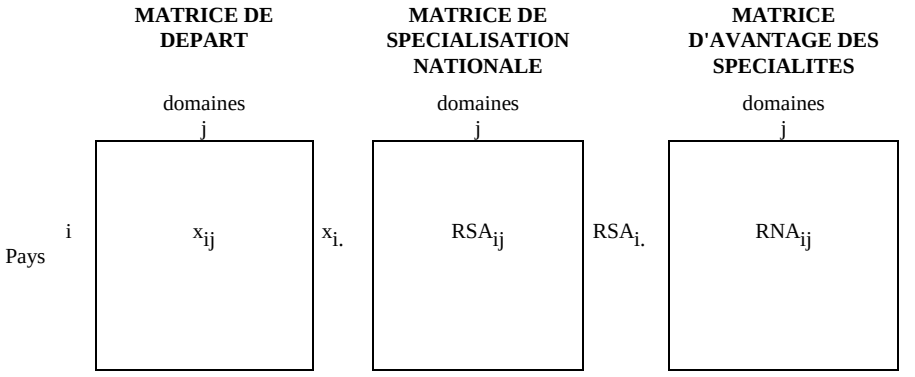
Il parait légitime que le premier exemple, choisi pour exposer ce genre de technique, soit issu des travaux réalisés par l'institut français d'évaluation de la recherche et des techniques, l'OST (BARRÉ, 1991). Le propos est de montrer comment dégager les caractéristiques des activités de recherche de onze pays. Lors d'une analyse des forces et des faiblesses des pays pour un ensemble de spécialités, les données doivent être normalisées. Pour répondre à cette exigence Barré a élaboré des indices qu'il a nommés *Revealed Scientific Advantages (RSA)* et *Revealed National Advantages (RNA),* en référence à la dénomination "*Revealed Technology Advantage*" employée par Patel et Pavitt dans l'analyse des processus de l'innovation. Le premier indice est comparable à l'indice d'activité utilisé par Price (voir p. 60). Dans le cas présent, il a servi à normaliser la matrice de départ qui récapitule l'ensemble des contributions scientifiques de chaque pays à chaque spécialité. Barré a appelé ce second tableau "*tableau de spécialisation nationale*" (tableau 12), puisque l'indice d'avantage (nommé RSA) représente le ratio de la performance du pays *i* dans le domaine *j* sur la performance de ce même pays dans tous domaines confondus. Ainsi, le vecteur RSA_i caractérise pour un pays *i* les points forts et les points faibles parmi ses spécialités.

Par la suite, l'auteur a voulu étudier les domaines pour les comparer à travers leurs développements pour ces pays. Il a alors calculé un nouvel indice, l'indice des *Revealed National Advantages* dont il s'est servi pour calculer les vecteurs RNA_j constituant le tableau "d'avantage des spécialités" (voir tableau 12).

Cette technique de pondération lui a permis d'étudier les activités scientifiques de 11 pays (France, RFA, Royaume-Uni, USA, Canada, Japon, Pays Bas, Suède, Italie, Inde, Australie) en attribuant à leurs publications une spécialité parmi 107. Les données ont été télédéchargées de *Pascal*. Les 107 spécialités non pas pu être complètement définies à partir du plan de classement documentaire de *Pascal*. Elles ont été attribuées à chaque référence par un groupe d'experts.

La détermination des profils de similarités entre les 107 spécialités selon leurs répartitions dans les pays a été analysée à l'aide d'une classification ascendante hiérarchique. Les experts, après interprétation des 9 regroupements de spécialités fournis par la classification automatique, ont dégagé en final 13 domaines majeurs.

Tabl. 12 - Récapitulation des tableaux et calculs employés dans l'analyse pays *x* domaines



x_{ij} = nombre de publications publiées par le pays *i* dans le domaine *j*

Notations : $i \in (1, ..., n)$ $j \in (1, ..., m)$

$$x_{i.} = \sum_{j} x_{ij} \qquad x_{.j} = \sum_{i} x_{ij} \qquad x_{..} = \sum_{i,j} x_{ij}$$

$$RSA_{ij} = \frac{x_{ij} / x_{.j}}{x_{i.} / x_{..}}$$ le vecteur $RSA_i = (RSA_{ij})_{j = 1,m}$ le vecteur $RNA_j = \left(\frac{RSA_{ij}}{RSA_{i.}} \right)_{i=1,n}$

Un nouveau tableau a été construit en réduisant à ces 13 domaines majeurs les facteurs décrivant les contributions scientifiques des pays. Après une nouvelle normalisation par les indices RSA et RNA, les deux tableaux associés sont de nouveau traités par classification automatique. Le tableau RSA est classifié pour regrouper les pays ayant le même comportement d'activité dans ces domaines. Pour le tableau RNA la classification regroupe les domaines étudiés de manière similaire par les 11 pays.

Barré, en proposant cette méthode, a recherché une qualité de résultat. Le fait de s'imposer *Pascal* comme source d'information, l'a obligé à redéfinir les 13 domaines qui vont permettre de caractériser de façon globale et suffisamment juste les contributions des pays à la science. Contrairement à la banque de données *SCI* (voir p 62), *Pascal* ne comporte pas ce genre de renseignements synthétiques mais elle a été choisie dans cette étude car elle a une meilleure couverture thématique et géographique de la science.

Habituellement, ces évaluations internationales sont effectuées uniquement à partir des données fournies par *SCI* (répartitions des publications des pays en une dizaine de domaines selon les études). Les auteurs appliquent ensuite la pondération de la mesure des relations par la normalisation de l'indice d'avantage. Puis ce tableau normalisé est injecté dans un logiciel d'analyse des données pour construire des cartes qui synthétisent les dépendances entre les pays et les domaines d'activités.

• *Analyse de la collaboration internationale (relations pays x domaines)*

La méthode applique la même série d'opérations que pour l'analyse de l'avantage national (voir p 86), excepté que le tableau de données initial ne contient plus, pour chaque pays, toutes les publications nationales réparties en une dizaine de domaines, mais uniquement les publications issues de collaborations internationales (tableau 13). Un tel tableau de relations permet de connaître les domaines scientifiques privilégiés par des collaborations internationales et ce pour chaque pays.

L'emploi d'une méthode d'analyse des données sur le tableau normalisé par l'indice d'avantage rapprochera les pays qui ont les mêmes profils de domaines privilégiés, donc ceux qui ont des politiques de collaborations internationales similaires. Comme la construction de ce tableau nécessite une source qui dispose à la fois d'une couverture multidisciplinaire et l'information sur les multi-affiliations des auteurs qui ont rédigé une publication, la base de prédilection est le *SCI*.

Tabl. 13 - Relations dans l'analyse des profils de collaboration internationale

	Domaines j
Pays i	x_{ij}

x_{ij} = nombre de publications co-signées avec un organisme étranger pour le pays *i* dans le domaine *j*

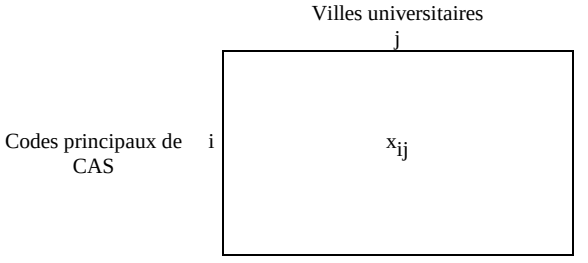
Cette méthode a été mise en pratique par plusieurs auteurs. Le LEPI (CNRS) a pu étudier la typologie des collaborations internationales pour toutes les publications du *SCI* entre 1981 et 1986, c'est-à-dire l'analyse d'un tableau comportant 98 pays décrits par 8 domaines. Ce tableau a été ensuite exploité et analysé par deux méthodes d'analyses des données, l'Analyse Factorielle de Correspondances (AFC) pour une présentation sous forme de carte des pays et des domaines et la méthode de l'Arbre à Longueur Minimale *(Minimum Spanning Tree, MST*) pour une disposition de pays sous forme arborescente (OKUBO et alii, 1992 ; MIQUEL et OKUBO, 1994).

Nagpul et Lalita Sharma ont aussi employé la même démarche mais sur un tableau plus restreint ne présentant que les relations concernant les publications internationales en physique. Ce tableau décompose l'activité de collaborations internationales de 36 pays en 10 domaines scientifiques. Ils ont ensuite cartographié le tableau normalisé par la méthode de l'AFC (NAGPUL et LALITA SHARMA, 1994).

• *Analyse de l'activité scientifique des Universités (relations villes x codes documentaires)*

Cet exemple d'analyse de tableaux de contingence, tente toujours de répondre à la question "qui fait quoi ?". Cette fois-ci, l'évaluation ne se situe plus à l'échelle internationale, mais nationale. Ce travail, exposé par le CRRM, présente comment classer les grands pôles universitaires français en chimie suivant leurs profils d'activités en recherche (DOU et alii, 1990d). Le tableau de relations exploitées a été construit grâce aux logiciels bibliométriques développés au CRRM à partir du télédéchargement de références bibliographiques de la banque de données du *Chemical Abstracts*. Ces références correspondent à toutes les publications scientifiques de 17 villes universitaires françaises répertoriées par le *Chemical Abstract Services* pour l'année 1985, soit près de 6500 références. Pour classer ces 17 villes selon leurs comportements de recherche en chimie, les auteurs ont choisi de décrire l'activité de ces villes en fonction du code principal du plan de classement de CAS affecté à chacune des références. Le tableau de relations analysé comporte le profil d'activité des 17 villes (en colonne) décrit par les 78 codes présents dans les références (tableau 14).

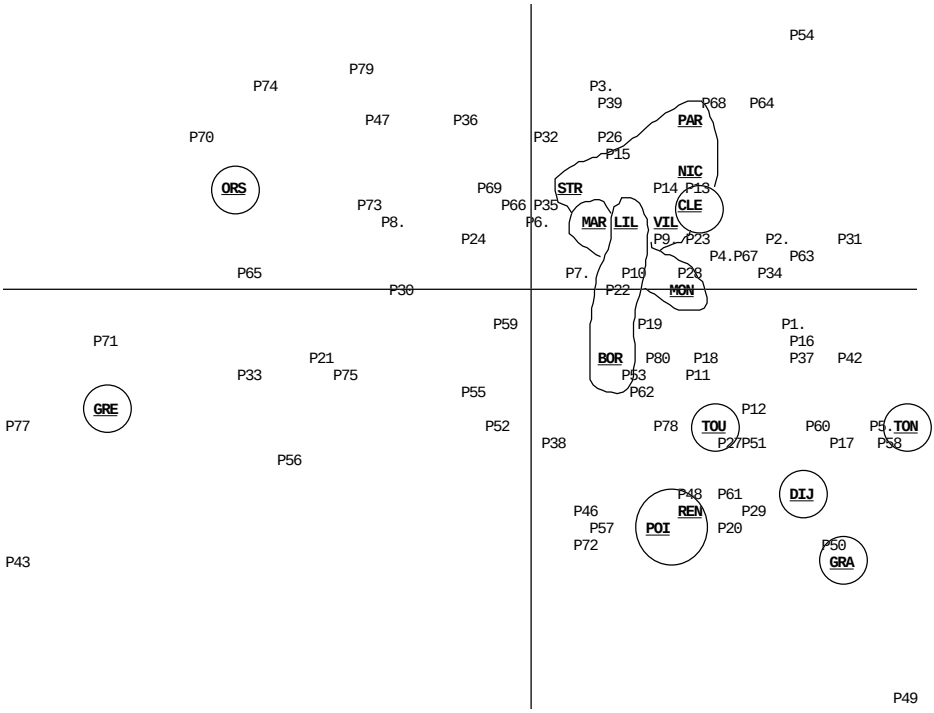
Tabl. 14 - Relations dans l'analyse de l'activité scientifique des Universités



x_{ij} = nombre de publications de la ville *j* comportant le code principal *i* de CAS

Ce tableau est soumis à plusieurs méthodes d'analyse statistiques des données, Classification Automatique Hiérarchique (CAH), Analyse Factorielle Discriminante (AFD) et Analyse Factorielle des Correspondances (AFC). Après une évaluation des résultats de chacune et de l'enchaînement de deux méthodes successives, les auteurs ont estimé que la combinaison AFC des données brutes suivie d'une CAH sur les coordonnées résultantes de l'AFC était la plus pertinente. Cette procédure d'analyse statistique offre beaucoup d'avantages. D'une part, la CAH finale donne un classement des villes selon leurs profils d'activité normalisés par la distance du Khi^2 (calculée par l'AFC). D'autre part, la lecture des cartes factorielles permet d'identifier les domaines d'activités (les codes) qui contribent à discriminer chacune des classes de la CAH (figure 9).

Fig. 9 - Plan 1-2 de l'analyse factorielle des correspondances du tableau 14



Les points commençant par un P symbolisent les codes principaux de CAS. Les points soulignés symbolisent les villes. Les classes obtenues par la CAH ont été superposées symboliquement sur cette carte en entourant les villes appartenant aux mêmes classes

• *Analyse pays x mots-clés*

Toujours pour répondre à la question "qui fait quoi ?" dans le domaine de l'état du système cardio-vasculaire de l'homme soumis à l'effort, Billard et alii (CEDOCAR-IRIT) ont étudié les corrélations entre les pays et les mots-clés pour un ensemble de références sélectionnées sur la banque Medline (BILLARD et alii, 1990). La matrice croisant ces deux unités bibliographiques a été analysée par la technique de l'Analyse Factorielle des Correspondances suivie d'une classification automatique.

CONCLUSION

Cet aperçu des techniques bibliométriques mettant en jeu la notion d'évaluation des relations intra ou inter éléments bibliographiques est loin d'être complet. Il donne tout de même une bonne idée de la richesse et de la qualité des résultats.

La richesse provient de l'union des multiples combinaisons d'analyses des relations entre entités bibliographiques avec les multiples techniques statistiques employées pour mettre en relief les dépendances entre ces entités. A chacune de

ces unions, une nouvelle vision des informations intrinsèques aux références bibliographiques se forge, offrant des renseignements globaux (indicateurs de tendance) mais aussi de nouveaux systèmes de relations entre les références d'où proviennent les données. Ainsi, les références ne sont plus des entités éparses et déconnectées, ces méthodes relationnelles les métamorphosant en parties élémentaires fortement connectées qui vont former un tout cohérent uniquement sur la base de règles de relations intrinsèques. Les règles qui serviront à la construction de cette cohérence ne sont pas des règles imposées par des a priori. Elles n'existent pas avant même que l'analyse des données soit appliquée. Elles se définiront d'elles mêmes selon le recensement de données intrinsèques aux références. C'est pour cette raison que les statisticiens classent les techniques d'analyse des données utilisées en bibliométrie dans la statistique descriptive et non dans la statistique établie sur des modèles.

Ce sont ces extraordinaires méthodes d'analyses de données qui procurent une telle qualité dans les résultats de la bibliométrie "relationnelle". Non seulement ces méthodes établies depuis de longues dates ont été soumises à de considérables validations mathématiques et statistiques, mais surtout les résultats qu'elles livrent sont d’une très grande accessibilité. Ce genre de technique ne contraint pas son utilisateur à être un statisticien de formation. Il est bien évident que pour être parfaitement maître des données fournies pas les analyses de données, le bibliomètre devra quand même connaître un minimum des fondements de la technique qu'il met en oeuvre. Mais le fait que la plupart des résultats soient livrés sous la forme de représentations graphiques est un avantage incontestable. L'oeil apprécie ces représentations de l'information. Non pas par jouissance intellectuelle, mais tout au contraire parce que le cerveau humain depuis l'origine des temps fait fonctionner les capacités visuelles alors que la conceptualisation purement intellectuelle est beaucoup plus récente, tout au moins pour le niveau de complexité que nécessiterait la synthèse de l'activité de recherche de milliers d'individus ! C'est pour cela que les résultats fournis par la bibliométrie sont à prendre avec prudence et réflexion. L’interprétation de ces résultats nécessite de solides connaissances dans la perception de l'utilisation des outils bibliométriques et d'incontestables notions de la pratique des analyses de données statistiques.

Il est utile de remarquer que ces outils statistiques sont employés bien souvent pour l'étude des relations entre des acteurs qui représentent des entités à l'échelle internationale ou nationale. Ces études mettent en évidence les grandes tendances de relations ou de similarités de relations entre ces acteurs en se désintéressant totalement des références et des documents qui ont servi de base de traitement. Ceci est vrai pour toutes les études répondant à des besoins d'évaluation comme aide à la gestion des programmes nationaux ou internationaux de la recherche. Par contre, ce désintérêt est moins vrai pour les études de l'histoire et de la sociologie de la science où l'oeuvre de chaque "individu-clé" impliqué dans l'évolution de la science reste constamment la base de l'organisation analysée. Mais les techniques bibliométriques développées

pour décrire cette structure n’ont pas été réfléchies pour garder un lien avec les références de départ. Pratiquement toutes ces méthodes sont fondées sur un tableau de relations ne conservant pas l'information sur la provenance des éléments bibliographiques corrélés. La dimension du document d'origine est donc perdue.

Jusqu'à présent, la notion de document avait été occultée car elle ne semblait pas indispensable, mais ce n'est plus vrai dès lors que la bibliométrie a voulu s'immiscer dans le monde de l'industrie. Les besoins en information des industriels ne répondent plus du tout aux mêmes préceptes. Leurs attentes peuvent être internationales lorsqu'ils ont un marché international mais certainement pas uniquement selon une approche globalisée. L'étude et la surveillance de la concurrence ne correspondent pas à une simple évaluation générale des tendances. Les industriels ont surtout besoin d’informations très précises dans les domaines d'innovation qui les concernent. La masse des documents se référant à ces domaines n'est pas forcément le principal facteur excluant une simple analyse par lecture. C'est plutôt la multiplicité des types de données présentes dans les documents brevets et donc la multiplicité des types de lecture qui doit en être faite. Cette complexité des relations entre les différentes entités de l'information brevet est l'une des principales raisons de l'intérêt suscité par les techniques bibliométriques. Par contre, le retour rapide à l'ensemble de documents à l'origine d'une structuration ou d'une organisation dégagée en fin d'étude bibliométrique est indispensable pour la finesse des analyses qu'impose la stratégie brevet en industrie.

CHAPITRE V
LES BREVETS

La maîtrise de l'information technique par des méthodes bibliométriques n'est apparue que très récemment. La bibliométrie a été au départ conçue pour subvenir à des besoins purement documentaires. Dans un second temps, les sociologues ont vu en ces données quantitatives un moyen pour mieux comprendre les phénomènes de la connaissance scientifique. Ce n'est que tout récemment que les instances dirigeantes ont ressenti le besoin d'appuyer leurs décisions sur des données quantitatives par souci d'une plus grande objectivité. Il est vraisemblable que pour les politiques, la recherche n'a d'intérêt que si elle se concrétise par des retombées industrielles. Les informations sur l'évaluation de la recherche ne peuvent les intéresser que si elles leur permettent d'évaluer les impacts sur le monde de l'industrie. C'est donc sous l'impulsion de stratégies politiques que sont nées les premières applications des techniques bibliométriques au brevet, celle-ci étant la seule parfaitement centralisée et concernée par la mise en application industrielle de la recherche.

Ce n'est qu'au début des années 1980 que les premières études sur l'information technique sont apparues. Cette éclosion peut vraisemblablement être attribuée aux travaux de Narin qui avait été chargé par l'institut américain *National Science Foundation* de créer une banque de données répertoriant tous les brevets déposés aux États Unis et permettant des exploitations statistiques. Cette banque de données, comme le *SCI*, a été construite spécialement pour des évaluations bibliométriques de l'innovation et de la propriété industrielle sur le sol américain. Elle contient les données de 59 pays protégeant des inventions par la procédure américaine. Dans l'optique de l'élaboration d'indicateurs macro-

bibliométriques, similaires à ceux utilisés pour l'évaluation de l'activité scientifique internationale (voir p 61 et p 86), ces brevets sont décrits par deux plans de classement. Le premier est constitué par 376 classes de brevets de l'office américain des brevets, le second par 57 classes de produits nommées *US Standard Industrial Classes* (SIC), découpées en environ 100 sous-classes par classe soit plus de 700 000 sous-classes au total (la classification SIC est réalisée par l'*Office of Technology Assessment of the US Patent Office*). Cette base brevet a surtout comme caractéristique intéressante la saisie des informations concernant les "citations" de brevets ou d'articles scientifiques antérieurs. Les chercheurs du *Computer Horizon Incorporated* (CHI) ont développé de nombreux indicateurs bibliométriques à partir de leurs banques de données qu'ils publient depuis 1972 dans un ouvrage intitulé *Science Indicator Reports*. A l'instar de la NSF, des instituts nationaux d'évaluation de la recherche (voir p 16) se sont eux aussi intéressés à l'information brevet. Ne disposant pas des mêmes moyens financiers que la NSF, ils n'ont pas constitué leurs propres fonds. Bien que des banques de données "brevets" mises à la disposition de tous sur des serveurs commerciaux soient d'une couverture internationale bien supérieure à celle du CHI (brevets uniquement déposés aux États Unis), la plupart de ces instituts nationaux d'évaluation font sous-traiter leurs études ou commandent leurs données au CHI.

A partir de ces exemples d'exploitation statistique des brevets, des entreprises se sont rapidement emparées du principe. Parce qu'elles ont de nouvelles exigences de surveillance de leur environnement, elles commencent à employer des techniques bibliométriques pour élaborer leurs propres indicateurs. Bien que ce nouvel axe soit devenu une spécialité pour certains centres de recherche en bibliométrie, le peu de travaux concernant les applications industrielles s'explique, tout d'abord, par le fait que nous n'en sommes qu'à la naissance de cette pratique mais surtout par l'exigence de confidentialité imposée par les entreprises. Mais pour qui travaille dans ce domaine d'application, l'intégration de la bibliométrie dans le monde de l'entreprise ne fait plus aucun doute. Cette intégration suit presque toujours la prise de conscience de la nécessité de mener une activité de veille technologique pour la pérennité de l’entreprise. Et elle s'accompagne presque toujours de l'implantation d'une cellule de spécialistes en veille technologique pour mettre en application les outils bibliométriques.

Comme pour les techniques appliquées à l'information scientifique, les techniques présentées dans ce chapitre feront référence à la fois à des domaines se référant davantage à la politique de l'évaluation des techniques par des instituts nationaux, et à des domaines moins globalisants qui répondent en fait à des problématiques industrielles.

LA REMISE EN CAUSE DES POSTULATS BIBLIOMETRIQUES

Il est indispensable, s'agissant du document brevet, de réévaluer les hypothèses de travail admises dans le cas des documents scientifiques. Le premier postulat admet que la publication scientifique est une représentation objective de l'activité de recherche de son auteur puisqu'elle est le passage obligé de la reconnaissance d'un chercheur dans la collectivité scientifique. Le second postulat suppose qu'il existe des liens intellectuels entre les publications qui définissent des structures consensuelles. En est-il de même pour l'information brevet ?

Pour essayer de répondre, nous allons tout d'abord rappeler brièvement quelle est la nature de cette information et si elle peut satisfaire aux deux postulats. Puis nous apprécierons les avantages que cette information brevet procure aux praticiens de la bibliométrie.

L'information brevet

L'acte de dépôt du brevet est soumis à un principe de réciprocité. L'entreprise ou l'organisme déposant prend le risque de porter à la connaissance de tous un savoir nouveau et inventif qui, jusque là, était gardé secret. En contrepartie, il dispose d'une protection juridique qui lui donne un droit d'exclusivité pour l'exploitation de cette invention (possibilité d'attaquer en justice le contrefacteur). Le fait de divulguer un savoir-faire lors d'un dépôt peut apparaître comme un obstacle à l'enthousiasme industriel pour cette procédure. On peut penser que les entreprises veulent conserver secrètes leurs compétences. En fait, les avantages d'une telle protection sont trop nombreux et trop importants pour que l'entreprise ne l'instaure pas dans sa stratégie (SOMMIER, 1992). Le dépôt de brevets est la seule pratique garantissant la sauvegarde du patrimoine technologique de l'entreprise.

Le contenu du document brevet présente un caractère de "garantie" sur le plan de l'intérêt qu'il recouvre. Parmi les conditions à remplir pour qu'un brevet soit accordé, deux d'entre elles exigent qu'il y ait dans la demande un aspect de nouveauté et d'activité inventive. Donc, à la différence des articles scientifiques une certaine "qualité d'innovation" est exigée dans le document brevet. Il faut être prudent sur ce critère de "qualité" car toutes les procédures de dépôt n'ont pas la même fermeté concernant la présence de ces deux conditions. Ainsi, il est bien connu que l'examen de la demande par un office français est bien moins strict que celui d'un office américain. Mais après la publication officielle de la demande de brevet un délai est laissé libre à toutes oppositions. Ainsi, les concurrents peuvent venir demander certaines révisions ou même le rejet du brevet. Donc, de même qu'il existe le terme de "libre concurrence" en économie on pourrait définir un "libre droit" pour le brevet. En laissant jouer cette liberté les caractères innovants sont valides dès lors que le brevet fait partie d'une stratégie industrielle et qu'il n'est pas remis en cause. Cette notion d'opposition est aussi à prendre en compte lors d'une analyse bibliométrique puisque

certaines banques de données introduisent les références brevets au moment des publications des demandes sans indiquer si les brevets ont été accordés ou non en fin de procédure !

Un brevet, après son dépôt national (ou régional), peut être étendu à d'autres pays. Cette information est bien souvent très utile à connaître car elle donne une idée de l'intérêt que l'entreprise porte à son invention et des perspectives internationales qu'elle veut lui offrir. L'étude de ces extensions, pour les portefeuilles de brevets des entreprises concurrentes, est une grande source d'information concernant les stratégies que ces entreprises vont mener pour leurs futurs produits. Là encore le brevet semble fournir des informations de grande qualité.

La liste des revendications, c'est-à-dire des points sur lesquels l'originalité du brevet repose et sur lesquels le déposant veut particulièrement être protégé, procure une information très technique et très précise. L'étude de son contenu est d'un apport considérable pour l'homme de l'art.

Les documents brevets entretiennent des liens entre eux par une pratique qui pourrait être assimilée à la pratique de la citation entre les articles scientifiques. Les déposants font référence aux brevets antérieurs plus ou moins proches de leur demande, de façon à préciser en quoi leur brevet en diffère. Ils font aussi référence aux éventuels articles scientifiques justifiant l'invention et sa priorité. Sur la base de ces références et d’une recherche d'antériorité, l'examinateur donne ses propres références de brevets ou publications s'approchant de l'invention soumise. Donc, ces citations portent moins sur des points conceptuels que sur des notions techniques. Mais plus important, elles ne sont pas régies par des phénomènes sociologiques comme l'auto-citation, les collèges invisibles, l'effet S^{t} Mathieu... Ces citations sont probablement plus légitimes que celles données par les chercheurs. De plus, on sait qu'un brevet très cité a souvent un rôle stratégique particulier, plus important en tout cas qu'un brevet isolé. Les autres brevets déposés à sa suite peuvent avoir pour objet de former une "ceinture technique" couvrant les différents domaines d'application du brevet.

A la suite de cette présentation, il est acceptable d’estimer que, bien que ce ne soit pas un point de passage obligé pour les entreprises (ou organismes), le document brevet peut répondre aux deux postulats. Il décrit pleinement l'activité de recherche et développement du déposant sur l'invention mais apporte en plus des informations sur les stratégies industrielles visées. Les documents brevets entretiennent entre eux des liens dont la nature est moins consensuelle mais tout aussi valable.

Les avantages du document brevet pour les traitements bibliométriques

Le document brevet a, par sa nature, une place unique parmi les supports écrits de l'information. Ce document faisant partie de procédures légalisées au

niveau national et international (tous les pays fortement industrialisés ont signé une convention commune), sa gestion lui confère des qualités que les documents scientifiques ne peuvent posséder.

Une production centralisée

Toutes les demandes de dépôt passent par des organismes officiels. Ces offices, producteurs de bases de données pour leur propre compte, ont mis ces données à disposition du public. L'exhaustivité y est donc parfaite par zone géographique ou type de procédure de dépôt (offices nationaux, office européen, office mondial). Pour obtenir des données recouvrant ces différentes zones ou procédures, des producteurs privés collectent et saisissent aussi de leur côté les données de dépôts de brevets, par exemple le producteur *Derwent*.

Une présentation formalisée

Une demande de dépôt doit satisfaire à des obligations, nationales et internationales, de mise en forme. Ce qui, repris sous forme informatisée, offre à l'utilisateur des données très bien structurées. Le traitement bibliométrique automatisé n'en sera que plus facile.

Un classement des documents brevets utilisé internationalement

Pour caractériser le contenu du brevet, les offices de dépôt affectent à chaque document des codes appartenant à la Classification Internationale des Brevets, CIB (*International Patent Classification, IPC*). Cette indexation est bien évidemment faite pour aider les offices à constituer des rapports de recherche d'antériorité. Le fait que cette indexation soit identique pour tous les brevets est un avantage important par rapport aux données scientifiques qui n'ont qu'une indexation dépendante du producteur de la banque de données. Au cours d'une étude bibliométrique, cette classification facilite la recherche et élimine les problèmes d'homogénéisation des données provenant de banques différentes (non automatisable pour l'indexation).

Le brevet a donc une homogénéité et une qualité de couverture que ne peut revendiquer aucun autre type de document. C'est pourquoi il constitue un fonds documentaire qui se prête bien à des opérations bibliométriques.

LES INDICATEURS UNIVARIES DE BREVETS

Comme pour les études bibliométriques relatives à la science, le simple dénombrement statistique est la mesure élémentaire de tout indicateur bibliométrique brevet. La plupart des études bibliométriques en information technique sont malheureusement restées à ce stade de la technique bibliométrique. Nous ne voulons pas dire par là que ces données n'ont aucune valeur, mais il est bon de garder à l'esprit qu'il ne faut pas les considérer comme des mesures absolues. Il est indispensable, pour donner un sens à ces données, de connaître leur situation dans le temps et dans leur domaine.

L'unique emploi du comptage de brevets ou de citations permet déjà d'acquérir une grande diversité d'informations. Une partie de ces dernières est récapitulée dans le tableau 15.

Tabl. 15 - Récapitulation des éléments bibliographiques utilisés en analyse de brevets

Unité(s) bibliographique(s) traitée(s)	**Pour le secteur technique étudié, information sur**
Date de priorité	évolution temporelle globale des dépôts
Pays de priorité	tendances des dépôts des pays
Pays / date de priorité	évolution des tendances des dépôts des pays dans le temps
Organisme déposant	principaux organismes concernés
Organisme / date de priorité	répartition des efforts des organismes dans le temps
Code documentaire	analyse "grossière" des domaines concernés
Code / date de priorité	évolution dans le temps des domaines concernés
Code / organisme déposant	organismes travaillant dans les mêmes domaines
Pays d'extension	marchés internationaux
Pays d'extension / priorité	nombre de brevets d'une famille (inventions stratégiques)
Pays d'extension / organisme	stratégie de dépôts pour chaque organisme
Citation / priorité	impact d'une invention
Citation / société	pionniers (les travaux sont souvent repris)

Les commandes en ligne de tri statistique

Les études menées par l'Institut Français du Pétrole (IFP) sont de très bons exemples de mise en pratique des comptages statistiques simples (MOUREAU et GIRARD, 1987a, 1987b ; MOUREAU et alii, 1992). Ces travaux expliquent comment la simple utilisation des commandes de tri statistique que mettent à leur disposition les serveurs commerciaux (Memt sur Questel, Get sur Orbit, Zoom sur ESA-IRS, Rank sur Dialog), permet à ces chercheurs de disposer de certains indicateurs. Ces commandes permettent de donner, par ordre décroissant de fréquence d'apparition, les éléments présents dans un champ pour un ensemble de références sélectionnées auparavant. Ce résultat représente tout simplement le début de la courbe de distribution de la loi de Zipf (voir p 42). La précision "le début" signifie que seuls les éléments les plus fréquents sont généralement employés dans ce genre de traitement.

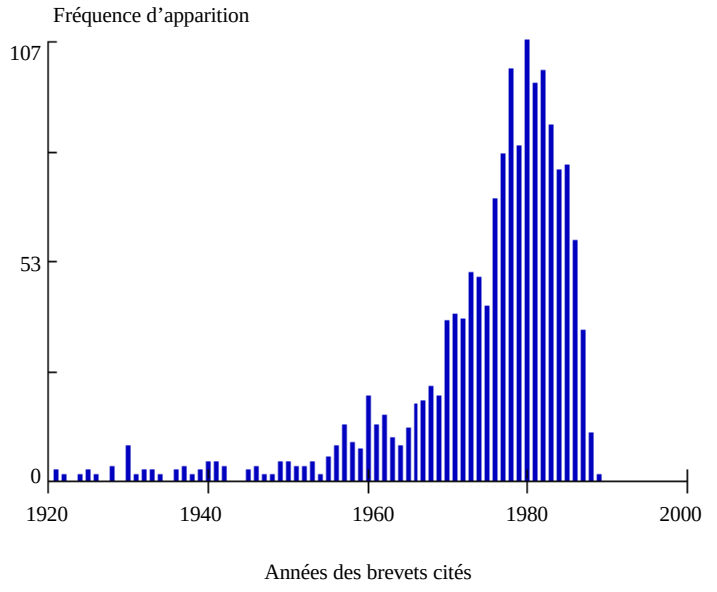
D'autres auteurs citent aussi leurs expériences dans la pratique de ces commandes accessibles en ligne sur les serveurs. Le dernier ouvrage de Jakobiak montre son utilisation comme outil d'aide à la surveillance de la concurrence en veille technologique (JAKOBIAK, 1994).

Certains auteurs ont aussi eu l'idée d'exploiter les résultats de ces commandes par des post-traitements informatiques. En France, le CRRM a développé plusieurs programmes informatiques qui réutilisent ces tris statistiques pour les mettre en valeur (DOU et HASSANALY, 1988 ; DOU et alii, 1989a) ou même pour en dégager des informations difficilement détectables par les langages d'interrogation des serveurs (DOU et alii, 1990b). Le dernier travail cité explique, par exemple, comment connaître toutes les entreprises qui possèdent trois domaines de compétences. Ce cas ne peut être résolu que partiellement par la logique booléenne des langages d'interrogation. Ainsi, l'utilisation de l'opérateur booléen "ET" avec les trois domaines de compétence ne retrouvera que les brevets où ces trois compétences sont signalées. Or il peut exister des sociétés qui protégent ces trois domaines mais en déposant chaque fois des demandes différentes. Les auteurs présentent le traitement informatique qui permet de connaître ces sociétés "récalcitrantes" à partir de résultats de tris statistiques.

Nous pouvons citer un deuxième exemple d'outil informatique développé par le CRRM exploitant de simples comptages. Ce logiciel nommé DATAGE a été spécialement développé pour traiter le champ citation des banques de données brevets. Il permet de comptabiliser l'âge des brevets cités dans un ensemble de référence (figure 10). La distribution de ces âges dessine un profil d'âge de la technique du domaine étudié (DOU et alii, 1990c).

Il est important de noter que, contrairement au télédéchargement des références brevets, les commandes offertes par les serveurs ne sont pas chères. Par contre, elles ne permettent absolument pas de maîtriser les erreurs persistant dans les banques de données brevets. Ces erreurs ne doivent pas être sous-

Fig. 10 - Profil de l'âge de la technologie pour un domaine précis



estimées lors d'une analyse statistique car leurs taux peuvent, selon les cas, être élevés. Au cours d'une évaluation du taux d'erreurs pour le champ "société déposante" de la banque WPI, Nivol a évoqué le cas d'une société dont le nom a été rédigé de 22 façons différentes et à qui correspondaient 7 acronymes Derwent, pour seulement 62 familles de brevets déposés (NIVOL, 1993). Ce cas extrême démontre que l'utilisation de l'information Derwent sans aucune correction provoquerait pour cette société un pourcentage d'erreur avoisinant 35%.

Les traitements croisés sur deux éléments brevets

Les tris simples fournissent des résultats très restreints. Le croisement entre deux entités de références brevets vient naturellement à l'esprit. Par exemple, pour comparer la politique de dépôt des sociétés au cours du temps, il faut pouvoir construire le tableau détaillant le nombre de brevets déposés par année pour chaque société. Ce nombre correspond tout simplement à la fréquence des co-occurrences nom de société et année de dépôt dans les références. Nous allons évoquer trois solutions pour calculer ces co-occurrences d'éléments de brevets.

La première solution ne nécessite, là encore, qu'une bonne connaissance du langage d'interrogation du serveur distribuant la banque de données brevets. La technique se déroule en trois phases. La première est la sélection de l'ensemble de références brevets à analyser. La seconde est l'activation de la commande de mise en mémoire des entités présentes dans l'ensemble des références et appartenant à la première catégorie d'information que l'on veut croiser dans le tableau (par exemple tous les noms de sociétés présents dans les champs des références). Il faut réaliser la même opération pour la deuxième catégorie d'information à croiser, par exemple toutes les dates de dépôts de brevets sélectionnés. Une fois ces deux listes constituées, il faut réutiliser l'un après l'autre tous les éléments conservés en mémoire en les combinant entre eux par couple avec l'opérateur booléen "ET" pour ré-interroger l'ensemble de références brevets sélectionné au départ. Le premier couple combinera le premier nom de société avec la première date de dépôt, puis le second couple, le premier nom de société avec la seconde date, ainsi de suite pour toutes les dates. Lorsque toutes les dates ont été utilisées, l'ensemble des réponses renvoyées par le serveur représente l'évolution des dépôts de la première société au cours du temps dans le domaine étudié. Il faut réaliser la même opération pour le second nom de société et ainsi de suite pour toutes les sociétés mémorisées. Cette technique est très peu coûteuse car elle n'impose pas le télédéchargement des références, mais elle est très longue à mettre en oeuvre puisque la construction d'un tableau qui croise vingt sociétés sur dix ans équivaut déjà à deux cents combinaisons, et donc autant d'interrogations en ligne. C'est pour cette raison qu'a été développé un programme effectuant la combinatoire et rédigeant les requêtes correspondantes déchargeant la personne qui interroge de cette lourde

tâche (DOU et alii, 1991). Cette collecte de valeurs d'un tableau de co-occurrences en ligne a deux avantages principaux, le coût relativement négligeable et la possibilité de connaître des valeurs de co-occurrence sans restreindre le nombre de références concernées (la banque complète peut être choisie comme étant l'ensemble de références sélectionné). Par contre, ce traitement ne peut s'appliquer à toutes les informations contenues dans une référence brevet, car toutes ne peuvent être chargées en mémoire.

Une seconde solution est l'achat du logiciel commercialisé PatStat+. Ce logiciel a été développé par *Derwent* pour permettre des traitements statistiques sur les données provenant de ses banques de données, c'est-à-dire principalement la base *WPI*. Le logiciel travaille à partir des références *WPI* télédéchargées. Outre les tris simples, ce logiciel construit quelques tableaux-types de croisement d'éléments de brevets. L'équipe de Moureau et Girard de l'IFP et Bernat (Elf Aquitaine) utilisent cet outil pour toutes les analyses de tris croisés.

Pour les deux précédentes solutions, deux problèmes restent posés : la limite des traitements à un nombre restreints d'informations brevets (quelques champs de notices brevets sont exploitables) et les erreurs introduites dans les banques (voir p. 99). Conscient de ces limitations, le CRRM a mis au point un logiciel bibliométrique totalement ouvert à la variété des données bibliographiques ainsi qu'à la diversité des traitements bibliométriques. Ce logiciel nommé DATAVIEW traite les notices télédéchargées provenant des banques de données. Il accepte tous les formats de notices quelles que soient leur origine, banques de données et serveurs. Une manipulation aisée de l'ensemble des données présentes dans ces notices permet à l'utilisateur de construire des résultats bibliométriques personnalisés (distributions bibliométriques, tris statistiques simples ou croisés et tableaux). La mise en oeuvre de cet outil sur l'information brevet pour l'élaboration d'indicateurs concurrentiels a été montrée dans plusieurs publications (NIVOL, 1993 ; DUMAS, 1994 ; ROSTAING, 1993 ; ROSTAING et alii, 1993b).

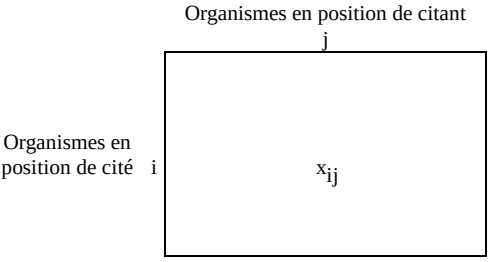
Les indicateurs de Battelle

La société américaine Battelle, consultant en stratégie pour la gestion de la technique, étaye ses dossiers par quelques indicateurs bibliométriques (ASHTON et alii, 1983 ; STACEY, 1992). Elle utilise particulièrement trois indicateurs : *activity*, *immediacy*, *dominance*.

- L'indicateur d'activité (*activity*) correspond au simple comptage de brevets par période, société, inventeur...
- L'indicateur d'immédiateté (*immediacy*) mesure l'âge de la technique en mesurant l'écart de temps entre les brevets citants et les brevets cités. Si les brevets citent le plus souvent des brevets récents, l'évolution de la technique du domaine est très rapide (technique identique à celle de l'âge de la technique, voir p 99).

- L'indicateur de dominance (*dominance*) mesure la pratique de la citation entre les principaux organismes déposants dans le domaine étudié. Cette évaluation passe par la construction d'un tableau croisant les organismes d'un côté en tant que citants et de l'autre en tant que cités (tableau 16). C'est une matrice analogue à celle des citations-croisées des études bibliométriques des revues.

Tabl. 16 - Indicateur de dominance de Battelle



x_{ij} = nombre de brevets de l'organisme *i* cité par l'organisme *j* citant

Ces études d'indicateurs permettent, entre autres, de classer les organismes déposants selon cinq catégories types d'activité brevet. Ces cinq catégories se définissent selon le tableau 17.

Tabl. 17 - Classement des sociétés selon leur activité brevet

Type de firme	Nombre de brevets	Citations reçues	Auto-citations	Citations données
Pionnier agressif	élevé	Elevé	élevé	faible
Leader indépendant	élevé	Faible	élevé	faible
Suiveur agressif	élevé	Faible	moyen	élevé
Pionnier non agressif	faible	Elevé	faible	faible
Médiocre	faible	Faible	faible	faible

Les indicateurs du CHI

Le CHI a développé un panel considérable d'indicateurs bibliométriques pour évaluer la science ou la technique. Sa recherche dans le domaine des indicateurs bibliométriques se détache de celle des autres par le fait qu'elle s'est très vite axée sur l'étude des techniques. Le portefeuille des indicateurs du CHI que Narin a relevés en 1989 lors du colloque des *Systèmes d'informations élaborées* de la *SFBA*[1] est inventorié dans l'encadré ci-contre.

Tous ces indicateurs sont des ratios établis à partir de comptages bibliographiques. Leur utilisation a été reprise par un grand nombre de chercheurs et de praticiens : certains pour les appliquer comme indicateurs pour

[1] Société Française de Bibliométrie Appliquée

la science fondamentale (ce sont essentiellement des laboratoires de recherche), et d'autres en tant qu'indicateurs de la technique (ce sont bien souvent des acteurs du monde industriel ou travaillant pour les entreprises).

Indicateurs de tendances d'activités
- A1 - nombre de brevets
- A2 - part d'une compagnie dans un domaine
- A3 - part d'un domaine dans une compagnie
- A4 - indicateur d'activité
- A5 - avantage concurrentiel
- A6 - évolution technologique
- A7 - distribution géographique
- A8 - indicateur de dispersion
- A9 - Identification de l'inventeur

Indicateurs d'impacts
- I1 - fréquence de citation de brevet
- I2 - ratios de performance de citation
- I3 - indicateur d'impact technique
- I4 - brevets les plus cités
- I5 - indicateur d'impact d'un domaine
- I6 - structure d'agrégats et d'auto-citation

Indicateurs de position
- P1 - Intensité de la science
- P2 - vitesse de référence
- P3 - concentration dans les domaines à fortes évolutions
- P4 - classification multiple
- P5 - trajectoires d'une compagnie

Indicateurs de liens
- L1 - compagnies et technologies "précurseurs" (citées)
- L2 - compagnies et technologies suiveuses (citants)
- L3 - corrélations (avec des compagnies liées)
- L4 - ratios d'attraction technologique
- L5 - statistiques de recouvrement
- L6 - carrefour de citation de compagnie à compagnie

L'exemple que nous présentons est l'application de l'indicateur d'activité pour les brevets. Cet indicateur a déjà été examiné pour son application à l'information scientifique (voir p. 60). Il sera calculé pour une série de pays et pour chaque domaine technique. Ces mesures permettent donc de comparer l'effort fourni par les pays pour chaque domaine. Pour ce faire, il faut normaliser les mesures de façon à limiter les effets de taille.

Une étude de ce type a été réalisée par la responsable de la direction scientifique de Total-CFP (DIMO, 1990) à partir des indicateurs normalisés mis au point par le CHI. La normalisation pondère les données selon la capacité technologique générale du pays et selon la capacité technologique du pays dans le domaine considéré (voir encadré ci-dessous). Nous retrouvons donc, à une soustraction près, l'indicateur d'activité.

Deux représentations graphiques mettent en valeur ces indices. Un graphe permet de comparer les pays selon l'évolution de leur indice E_a en fonction du temps, pour chaque secteur. Pour replacer les écarts d'activité de chaque pays selon leurs indices d'activités, une seconde représentation est réalisable. Pour

chaque section, on dispose la série des années pour chaque pays selon les axes I_{ag} en fonction de I_{as} (la série pour un pays est reliée par une ligne continue).

Soit

N_{sp} = Nombre de brevets d'un secteur pour le pays et par année
N_s = Nombre de brevets d'un secteur pour l'ensemble des pays pour l'année
N_p = Nombre de brevets tous secteurs pour le pays par année
N = Totalité des brevets pour une année tous secteurs tous pays

On peut définir:

Un *indice d'activité "sectorielle"* du pays $I_{as} = N_{sp} / N_s$
Un *Indice d'activité "générale"* du pays $I_{ag} = N_p / N$

L'*indicateur d'écart d'activité* est défini par le calcul:

$$Ea = \frac{(Ias - Iag)}{Iag} = \frac{Ias}{Iag} - 1$$

Pour un pays, une valeur de Ea > 0 signifie que son activité dépasse la moyenne nationale tandis que Ea < 0 signifie l'inverse.

La balance scientifique (relations science - technologie)

L'indicateur présenté ici a été élaboré par l'OST avec la collaboration de Zitt (BARRÉ et ZITT, 1993). Il met en jeu à la fois des données scientifiques et des données brevets. Son objectif est justement l'analyse de la balance entre la contribution et l'exploitation des ressources scientifiques pour chaque pays. D'un côté chaque pays contribue à la constitution d'un bien collectif mondial (la connaissance scientifique) et, de l'autre, il utilise ce bien collectif pour renforcer la compétitivité via ses entreprises (le dépôt de brevet).

La balance scientifique d'un pays est le rapport de la contribution au stock mondial de connaissances scientifiques à l'utilisation de cette science mondiale. Les postulats de travail pour calculer cet indicateur sont :

1) la production scientifique d'un pays est égale à la part mondiale de ce pays dans la publication scientifique répertoriée par l'ISI
2) la production technologique d'un pays est égale au poids mondial de ce pays dans les brevets accordés aux États Unis
3) la citation par un brevet d'une publication scientifique est la marque de l'utilisation par ce brevet de la connaissance scientifique.
4) en connaissant pour chaque champ technologique les brevets citants et les disciplines citées, il est possible de caractériser l'intensité scientifique des différents champs technologiques et la contribution technologique des disciplines scientifiques.

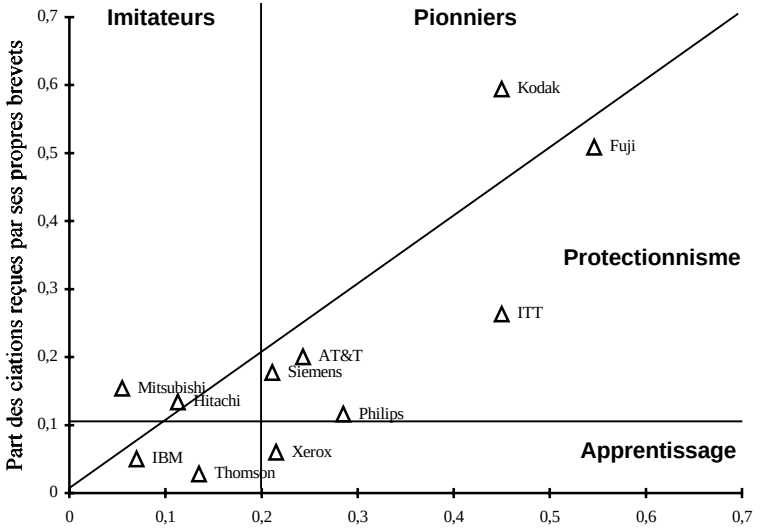
Sur la base de ces principes, une série d'indicateurs normalisés pour évaluer cette balance scientifique a été définie. L'application à cinq pays (Grande Bretagne, Allemagne, Japon, États-Unis et France) s'appuie sur des données fournies par le CHI.

Le graphe de l'avantage stratégique

Ce graphe met en jeu les mêmes données que celles de la société Battelle dans le calcul de son indicateur de dominance (voir p 102). Dans ce graphe, le phénomène de "citation" entre brevets est considéré comme une mesure des dépendances existant entre les sociétés dépositaires de ces brevets. Ainsi, le graphe va consister à situer les sociétés selon deux axes. Le premier axe positionne la société en fonction de son aptitude à ré-exploiter ses propres inventions tandis que l'autre estime la capacité de la société à protéger ses propres inventions.

Ces deux mesures sont tout simplement, pour l'axe des abscisses, la part des brevets déposés par la société qui cite des brevets antérieurs lui appartenant, et pour l'axe des ordonnées, la part des citations reçues par la société provenant de ses propres brevets. Ces deux pourcentages n'utilisent en fait qu'une petite partie des données présentées dans le tableau 16. Le positionnement de chaque société ne dépendra que de 3 valeurs : le nombre de brevets qu'une société cite et qui lui appartiennent (pour la société *i* c'est la valeur x_{ii}), le nombre de brevets que la société a cités (pour la société *i* c'est la valeur $x_{.i}$ = somme de la colonne *i*), et le nombre de brevets qui citent la société (pour la société i c'est la valeur $x_{i.}$ = somme de la ligne *i*).

Fig. 11 - Graphe de l'avantage stratégique selon Mogee (1994)



La mesure de l'aptitude à ré-exploiter ses propres inventions est représentée par le ratio $x_{ii}/x_{.i}$ et est reportée sur l'axe des abscisses, tandis la capacité à protéger ses propres inventions est représentée par le ratio $x_{ii}/x_{i.}$ et est reportée sur l'axe des ordonnées (figure 11). Les sociétés situées au-dessus de la bissectrice sont davantage citées par leurs propres brevets que par les brevets

des concurrents, elles sont dans une situation d'autosuffisance. Tandis que les sociétés situées en dessous de la bissectrice sont en position où elles évoluent plus par ce qu'elles apprennent chez les autres plutôt que chez elles. Les sociétés peuvent finalement être classées dans quatre quadrants formés par l'intersection entre deux zones qui découpent l'axe des abscisses entre les sociétés qui imitent (à gauche) et les pionnières (à droite) avec deux zones qui découpent l'axe des ordonnées entre les sociétés qui apprennent (en bas) et celles qui se protègent (en haut).

Le graphe BCG

Un autre type de graphe stratégique a été mis au point récemment par Nivol. Ce graphe n'est plus construit à partir des dépendances entre sociétés traduites par des citations, mais de l'analyse globale du portefeuille de brevets des entreprises (NIVOL, 1993 ; ROSTAING et alii, 1993b).

Ce portefeuille peut être analysé selon différents critères d'observation : domaine d'activité (représenté par différents codes de classifications documentaires), années de priorité, pays d'extension, sociétés déposantes, famille brevets. Nivol a remarqué que la plupart des traitements sur l'information brevet ne fournissent que des représentations graphiques schématisant des vues parcellaires des politiques générales de dépôts et de recherches développées par société. Or il tenait à obtenir une représentation synthétique, un résumé des principales caractéristiques de chacune des politiques. Pour cela, il s'est inspiré du graphe B.C.G. développé par le cabinet américain *Boston Consulting Group* et l'a adapté aux données stratégiques de la propriété industrielle.

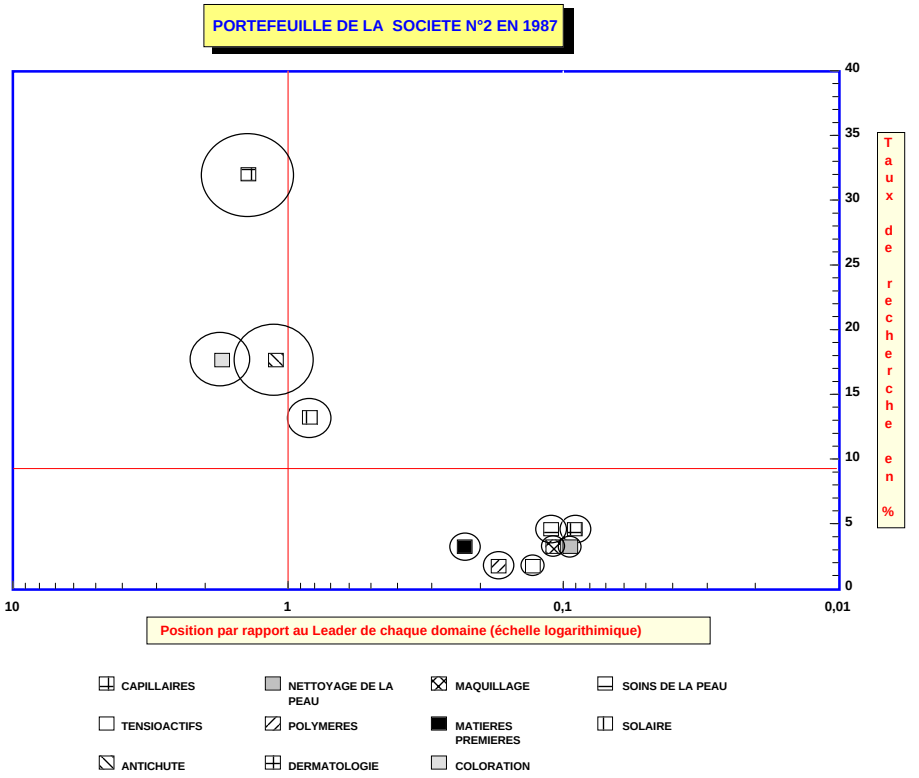
Ces graphes (figure 12) exposent en une seule vision plusieurs facteurs importants en analyse brevet :

- *La répartition des activités*
  Les ordonnées définissent le taux de recherche comme pourcentage de brevets prioritaires déposés dans un domaine considéré par rapport au total des brevets prioritaires déposés tous domaines confondus.

- *La position de l'activité de la société par rapport à sa concurrente immédiate*
  La ligne verticale d'abscisse 1, dite "barre du *leader*", délimite la zone de gauche où la société considérée est *leader* dans le domaine en question. Notons que, pour des raisons de notation, l'échelle horizontale est de type logarithmique. Pour chaque domaine situé sur la barre du *leader* (abscisse = 1), la société étudiée est *leader* ex aequo avec un autre concurrent. De part et d'autre de cette verticale la société a, soit un nombre de brevets déposés dans le domaine considéré de plus en plus éloigné de celui du *leader* (déplacement vers la droite) soit, elle est en tête dans ce domaine et a un nombre de brevets de plus en plus élevé par rapport à son concurrent le plus direct dans le domaine (déplacement vers la gauche).

- *La mesure du monopole d'exploitation de chaque domaine d'activité considéré*
  Les surfaces des cercles sont proportionnelles au nombre de brevets d'extension déposés. Elles représentent la mesure du monopole d'exploitation pour chaque domaine considéré. Ces cercles induisent des gradients qui permettent, à la lecture des graphes, de relever très facilement différentes anomalies dans les politiques d'extension par société et par domaine.

Fig. 12 - Graphe BCG adapté au portefeuille de brevets par Nivol



Comme pour les graphes BCG, Nivol a divisé la surface de ce graphe en quatre parties :

1) la zone vedette (en haut à gauche) : dans cette zone, l'entreprise travaille beaucoup dans le domaine et elle est payée en retour par sa position de *leader*.

2) la zone dilemme (en haut à droite) : les efforts dans le domaine sont aussi intenses que dans la zone vedette mais ici la société n'est pas *leader*.

3) la zone vache à lait (en bas à gauche) : la société domine sans effort dans ce domaine.

4) la zone poids mort (en bas à droite) : la société ne fait pas grand chose dans le domaine et elle est largement dominée.

Ces graphes stratégiques sur la position par domaine d'une société livrent une image synthétique du portefeuille de brevets de chaque société concurrente. La comparaison des différentes situations pour chaque société en est donc grandement facilitée. Ce type de résultats bibliométriques sur l'information brevets ainsi que beaucoup d'autres indicateurs brevets ont été élaborés par Nivol dans un contexte de surveillance de la concurrence pour une industrie française. La contrainte de temps étant primordiale dans une activité de veille technologique, Nivol s'est doté d'outils informatiques adaptés à ses besoins bibliométriques très variés en collaboration active avec le CRRM.

LES CARTES RELATIONNELLES DE BREVETS

Les études mettant en oeuvre des techniques statistiques d'analyse des données pour présenter les structures sous-jacentes à des tableaux de relations sont bien moins nombreuses dans l'évaluation de la technique que dans l'évaluation de la science. Nous présenterons ici uniquement les équipes françaises qui travaillent dans cette voie. Les travaux étrangers sont presque inexistants et sont souvent bien moins pertinents que ceux menés en France (tout au moins pour ceux qui ont été portés à notre connaissance).

Deux approches se dégagent. Les premiers travaux évoqués n'ont fait qu'appliquer aux données brevets des techniques bibliométriques mises au point pour les références scientifiques. La seconde approche se compose de deux techniques nouvelles d'analyse de données qui n'ont encore jamais été mises en application sur des références scientifiques (l'analyse relationnelle et la sériation). Bien qu'elles n'aient pas été développées spécifiquement pour la bibliométrie appliquée aux brevets, ces techniques y trouvent un champ d'application particulièrement propice. Outre le recours à la statistique, ces deux applications se distinguent aussi par le choix des éléments de description du contenu technique des brevets. Les techniques bibliométriques pour évaluer les liens consensuels entre les documents scientifiques font intervenir soit les mots du titre ou des descripteurs soit les codes de classification documentaire. Dans le cas du brevet, seules ces deux techniques emploient les codes de la Classification Internationale des Brevets comme éléments de description du contenu technique et non les mots du titre.

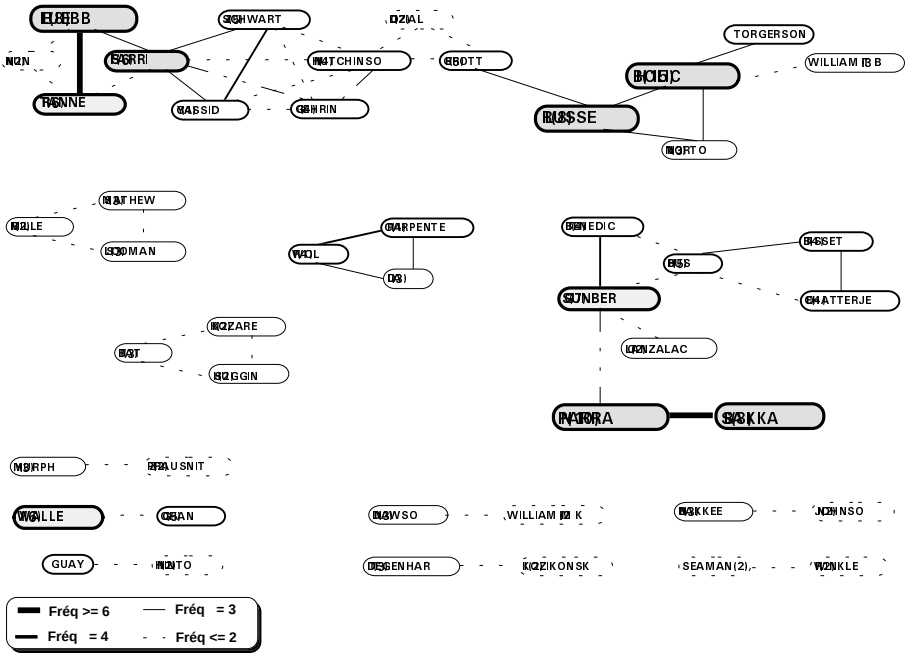
Les réseaux de relations

Les premiers travaux qui ont mis en évidence la dimension relationnelle entre les données brevets sont ceux du CRRM. Ils ont appliqué la même technique de réseaux de relations de codes qu'ils employaient pour l'information scientifique (voir p 79 et 84), mais cette fois-ci sur les codes documentaires attribués par le producteur *CAS* aux références brevets (comme aux références scientifiques) de la banque *CAS* (DOU et HASSANALY, 1988) ou bien sur les codes documentaires attribués par le producteur *Derwent* aux références brevets présentes dans la banque *WPI* (DOU et alii, 1989a). Dans ce dernier article, les

auteurs estiment que ces réseaux pourraient être établis sur toutes les autres entités brevets présentes dans les références, le nom de société, les noms d'inventeurs, les pays...

Tout récemment, cette technique de réseau de relations a été mise en application sur les noms d'inventeurs (NIVOL, 1993) (figure 13) . Nivol a montré que la construction des réseaux de relations de co-dépôts de brevets pour les inventeurs donne des informations utiles à l'identification des structures générales de recherche chez les concurrents directs d'une entreprise. Il a pu caractériser la structure et le fonctionnement de la recherche dans différentes sociétés en fonction de la physionomie générale de ces réseaux de travail.

Fig. 13 - Réseaux de relation des inventeurs de brevets



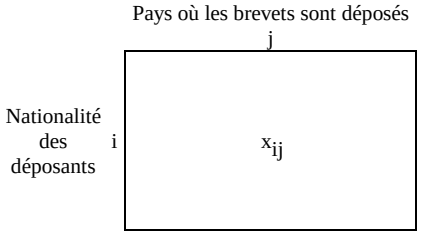
Etude de la nationalité des déposants par pays de dépôt

Certains ont cherché tout d'abord à connaître la fréquence de dépôts de brevets en chimie pour les différentes nationalités déposantes et pour les différents pays où ces brevets sont déposés ou étendus ou désignés. Ces fréquences analysées par des traitements statistiques leur ont permis de dégager les corrélations significatives entre ces deux ensembles (DORÉ et alii, 1987).

Ils ont considéré comme source de brevets en chimie la base *Chemical Abstracts* qui couvre 95% des dépôts de brevets mondiaux en matière de chimie. L'étude n'a tenu compte que des brevets recensés en 81 par *Chemical Abstract Services*, soit 71 770 brevets. La construction de la matrice de contingence analysée croisait les 11 premiers pays déposants (plus la réunion de tous les

autres sur une ligne supplémentaire) avec les 10 principaux pays de dépôts (plus une colonne "divers" et deux colonnes correspondant aux procédures régionales européennes et mondiales) (tableau 18). Les traitements statistiques exécutés sur ce tableau ont fait appel à des méthodes d'analyses de données déjà rencontrées : Analyse Factorielle des Correspondances, Classification Automatique Hiérarchique et analyse typologique par construction d'un Arbre de longueur minimale.

Tabl. 18 - Relations nationalité de dépôt *x* pays de dépôts



x_{ij} = nombre de brevets déposés dans le pays *j* par les déposants de nationalité *i*

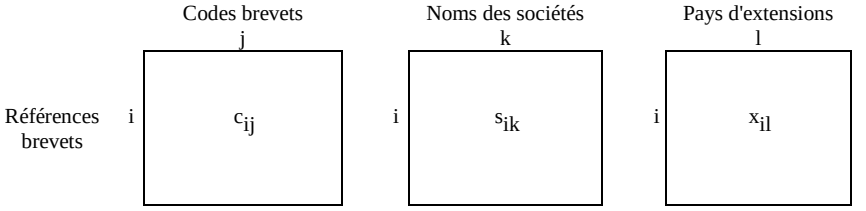
L'analyse relationnelle appliquée aux brevets

En 1990, le centre de mathématiques appliquées d'IBM France (CEMAP) a trouvé dans la bibliométrie un bon domaine d'application des méthodes statistiques conçues antérieurement. Ces méthodes statistiques, qui s'inscrivent dans le cadre méthodologique général de l'Analyse Relationnelle, sont parfaitement adaptées aux caractéristiques des données bibliographiques (loi de Zipf) et permettent de prendre en considération une plus grande part de l'information.

Une étude de Bédécarrax et Huot (BÉDÉCARRAX et HUOT, 1992), présente la méthodologie mise en oeuvre dans l'application de l'analyse relationnelle pour le traitement des corpus provenant de la base *Derwent*. Le traitement statistique porte sur l'exploitation des relations entre les numéros de brevets et les champs de noms de sociétés, les pays d'extension et les codes CIB décrivant le brevet. Ces relations sont recensées dans des tableaux (tableau 19).

Plusieurs traitements sont possibles, soit à partir de ces matrices de départ soit des matrices de similarités relationnelles bâties à partir de ces dernières. Les méthodes de classification automatique proposées vont réorganiser les données contenues dans ces matrices sans élimination ni déformation des données (rencontrées dans toutes les autres analyses des données) mais par simple permutation des lignes ou des colonnes de la matrice initiale. L'objectif de ces réorganisations est d'obtenir une correspondance optimale entre les ensembles d'éléments croisés selon une perspective visée initialement (critère à optimiser). Un autre article (HUOT et alii, 1992) présente de façon didactique une étude employant l'analyse relationnelle suivie de sa représentation factorielle des codes.

Tabl. 19 - Relations dans l'analyse relationnelle des références *Derwent*



c_{ij} = 1 si le code *j* est présent dans le brevet *i* sinon 0
s_{ik} = 1 si la société *k* a déposé le brevet *i* sinon 0
x_{il} = 1 si le pays d'extension *l* est présent dans le brevet *i* sinon 0

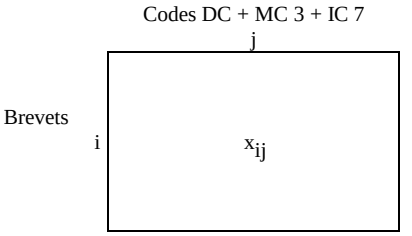
Une seconde étude a été l'aboutissement d'une collaboration entre le CEMAP et le CRRM (ROSTAING et alii, 1993a). Les auteurs ont cherché à estimer les avantages que pourrait apporter une utilisation simultanée de plusieurs catégories d'éléments bibliographiques qui décrivent le contenu technique des brevets. Il existe plusieurs champs dans une référence brevet qui expriment de façon contrôlée le texte technique signalé dans la référence. Pour la référence issue de la banque de données *WPI* de *Derwent*, on peut en dénombrer au moins cinq : le titre Derwent (le titre original est remplacé par un titre donné par un expert Derwent), le titre normalisé (c'est le titre Derwent après un léger traitement automatique lexical), la codification documentaire Derwent (DC, affecté par Derwent), la codification des manuels codes de Derwent (MC, affecté par Derwent, plus précis que les premiers), la codification internationale des brevets (IC, affecté pendant la procédure de dépôts par les offices spécialisés). Les mots du titre étant des données moins bien contrôlées que les codifications documentaires, les auteurs ont voulu évaluer l'apport spécifique de chacune des trois classifications documentaires comme éléments de description des brevets.

A partir d'un ensemble de références correspondant au thème des "systèmes transdermiques thérapeutiques sous forme de timbre", la première étape est de déterminer les niveaux de chacune des hiérarchies des plans de classement à traiter. Sur la base de critères purement statistiques, ce niveau pour chaque plan de classement a été défini : la hiérarchie la plus complète pour les DC, le second niveau de la hiérarchie pour les MC (3 caractères = MC3) et le quatrième niveau de la hiérarchie des IC (7 caractères = IC7). Le tableau des relations intriquant ces trois codifications a été construit (tableau 20). Il est classifié par l'analyse relationnelle pour agréger les lignes, puis les colonnes.

La classification des lignes a permis de regrouper tous les brevets appartenant à la même famille d'inventions selon leur similitude de description par les trois codifications. Les auteurs démontrent, par un exemple concret, que l'utilisation simultanée des trois codifications permet de mieux classer les références brevets et d'offrir une meilleure interprétation des agrégats par la complémentarité des sens de chaque codification (les MC expriment mieux les

notions de produits chimiques tandis que les IC décrivent mieux les spécificités techniques).

Tabl. 20 - Relations dans l'analyse de la complémentarité de codifications brevets



x_{ij} = 1 si le brevet *i* comporte le code *j* (un code soit DC soit MC3 soit IC7), sinon 0

La classification des colonnes regroupe les codes selon leurs profils de présence dans l'ensemble des brevets. Initialement, cette classification avait pour but de détecter automatiquement à la fois la synonymie et les déficiences entre les trois plans de classement. Les résultats livrés ne permettent pas de conclure sur ces points, mais par contre, cette classification s'avère parfaitement adaptée pour la détection de toutes les irrégularités de description de certains brevets. En effet, cette méthode désolidarise immédiatement tous les codes employés de manière marginale. Les documents qui possèdent ces codes sont donc décrits de façon atypique. Ainsi, deux types de documents atypiques sont mis en évidence, ceux qui correspondent à des erreurs lors de la collecte des références, et ceux qui possèdent des revendications à protéger très originales et donc probablement très innovantes.

En conclusion, l'analyse relationnelle couplée à la technique de complémentarité de codes permet, pour la classification des lignes, d'obtenir une structure finale plus riche, et pour la classification des colonnes, la détection des documents bruits et des documents potentiellement stratégiques.

La *"sériation des similarités spécifiques"*

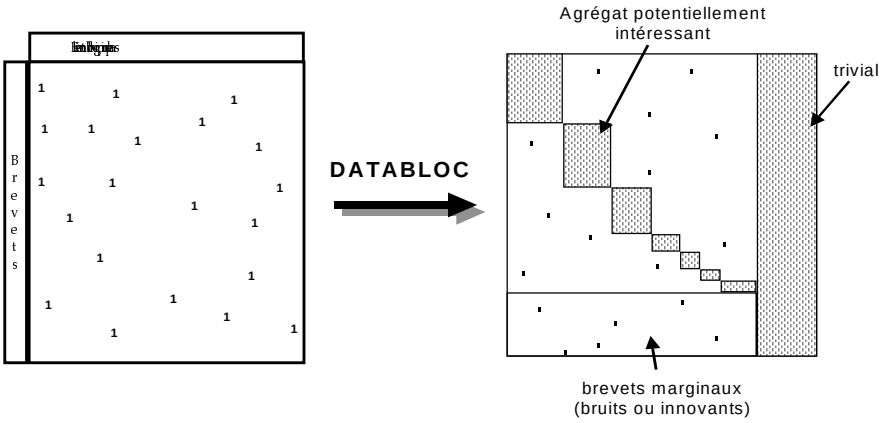
La méthode d'analyse des données nommée sériation est une méthode de classification automatique. En France, cette méthode a été mise au point, entre autres, par Marcotorchino (CEMAP) pour des données tirées de domaines autres que ceux de la bibliométrie. Or les distributions des données rencontrées en bibliométrie sont de type hyperbolique - zipfienne et donc ne respectent pas les mêmes répartitions que dans les domaines traditionnels de l'analyse des données (voir p 46). Cette caractéristique des données de notre discipline a dissuadé l'équipe du CEMAP de l'inadéquation de la sériation à ces données.

Le CRRM a repris cette technique pour l'adapter aux données bibliométriques. Ainsi, Baldit, dans sa thèse, a-t-il pu parfaire un nouvel algorithme de sériation adéquat à l'analyse bibliométrique (BALDIT, 1994). Ce travail ne s'est pas limité au perfectionnement de la technique de classification

statistique, il a aussi permis une amélioration considérable dans l'interprétation des résultats par une "navigation" hypertexte construite sur les résultats de la sériation. L'aboutissement de ces deux axes de recherche a été le développement d'un logiciel qui intègre le tout, DATABLOC.

Cette nouvelle méthode d'analyse des données, nommée "*sériation des similarités spécifiques*", a été présentée dans une étude traitant de l'information brevet (QUONIAM et alii, 1993). L'algorithme est conçu pour analyser des tableaux de présence/absence identiques à ceux exploités par l'analyse relationnelle (tableau 19). Les avantages de cette sériation sont nombreux. Premièrement, c'est une classification automatique qui n'impose pas la définition du nombre d'agrégats formés ni le nombre d'éléments composant ces agrégats. Deuxièmement, elle détecte automatiquement les éléments bibliométriques indésirables à la technique de sériation : le peu d'éléments à très fortes fréquences que les auteurs reconnaissent sous le nom d'ensemble trivial (voir p. 34) et isole pour qu'ils ne perturbent pas les autres agrégats (figure 14). Ces éléments n'apportent aucun intérêt statistique, puisqu'ils sont présents dans pratiquement toutes les références et devraient donc être liés à tous les autres éléments. Troisièmement, la sériation détecte automatiquement les références ne pouvant appartenir à aucun des agrégats car elles sont décrites par des éléments beaucoup trop marginaux.

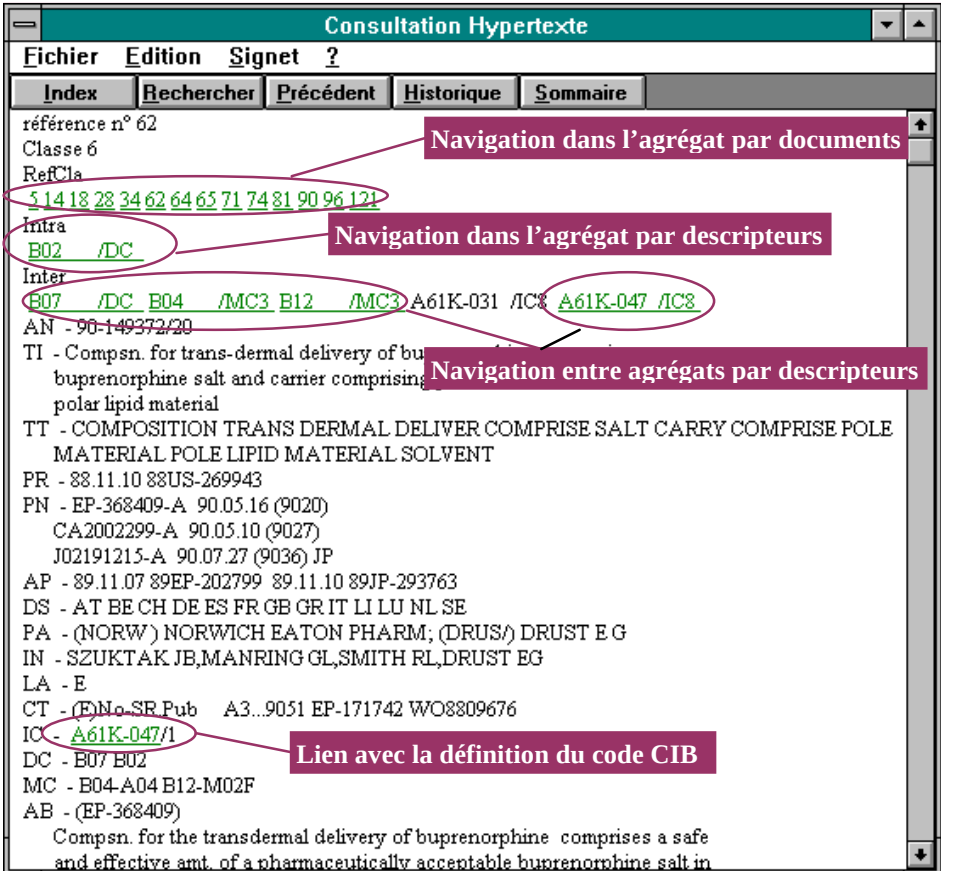
Fig. 14 - Résultats statistiques schématiques créés par DATABLOC



Et pour finir, cette sériation a le considérable avantage de classifier simultanément les deux dimensions du tableau de départ. Dans le cas d'un tableau croisant les références de brevets avec les codes de classification qui expriment les revendications protégées par ces brevets, la sériation optimise la combinatoire des permutations lignes et colonnes de façon à obtenir des agrégats discriminés à la fois par des brevets et des codes. Ces agrégats rassemblent les brevets comportant la même combinaison de codes mais aussi les codes selon leurs profils de similarité dans l'ensemble des brevets. On obtient donc un agrégat parfaitement homogène et facile à interpréter par simple lecture des intitulés de lignes et de colonnes.

L'objectif d'un tel développement informatique est la structuration d'une grande masse de brevets comme outil d'aide à la veille technologique. Dans ce contexte, le résultat statistique en lui-même n'est pas d'une grande utilité. Par contre le fait que les brevets soient classés selon leurs spécificités techniques facilite grandement l'interprétation. Un expert dispose alors d'une "grille de lecture" offrant un accès plus rapide à l'essentiel des caractéristiques des brevets. Pour parachever cette aide, il fallait que l'expert puisse s'en servir sans aucune entrave informatique. Baldit a proposé une navigation dans les références des brevets avec des liens hypertextes construits en fonction des résultats statistiques de l'agrégation : navigation à l'intérieur d'un agrégat, passage d'un agrégat à l'autre par les éléments bibliographiques charnières à ces agrégats, navigation par le réseau des éléments bibliographiques d'un même agrégat et navigation entre les codes et leurs significations dans les plans de classement (figure 15).

Fig. 15 - Navigation hypertexte selon les agrégats statistiques de DATABLOC



Cet outil de navigation, qui fait appel pour tous à des réflexes intuitifs, favorise ainsi la richesse d'interprétation pour les experts et par là même la qualité des résultats dans un processus de veille technologique.

CONCLUSION

Nous venons de le constater, la bibliométrie appliquée aux brevets est un outil parfaitement adapté à l'activité de surveillance de la concurrence. Ce n'est pas une fin en soi, mais la bibliométrie constitue un outil supplémentaire à introduire dans la panoplie des techniques de veille industrielle. Elle se trouve à la conjonction de deux constats. D'un côté, une industrie qui a l'intention de pérenniser son métier se doit d'avoir une stratégie de protection de son patrimoine technologique et donc une politique de propriété industrielle. D'un autre, la masse et la complexité[1] des informations brevets font qu'il n'est plus envisageable de continuer une activité de surveillance sans des outils performants pour apporter une aide considérable dans le choix des informations à considérer et à expertiser. La bibliométrie appliquée aux brevets répond à ces deux critères. Elle sait construire des aides à la lecture et à la compréhension des influences entre les différents facteurs importants dans la stratégie brevet sans aucune limitation du nombre de brevets analysés.

Ce constat établi, on peut s'étonner que si peu d'études bibliométriques soient exposées ici. Deux explications peuvent être formulées. La première est liée au coût de l'information brevet. Cette information étant critique pour les entreprises, celles-ci sont prêtes à payer cher pour acquérir ces données, si bien que les producteurs et serveurs de banques de données n'hésitent pas à vendre cette information au prix fort. Cela explique que les serveurs abandonnent petit à petit les répertoires scientifiques au profit des répertoires brevets et entreprises. Ceci précisé, il parait compréhensible que les centres de recherches en bibliométrie n'exploitent ce type de données que lorsqu'ils réalisent une étude en collaboration avec une entreprise prête à financer la constitution du corpus à analyser. C'est alors que la seconde raison entre en jeu. Les entreprises sont prêtes à financer des études, mais certainement pas pour voir les résultats et les méthodes communiqués au public et à la concurrence. Les centres de recherche sont donc contraints de signer des clauses de confidentialité qui leur imposent le secret. Les travaux bibliométriques publiés sont donc principalement des études réalisées par des centres nationaux d'évaluation de la recherche et des techniques. Ces études financées par les gouvernements, les instituts nationaux ou les groupes d'intérêts ont pour volonté de dégager des informations générales sur la position et les relations entretenues par des entités d'envergure nationale ou internationale. Ces résultats sont bien souvent peu utiles pour l'aide à l'innovation dans une industrie.

[1] Le nombre de facteurs à considérer dans une analyse brevets est nettement plus important que dans une analyse de références scientifiques. Pour s'en rendre compte, il suffit de consulter la liste des éléments bibliographiques présents dans une référence brevet et d'imaginer le nombre de combinaisons entre ces éléments qui pourraient être pertinentes pour une parfaite connaissance des stratégies de brevets dans le métier d'une entreprise.

CONCLUSION
LA BIBLIOMÉTRIE EN MOUVEMENT

Bien que méconnue d'un grand nombre de chercheurs, la bibliométrie arrive à maturité. Son évolution a été considérable tant au niveau des concepts que des techniques. Elle a vu le jour sous l'apparence d'un outil de gestion des bibliothèques puis s'est transformée en un instrument de mesure dédié aux sciences de l'information, pour devenir finalement la principale méthode d'évaluation quantitative des sciences et des techniques. Ce dernier virage n'est pas pour plaire à tous. L'abandon de la recherche sociologique et cognitive des sciences de l'information, au profit du simple perfectionnement de techniques opérationnelles, désole nombre de spécialistes de la bibliométrie.

Cette évolution a une explication, le besoin de financement. La collecte des informations à partir des banques données internationales puis leur traitement par des méthodes statistiques perfectionnées a fait de la bibliométrie une activité coûteuse. L'investissement en matériel informatique performant pour le stockage de grandes masses de données et leur exploitation rapide est devenu indispensable. Le maintien d'une activité bibliométrique de qualité impose donc des investissements non négligeables. Seuls les instituts nationaux et les industries peuvent se permettre de telles dépenses, dans la mesure où les renseignements fournis sont très précieux pour l'élaboration de leurs stratégies. La majeure partie des chercheurs en bibliométrie a probablement quitté durablement le domaine de la connaissance, de la communication et de l'évolution scientifiques pour s'introduire dans le secteur de la recherche opérationnelle.

Lorsqu'elle est au service des politiques de la recherche publique, la bibliométrie est source de nombreuses interrogations. Cet outil quantitatif,

introduit dans le processus de prise de décision de la politique de recherche, est à double tranchant. En effet, les indicateurs ont été pensés initialement pour disposer de renseignements plus objectifs que le simple avis des experts. Aussi, certains responsables peuvent estimer que la seule connaissance de ces indicateurs suffit pour se forger des opinions. Or comme toute donnée statistique, ces indicateurs peuvent être interprétés de multiples façons s'ils sont sortis de leurs contextes. Il ne faut absolument pas considérer les données bibliométriques comme des substituts à l'avis d'experts, mais tout au contraire, comme des outils de travail dans l'élaboration du consensus parmi des experts. Une information non vérifiée peut se transformer en désinformation. De la même façon, un indicateur bibliométrique non validé par un groupe d'experts peut provoquer de graves contresens. Ces données statistiques n'ont de sens que par l'interprétation qui peut en être faite. Et qui est le mieux placé pour déchiffrer ces indices, si ce n'est le spécialiste du thème étudié ? Cette condition de validité des informations fait l'unanimité dans la communauté des bibliomètres. Cependant, la crainte d'une utilisation abusive par les pouvoirs publics d'indicateurs trop simplistes pour être significatifs reste présente.

Pour ne pas entrer dans cette exploitation bien trop périlleuse de l'outil bibliométrique et pour respecter par là même une certaine éthique, certains centres de recherches ont préféré mettre leur savoir au service de l'industrie. Ce nouveau domaine d'application statistique est ressenti comme une nécessité dans le milieu industriel. Au même titre que le marketing ou l'analyse financière, la bibliométrie a sa place dans l'entreprise comme méthode statistique. Son intérêt pour estimer la position de la R&D et du portefeuille brevets d'une entreprise par rapport à sa concurrence ne fait plus aucun doute. Le suivi systématique de l'activité technologique des concurrents est à l'ordre du jour dans les industries. L'utilisation des termes "veille technologique", "veille industrielle" et "intelligence économique" est devenue monnaie courante dans les discours industriels. Mais les outils spécifiques à cette activité font fortement défaut. Les techniques bibliométriques paraissent parfaitement répondre à cette attente.

De longue date, les méthodes bibliométriques sont expérimentées allant vers toujours plus de pertinence et de perfectionnement dans l'analyse de la complexité. Ces méthodes sont arrivées à un niveau de mise au point suffisant pour concevoir des applications dans le contexte industriel. Mais pour une parfaite intégration, elles doivent répondre à de nouvelles contraintes propres aux exigences industrielles. La bibliométrie se doit de garantir au moins quatre critères pour la réussite de son intégration. Tout d'abord, s'adapter à la variété des types de données intéressant l'industrie. Il n'est pas imaginable, pour une entreprise, de limiter ses analyses à une seule source d'information lors d'une large surveillance de sa concurrence. De plus, chaque métier fait appel à des spécialités plus ou moins bien représentées selon la banque de données consultée. Les techniques bibliométriques doivent prendre en compte cette variété de sources et exploiter au mieux la spécificité de chacune d'elles. En

second lieu, les analyses bibliométriques doivent pouvoir traiter des volumes de données considérables pour répondre à des exigences d'exhaustivité ou à des désirs de vision globale d'un domaine. Le troisième critère concerne la reproductibilité et la réalisation systématique des traitements. La bibliométrie ne peut s'envisager dans un processus de surveillance que si les résultats fournis sont reproductibles sur des jeux de données différents et si certains indicateurs sont produits avec régularité. Elle doit servir non seulement à construire des dossiers de synthèse mais aussi à l'élaboration de systèmes d'alerte basés sur des comparaisons d'indices prélevés à un intervalle de temps régulier. La dernière caractéristique indispensable aux outils bibliométriques est la notion de rapidité d'exécution. L'activité de veille industrielle a pour principe de donner la bonne information au bon moment. Aussi, la bibliométrie ne peut être utile à l'entreprise que si elle respecte cette contrainte. Pour répondre à toutes ces exigences, seule une bibliométrie informatisée offrira une totale opérationnalité, si elle est en synergie avec les nouvelles technologies de l'information.

Pour l'obtention de résultats concluants, il faut impérativement accompagner cette informatisation par une parfaite maîtrise des outils. Plus les outils sont faciles à manipuler et sont performants, plus l'erreur est facilement générée. L'utilisation de tels moyens en entreprise impose une professionnalisation de cette activité. Les compétences requises sont multiples et spécifiques. Une très bonne connaissance des banques de données et des politiques de leurs producteurs parait indispensable. Il faudra y adjoindre, bien évidemment, la connaissance des données propres aux procédures de dépôts de brevets. Être formé à la statistique et aux méthodes de l'analyse des données parait essentiel pour une compréhension et une étude parfaite des résultats élaborés. Une très bonne culture des fondements et des concepts de la bibliométrie permet d'assurer le contrôle et la validité des traitements dans leur ensemble. Il est déconcertant de remarquer que la bibliométrie permet de transformer l'information purement intellectuelle, formalisée grâce au support de l'écrit, en une information soit quantitative soit cartographique. Il faut être bien conscient de cette impressionnante condensation de l'information : recherches ⇨ articles ⇨ références bibliographiques⇨ tableaux croisant 1 ou 2 champs bibliographiques ⇨ exploitation statistique des tableaux ⇨ indicateurs. La bibliométrie met en oeuvre des techniques fortement "réductionnistes". Le praticien doit constamment garder cela à l'esprit pour ne pas se laisser aller à de trop hâtives et trop simplistes conclusions. Les techniques bibliométriques ne seront admises par les industriels que si elles sont utilisées avec professionnalisme et donc mises entre les mains de spécialistes.

Les progrès de l'informatique et des nouvelles technologies de l'information laissent présager de grandes améliorations des outils bibliométriques. La méthodologie bibliométrique étant arrivée à maturité, on peut espérer qu'une partie de l'énergie et l'ingéniosité de la recherche se dirigera vers de nouveaux produits reposant sur ces progrès technologiques. À court terme, ces technologies devraient être intégrées dans des développements

informatiques orientés vers la bibliométrie. Des travaux prennent déjà en compte les nouveaux médias de stockage de l'information comme le CD-ROM. Certains centres de recherche, bien souvent français, ont rapidement vu dans les systèmes hypertextes des outils de navigation tout à fait adaptés à la conciliation des résultats statistiques avec les données initiales, les références. La mise en oeuvre de ce nouveau concept informatique, dans la phase d'expertise des résultats bibliométriques, paraît très prometteuse, surtout pour les applications de veille technologique.

A moyen terme, on peut penser que les techniques linguistiques auront été suffisamment adaptées à l'information scientifique et technique. Alors, un nouveau champ d'investigation verra le jour, car les analyses bibliométriques ne se limiteront plus uniquement aux traitements des données produites par des processus d'analyse documentaire et d'indexation (phase de création d'une notice bibliographique). Les textes mêmes des auteurs ou des brevets offriront l'opportunité d'analyses beaucoup plus fines. Tout d'abord, les bibliomètres s'attaqueront aux résumés ou aux principales revendications des brevets pour ensuite lancer leur dévolu sur les textes complets. Les techniques linguistiques permettent de transformer les textes rédigés avec la subtilité et la complexité du langage naturel en un texte constitué d'un vocabulaire "réduit". Cette intervention des méthodes linguistiques pourrait finalement se rapprocher d'une indexation automatisée. Le vocabulaire "réduit" fera perdre forcément une part importante du sens du texte d'origine, mais par contre, il sera garant d'un gain de signification statistique (réduction de la diversité du vocabulaire donc augmentation du signal de chaque terme du vocabulaire). Ce nouvel outil intégré dans la chaîne des traitements bibliométriques laissera envisager l'analyse et la compréhension des sources construites autour d'une diffusion de la connaissance en langage naturel. Ainsi, les nouveaux vecteurs de communication scientifique et technique tels que la messagerie électronique et les forums de discussion sur le réseau international Internet ne seront peut-être plus aussi mystérieux...

BIBLIOGRAPHIE

ADVISORY BOARD FOR THE RESEARCH COUNCILS, *Evaluation of the National Performance in Basic Research*, The Royal Society, Economic and Social Research Council, Londres, 1986

AGIRRE K FZ, PIRIS J M, TUSELL F, "An Analysis of Citations in Statistical Journals", Actes du colloque : *First International Symposium of Applied Stochastic Models and Data Analysis*, 23-26 avril, Grenade, Espagne, 1991

AIYEPEKU W O, "Bradford Distribution Theory - Compounding of Bradford Periodical Literatures in Geography", *Journal of Documentation*, Vol 33, N°3, p. 210-219, 1977

ALABI G, "Bradford's Law and its Application", *International Library Review*, Vol 11, p. 151-158, 1979

ASHOK J, GARG K C, "Laser Research in India : Scientometric Study and Model Projections", *Scientometrics*, Vol 23, N°3, p. 395-415, 1992

ASHTON W B, CAMPBELL R S, LEVINE L O, "Patent Analysis as a Technology Forecasting Tool", Actes du colloque : *Fall Conference*, Atlantic City, NJ, 18-21 sept, 1983

BALDIT P, *La sériation des similarités spécifiques dans la recherche de l'information stratégique*, Thèse : Aix-Marseille III, Déc., 1994

BARRÉ R, "Clustering Research Fields for Macro-strategic Analysis : a Comparative Specialisation Approach", *Scientometrics*, Vol 22, N°1, p. 95-112, 1991

BARRÉ R, LAVILLE F, "Macrobibliométrie sur les brevets européens : des données brevets aux indicateurs technologiques", *Revue française de bibliométrie*, Vol 12, p. 346-359, 1993

BARRÉ R, ZITT M, "Science applicable et technologies intensives en science : indicateurs globaux pour quelques grands pays", *Les Cahiers de l'ADEST*, Numéro spécial, p. 117-130, 1993

BASSECOULARD-ZITT E, ZITT M, "Une chaîne d'analyse scientométrique dynamique. Application à l'étude en longue période d'une revue scientifique", *Revue française de bibliométrie*, Vol 12, p 93-117, 1993

BAUIN S, CRANCE M, SIGOGNEAU M, QUINAULT L, "Les français dans la base de publications scientifiques SCI de l'ISI", *Les Cahiers de L'ADEST*, Numéro spécial, p. 40-43, 1993

BÉDÉCARRAX C, HUOT C, "L'application de l'analyse relationnelle à la veille technologique", *Revue française de bibliométrie*, Vol 9, p. 64-80, 1992

BENNION B, KARSCHAMROON S, "Multivariate Regression Models for Estimating Journal Usefulness in Physics", *Journal of Documentation*, Vol 40, p. 217-227, 1984

BENZÉCRI JP, *L'analyse des données*, Tome 1 : La taxinomie, Tome 2 : L'analyse des correspondances, Editions Dunod, Paris, 1973

BERNAT J-P, "Approches préliminaires pour l'élaboration d'une technique de suivi scientométrique", *Les Cahiers de l'ADEST*, Numéro spécial, p. 74-79, 1993

BILLARD P, DOUSSET B, HILAIRE A, LAURENT D, PAOLI C, LONGEVIALLE C, "Medline vue par l'analyse factorielle et la classification automatique", *Revue française de bibliométrie*, Vol 7, p. 61-74, 1990

BRADFORD S C, "Sources of Information on Specific Subjects", *Engineering*, Vol 137, p. 85-86, 1934

BRADFORD S C, *Documentation*, Crosby Lockwood & Son, London, 156 p., 1948

BRAUN T, GLANZEL W, SHUBERT A, "Scientometric Indicators Datafiles", *Scientometrics*, Vol 28, p. 137-150, 1993

BROOKES B C, "The Derivation and the Application of the Bradford-Zipf Distribution", *Journal of* D*ocumentation*, Vol 24, N°4, p. 247-267, 1968

BROOKES B C, "Biblio-, Sciento-, Info-métrics ??? What are we talking about ?", *First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*, August 24-28, Diepenbeek, Belgium, 1987

BUDD J, HURT C D, "Superstring Theory : Information Transfert in an Emerging Field", *Scientometrics*, Vol 21, N°1, p. 87-98, 1991

BUFFETEAU A, "Etude de l'impact de la revue de l'Institut Français du Pétrole par des méthodes bibliométriques", *Revue française de bibliométrie*, Vol 9, p. 293-323, 1991

BURREL Q L, "The Bradford Distribution and the Gini Index", *Scientometrics*, Vol 21, N°2, p. 181-194, 1991

CALLON M, COURTIAL J-P, LAVILLE F, "Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research : the Case of Polymer Chemistry", *Scientometrics*, Vol 22, N°1, p. 155-205, 1991

CALLON M, COURTIAL J P, PENAN H, *La scientométrie*, Edition Presses universitaires de France, Paris, 126 p., 1993

CALLON M, LEYDESDORFF L, "La recherche française est-elle en bonne santé ?", *La Recherche*, N°186, p. 412-419, 1987

CARDINE P, MULLER E, TURNER W A, "Une station de travail de lecture des contenus documentaires pour la veille scientifique et technique", *Les Cahiers de L'ADEST*, Numéro spécial, p. 34-39, 1993

CARPENTER M, "Similarity of Pratt's Measure of Class Concentration to the Gini Index", *Journal of the American Society for Information Science*, Vol 30, p. 108-110, 1979

CARPENTER M P, NARIN F, "Clustering of Scientific Journals", *Journal of the American Society for Information Science*, Vol 24, p. 425-436, 1973

CARPENTER M P, NARIN F, "The Adequacy of the Science Citation Index (SCI) as an Indicator of International Scientific Activity", *Journal of American Society for Information Science*, Vol 32, N°6, p. 430-439, 1981

CHEN Y S, *Statistical Models of Text : a System Theory Approach*, Thèse : Purdue University, 1985

CHUNG Y-K, "Bradford Distribution and Core Authors in Classification Systems Literature", *Scientometrics*, Vol 29, N°2, p. 253-269, 1994

COILE R C, "Lotka's Frequency Distribution of Scientific Productivity", *Journal of the American Society for the Information Science*, Vol 28, N°6, p. 366-370, 1977

COLE F J, EALES N B, "The History of Comparative Anatomy. Part I : a Statistical Analysis of the Literature", *Science Progress*, Vol 11, p. 578-596, 1917

COURTIAL J-P, *Introduction à la scientométrie*, Editions Anthropos - Economica, 137 p., 1990

COURTIAL J-P, "Comments on Leydesdorff's : a Validation Study of Leximappe", *Scientometrics*, Vol 25, N°2, p. 313-316, 1992

DE LOOZE M-A, JOLY P-B, "Utilisation d'outils bibliométriques appliqués aux biotechnologies végétales", *Les Cahiers de l'ADEST*, Numéro spécial, p. 88-93, 1993

DEMAZURE M**,** "De la pratique et du bon usage des processus d'évaluation des chercheurs", *Pour la science*, N°117, p. 7, 1992

DESVALS H, DOU H (éds), *La veille technologique*, Edition Dunod, Paris, 436 p., 1992

DEVALAN P, BELLE F, "Veille technologique par la bibliométrie : une image statistique de la banque de données bibliographique du CETIM", *Technologie mécanique*, N°9, p. I-XI, 1989

DEVALAN P, BELOT J-M, FREMAUX P, "Les marchés de la productique", *CETIM-Informations*, N°124, p. 35-41, 1991

DEVALAN P, CANDORET J P, BOUVET C, LION J C, "La bibliométrie. Un outil de veille technologique pour l'entreprise", *CETIM-informations*, N°116, p. 89-95, 1990

DIMO I, "Méthodologie pour l'étude de l'évolution scientifique et technologique", *Revue française de bibliométrie*, Vol 6, p. 302-330, 1990

DOBOROV G M, KORENNOI A A, "The Information Basis of Scientometrics", in *On Theoretical Problems of Informatics*, Moscow VINITI for FID, p. 165-191, 1969

DORÉ J C, GILBERT J, MIQUEL J-F, DUTHEUIL C, "Banques de données et analyses multivariables", *Revue française de bibliométrie*, Vol 1, p14-25, 1987

DOREIAN P, "Structure Equivalence in Psychology Journal Network", *Journal of American Society for Information Science*, Vol 36, N°6, p. 411-417, 1985

DOU H, HASSANALY P, "Mapping the Scientific Network of Patent and Non-patent Documents from Chemical Abstracts for a fast Scientific Analysis", *World Patent Information*, Vol 10, N°2, p. 133-149, 1988

DOU H, HASSANLY P, QUONIAM L, "Easy Mapping Classification of Patent References with Microcomputers", Actes du colloques : *The Montreux International Chemical Conference*, Edition Harry Collier, Infornortics Ltd, Calne, Suisse, p. 283-309, 1989a

DOU H, HASSANALY P, QUONIAM L,"Infographics Analytical Tools for Decision Makers", *Scientometrics*, Vol 17, N°1-2, p. 61-70, 1989b

DOU H, HASSANALY P, QUONIAM L, "Informations stratégiques en chimie. Analyse topologique automatique de la base Chemical Abstract", *Revue française de bibliométrie*, Vol 7, p. 14-45, 1990a

DOU H, HASSANLY P, QUONIAM L, LA TELA A, "Competitive Technology Assessment. Strategic Patent Clusters obtained with Non-boolean Logic. New Applications of the GET Command", *World Patent Information*, Vol 12, N°4, p. 222-229, 1990b

DOU H, HASSANALY P, QUONIAM L, LA TELA A, "La veille technologique et l'information documentaire", *Le Documentaliste*, Vol 27, N°3, 1990c

DOU H, HASSANALY P, SNEE S, "Automatic Generation of Strategic Matrices from Online Databases", *World Patent Information*, Vol 13, N°4, p. 223-229, 1991

DOU H, QUONIAM L, HASSANALY P, "Etude de la chimie à Marseille de 1981 à nos jours", *Science et Technologie*, N°9, p. 34-38, 1989c

DOU H, QUONIAM L, ROSTAING H, NIVOL W, "L'analyse des données au service de la bibliométrie", *Revue française de Bibliométrie*, Vol 8, p. 27-67, 1990d

DOUSSET B, DKAKI T, LONGEVIALLE C, "Qualité de l'information et analyse des données", *Revue française de bibliométrie*, Vol 12, p. 198-204, 1993

DOUSSET B, DKAKI T, KOUSSOUBE S, LONGEVIALLE C, HILAIRE A, "Qu'apporte la représentation de la 4ème dimension en analyse de données multidimensionnelles", Actes du colloque : *Les systèmes d'informations élaborées*, Ile Rousse, Juin, 1991

DUMAS S, *Développement d'un système de veille stratégique dans un Centre Technique*, Thèse : Aix-Marseille III, 199 p., 1994

DUTHEUIL C, *L'état de l'art de la bibliométrie et de la scientométrie en France et à l'étranger*, Rapport pour le compte du SGDN n° 24/SGDN/STS/VST/5, 64 p., 1991

DUTHEUIL C, GRANDJEAN N, PETITJEAN TESTA F, "Qualification de l'information à traiter en bibliométrie", *Revue française de bibliométrie*, Vol 12, p. 158-171, 1993

EGGHE L, "Pratt's Measure for some Bibliometric Distributions and its Relation with the 80/20 Rule", *Journal of the American Society for Information Science*, Vol 38, N°4, p. 288-297, 1987

EGGHE L, "The relative Concentration of a Journal with the Respect to a Subject and the Use of Online Services in Calculating it", *Journal of the American Society for Information Science*, Vol 39, N°4, p. 281-284, 1988

EGGHE L, "The Exact Place of Zipf's and Pareto's Law Amongst the Classical Informatrics Laws", *Scientometrics*, Vol 20, N°1, p. 93-106, 1991

EGGHE L, ROUSSEAU R, "A Characterization of Distribution which Satisfies Price's Law and Consequences for the Laws of Zipf and Mandelbrot", *Journal of Information Science*, Vol 12, p. 193-197, 1986

EGGHE L, ROUSSEAU R, "Elements of Concentrations Theory", Actes du colloque : *Informetrics 89/90*, Editions Elsevier, Amsterdam, 1990

ESTIVALS R, "La statistique bibliographique", *Bulletin des bibliothèques de France*, N°12, p. 481-502, 1969

FAIRTHORNE R A, "Progress in Documentation", *Journal of Documentation*, Vol 25, N°4, p. 319-343, 1969

FEDOROWICZ J, "A Zipfian Model of an Automatic Bibliographic System : an Application to MEDLINE", *Journal of the American Society for Information Science*, p. 223-232, 1982

GARFIELD E, *Citation Indexing - its Theory and Application in Science, Technology, and Humanities*, John Willey & sons, New York, 274 p., 1979

GARFIELD E, "The 1,000 Contemporary Scientists most-cited 1957-1978. Part I. The Basic List and Introduction", *Current Contents*, N°41, p. 5-14, 1981

GARFIELD E, "Journal Citation Studies : 36 Pure and Applied Mathematics Journals : what theys cite and vice-versa", *Current Contents*, Vol 15, p. 5-13, 1982

GEORGEL A, *Classification statistique et réseau de neurones formels pour la représentation des banques de données documentaires*, Thèse : Université Paris VII, Juin, 1992

GEORGEL A, BAOUM D, TURNER W, "L'analyse statistique au service de la navigation hypertextuelle : la constructiond'un modèle neuronal d'accès aux informations documentaires", *Revue française de bibliométrie*, Vol 12, p. 29-41, 1993

GOFFMAN W, NEWILL V A, "Generalisation of Epidemic Theory ; an Application to the Transmission of Ideas", *Nature*, Vol 204, p225-228, 1964

GROOS O V, "Bradford's Law and the Keenan-Atherton Data", *American Documentation*, Vol 18, p. 46, 1967

GROSS P L K, GROSS E M, "College Libraries and Chemical Education", *Science*, Vol 66, p. 1229-1234, 1927

HAITUN S D, "Stationary Scientometrics Distributions", *Scientometrics*, Vol 4, part I, p. 5-25, part II, p. 89-104, part III, p. 181-194, 1982

HAON H, PAOLI C, ROSTAING H, "Perception d'un programme de R & D à travers l'analyse bibliométrique des banques de données d'origine japonaise", Actes du colloque : *L'information, intelligence de l'entreprise, IDT 93*, 22-24 juin, Paris, 1993

HAWKINS D T, "Unconventional Use of One-line Information Retrieval Systems : One-line Bibliometrics Studies", *Journal of American Society for Information Science*, Vol 28, N°1, p. 13-18, 1977

HE C, PAO M, "A Discipline-specific Journal Selection Algorithme", *Information Processing and Management*, Vol 22, p. 405-416, 1986

HEALEY P, ROTHMAN H, HOCH P K, "An Experiment in Science Mapping for Research Planning", *Research Policy*, Vol 15, p. 233-251, 1986

HUBERT J J, "A Relationship between two Forms of Bradford's Law", *Journal of the American Society for Information Science*, Vol 29, N°2, p. 159-161, 1978

HULME E W, *Statistical Bibliography in Relation to the Growth of Modern Civilization*, Edition Grafton, London, 44 p., 1923

HUNT C, ZARTARIAN V, *Le renseignement stratégique au service de l'entreprise*, Edition First, 245 p., 1990

HUOT C, QUONIAM L, DOU H, "New Method concerning Analysis of Downloaded Data for Strategic Decison", *Scientometrics*, Vol 25, N°2, p. 279-294, 1992

HUSTOPECKY J, VLACHY J, "Identifying a Set of Inequality Measures for Science Studies", *Scientometrics*, Vol 1, p. 85-98, 1978

JAKOBIAK F, *Pratique de la veille technologique*, Les Éditions d'Organisation, Paris, 232 p., 1991

JAKOBIAK F, *Les brevets source d'information*, Editions Dunod, Paris, 188 p., 1994

KATZ J S, "Geographical Proximity and Scientific Collaboration", *Scientometrics*, Vol 31, N°1, p. 31-34, 1994

KENDALL M G, "The Bibliography of Operational Research", *Operational Research Quarterly*, Vol 2, p. 31-36, 1960

KESSLER M M, "Bibliographic Coupling between Scientific Papers", *American Documentation*, Vol 14, p. 10-15, 1963

KESSLER M M, "Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing", *American Documention*, Vol 16, p. 223-233, 1965

LAFOUGE T, "Problématique de la circulation de l'information", *Documentaliste*, Vol 28, N°3, p. 132-134, 1991

LAFOUGE T, "Circulation des documents dans un système d'information documentaire", Actes du colloque : *Les systèmes d'information élaborée*, 9-11 juin, 1993

LAFOUGE T, QUONIAM L, "Les distributions bibliométriques", *Revue française de bibliométrie*, Vol 9, p. 128-138, 1992

LAINE F, *La veille technologique - De l'amateurisme au professionnalisme*, Editions Eyrolles, 138 p., 1991

LAREDO P, MUSTAR P, CALLON M, "Caractériser le profil stratégique des laboratoires de recherche : la méthode de la “rose des vents””, *Les Cahiers de l'ADEST*, Numéro spécial, p. 141-149, 1993

LAW J, BAUIN S, COURTIAL J P, WHITTAKER J, "Policy and the Mapping of Scientific Change : a Co-word Analysis of Research into Environmental Acidification", *Scientometrics*, Vol 14, p. 251-264, 1988

LAW J, WHITTAKER J, "Mapping Acidification Research : a Test of the Co-word Method", *Scientometrics*, Vol 23, N°3, p. 417-461, 1992

LAWANI S M, "Bradford's Law and the Literature of Agriculture", *International Library Review*, Vol 5, p. 341-350, 1973

LAWANI S M, "On the Heterogeneity and Classification of Author Self-citations", *Journal of the American Society for Information Science*, Vol 33, N°5, p. 281-284, 1982

LAWSON J, KORSTREWSKI B, OPPENHEIM C, "A Bibliometric Study of a new Subject Field : Energy Analysis", *Scientometrics*, Vol 2 , N°3, p 227-237, 1980

LEGENDRE L, LEGENDRE P, *Écologie numérique*, Editions Masson, Presses de l'Université du Québec, tome I : Le traitement multiple des données écologiques, tome II : La structure des données écologiques, 1984

LEIMKUHLER, "The Bradford Distribution", *Journal of Documentation*, Vol 23, N°3, p 197-207, 1967

LESCA H, *Information et adaptation de l'entreprise*, Edition Masson, 220 p., 1988

LEYDESDORFF L, "The Development of Frames of Reference", *Scientometrics*, Vol 9, N° 3-4, p. 103-125, 1986

LEYDESDORFF L, "Various Methods for Mapping of Science", *Scientometrics*, Vol 11, N°5-6, p. 295-324, 1987a

LEYDERSDORFF L, "Co-words and Citations Relations between Document Sets and Environments", *First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*, August 24-28, Diepenbeek, Belgium, 1987b

LEYDESDORFF L, "A Validation Study of Leximappe", et "A Reply to Courtial's Comments", *Scientometrics*, Vol 25, N°2, p. 295-312 et p 317-319, 1992

LOTKA A J, "The Frequency Distribution of Scientific Productivity", *Journal of the Washington Academy of Sciences*, Vol 16, N° 12, p. 317-323, 1926

MCCAIN K W, TURNER W A, "Citation Context Analysis in Genetics", *Scientometrics*, Vol 17, 1989

MCCREERY L S, PAO M L, "Bibliometric Analysis of Ethnomusicology", Actes du colloque : *The American Society for Information Science (ASIS) 47th Annual Meeting*, NY : Knowledge Industry Publications, 21-25 oct, p. 212-216, 1984

MCROBERTS M, MCROBERTS B, "Problems of Citation Analysis : a Critical Review", *Journal of the American Society for Information Science*, Vol 40, N°5, p. 342-349, 1989

MAIA M F, MAIA M D, "On the Unity of Bradford's Law", *Journal of Documentation*, Vol 40, N°3, p. 206-216, 1984

MANDELBROT B, "An Information Theory of the Statistical Structure of Language", Actes du colloque : *Symposium on Applications of Communication Theory*, Butterworth, London, p. 486-500, 1953

MARTIN W A, "Methods for Evaluating the Number of Relevant Documents in a Collection", *Journal of Information Science*, Vol 6, N°5, p. 173-177, 1983

MARTINET B, RIBAULT J-M, LEBIDOIS D, *Le management des technologies*, Les Éditions d'Organisation, Paris, 309 p., 1991

MICHELET B, *L'analyse des associations*, Thèse : Université de Paris VII, 26 oct, 1988

MIQUEL J F, OKUBO Y, "Structure of International Collaboration in Science - Part II: Comparisons of Profiles in Countries using a LINK Indicator", *Scientometrics*, Vol 29, N°2, 1994

MOED H, BURGER W J M, FRANKFORT J G, VAN RAAN A F J, "The Application of Bibliometrics Indicators : Important Field and Time Dependant Factors to be Considered", *Scientometrics*, Vol 8, N°3-4, p. 177-204, 1985

MOED H F, DE BRUIN R E, NEDERHOF A J, TIJSSEN R J W, "International Scientific Co-operation and Awareness within the European Community : Problems and Perspectives", *Scientometrics*, Vol 21, N°3, p. 291-311, 1991

MOED H F, VAN RAAN A F J, "Observations and Hypotheses on the Phenomenon of Multiple Citation to a Research Group's Oeuvre", *Scientometrics*, Vol 10, N°1-2, p. 17-34, 1986

MOGEE M E, "Patent Analysis for Strategic Advantage : using International Patents Records", *Competitive Intelligence Review*, Vol 5, N°1, p. 27-35, 1994

MORIN J, *L'excellence technologique*, Edition Publi Union, Paris, 253 p., 1985

MOUREAU M, GIRARD A, "Utilisation des banques de données sur les brevets", *Revue française de bibliométrie*, Vol 1, N°2, p. 9-20, 1987a

MOUREAU M, GIRARD A, "Patents and Statistical Analysis; a User View", *Derwent Online News*, N°4, p. 5-8, 1987b

MOUREAU M, GIRARD A, BUFFETEAU A, "Données bibliométriques à l'Institut Français du Pétrole", dans : DESVALS H, DOU H (éds), *La veille technologique*, Editions Dunod, p. 106-118, 1992

MURPHY L J, "Lotka's Law in the Humanities?", *Journal of the American Society for Information Science*, Vol 24, p. 461-462, 1973

NADEL E, "Commitment and Co-citation - an Indicator of Incommensurabiblity in Patterns of Formal Communication", *Social Studies of Science*, Vol 13, N°2, p. 255-283, 1983

NAGPUL P S, LALITA SHARMA, "Research Output and Transnational Cooperation in Physics Subfields : a Multidimensional Analysis", *Scientometrics*, Vol 31, N°1, p. 97-122, 1994

NIVOL W, *Système de surveillance systématique pour le management stratégique de l'entreprise*, Thèse : Université Aix-Marseille III, 10 mai, 333p., 1993

NIYAMOTO S, NAKAYAMA K, "A Technique of Two-stage Clustering Applied to Environmental and Civil Engineering and Related Methods of Citation Analysis", *Journal of the Society for Information Science*, Vol 34, p. 192-201, 1983

OBERSKI J E L**,** "Some Statistical Aspects of Co-citation Cluster Analysis and a Judgement by Physicists", dans : VAN RAAN (ed), *Handbook of Quantitative Studies of Science and Technology*, Editions Elsevier Science Publishers B.V., North-Holland, p. 431-462, 1988

OKUBO Y, MIQUEL J F, FRIGOLETTO L, DORE J C, "Structure of the International Collaboration in Science Typology of Countries through Multivariate Techniques using a Link Indicator", *Scientometrics*, Vol 25, N°2, p. 321-351, 1992

OST, *Science et technologie, indicateurs 1992*, Editions Economica, 1992

PAISLEY W, "The Convergence of Communication and Information Science", dans : Eldeman, Hendrik, (ed), *Librairies and Information Science in the Electronic Age*, Philadephia, PA : ISI Press, p. 122-153, 1986

PAOLI C, LAVILLE F, LONGEVIALLE C, "La station d'analyse bibliométrique ATLAS", *Les Cahiers de L'ADEST*, Numéro Spécial, p. 83-87, 1993

PARKER-RHODES A F, JOYCE T, "A Theory of Word-frequency Distribution", *Nature*, Vol 178, p. 1308, 1956

PENAN H, "Analyse des citations : application à la théorie microéconomique", dans : DESVALS H, DOU H (éds), *La veille technologique*, Editions Dunod, p. 313-330, 1992

PERITZ B C, "Are Methodological Papers more Cited than Theoretical or Empirical ones ? The Case of Sociology", *Scientometrics*, Vol 5, N°4, p. 211-218, 1983

PETERS H P J, VAN RAAN A F J, "Structuring Scientific Activities by Co-author Analysis. An Exercice on a University Faculty Level", *Scientometrics*, Vol 20, N°1, p. 235-255, 1991

POLANCO X, GRIVEL L, "Mapping knowledge", Actes du colloque : *Fourth international conference on bibliometrics, informetrics and scientometrics*, 11-15 sept., Berlin, Allemagne, 1993

POLANCO X, GRIVEL L, FRANÇOIS C, BESAGNI D, "L'infométrie, un programme de recherche", *Revue française de bibliométrie*, Vol 12, p. 20-28, 1993

PONTIGO J, LANCASTER F W, "Qualitative Aspects of the Bradford Distribution", *Scientometrics*, Vol 9, N°1-2, p. 59-70, 1986

PRATT A D, "A Measure of Class Concentration on Bibliometrics", *Journal of the American Society for Information Science*, Vol 28, p. 285-292, 1977

PRAVDIC N, OLUIC-VUKOVIC V, "Distribution of Scientific Productivity : Ambiguities in the Assignement of Author Rank", *Scientometrics*, Vol 20, N°1, p. 131-144, 1991

PRICE D, *Little Science, big Science*, Columbia, New York, 118 p., 1963

PRICE D, "A General Theory of Bibliometrics and other Cumulative Advantage Processes", *Journal of the American Society for Information Science*, Vol 27, p. 292-306, 1976

PRICE D, "The Analysis of Scientometric Matrices for Policy Implications", *Scientometrics*, Vol 3, N°1, p. 47-54, 1981

PRICE D, BEAVER D, "Collaboration in an Invisible College", *American Psychologist*, Vol 21, p. 1011-1018, 1966

PRITCHARD A, "Statistical Bibliography or Bibliometrics ?", *Journal of Publication*, Vol 25, p. 348-349, 1969

QUONIAM L, "Bibliométrie sur des références bibliographiques : méthodologie", dans : DESVALS H, DOU H (éds), *La veille technologique*, Edition Dunod, p. 244-262, 1992

QUONIAM L, DOU H, HASSANALY P, MILLE G "Bibliométrie et chimie. Exemple sur les acides gras phospholipides", *Analusis*, Vol 19, N°1,p 48-52, 1991

QUONIAM L, HASSANALY P, BALDIT P, ROSTAING H, DOU H, "Bibliometric Analysis of Patent Documents for R&D Management", *Research Evaluation*, Vol 3, N°1, p. 13-18, 1993

RADHAKRISHNAN T, KERNIZAN R, "Lotka's Law and Computer Science Literature", *Journal of the American Society for Information Science*, Vol 30, p. 51-54, 1979

RAISING L M, "Statistical Bibliography in the Health Science", *Bulletin of Medical Library Association*, Vol 50, N°3, p. 450-461, 1962

REMY D, VERGNES G, MOSSETTI M, "Analyse bibliométrique appliquée à la recherche fondamentale : une expérience sur 10 ans", *Revue française de bibliométrie*, Vol 9, p. 227-237, 1991

RICE R, BORGMAN C L, BEDNARSKI D, HART P J, "Journal-to-journal Citation Data : Issues of Validity and Reliability", *Scientometrics*, Vol 15, N°3-4, p. 257-282, 1989

ROSTAING H, *Veille technologique et bibliométrie : concepts, outils, applications*, Thèse : Aix-Marseille III, 13 janv, 353 p., 1993

ROSTAING H, NIVOL W, QUONIAM L, BÉDÉCARRAX C, HUOT C, "L'exploitation systématique des bases de données : des analyses stratégiques pour l'entreprise", *Les Cahiers de l'ADEST*, Numéro spécial, p. 7-22, 1993a

ROSTAING H, NIVOL W, QUONIAM L, LA TELA A, "Le logiciel DATAVIEW et son application comme outil d'aide à l'évaluation de la concurrence", *Revue française de bibliométrie*, Vol 12, p. 360-388, 1993b

SCHUBERT A, BRAUN T, "Relative Indicators and Relational Charts for Comparative Assessement of Publication Output and Citation Impact", *Scientometrics*, Vol 9, N°5-6, p. 281-291, 1986

SIGONEAU M, BAUIN S, COURTIAL J-P, TURNER W A, "Etude d'un front de recherche identifié par les co-citations à partir de la base *PASCAL*", *Les Cahiers de l'ADEST*, 1990

SMALL H G, "Co-citation in the Scientific Literature : a new Measure of the Relationship between two Documents", *Journal of the American Society for Information Science*, Vol 24, N°4, p. 265-269, 1973

SMALL H G, GRIFFITH B C, "The Structure of Scientific Literature I **:** Identifying and Graphing Specialities", *Science Studies*, Vol 4, N°1, p. 17-40, 1974

SMALL H G, SWEENEY E, GREENLEE E, "Clustering the Science Citation Index using Co-citations. II. Mapping Science", *Scientometrics*, Vol 8, N°5-6, p. 321-340, 1985

SOMMIER J-L, "La propriété industrielle, outil de management pour la stratégie de l'entreprise", dans : DESVALS H, DOU H (éds), *La veille technologique*, Editions Dunod, 436 p., 1992

STACEY G, "Méthodes pratiques de mesure des performances dans la gestion de la technologie", Actes du colloque : *Les journées de l'ADEST*, 1-2 juin, p. 99-110, 1992

STEVENS K, NARIN F, *National citation indicators based on citing year : the citation time anomaly*, CHI, Haddon Heights, New Jersey, 1989

SUBRAMANYAM K, "Lotka's Law and the Literature of Computer Science", *IEEE Transactions of Professional Communications*, Vol 22, p. 187-189, 1979

SURAUD M-G, QUONIAM L, ROSTAING H, DOU H, "Analyse bibliométrique comme outil d'aide à la mise en place d'un groupe de recherche en physique fondamentale", *Revue française de bibliométrie*, Vol 13, p. 100-120, 1994

SURAUD M-G, QUONIAM L, ROSTAING H, DOU H, "On the Significance of Data Bases Keywords for a Large Scale Bibliometric Investigation in Fundamental Physics", *Scientometrics*, Vol 33, N°1, p 41-63, 1995

TAGUE J, "The Law of Exponential Growth : Evidence, Implications and Forecasts", *Library Trends, Bibliometrics*, Vol 30, 1981

TODOROV R, "Displaying Content of Scientific Journal : a Co-heading analysis", *Scientometrics*, Vol 23, N°2, p. 317-334, 1992

TODOROV R, GLÄNZEL W, "Journal Citation Measures : a Concise Review", *North-Holland Information & Business*, 1987

TODOROV R, WINTERHAGER M, "Mapping Australian Geographic : a Co-heading Analysis", *Scientometrics*, Vol 19, N°1-2, p. 35-56, 1990

TODOROV R, WINTERHAGER M, "An Overview of Mike Moravcsik's Publication Activity in Physics", *Scientometrics*, Vol 20, N°1, p. 163-172, 1991

TURNER W A, "De la bibliométrie à l'infométrie : des axes de recherche nouveaux pour la veille scientifique et technologique", *Revue française de bibliométrie*, Vol 6, p. 161-179, 1990

VAN RAAN A F J (ed), *Handbook of Quantitative Studies of Science and Technology*, Editions Elsevier, 774 p., 1988

VAN RAAN A F J, PETERS H P F, "Dynamic of a Scientific Field Analysed by Co-subfield Structures", *Scientometrics*, Vol 15, N°5-6, p. 607-620, 1989

VICKERY B C, "Bradford's Law of Scattering", *Journal of Documentation*, Vol 4, N°3, p. 198-203, 1948

VILLAIN J, *L'entreprise aux aguets*, Editions Masson, Collection Le nouvel ordre économique, Paris, 192 p., 1990

VINKLER P, "An Attempt of Surveying and Classifying Bibliometric Indicators for Scientometric Purposes", *Scientometrics*, Vol 13, N°5-6, p. 239-259, 1988

VLACHY J, "Citation Histories of Scientific Publications. The Data Sources", *Scientometrics*, Vol 7, N°3-6, 1985

VOOHS H, "Lotka and Information Science", *Journal of the American Society for Information Science*, Vol 25, p. 270-272, 1974

WALLACE D P, "The Relationship between Journal Productivity and Obsolescence", *Journal of the American Society for Information Science*, Vol 37, N°3, p. 136-145, 1986

WARMESSON I, PARISOT P, BÉDÉCARRAX C, HUOT C, "Traitements linguistiques et analyse des données pour une exploitation systématique des banques de données", *Revue française de bibliométrie*, Vol 12, p. 281-294, 1993

WEAVER W, SHANNON C E, *Théorie mathématique de la communication*, Editions Bibliothèque du CEPL, 188 p., 1975

WHITE H D, "Bradfordizing Search Output - How it would help Online Users", *Online Review*, Vol 5, N°1, p. 47-54, 1981a

WHITE H D, GRIFFITH B C, "Author Cocitation : a Literature Measure of Intellectual Structure", *Journal of the American Society for Information Science*, Vol 32, N°3, p. 163-172, 1981b

WHITE H D, GRIFFITH B C, "Quality of Indexing in Online Data Bases", *Information Processing & Management*, Vol 23, N°3, p. 211-224, 1987

WHITE H D, MCCAIN K W, "Bibliometrics", *Annual Review of Information Science and Technology (ARIST)*, Vol. 24, p. 119-186, 1989

WILKINSON E A, "The Ambiguity of Bradford's Law", *Journal of Documentation*, Vol 28, p122-130, 1972

ZIPF G K, *Human Behaviour and the Principle of Least Effort*, Editions Addison Wesley, 257 p.,1949

ZITT M, BASSECOULARD E, "Development of a Method for Detection and Trend Analysis of Research Fronts built by Lexical or Cocitation Analysis", *Fourth International Conference on Bibliometrics, Informetrics and Scientometrics*, Berlin, Allemagne, 11-15 Sept, 1993