Veille Technologique et Bibliométrie : concepts, outils, applications

Hervé Rostaing

UNIVERSITE DE DROIT ET DES SCIENCES D'AIX-MARSEILLE

Faculté des Sciences et Techniques de Saint Jérôme

_____

Veille Technologique et Bibliométrie : Concepts, Outils, Applications.

THESE

Présentée et soutenue publiquement par:

ROSTAING Hervé

le 13 Janvier 1993

pour obtenir le grade de
Docteur en Sciences

Spécialité:
Sciences de l'Information et de la Communication

Président du Jury:	Rouault	(Professeur Grenoble)
Autres membres du Jury:	Carpentier	(Professeur CELSA)
	Courtial	(Professeur Nantes)
	Dou	(Professeur CRRM)
	Jakobiak	(Responsable VT à ATOCHEM)
	Quoniam	(Maître de Conférence CRRM)

Résumé en français:

Dans la compétition économique actuelle l'information est une composante essentielle de la réussite. Il est vital pour l'entreprise d'être constamment informée sur son environnement. Pour qu'une société soit compétitive elle doit faire preuve d'une forte activité innovatrice. Ce mémoire expose la réflexion menée pour concevoir et développer un logiciel d'aide à l'élaboration des stratégies des programmes en Recherche et Développement. Le premier volet traite du besoin en information scientifique et technique des entreprises, les structures et les principes à mettre en place pour une gestion efficace de cette information. Le système de Veille Technologique préconisé doit permettre l'élaboration d'indicateurs de tendances qui aideront les décideurs à diagnostiquer l'état des activités scientifiques et techniques. L'outil informatique proposé repose sur une technique de traitement automatique de l'information publiée: la bibliométrie. Après un état de l'art des méthodes bibliométriques et une analyse critique de leurs applications à la Veille Technologique, on définit un nouvel outil plus adapté à l'élaboration d'indicateurs en Veille Technologique. Ce logiciel et ses multiples traitements sont présentés ainsi que des cas pratiques d'exploitations de bases de données accessibles en ligne. La performance de cet outil repose sur ses capacités à intégrer la diversité des sources d'information, à manipuler avec souplesse les données textuelles et à livrer de multiples résultats statistiques en bibliométrie.

Résumé en anglais:

Nowadays, in the current economic competition, information is the key of success. A company needs to be continuously well informed about its environment in order to have a high innovating activity and to stay competitive. This thesis explains the thinking process which is followed in order to design and develop a software to help the elaboration of Research & Development programs. First we show industry needs in scientific and technical information. Then, we study structures and methods which must be apply to manage this information in the best possible way. Technology monitoring must allow to elaborate tendancy indicators to help decision makers to diagnose the standing of scientific and technical activities. The suggested software tool is based on an automatic treatment of published information: the bibliometrics. After the review of the various bibliometrics methods and a critical analysis of their applications in strategic management of technology, we define a new method and a new tool which fits with the creation of indicators in technological management. The performance of this software results from its capacity to use various data formats, to treat these data easily and to give multiple bibliometric statistical results.

Je tiens à remercier en premier mon laboratoire d'accueil, le Centre de Recherche Rétrospective de Marseille, et plus particulièrement Monsieur le **Professeur Henri Dou** et Madame **Parina Hassanaly** qui ont su m'accorder toute leur confiance ainsi que leurs précieux enseignements.

Je tiens aussi à exprimer toute ma reconnaissance à **Luc Quoniam** sans qui mon travail ne serait ce qu'il est. J'ai trouvé en lui une compétence et un soutien de tous les jours. Je l'en remercie.

J'ai le plaisir de remercier **Albert La Tela** qui m'a apporté son expérience en informatique ainsi qu'une collaboration de tous les instants.

Je remercie **les membres du jury** et **les rapporteurs** de ma thèse de bien avoir voulu consacrer leur temps à la lecture de mon mémoire.

Je dois aussi remercier mes premières lectrices, pour leur aide méticuleuse: ma mère **Yvette** et ma compagne **Magali**. Je remercie aussi les **membres de ma famille** qui m'ont toujours accompagné durant toutes ces longues études.

Je ne pourrais terminer sans remercier **Cynthia** chez qui j'ai retrouvé la chaleur d'un foyer lorsque j'étais éloigné du mien et **William** dont la complicité dans notre travail en commun à toujours été d'une grande stimulation.

Sommaire

Sommaire

Introduction

I. Introduction

Le monde de la science et de la technique se transforme rapidement et profondément, allant toujours vers plus de complexité. Les spécialités se multiplient, et les frontières qui les délimitent sont de plus en plus floues. La prise de connaissance de l'existence de ces domaines et la compréhension de leurs activités sont de plus en plus difficiles à maîtriser.

Dans ce décor changeant, l'**exploitation de méthodes, de structures et d'outils est devenue indispensable** pour mieux appréhender cette complexité. Les entreprises sont tout particulièrement concernées par ces rapides changements. Pour qu'une entreprise soit compétitive elle doit être constamment **informée des dernières découvertes, inventions ou innovations**. Elle doit pour cela s'imposer une constante observation des mutations scientifiques, techniques et technologiques.

Les entreprises, sous cet incessant besoin d'information qui les environne, mettent en place de nouvelles structures spécialisées dans la gestion de la collecte et du traitement de l'information venant de l'extérieur. Cette nouvelle activité nécessite l'application de méthodes adaptées. Des réflexions ont déjà été menées pour échafauder des méthodes de surveillance performantes et adaptées au monde industriel. Par contre, l'absence d'outils d'aide dans cette activité est un obstacle au bon fonctionnement de ces méthodes.

L'étude exposée dans ce mémoire a précisément pour objectif la conception et le développement d'un outil informatique qui répond aux exigences de cette activité de surveillance. Il est conçu pour offrir une aide considérable dans l'élaboration des stratégies des programmes de recherche et de développement.

Le premier chapitre rappelle la problématique dont souffrent les entreprises en ce qui concerne la gestion de l'information scientifique et technique. Il aborde successivement le besoin en information des entreprises, le concept de la gestion du flux d'information extérieur à l'entreprise et les systèmes de surveillance industrielle.
Puis nous déterminons l'activité de surveillance à laquelle ce mémoire se restreint: la **veille technologique**. La finalité de la veille technologique dans le processus d'innovation de l'entreprise est rappelée. On insiste tout particulièrement sur les destinataires des renseignements fournis par un service de veille technologique. Les principes de

fonctionnement et les actions à mener par ce service sont ensuite abordés. Deux principales fonctions sont préconisées dans ce mémoire:

- **le renseignement en continu**
- **l'élaboration d'indicateurs.**

La première fonction a déjà été étudiée par des auteurs; elle n'est que brièvement évoquée. Par contre, les modalités de réalisation de la seconde fonction sont approfondies. Une technique de traitement automatique de l'information scientifique et technique est proposée. Cette technique, basée sur le principe de la **bibliométrie,** permet de dégager des indicateurs de tendances générales concernant les activités de recherches scientifiques et techniques. La parfaite application de cette technique dans le processus de veille technologique est finalement discutée. Plusieurs raisons sont avancées pour justifier cet emploi.
Cette première partie nous permet de définir le cadre de l'étude menée lors de cette thèse ainsi que les raisons de la conception d'un outil informatique bibliométrique.

Avant d'exposer les traitements bibliométriques, un chapitre rappelle les données de base auxquelles ces traitements font appels. **Ce second chapitre fait une rapide présentation des sources d'information exploitées en bibliométrie: les bases de données accessibles par communication informatique**. La fonction de ces bases de données informatisées ainsi que les divers acteurs intervenant dans cette chaîne de diffusion d'information sont abordés. Ensuite, l'énumération des différents types de bases d'information intéressant la veille technologique est réalisée. Finalement ce chapitre se termine par une catégorie de renseignements, fournis par ces bases de données, qui est plus particulièrement utile aux traitements bibliométriques: l'ensemble des "champs indexés". **Le lecteur est mis en garde sur le fait qu'il faut une parfaite connaissance de la création de ces données lors de l'interprétation des résultats statistiques obtenus par les méthodes bibliométriques.**

Le troisième chapitre est consacré à la bibliométrie. Il présente les différentes méthodes de traitement et les grandes écoles de pensée en bibliométrie. **Ceci permet d'établir les fondements communs à toutes ces méthodes**.
En premier lieu, les postulats de départ de la bibliométrie sont énoncés. Une définition qui permet de couvrir la diversité des pratiques bibliométriques est ensuite discutée. Un bref historique est décrit, où les principaux centres français de recherche ayant une activité bibliométrique sont répertoriés.

Puis, les différentes méthodes bibliométriques ont été distribuées en deux parties distinctes selon qu'elles s'appliquent aux domaines de la science ou aux domaines de la technique et de la technologie.
La **première partie consacrée à la mesure de la science** aborde:

- les lois et les modèles mathématiques:
  courbes logistiques, loi de Bradford, loi de Lotka, loi de Zipf-Mandelbrot, théorie de la communication
- les indicateurs univariés:
  dénombrement des publications et des citations, indicateurs concernant les journaux, les chercheurs, les laboratoires, les pays ou un domaine
- les cartes relationnelles:
  méthodes des co-citations, des citations-croisées, des cooccurrences de mots et autres méthodes d'analyse de relations

La seconde **partie consacrée à la mesure des techniques et des technologies** traite:

- tout d'abord de la remise en cause de la validité des postulats bibliométriques pour l'information technique et technologique
- des indicateurs univariés:
  comptages statistiques simples et indicateurs proposés par des sociétés consultantes en stratégie technologique
- des cartes relationnelles:
  rares exemples appliquant ces techniques aux informations brevets.

Ce troisième chapitre nous permet de faire l'état de l'existant en ce qui concerne les méthodes et les outils développés en bibliométrie. Il met en évidence deux points fondamentaux:

- **l'inadéquation de ces méthodes aux exigences de la veille technologique**
- **une absence pratiquement totale d'outils informatiques bibliométriques.**

Ces deux points sont repris dans le quatrième chapitre **pour étayer la réflexion menée lors de la conception de l'outil bibliométrique développé pendant cette thèse**. Pour que cet outil soit parfaitement adapté à l'élaboration d'indicateurs stratégiques dans le monde industriel il doit remplir trois principaux critères:

- accepter la diversité des sources d'information

- ❑ accepter la diversité des éléments bibliométriques traités
- ❑ accepter la diversité des méthodes mathématiques appliquées en bibliométrie

Le logiciel DATAVIEW, développé pendant cette thèse, a été conçu dans le respect de ces trois critères. **Ce chapitre présente les concepts bibliométriques développés pour cet outil informatique et justifie leur emploi :** notions de forme, d'occurrence de forme, de fréquence de forme, de fréquence de paire et d'indice d'association.
Ce logiciel est présenté comme une plate-forme de traitements s'insérant dans la chaîne des opérations à mener lors d'un traitement bibliométrique. Cette chaîne de traitement est décomposée pour être expliquée étape par étape. **Une partie des fonctions de DATAVIEW est explicitée** par l'intermédiaire d'exemples de traitements effectués sur un échantillon de données. Le panel des résultats bibliométriques créé par ce logiciel est présenté ainsi que des exemples de représentations infographiques. L'aspect infographique des données bibliométriques est important car il offre les avantages suivants:

- ❑ caractère synthétique des résultats
- ❑ aide à l'interprétation des résultats
- ❑ présentation adaptée à la prise de décision.

L'exposé de ce mémoire se termine par un chapitre contenant deux cas pratiques d'études bibliométriques réalisées à l'aide de l'outil bibliométrique développé dans le cadre de cette thèse.

❑ Le premier cas met l'accent sur la nécessité de connaître parfaitement les méthodes statistiques et leur mise en application lors de traitements bibliométriques pour obtenir une interprétation exacte des résultats. D'où le besoin d'offrir des formations spécifiques au nouveaux métiers que sont ceux de la Veille Technologique et de la gestion des informations scientifiques et techniques.

❑ La seconde étude menée en collaboration avec le CESMAP d'IBM montre l'intérêt qu'apporte l'analyse simultanée de plusieurs catégories de données bibliographiques. Une méthode d'analyse bibliométrique a été élaborée pour permettre une meilleure retranscription du contenu des innovations présentes dans les brevets. L'emploi de l'analyse relationnelle pour une étude simultanée de trois champs brevets codifiés a permis non seulement de dégager les grandes tendances pour le thème étudié, mais auusi de focaliser l'attention sur certaines innovations comportant des caractères marginaux par rapport à l'ensemble des brevets.

Pour conclure ce mémoire, nous replacerons le travail réalisé pendant cette thèse dans deux contextes. La contribution de ce travail peut s'interpréter: primo en tant que développement d'un outil d'aide dans le processus de Veille Technologique; secundo, cet outil offre aussi une plate-forme de recherche idéale pour concevoir de nouveaux traitements de données spécifiques à l'information scientifique et technique.

L'information demeure le seul bien dont la gestion est des plus problématiques car son flux est grandissant. La création d'un outil d'aide à la maîtrise de cette information doit être d'un grand soutien dans ce défi quotidien.

La veille technologique et le besoin d'information élaborée en entreprise

II. La veille technologique et le besoin d'information élaborée en entreprise

A. L'incontournable besoin en information pour les entreprises

Dans la compétition économique actuelle l'information est une composante essentielle de la réussite.

Notre société fait état de nombreux paradoxes. Sous l'évolution galopante des techniques de communication, l'information est sujette à l'un d'eux. Nous sommes noyés sous un flot d'informations diffusées par de multiples vecteurs de communication (presse, radio, télévision, publicité, affichage...) de plus en plus performants. Mais paradoxalement, une grande partie de cette information ne présente aucun intérêt. Alors que l'information devrait être source de renseignements, de connaissances donc possédant la vertu d'être formatrice, bien trop souvent elle est en majeure partie proprement stérile.

L'entreprise vit au même titre ce phénomène. Elle est soumise à un flux continu d'information *fatale* selon la dénomination donnée par Jakobiak dans [JAKO91]. Il définit sous cette appellation, à connotation fataliste, la masse d'information que tout professionnel reçoit quotidiennement et qui n'est pas de nature utile pour le travail.
Par contre, les décideurs dans les entreprises ont besoin, pour étayer leurs stratégies, d'une information ciblée indispensable pour agir et décider. François Jakobiak, pour spécifier le caractère indispensable de cette information, l'a nommée information *critique*. Dans ce contexte, la mise en place de systèmes de gestion d'information qui auraient pour objectif de trouver et diffuser l'information *critique* à l'entreprise paraît crucial.

Depuis les années 80, ce besoin d'information de qualité pour le bon fonctionnement d'une entreprise se fait ressentir plus précisément. Le monde politique et économique français commence à s'en préoccuper de façon sérieuse. Des déclarations de personnalités dirigeantes, comme "*L'information est la clé de l'élaboration des stratégies*" (Jacques Delors), se font de plus en plus nombreuses. Mais plus que des déclarations, des actions de sensibilisation auprès des structures les plus concernées, c'est à dire les entreprises, ont été lancées par le gouvernement. La commission "Europe technologique, industrielle et commerciale" présidée par Antoine Riboux, pour le Commissariat général du X^{ème} Plan, a constitué un groupe "Veille technologique & politique de brevets". Les principales recommandations de la commission étaient la sensibilisation, la mobilisation, l'incitation et la formation pour un développement prometteur de la veille technologique [COMM89].

A la suite de ce rapport, de nombreux programmes d'études et de recherches ont été menés pour améliorer la connaissance et les processus concernant la maîtrise de l'information stratégique pour le monde industriel. Une ferme volonté d'encourager l'insertion de ce concept dans l'enseignement des futurs cadres a fait émerger des formations spécifiques à ce nouveau métier et insérer dans les programmes des écoles des modules concernant le management des technologies.

La grande presse n'a ressenti qu'assez récemment l'attrait des industries pour ce nouveau centre d'intérêt et le public a eu droit à une profusion d'articles dont voilà quelques titres:

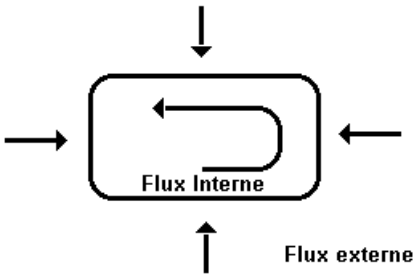
- *L'éveil des entreprises à la veille technologique,*	L'industrie,	22 Fév 89
- *Le veilleur, un espion en col blanc,*	Libération,	5 Juin 90
- *Profession: veilleur technologique,*	Science et technologie,	Juil-Août 90
- *La chasse à l'information est ouverte,*	Le monde,	26 Sep 90
- *Entreprise: la guerre de l'information,*	Le nouvel observateur,	21 Nov 90
- *Les entreprises se mettent en état de veille,*	L'usine nouvelle,	29 Nov 90
- *L'indispensable vigie,*	Le monde informatique,	28 Jan 91
- *La veille dans tous ses états,*	01 informatique,	17 Mai 91

Tout laisse croire que l'instauration de systèmes spécialisés dans la surveillance de l'information environnant le monde industriel, à l'instar des japonais et des américains, est devenue incontournable. Pour certains économistes, l'information est désormais le troisième facteur de production au même titre que la main d'oeuvre et le capital.

B. Le concept de veille industrielle

1. La notion de flux d'information

Considérons l'entreprise comme un système à part entière; étudions le flux d'information pour ce système. Ce système est soumis à deux principaux flux d'information: l'information produite par le système et l'information reçue de l'extérieur par le système.



Les structures de l'entreprise sont parfaitement bien organisées pour la gestion du flux d'information interne. Mais cette gestion opérationnelle indispensable au bon fonctionnement de l'entreprise livre une information de faible intérêt pour diagnostiquer la position de l'entreprise par rapport à son environnement extérieur. Cette information peut donner l'alerte sur un dysfonctionnement interne mais elle ne permettra certainement pas de positionner l'entreprise par rapport à sa concurrence [LESC86].

C'est donc le flux d'information extérieur à l'entreprise qu'il faut savoir canaliser. La gestion de l'information environnant l'entreprise est le seul moyen pour le décideur d'estimer les évolutions des marchés, des produits, des technologies, de l'économie, des tensions sociales... etc...

Les événements extérieurs modèlent l'avenir de toute institution il faut donc être capable de détecter les modifications du monde extérieur.

2. La gestion des flux d'information: la veille industrielle

Le terme *veille* est souvent employé pour désigner cette activité de surveillance de l'environnement des entreprises. Sous un effet de mode, une récente prolifération d'adjectifs est venue qualifier le terme de *veille*. On trouve ainsi les appellations *veille industrielle, veille globale, veille environnementale, veille stratégique, veille informative, veille technologique, veille concurrentielle, veille commerciale, veille d'acquisition, veille des ressources humaines*...Cette terminologie est employée selon des pratiques assez confuses. Toutes ces veilles ne sont pas censées couvrir les mêmes activités.

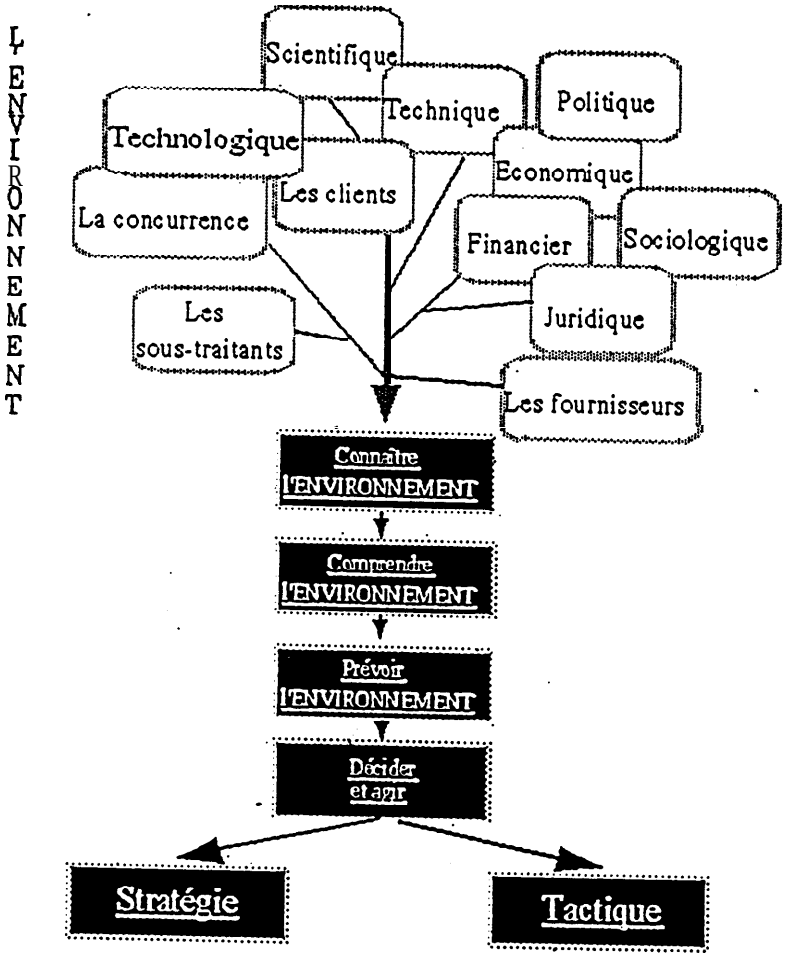
Nous allons, dans ce mémoire, uniquement nous concentrer sur l'activité reconnue sous le nom de *veille technologique*. Nous allons tout d'abord la replacer par rapport à ce que l'on a appelé la *veille industrielle*.

On peut définir la veille générale comme l'ensemble des activités de surveillance de l'environnement d'une entreprise pour fournir des données utiles à la définition de ses stratégies d'évolution. Cette surveillance récolte donc des informations de natures très variées: économique , financière, commerciale, scientifique, technique, technologique, sociologique, politique, juridique, les clients, les sous-traitants, les fournisseurs... Nous avons préféré l'appellation *veille industrielle* (utilisée par Martinet et Ribault dans [MART89]) à ses synonymes *veille globale, veille environnementale, veille informative, veille stratégique* parce qu'elle paraît mieux retranscrire la réunion de l'ensemble de ces activités et surtout parce qu'elle précise à qui elle s'adresse.
Une présentation schématique de cette veille industrielle est donnée par Villain dans son ouvrage [VILA89] selon la figure 1. Cette représentation arborescente est dénommée le *Bonsaï de la surveillance et de l'intelligence de l'environnement* en raison de l'origine japonaise de cette activité.

Il faut tout d'abord remarquer que parmi la liste, non-exhaustive, des divers secteurs de surveillance de l'environnement seuls ceux concernant l'économie, la finance et le commercial sont depuis longtemps pris en considération par les industriels. Nous observons depuis peu un intérêt grandissant pour l'information technologique. Le monde industriel prend conscience de l'importance de la dimension technologique dans l'élaboration des stratégies industrielles [MORI85].

Il n'existe pas de séparation nette entre ces différents secteurs de surveillance. Une même information peut avoir un caractère critique dans plusieurs secteurs de veilles.

Le "Bonzai" de la surveillance et de l' "intelligence" de l'ENVIRONNEMENT



source: "l'Entreprise aux aguets" J.Villain ed.MASSON

Figure 1

C. La veille technologique: la gestion des flux d'information scientifique, technique et technologique

Avant d'exposer la nature de l'activité de la veille technologique, nous allons donner une définition de ce que l'on entend par *technologie*. Je reprendrai simplement l'énoncé qu'en a fait F Lainé dans [LAIN91]:

> *"La technologie, c'est l'ensemble des connaissances scientifiques et des savoir-faire applicables aux arts industriels."*

On peut la différencier de la technique par le simple fait qu'elle rentre dans un processus industriel, c'est à dire dans un processus de transformation d'un produit de façon à lui donner de la valeur ajoutée.

1. La finalité: aide à l'innovation

Une fois cette définition posée, on est en droit de se demander ce que sont les arts industriels. Les arts industriels concernent tout ce qui touche le métier de l'entreprise, c'est à dire les connaissances permettant la maîtrise de la production. **Une entreprise reste productive non seulement si elle sait contrôler les technologies de production mais surtout si elle sait conserver son avance technologique vis à vis de ses concurrents. Pour rester compétitive face à la concurrence, l'entreprise n'a qu'une solution: *innover*(1).**

La réussite exige aujourd'hui de chaque entreprise qu'elle soit innovatrice pour créer des produits nouveaux. Pour tous les grands stratèges industriels, cette exigence dans leur discours est un dénominateur commun. L'un d'eux, Kami [KAMI89], donne trois raisons à cela:

- une technologie qui évolue très vite:
  il faut impérativement ne pas se faire dépasser par de nouvelles technologies plus performantes et mieux adaptées

- une saturation du marché plus rapide:
  la meilleure distribution de masse et la communication plus large accélèrent l'offre de produits au marché tout entier

- une concurrence accélérée:
  la copie des produits est de plus en plus rapide.

---

(1) L'innovation selon Mensch est le matériau ou la technique, en production ou en utilisation régulière, aboutissant pour la première fois à un marché organisé [MENSC88]. Innover est donc l'introduction dans le circuit économique d'une invention ou d'une découverte.

Comment aider l'innovation dans son entreprise? Voilà la réelle question qu'il faut se poser après un tel constat.

Il va d'abord falloir être original. Kami estime que pour devenir original quatre solutions sont envisageables pour les entreprises:

- ☞ rechercher scientifiquement
- ☞ adapter
- ☞ acquérir
- ☞ créer.

☞ rechercher scientifiquement:

Le processus de recherche dans une entreprise nécessite la mise en place de programmes de recherche. L'établissement de ces programmes passe nécessairement par une phase décisionnelle. Ces décisions ne sont pas prises au hasard mais bien évidemment en ayant une parfaite connaissance de ce que l'on sait faire et de ce que les autres savent et programment de faire.

Ensuite, le programme accepté est mis en oeuvre, une constante information sur les activités scientifiques et techniques extérieures est à diffuser pour aider les chercheurs, savoir si de nouvelles découvertes n'accéléreraient pas leurs recherches.

Il faut aussi informer les décideurs sur la concurrence pour fournir des indications sur l'avancement des recherches des autres entreprises ainsi que des informations provenant d'autres secteurs de surveillance: économie, marché, finance... pour pouvoir moduler les programmes selon les évolutions externes à l'entreprise.

☞ adapter:

Par adapter, Kami sous-entend: utiliser des procédés existant dans des domaines non-concurrents. Pour cela il faut précisément mettre en place un système de gestion des informations scientifiques, techniques et technologiques qui soit capable de détecter des opportunités. Ensuite, dans le cas de leur mise en oeuvre dans le circuit de production, les renseignements qu'il est utile d'acquérir ont une nature technique plus précise.

☞ acquérir:

Le processus d'acquisition d'innovation répond aux mêmes caractéristiques que dans le cas précédent. Le système de gestion des informations doit ici chercher les opportunités dans le propre domaine de compétence de l'entreprise. Ceci ne concerne que les informations techniques et technologiques pour aboutir à des achats de brevets ou de licences déjà en exploitation ou non.

☞ <u>créer:</u>

Cette dernière solution est un peu plus hasardeuse car le facteur principal entrant en jeu est un facteur humain. Les deux ingrédients indispensables pour la création sont la créativité et l'environnement. La créativité est difficile à gérer car elle dépend prioritairement des dispositions mentales des personnes qui sont censées être les sources de création. L'environnement de travail doit être propice à l'émulsion créative (on reconnaît que l'efficacité de la création est améliorée en collectif: brainstorming, matrice de créativité, réunion qualité...).

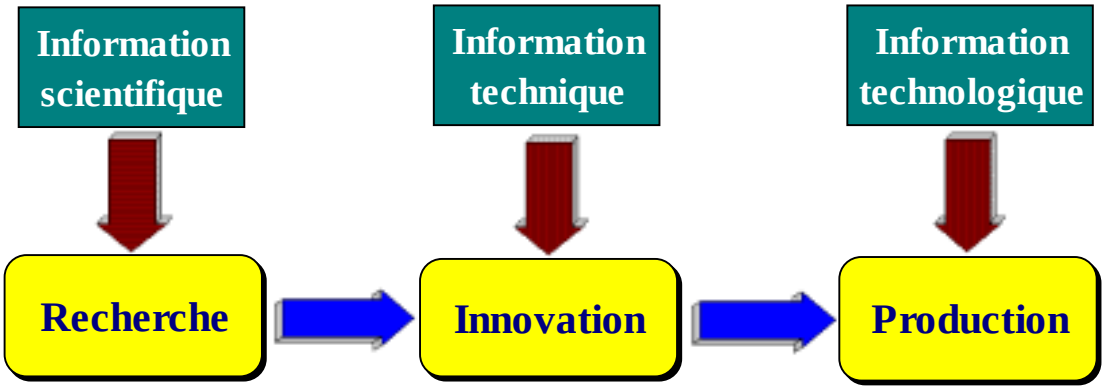
Mais de toute évidence pour pouvoir émettre des idées il faut forcément à un moment ou un autre avoir été informé de l'existant dans le domaine et de plus avoir une connaissance générale des autres domaines. Car c'est bien la réunion de ces deux ensembles de données qui, dans une "lueur d'esprit" de génie, feront émerger l'idée novatrice (c'est la phase "illumination" de l'innovation).

Le dénominateur commun à tous ces procédés d'innovation est l'information. Mais pour chaque acteur, entrant en jeu dans le processus d'innovation, cette information n'est pas de même nature. Tous les acteurs dans la chaîne des tâches de l'innovation sont importants mais ceux qui seront les plus influents sur le bon fonctionnement du système sont les décideurs. Les décideurs sont les coordinateurs du système. La bonne gestion de l'innovation dépend des décisions opportunes d'investissement et par conséquent dépend de la qualité des renseignements fournis aux décideurs.

Les principaux clients d'un système de veille technologique sont les décideurs, non pas par la quantité mais par la qualité des informations que l'on doit leur livrer. Nous verrons plus loin quelles sont les caractéristiques de ces informations de qualité.

On peut récapituler la finalité de la veille technologique en trois points:

⇨ L'aide à l'innovation dans une entreprise est la recherche du bon fonctionnement de la chaîne suivante:



⇨ On a vu aussi que les idées et les concepts de l'innovation sont généralement inspirés par des événements extérieurs et des modifications subtiles de l'environnement externe. Il faut donc préconiser que cette aide impose une très bonne connaissance de l'environnement scientifique, technique, technologique extérieur à l'entreprise. Il est à noter que l'adjectif *technologique* dans le terme *veille technologique* est donc réductionniste par rapport à sa réelle fonction.

⇨ On peut aussi dissocier la nature de l'information en sortie du processus de veille technologique en deux grandes catégories:

- ❍ des informations pour les acteurs dans la chaîne de l'innovation
- ❍ des informations pour les gestionnaires de la chaîne de l'innovation

2. Les principes de fonctionnement

Maintenant que les besoins en veille technologique sont établis, on en vient à envisager les principes qu'il faut imposer pour réussir au mieux cette activité. Les principes qui vont être énoncés par la suite ne sont pas réservés à la veille technologique. Ils pourraient aussi bien régir une tout autre surveillance dans un autre domaine que l'innovation. Ce sont des principes généraux propres à l'élaboration d'informations stratégiques dans un environnement industriel.

a) L'exemple du Japon

On prend souvent le Japon comme modèle de veille industrielle. La réussite économique de ce pays est fondamentalement associée à son comportement de traqueur d'information.

Le redressement fulgurant que ce pays humilié a montré après la seconde guerre mondiale est en grande partie imputable à son aptitude élevée à recueillir et à exploiter l'information pour les actions commerciales et industrielles. Cette surveillance exacerbée que les entreprises effectuent, a été dans un premier temps à l'origine de leur surnom de "copieurs" mais de nos jours comment appliquer ce qualificatif à ceux qui sont les premiers mondiaux dans des domaines de pointe tels que l'électronique.
L'organisation que ce pays a développée pour la surveillance industrielle n'est pas confinée dans une structure d'entreprise mais prend une dimension nationale. L'attention accordée par les japonais à l'information est une caractéristique culturelle, un état d'esprit qui remonte à des temps historiques puisqu'en 1868 figurait dans la constitution japonaise la phrase suivante :

> *"Nous irons chercher la connaissance dans le monde entier afin de renforcer les fondements du pouvoir impérial".*

L'infrastructure instituée, dans ce sens, par l'état japonais est lourde de signification. Le MITI (Ministry of International Trade and Industry) joue un rôle central autour duquel gravitent de nombreux organismes d'état comme le JETRO (Japan External Trade Organisation), le JICST (Japan Information Center of Science and Technology) contrôlé par l'AIST (Agency of Industrial Science and Technology), le JAPATIC (Institut japonais de la production industrielle). Tous offrent des services de collecte, de stockage et de diffusion auprès des entreprises. Cette mentalité les pousse à une entraide et à un partage de l'information à une échelle nationale.

La caractéristique majeure de ce système est que la collecte de l'information est réalisée tous azimuts. Chaque individu en voyage à l'étranger est moralement tenu de rapporter toute information qui a été, à ses yeux, source d'étonnement ou d'intérêt. **Ainsi cette collecte**

généralisée et touchant tous les domaines est centralisée dans les instituts nationaux pour y être traitée. Ceci implique bien évidemment une main d'oeuvre que ces quelques chiffres confirment: le JAPATIC emploie 2300 personnes, le JICST utilise près de 2300 permanents et 5000 scientifiques qui réalisent des analyses d'articles.

b) Méthode adaptée à la mentalité occidentale

La France est bien loin de posséder de telles infrastructures. Elle ne sera bien évidemment jamais le berceau d'un tel "monstre" car nos mentalités ne s'y prêtent pas. Aussi, il va falloir faire preuve de modestie et d'intelligence.

Un concept fondamental pour la veille technologique *à l'occidentale* fait l'unanimité chez les professionnels de cette nouvelle activité: segmenter la surveillance. L'entreprise va devoir se restreindre à la seule maîtrise des domaines d'activités préférentiels. Cette sectorisation de la surveillance ne concerne naturellement pas des restrictions géographiques mais uniquement des restrictions par secteurs.

Quels sont les sujets à guetter en priorité? Pour répondre à cette question il faut connaître les domaines qui pourraient, dans un futur plus ou moins proche, être critiques à la prospérité de l'entreprise. Une méthodologie, présentée par plusieurs auteurs [JAKO91] [MART89] [LESC86] [MORI85], parait particulièrement adaptée à la détermination de ces sujets critiques: **les Facteurs Critiques de Succès** (Critical Success Factors) développé par Rockart [ROCK79]. Ces facteurs critiques sont en fait les secteurs d'activité de l'entreprise où tout doit se passer parfaitement pour continuer à être prospère.

Appliqués à la veille technologique ces facteurs critiques concernent les secteurs de l'entreprise où il faut être en permanence bien informé (scientifiquement, techniquement et technologiquement) pour ne pas mettre en jeu la pérennité de son métier.

Ce principe dans le contexte d'une veille industrielle doit s'appliquer à tous les secteurs de la surveillance industrielle. En ce qui nous concerne, les sujets clés à lister ici ne concernent que les domaines qui risquent de mettre en péril la performance d'innovation de l'entreprise si on n'y prend pas garde.

Cette première phase (détermination des FCS) constitue probablement la plus importante dans le processus de veille technologique. Elle peut finalement être assimilée à

la première étape dans les modèles d'aide à la prise de décision: définition des objectifs (voir plus loin le paragraphe *L'information à fournir*). Cette définition de facteurs critiques correspond finalement à la définition des objectifs à se fixer pour le système de surveillance de l'environnement industriel.

Ensuite, viennent les étapes de programmation des actions à mener et de mise en oeuvre de ces actions:

☞ **Etablir le plan d'action de surveillance:**
Pour chaque facteur critique retenu, il va falloir établir les axes et les actions de recherche à réaliser pour être sûr de renseigner au mieux ce facteur critique. Cette étape fait appel aux compétences propres à l'activité de la veille technologique: savoir où chercher, comment trouver et comment traiter l'information pour élaborer l'information critique.

☞ **Mettre en application le plan et contrôler son bon fonctionnement:**
Cette phase est purement opérationnelle et va mettre en jeu tous les acteurs de la veille technologique. Nous verrons dans le paragraphe suivant qui sont ces acteurs et à quel moment ils interviennent dans le plan d'action. Les compétences et le savoir-faire propre à l'activité de chaque acteur sont les clés du bon fonctionnement du plan d'action.

La méthode de veille technologique peut se résumer succinctement par la figure 2.

Cette méthodologie, avec certaines divergences pour chaque auteur, est bien expliquée dans les ouvrages déjà cités ci-dessus.

METHODE DE SURVEILLANCE

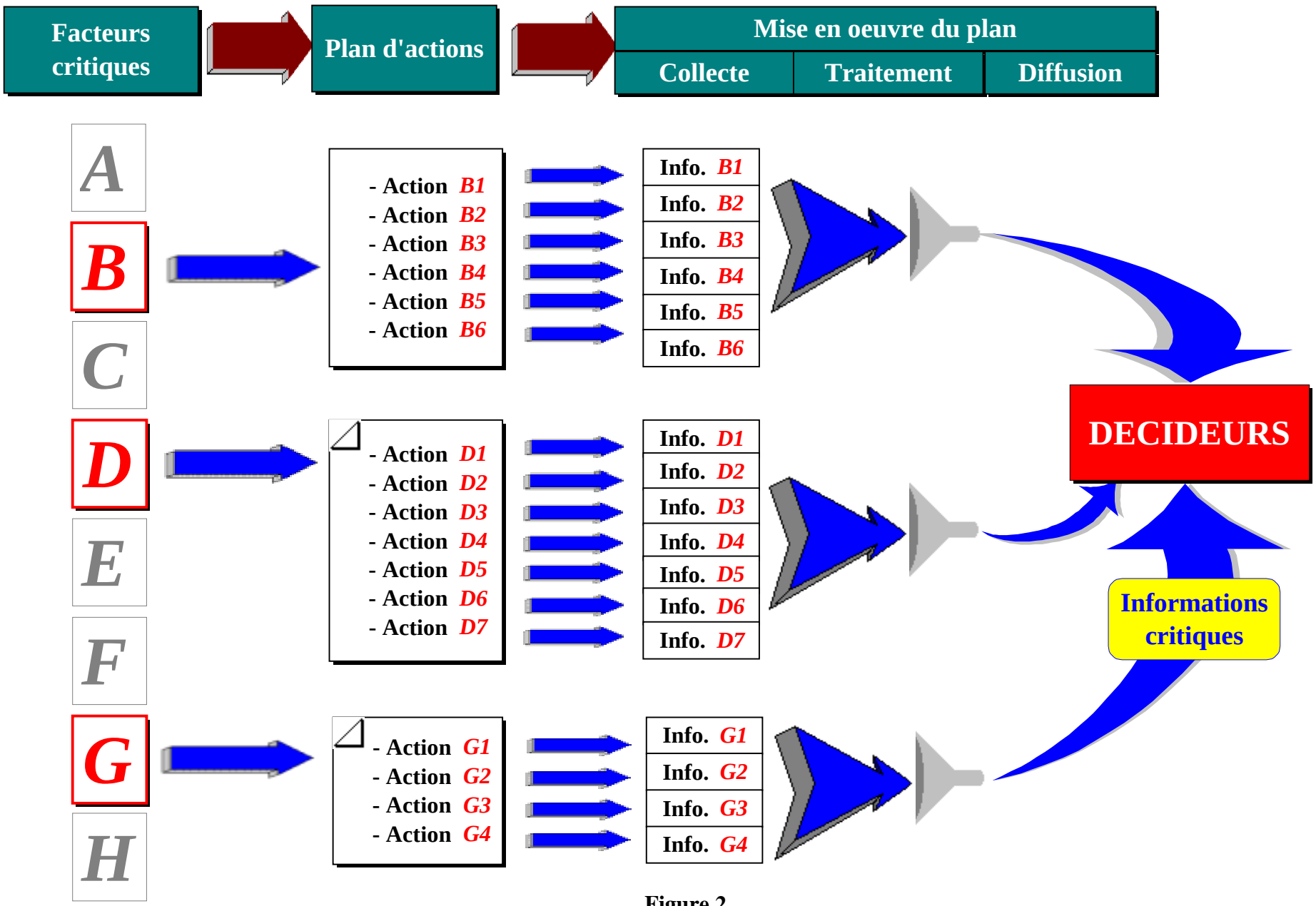


Figure 2

3. La mise en application des principes

Maintenant que les grands principes pour implanter un système de veille technologique sont connus, on est en mesure de se poser la question: comment sont-ils mis en pratique dans le contexte industriel? Sur ce point les divergences commencent à être considérables selon les cultures d'entreprise. Il n'y a probablement pas, comme pour tout problème, une solution unique. Les conditions et les contraintes pour l'installation d'une veille en entreprise sont variées et complexes.

Nous ne présenterons pas dans ce mémoire les différentes interprétations des précédents principes. **Nous exposerons simplement ce qui se fait depuis longtemps dans les entreprises et qui correspond, pour certains, à une veille technologique. Puis nous indiquerons en quoi nous nous distinguons de cette vision et nous exposerons notre perception du système de veille technologique.**

a) La situation courante du système d'information

Spontanément les entreprises ont perçu qu'il était utile que l'information créée à l'extérieur de l'entreprise y soit introduite. Grâce aux grands progrès dans la communication, l'acquisition de l'information brute est devenue assez facile. **Aussi, des systèmes de gestion d'information se sont mis en place avec pour fonction, la collecte de cette information brute selon la demande du personnel. Ces structures de gestion de l'information sont les centres de documentation.**

Leur fonction est indispensable. On a vu plus haut qu'il fallait constamment informer les différents acteurs de la chaîne de l'innovation pour que l'innovation dans l'entreprise reste performante. Ceci correspond exactement à l'activité réalisée quotidiennement par les centres de documentation.

D'un certain côté, on pourrait considérer que cette activité correspond à celle de la veille technologique. **Mais l'information que fournissent ces centres de documentation n'est pas suffisante pour les gestionnaires de la chaîne de l'innovation: les décideurs.** Or ce sont précisément les principales personnes qui ont besoin d'être correctement informées puisque qu'elles décident des futures stratégies.

Les centres de documentation "abreuvent" particulièrement bien les acteurs qui ont une tâche scientifique ou technique dans l'entreprise. Les personnes qui sont particulièrement friandes de l'information brute sont les spécialistes. Ils ont la possibilité de suivre continuellement les

évolutions de leurs domaines par les informations telles quelles sont publiées dans les revues spécialisées. L'information dont ils sont demandeurs ne nécessite aucun traitement avant sa mise à disposition.

Dans ces conditions, comment les gestionnaires font-ils pour prendre leur décision? Ils font directement appel à ces spécialistes. Ceux-ci experts dans leurs domaines vont informer les décideurs de la situation actuelle dans les divers domaines.

Si on essaie de calquer ce processus sur le schéma de la méthode de veille technologique qui a été établi précédemment on obtient la figure 3. **Deux principales étapes ne sont pas conformes aux principes généraux de la veille technologique:**

- ☞ Premièrement, c'est l'expert lui même, suivant sa curiosité personnelle, qui définit des profils de recherche qui permettront la sélection des articles par les services documentaires.

- ☞ Deuxièmement, la vision de la situation d'un domaine dépend principalement d'un seul individu. L'expert va faire autorité auprès du décideur. Un tel système est inacceptable car il repose sur l'opinion d'un seul individu.

L'amélioration évidente est de ne plus considérer l'avis d'un expert mais celui de plusieurs. Cette solution ne fait que déplacer le problème car ce sont toujours les experts qui ont la main mise sur la connaissance. De plus l'établissement de consensus parmi un groupe d'experts introduit une bonne dose d'aléas. "*Comme l'ont montré de nombreuses études de sociologie des sciences, le consensus des experts est un résultat construit dans la douleur et la passion."* Cette phrase est tirée de l'article [CALL87] de Callon. Cette pratique de se remettre à des groupes d'experts pour aider la prise de décision n'est pas le propre de la veille technologique. Nombreuses sont les pratiques de prise de décision qui dépendent des seuls points de vue d'expert. Dans le domaine de la stratégie Kami [KAMI89] en parle en ces termes: "*La prise de décision repose bien souvent sur des opinions et des consensus d'experts, c'est se reposer sur une sagesse traditionnelle et sur la loi des moyennes. Cela ne conduit à aucune innovation parce que cette démarche est étroitement liée au statu quo."*

Une autre faiblesse émerge de cette description du système de gestion de l'information: son inertie. Le décideur n'est pas suffisamment rapidement informé des dernières tendances. La fonction de ces experts, définie dans la structure de l'entreprise, n'est pas de renseigner le décideur. Il est probable que leurs tâches sont suffisamment prenantes pour ne pas leur en affliger une seconde.

PROCESSUS DU SYSTEME DE RENSEIGNEMENT EN CONTINU

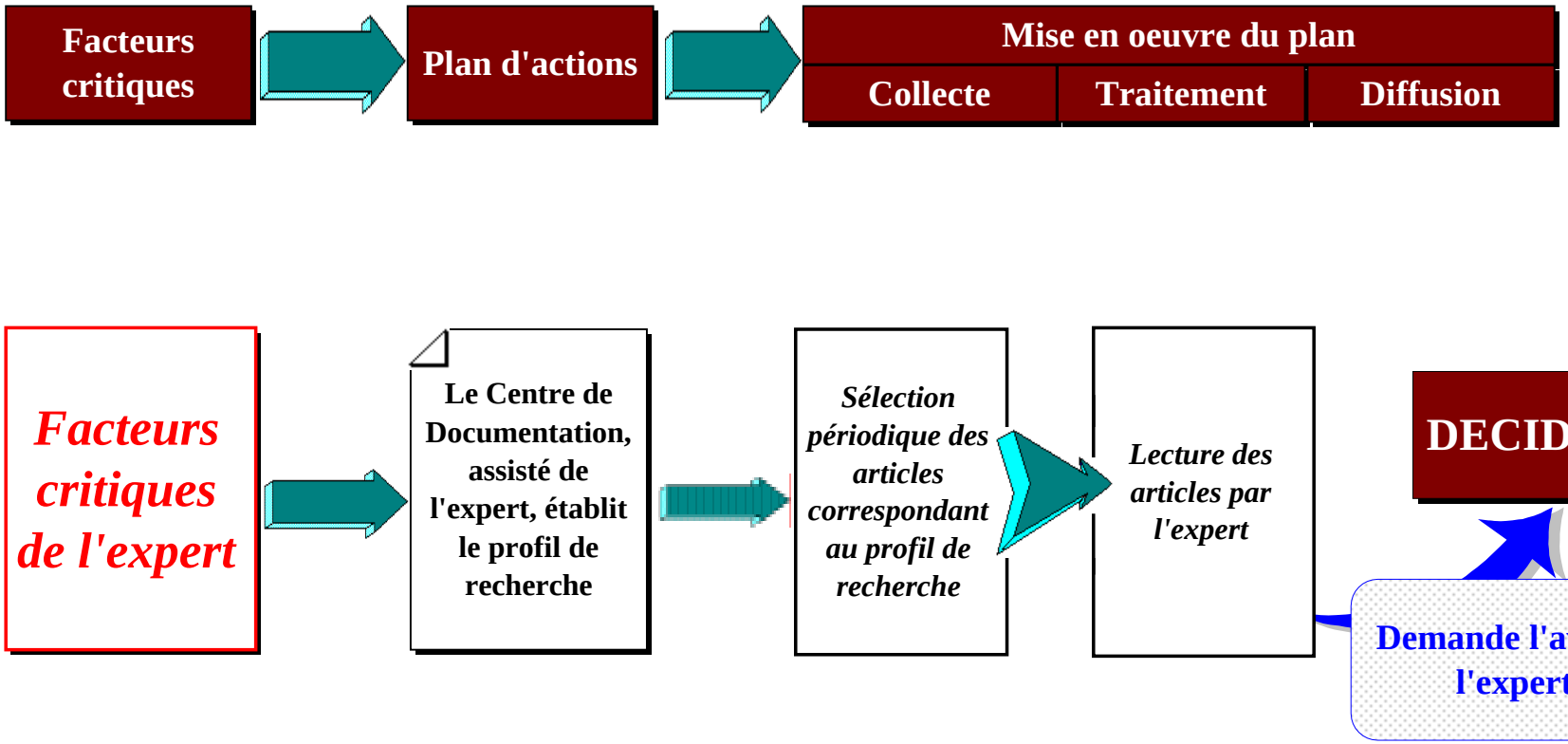


Figure 3

b) Le système à préconiser

La volonté d'un système de veille technologique est donc de produire une information destinée aux décideurs qui leur permet de construire par eux-mêmes leur propre point de vue sur la situation.

Pour que ces informations leur soient accessibles il va falloir traiter l'information brute de manière à procurer des renseignements décrivant les mutations. De plus, pour que ces informations soient pertinentes, il est indispensable que les décideurs eux-mêmes établissent les facteurs critiques à surveiller.

Nous envisageons pour réussir dans cette mission d'installer deux dispositifs en parallèle:

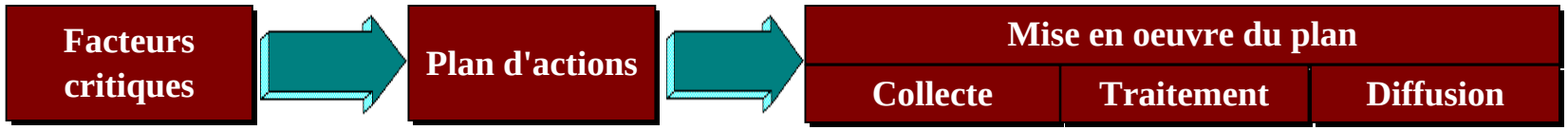
- ☞ Le renseignement systématisé
- ☞ L'élaboration d'indicateurs de tendances

(1) Le renseignement systématisé

Cette première structure est là comme système d'alerte au brusque changement dans l'environnement. Elle est donc la réponse à la trop grande inertie présente dans le système que l'on a exposé auparavant. **Nous l'avons nommé *renseignement* car cette activité va s'attacher tout particulièrement à la collecte de l'information informelle.** L'information informelle désigne tous les types de communication d'information qui ne se font pas par l'intermédiaire d'un support physique institutionnalisé (livres, articles, brevets, banques de données...). Elle fait essentiellement référence à tout ce qui est communiqué par voie orale. La qualité primordiale de cette information est sa primeur. Ce sont soit des informations qui seront formalisées plus tard soit des informations censées restées secrètes pour les concurrents. Bien sûr, l'information formalisée n'est pas à exclure de la collecte mais celle-ci est déjà très bien gérée par les centres documentaires traditionnels. Une présentation assez exhaustive des sources et des moyens de collecte propre à cette activité de renseignement est exposée dans [MART89].

La structure opérationnelle qui paraît la plus adaptée à cette mission est celle exposée par Jakobiak [JAKO91]. **Elle est bâtie à partir d'un réseau d'observateurs pour la collecte et des groupes de travail d'analyseurs pour la validation et l'analyse des informations.** En calquant cette structure sur le schéma des principes de la veille technologique on obtient alors la figure 4.

PROCESSUS DU SYSTEME DE RENSEIGNEMENT EN CONTINU



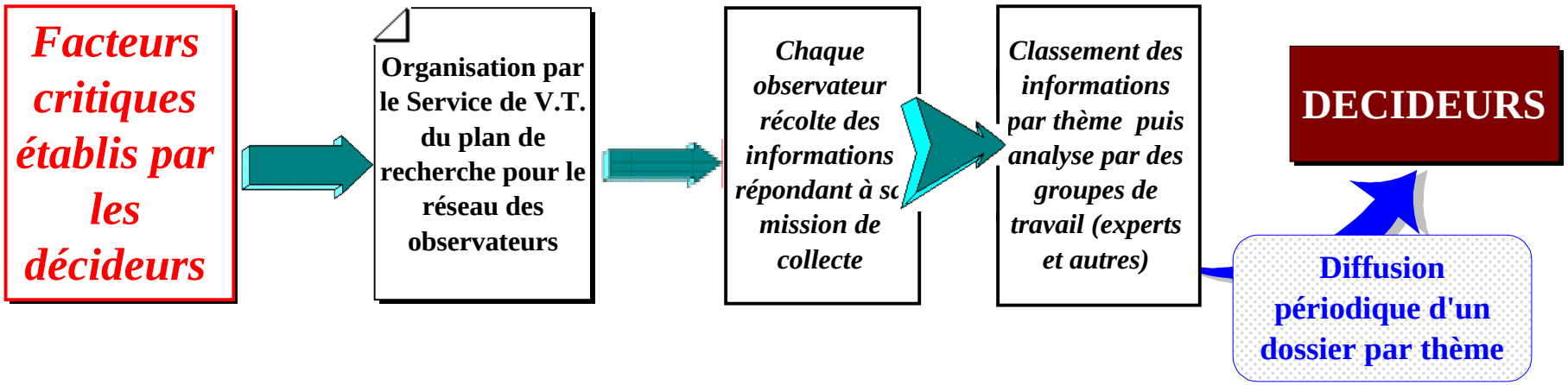


Figure 4

Par sa forme légère cette structure doit pouvoir drainer rapidement les informations. Les dossiers périodiques sont vite constitués et gardent ainsi toute la primeur de l'information. Dans cette structure on peut remarquer que le service de veille technologique n'intervient à aucun moment dans la phase opérationnelle. Le service de veille technologique selon Jakobiak a pour unique mission de veiller au bon fonctionnement du processus en animant et contrôlant chaque étape.

(2) L'élaboration d'indicateurs de tendance

Le renseignement en continu va permettre d'alerter les décideurs sur l'émergence de dangers non soupçonnés jusqu'à ce jour. Mais cette information ne leur donne pas assez de recul par rapport aux événements. Il va falloir pour décider des nouveaux programmes non plus des informations aussi pointues mais au contraire essayer de repositionner ces renseignements dans un contexte plus général.

Il paraît envisageable que ces dossiers d'alerte suscitent de nouvelles questions qui obligent à engager des investigations sur des secteurs de compétence inconnus au métier de l'entreprise. Les décideurs pour se faire une idée générale de ce nouveau domaine auront besoin de données qui font l'état des lieux.
On peut penser aussi qu'il soit indispensable de passer régulièrement au crible un secteur d'activité de l'entreprise pour estimer s'il souffre de quelques maux. Il est bon alors de replacer son activité par rapport à celle de la concurrence, de connaître les évolutions et les tendances pour l'avenir.

De manière générale, les décideurs ont l'habitude de se reporter à ce genre d'indicateurs pour établir leur stratégie. Lorsque l'on étudie par exemple *l'analyse de la discontinuité,* une théorie d'aide à la planification stratégique exposée par Kami, on peut distinguer quatre grandes phases:

- ⇨ Définir des objectifs
- ⇨ Etablir les nouveaux programmes
- ⇨ Mettre en oeuvre ces programmes
- ⇨ Contrôler l'évolution des programmes

Remarque: Il intéressant de retrouver en grande partie les mêmes étapes que celle énoncées en tant que principes de la veille technologique. Ce qui montre que les principes de veille technologique font partie intégrante d'un modèle de prise de décision stratégique.

La définition des objectifs est la phase cruciale de cette théorie puisque toutes les autres en découlent. La définition des objectifs est principalement étayée sur l'examen détaillé de la position de l'entreprise dans son environnement. Cet examen se présente sous la forme de réponses à une succession de questions concernant:

- le profil de l'environnement extérieur:
  Quels sont les facteurs-clés de notre environnement et dans quelle mesure pouvons-nous les maîtriser? Chercher à déterminer la position de l'entreprise dans le cadre des réalités externes actuelles, mieux comprendre l'existant.

- le profil de l'environnement intérieur:
  Où en sommes-nous à cet instant précis? Chercher des "instantanés" pour les activités de l'entreprise telles qu'elles sont à présent.

- l'environnement extérieur futur:
  Où allons-nous? Envisager l'influence des facteurs externes futurs. Faire des hypothèses sur les forces-clés de demain. Ce n'est pas là une analyse statique à un moment donné. Ce doit être un processus continu d'analyse et de recherche dynamique: affiner, ajuster, réfléchir à nouveau. Estimer les défis et les menaces les plus importantes qui nous attendent.

- les capacités:
  Où pouvons-nous aller? Connaître les forces et les besoins de l'entreprise. Apprécier ses capacités avec réalisme.

- l'environnement intérieur futur:
  Où pourrions-nous aller? Estimer l'adéquation entre les produits ou les services que l'entreprise pourrait proposer et les probables secteurs de marchés.

Comme on peut le voir à la lecture de ces lignes, **primo le besoin d'information est constant, secundo l'information doit être très diversifiée**. Tous les secteurs d'activité de l'entreprise peuvent être pris en compte si ils représentent des facteurs prépondérants sur les objectifs à atteindre. Les décideurs ont donc non seulement besoin des données fournies par le système veille technologique mais aussi de celles des autres services de l'entreprise: service commercial, service marketing, service financier, service prospective...

Ici l'information utile n'est pas une information brute mais une information indiquant les tendances générales. C'est justement sur ce point précis qu'il y a une pénurie de

données dans le domaine scientifique, technique et technologique. Les services commerciaux, marketing et financier ont développé depuis longtemps des outils répondant à cette attente. Ces outils fournissent des données décrivant les évolutions, les tendances, les émergences dans leurs secteurs. Ils offrent donc des idées sur la position de l'entreprise par rapport à la situation générale dans un domaine. Les mêmes outils pour la science, la technique et la technologie sont inexistants dans les entreprises.

Voilà quelques unes des multiples raisons pour lesquelles il faut instaurer une seconde tâche à la veille technologique. Pour cette nouvelle tâche, le service veille technologique va devoir s'investir dans la phase opérationnelle car les compétences que requièrent ces nouvelles investigations n'existent nulle part ailleurs dans l'entreprise. Ces personnes seront reconnues par la suite sous la désignation de *spécialistes VT*.

Comme précédemment, on peut schématiser cette nouvelle fonction avec la figure 5.
Le thème de l'étude est défini par les besoins des décideurs.
La définition de la stratégie de l'étude exige de nombreuses compétences dont particulièrement une très bonne connaissance du domaine concerné par le thème, d'où le concours d'experts. Ces experts interviennent une seconde fois dans la phase opérationnelle au niveau de l'aide à l'interprétation des résultats des analyses statistiques.
Mais le point le plus important à noter dans cette figure est justement la présence d'un traitement spécifique à l'induction des tendances.
En final, un dossier est constitué pour rassembler la quintessence des résultats des interprétations. Mais peut-être encore plus appréciée, est la présentation des résultats dans une réunion. L'avantage principal des indices de tendances est qu'ils sont produits sous des formes graphiques toujours plus agréables et plus convaincantes que de longs discours.

C'est précisément cette élaboration de tendances dans le domaine de la veille technologique qui sera étudiée dans la suite de ce mémoire.

PROCESSUS DU SYSTEME D'ELABORATION D'INDICATEURS DE TENDANCES

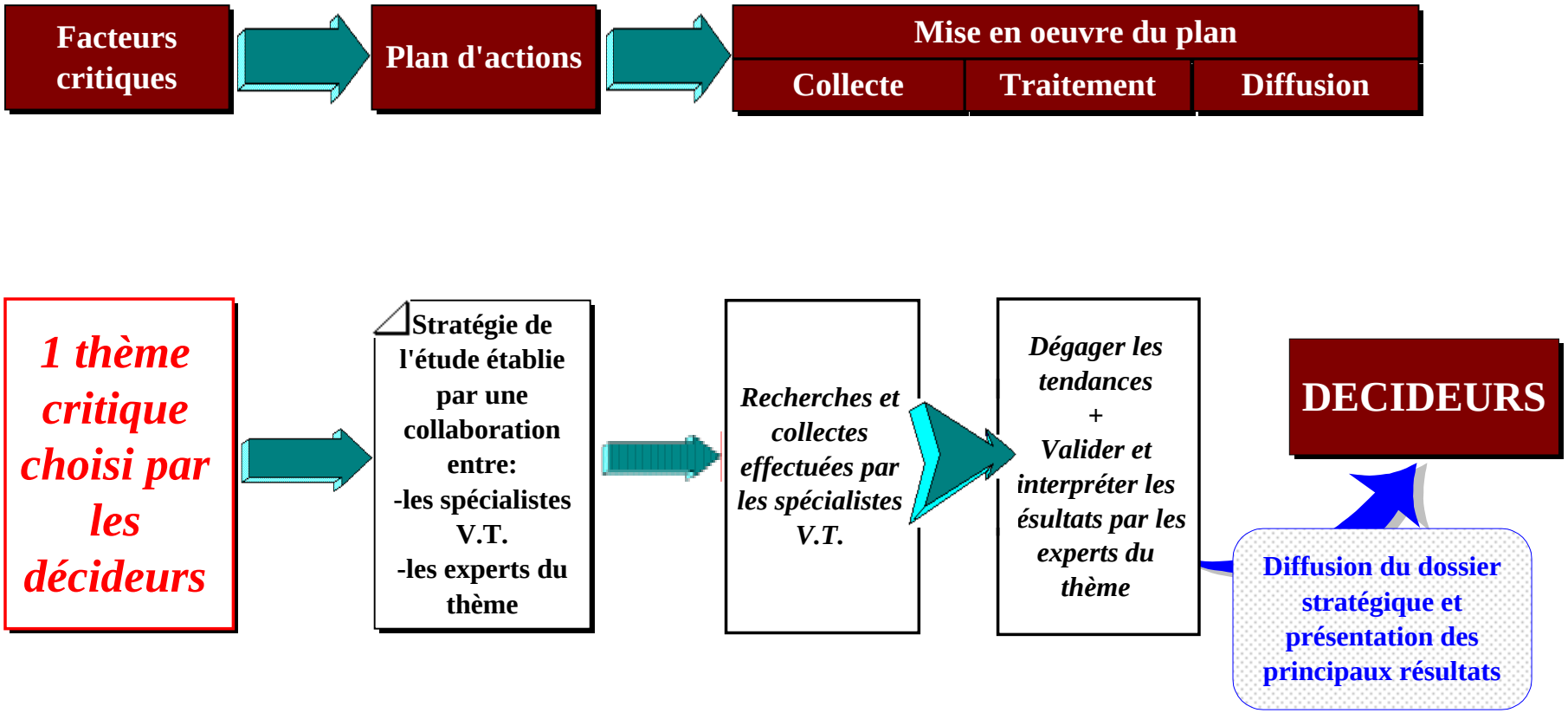


Figure 5

(3) Le dispositif complet

La description du dispositif au complet de veille technologique peut se représenter selon la figure 6. Cette figure nous montre que **les deux tâches de la veille technologique sont accomplies par deux structures séparées.**

Dans la première, celle du renseignement systématique, le service de veille technologique se restreint à la fonction d'animateur. Il est simplement le moteur du processus. Il exploite pour la phase opérationnelle les compétences et les personnes déjà présentes dans l'entreprise. Cette première activité est la principale abordée par les ouvrages de veille technologique. Les auteurs ne sont pas toujours d'accord sur les structures et les acteurs à mettre en place mais les principes sont dans les grandes lignes très similaires.

Par contre, la seconde activité n'est pratiquement pas développée chez les auteurs. Les raisons de cette absence sont assez simples. **Pour qu'elle soit performante, cette activité doit disposer de deux éléments fondamentaux: des personnes qualifiées et des outils spécifiques.** Or pour l'instant ces deux éléments sont déficients dans les entreprises. Les outils ne sont pas introduits dans le monde industriel car ils sont encore bien souvent à l'état de "recherche améliorée". Et donc le personnel qualifié n'a pas lieu d'être. Nous exposerons dans la suite quel outil nous préconisons pour cette tâche et quelles compétences il mettra en jeu.

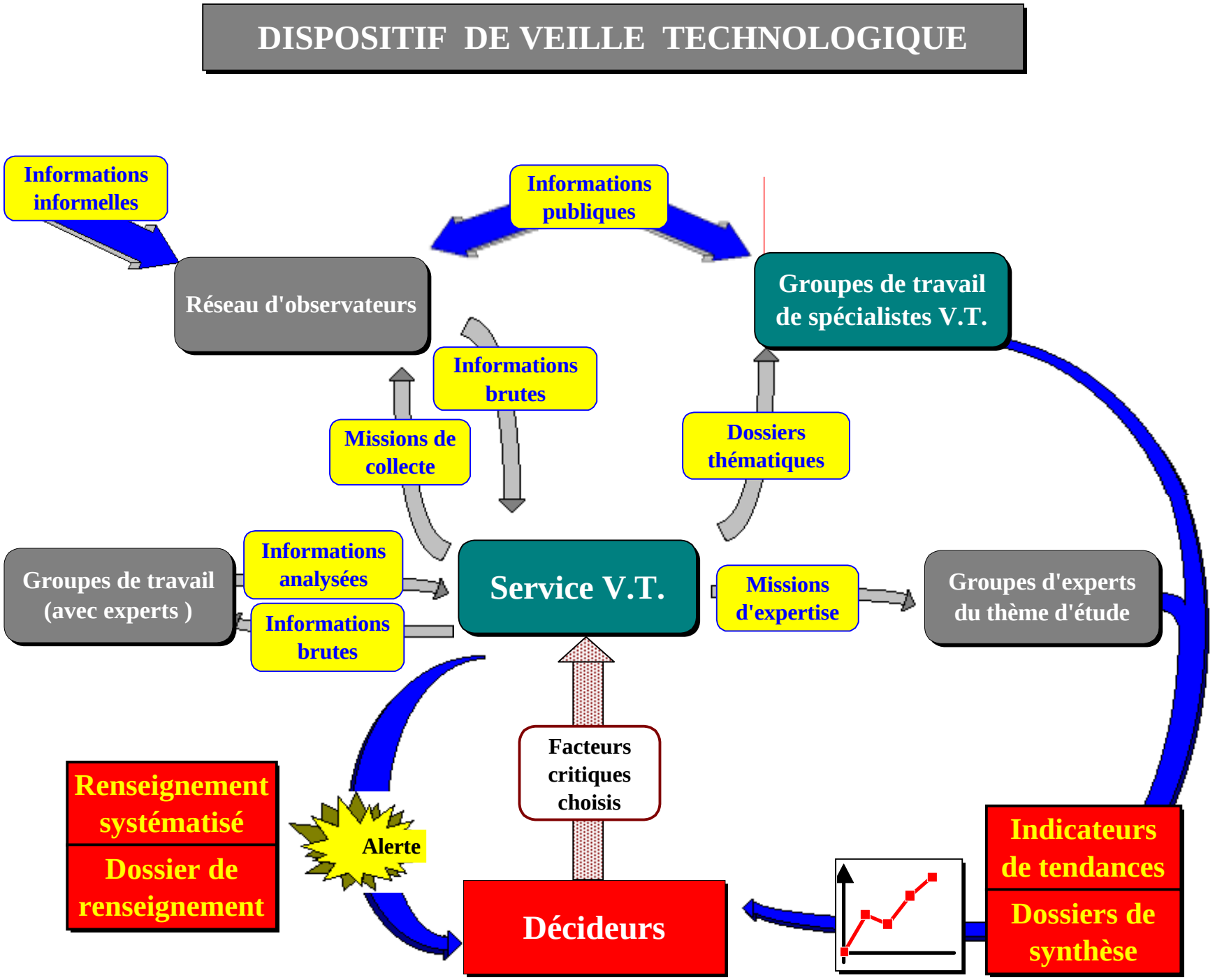

DISPOSITIF DE VEILLE TECHNOLOGIQUE
Informations informelles
Informations publiques
Réseau d'observateurs
Groupes de travail de spécialistes V.T.
Informations brutes
Missions de collecte
Dossiers thématiques
Groupes de travail (avec experts)
Informations analysées
Informations brutes
Service V.T.
Missions d'expertise
Groupes d'experts du thème d'étude
Facteurs critiques choisis
Renseignement systématisé
Dossier de renseignement
Alerte
Décideurs
Indicateurs de tendances
Dossiers de synthèse

Figure 6

D. La bibliométrie: technique d'élaboration d'indicateurs de tendances en veille technologique

La *bibliométrie* est l'outil que l'on déclare apte à répondre aux lacunes, exposées précédemment, du système de veille technologique. Il n'est probablement pas le seul instrument mais il parait être très adapté aux exigences de la veille technologique. Nous allons essayer de justifier cette proposition dans ce paragraphe.

Pour l'instant, le décideur a trois solutions pour évaluer les tendances dans un des domaines du monde scientifique ou technique:

- ☞ demander l'avis d'un expert ou d'un groupe d'expert et nous avons vu les problèmes que cela induit.

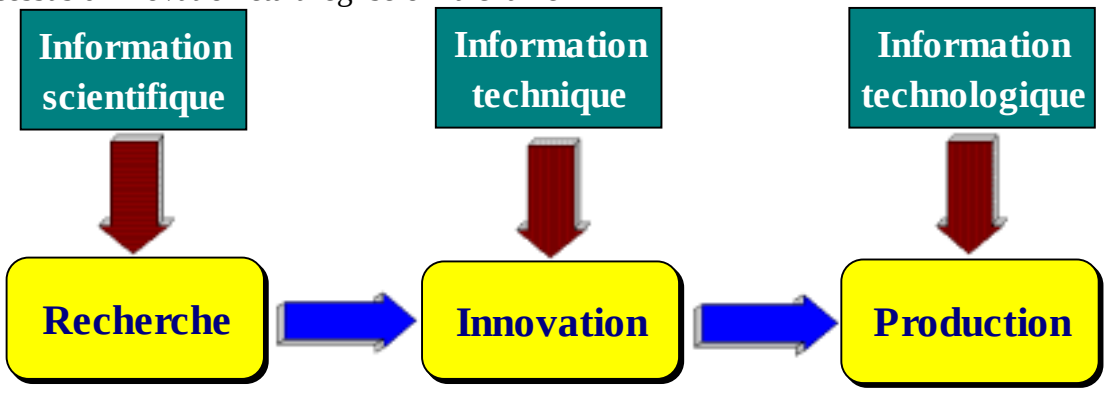
- ☞ faire réaliser des dossiers de synthèse. Ce procédé est courant mais il n'est pas adapté au système de veille technologique. Sa mise en oeuvre impose la collecte, la lecture et l'analyse d'un trop grand nombre de documents pour que cela puisse être réalisé dans les délais exigés par une veille technologique efficace. Il faut aussi prendre en considération que l'influence des compétences des personnes, qui réalisent ce dossier, n'est pas négligeable sur les résultats.

- ☞ suivre son propre instinct. Si l'entreprise a déjà instauré un système de renseignement systématisé cet instinct sera influencé avantageusement par les dossiers de renseignements qu'il reçoit régulièrement. Mais ceci ne lui donne qu'une vision parcellaire de la situation générale.

Il serait bon que le jugement du décideur soit complété par une opinion plus large et plus objective que celles qui lui sont proposées. La bibliométrie, concept qui sera développé dans le paragraphe suivant, est un outil statistique de mesure des tendances de la science, des techniques et des technologies.

1. L'aide à l'innovation

La finalité de la veille technologique, on l'a vu, est l'aide à l'innovation. On a aussi vu que le processus d'innovation était régi selon la chaîne



La première phase, purement scientifique, utilise par conséquent l'information scientifique. Les articles, colloques et rapports sont les principaux documents scientifiques dont le contenu renseigne parfaitement sur l'état des recherches fondamentales ou appliquées. La bibliométrie a été justement développée pour mieux cerner la science et ses programmes de recherche (Cf chapitre IV la partie *La mesure de la science*).

La seconde phase concerne la mise en pratique des recherches fondamentales à l'échelle du laboratoire. Elle prend principalement un caractère technique. La technique étant un des éléments déterminants de la production, les entreprises ont l'habitude de se protéger juridiquement par des brevets. Ces documents n'ont pas uniquement un caractère juridique puisqu'ils forcent l'entreprise à décrire précisément le procédé à protéger. Ils sont donc la source d'une information très détaillée et très complète des techniques employées. La bibliométrie a intégré plus récemment le traitement des documents brevets. Depuis ils sont devenus les sources privilégiées de certains auteurs en traitement bibliométrique (Cf chapitre IV la partie *La mesure des techniques et technologies*).

La phase finale est le développement. Elle correspond à la mise en pratique, à l'échelle de la production industrielle, des techniques du laboratoire. Les informations impliquées par le développement sont couvertes en partie par les brevets qui permettent, outre les techniques, de protéger les procédés industriels. Les publications scientifiques rédigées dans le domaine de l'ingénierie peuvent être prises aussi en compte. En fait, les données les plus importantes sont celles couvrant le savoir-faire. Celles-ci restent malheureusement sous forme d'information informelle, le bouche-à-oreille étant le principal vecteur de communication du savoir-faire. Ce type d'information est donc très difficilement évaluable. La veille technologique peut fournir sur le savoir-faire uniquement des informations ponctuelles et parcimonieuses par la structure du renseignement systématisé.

On peut conclure que la bibliométrie répond totalement au besoin en ce qui concerne le traitement d'informations de type texte publié et couvrant à la fois les domaines des sciences et de techniques.

2. La source de l'information

La méthodologie de la bibliométrie est basée sur le traitement de références de documents. Depuis les années 70 il est possible d'accéder aux versions informatiques des plus grands bulletins signalétiques par la télécommunication. Ces répertoires informatisés de références sont appelés *bases de données*. Ces bases de données sont de colossaux gisements d'information. Les domaines couverts par ces bases de données sont très divers: sciences, techniques, technologies, économies, juridiques, normes, finances, environnement, science humaine, science sociale... (Cf chapitre III *Les bases de données*)

Les avantages de l'utilisation des bases de données comme source d'information sont nombreux. Parmi ceux-ci certains déterminent leur emploi dans un système de veille technologique:

- l'exhaustivité:
  Les bases de données offrent une exhaustivité d'information sur le plan de la couverture thématique, géographique et temporelle. Jakobiak estime que ce sont les premières sources d'information en science, technique et technologie. Dans son premier ouvrage [JAKO88], il qualifie la couverture des bases scientifiques comme *excellente*, des bases techniques (brevets) comme *très bonne*, des bases technologiques (ingénierie) comme *nettement insuffisante*. A part la troisième catégorie - l'explication de ce jugement a été exposée juste auparavant - la qualité d'exhaustivité de ces bases est remarquable. Cet attachement aux bases de données est vérifiable dans son dernier ouvrage [JAKO92] où les exemples de dossiers stratégiques sont exclusivement constitués à partir de ce type de source.

- la qualité du contenu:
  Ces références constituent déjà une étape vers une information élaborée. En effet, l'indexation, dont les références sont l'aboutissement, est un traitement d'analyse des documents originaux. Cette analyse des documents est effectuée selon deux objectifs: l'extraction des concepts majeurs exprimés dans le document pour les transcrire en descripteurs de recherche, et la réduction du texte (résumé) pour la diffusion de l'information. Ce traitement typiquement documentaire est très bien expliqué dans l'ouvrage [CHAU88] de Chaumier.

- ☞ la qualité de la mise en forme:
  Cet avantage est primordial pour le traitement bibliométrique. Le fait que les références soient saisies selon une structure rigide va nous permettre d'automatiser les traitements. Le fait de savoir, où se situent les différentes données dans la référence et selon quelles règles elles y sont rédigées, va permettre de tendre vers un traitement bibliométrique totalement automatisé, le "zéro manuel" du traitement bibliométrique. Ceci est fondamental dans le processus veille technologique: réduire au minimum le temps d'exécution de chaque étape du plan d'action sans perdre en qualité.

- ☞ le temps d'accès:
  Dans le même état d'esprit, il est très intéressant que la phase de collecte puisse se faire pratiquement instantanément.

En contre-partie de ces formidables avantages, la consultation des bases de données est coûteuse et nécessite une qualification, une expérience et une grande rigueur.

3. L'aide à la prise de décision

Une remarque sur l'exploitation des bases de données comme source d'information en veille technologique revient souvent sous une forme semblable à *"plus l'information est formalisée, plus elle date et donc moins elle a d'intérêt"*. Cette remarque est judicieuse mais elle ne concerne pas le même type d'information.

Il faut connaître le plus rapidement les nouveaux projets de recherche des principaux concurrents pour ne pas perdre de l'avance technologique. Ceci est la première tâche du service de veille technologique: le renseignement systématisé.

Mais, il est bien évident qu'un décideur ne va pas s'engager dans un nouveau projet sur un hypothétique renseignement. Il va vouloir connaître l'évolution des recherches du concurrent pour estimer si ce nouveau projet est plausible et ainsi essayer de confirmer le renseignement. Il peut encore vouloir replacer ce nouveau projet dans la conjoncture générale, situer ses propres recherches en cours par rapport à ce que font les autres...
Pour cela, il faut pouvoir se reposer sur des indices sûrs. Le lancement d'un nouveau projet est un lourd investissement qui ne peut se prendre à la va-vite. Le changement d'un axe de recherche met en jeu l'avenir de l'entreprise puisqu'il doit aboutir à des innovations (et donc à une pérennité momentanée). L'information qui doit servir à prendre de telles décisions doit avoir une validité plus que certaine. Par sa nature même, l'information informelle n'a justement pas cette qualité. **Par contre, les informations formalisées peuvent servir sans**

crainte (si on maîtrise leurs limites) à l'élaboration d'indicateurs. Suivant ce raisonnement, l'emploi des bases de données en bibliométrie me paraît totalement justifié et approprié à l'élaboration d'indicateurs de tendance.

Nous nous sommes déjà longuement exprimés sur le besoin d'indicateurs de tendance dans le processus de la prise de décision. Nous y ajouterons ces quelques réflexions tirées de l'article [CALL87] de Callon où il décrit l'avantage qu'il y a de reposer le choix de programmes de recherche nationaux sur des indicateurs bibliométriques:

> *"... toute décision, à moins qu'elle ne se veuille secrète, s'appuie sur des évaluations explicites, même si celles-ci sont le plus souvent lapidaires et imprécises.*
>
> *C'est parce que les évaluations jouent un rôle important dans la prise de décision, qu'il est apparu souhaitable au fil des ans d'élaborer des outils, appelés indicateurs, destinés à fonder les jugements sur des données observables, vérifiables et contrôlables. Affirmer que la recherche fondamentale est en bonne santé est une chose, imaginer et calculer des indicateurs étayant cette thèse en est une autre. Sans données pour la soutenir, cette proposition sera écartée sans difficulté; appuyée sur des indicateurs qui la confirment elle devient plus difficilement contournable. C'est dans la discussion qui en résulte, dans le jeu des preuves et des contre-preuves qui sont produites et calculées, que s'engendre ce qu'il est alors possible d'appeler la décision rationnelle. Les évaluations ne déterminent pas mécaniquement les choix; elles les rendent plus rigoureux parce que plus difficiles à justifier et à imposer. C'est à ce prix, et à ce prix seulement, qu'au règne de l'arbitraire se substituera progressivement celui de l'arbitrage.*
>
> *Pour faciliter ces évaluations contradictoires, les indicateurs doivent être multiples. D'abord parce qu'un jugement ou une appréciation, qui tend toujours à être synthétique, s'applique en fait à des réalités multiformes. Ensuite parce que les différents acteurs impliqués ne sont pas nécessairement d'accord pour privilégier les mêmes aspects et les mêmes questions. Il faut donc prévoir des indicateurs suffisamment nombreux pour accroître les informations et les angles d'analyse, mais aussi pour que chacun soit en mesure d'y puiser les données qu'il juge pertinentes."*

On peut encore ajouter à tous ces avantages que les indicateurs de tendance sont presque toujours représentés sous forme graphique. Il est bien connu que les décisions sont toujours mieux établies lorsqu'elles sont étayées sur des graphes représentatifs [BERT77].

4. La détermination des premiers facteurs critiques

La phase la plus importante dans la veille technologique est l'établissement de la liste des facteurs critiques de l'entreprise. Etablir une liste exhaustive des dangers probables dans l'avenir n'est déjà pas chose facile. A cette difficulté vient s'ajouter la sélection des facteurs critiques prioritaires parmi ceux de cette énumération. Jakobiak prévoit au maximum une dizaine de facteurs critiques à surveiller pour que la veille soit performante. Il ne faut absolument pas se tromper lors de ce choix car alors le risque serait grand de passer à côté d'opportunités. Nous sommes là encore en présence d'une prise de décision. Il faut pouvoir juger et décider des futurs objectifs de veille avec le plus d'objectivité possible.

En réalité, les facteurs critiques sont soumis à une remise en cause périodique qui peut être représentée par la figure 7. Les indices significatifs décelés par les alertes du renseignement systématisé et confirmés par les indicateurs de tendances vont être la source de la réflexion pour le réajustement des prochains facteurs critiques.

Réfléchissons maintenant au démarrage du processus. On se trouve face au fameux problème de l'oeuf et de la poule: Qui est apparu le premier? Dans le cas présent, ce sont forcément les signes de menace qui poussent les décideurs à vouloir être mieux renseignés. Mais vont-ils faire les bons choix? N'est-il pas possible de les aider dans cette décision?

Pour pouvoir décider d'un axe de surveillance accrue sur un sujet ou sur une société concurrente, il faut avoir vu le danger "poindre à l'horizon". Il faut donc disposer de moyens d'analyse rétrospective pour retracer les situations passées, leurs émergences, leurs évolutions, leurs disparitions... **A la vue et à l'analyse de ces images rétrospectives les facteurs d'étonnement et les indices pourront induire des sujets où il semblerait intéressant d'être mieux informer.** La bibliométrie a été bien trop souvent qualifiée de méthode d'analyse rétrospective (et non prospective[1]) pour qu'elle ne puisse pas s'avérer une aide efficace dans la détermination des premiers facteurs critiques.

(1) à ma connaissance, il n'existe de toute manière aucune méthode quantitative qui puisse décemment se qualifier de prospective dans un domaine aussi complexe que la connaissance scientifique.

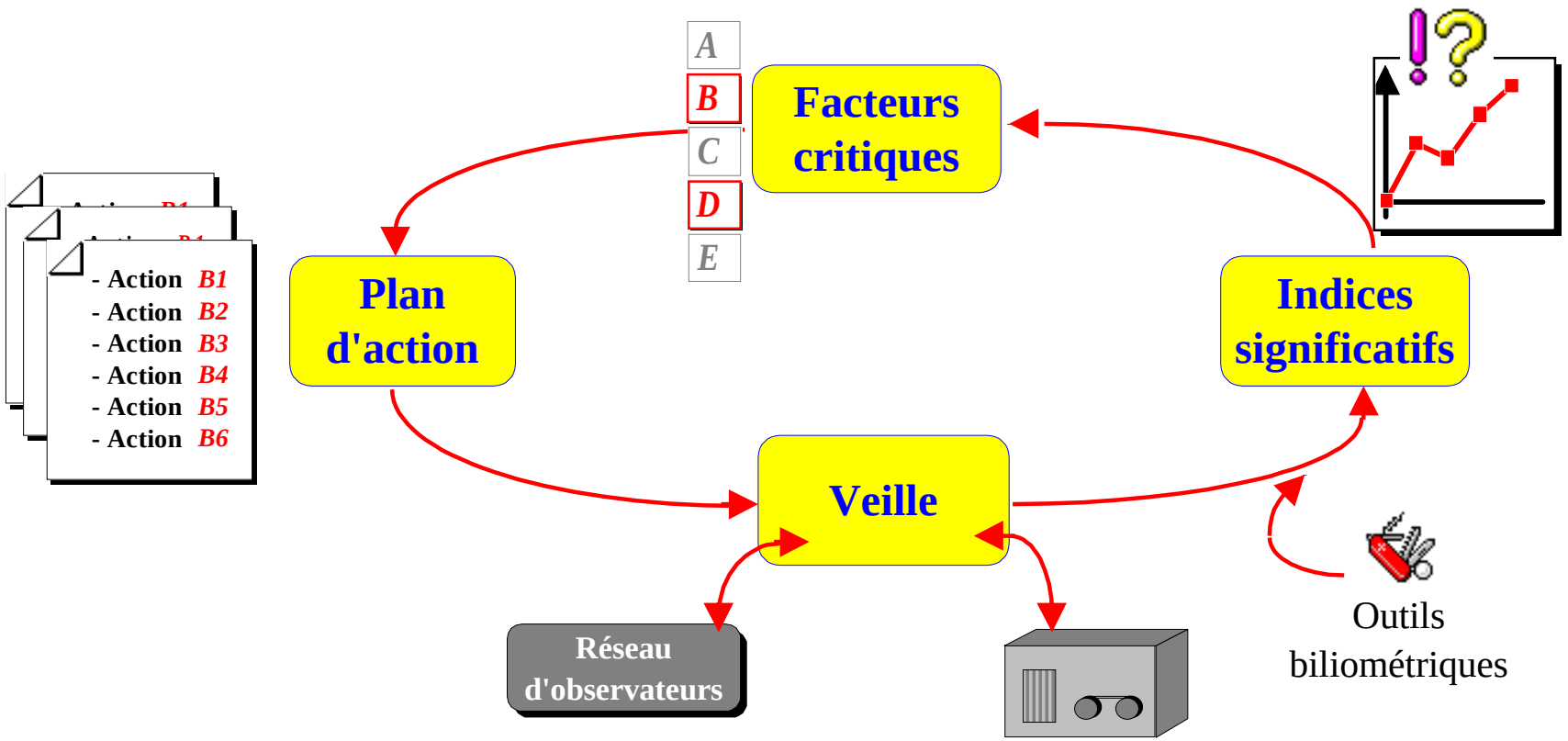

A
B
C
D
E
Facteurs critiques
Plan d'action
Indices significatifs
- Action B1
- Action B2
- Action B3
- Action B4
- Action B5
- Action B6
Veille
Réseau d'observateurs
Outils biliométriques

Figure 7

5. Le caractère dynamique

Comme le système économique actuel a des durées d'obsolescence technologique trois fois plus rapide que dans le passé (selon Kami [KAMI89]), il devient par conséquent nécessaire d'avoir des informations de meilleure qualité, plus pertinentes et surtout plus rapidement.

Le système de veille technologique doit procurer des informations élaborées rapidement. Il n'est donc pas imaginable de demander à des groupes de travail de constituer des dossiers de synthèse par la lecture d'un trop grand nombre de documents.
De nos jours, l'accès à l'information est devenu pratiquement instantané avec l'apparition des bases de données. Par la suite, il ne faut surtout pas ralentir ce flux par des traitements manuels surtout s'ils peuvent s'effectuer en bonne partie automatiquement.

L'outil informatique qui sera décrit plus loin est justement développé pour éviter tous les traitements manuels répétitifs et laborieux. Il tend à réduire la lecture des références au strict minimum, c'est-à-dire après avoir dégagé de l'ensemble celles qui nécessitent une attention particulière parce qu'elles ont des caractéristiques originales.

La veille technologique n'est performante que lorsqu'elle est dynamique et itérative, ceci dans le but de dégager des gradients d'information. Comment envisager l'étude comparative dynamique d'un axe de surveillance si ce n'est par un système spécifique informatisé. Admettons qu'un dossier sur un thème sensible soit effectué à une période t et qu'à la suite de la présentation de ce dossier il ait été décidé de mettre en place une surveillance sur ce thème. Est-il concevable qu'à chaque $t+dt$, correspondant à la période fixée entre deux présentations de ce dossier, il faille comparer manuellement les nouveaux documents collectés avec ceux déjà récoltés pour estimer ce qu'ils apportent de nouveau? Imaginons que ce dossier concerne un sujet qui couvre un grand nombre de documents. Comment faire sans un système informatique automatisé?

6. La fonction des experts

Si on a établi un système de veille technologique c'est pour éviter que les renseignements livrés aux décideurs soient uniquement dépendants de l'opinion des hommes de l'art.

Ainsi, l'intervention des experts n'est là que pour valider deux étapes (se reporter à la figure 5), non pour décider de celles-ci. Ceci n'est pas pour dévaloriser leurs fonctions dans l'entreprise mais uniquement par souci d'une plus grande objectivité des résultats. Leur présence dans l'édification de l'information critique finale est toujours aussi indispensable car

ce seront pratiquement les seuls qui pourront décrypter des signes inattendus parmi les résultats.

La fonction d'un expert dans ce système de veille technologique n'est plus de lire, classifier et synthétiser une masse de documents mais de valider ou invalider les résultats d'un traitement automatique qu'on lui expose. A la suite de la confirmation des grandes tendances, il pourra alors s'étonner de certaines corrélations plus fines et inattendues. Ce sont ces facteurs d'étonnement qui impliqueront, eux, certainement une lecture plus approfondie des références et même imposeront bien souvent la commande des documents originaux. Il paraît tout à fait normal de ne faire intervenir les experts qu'à ce niveau dans le traitement des informations. De simples travaux de lecture et de synthèse pourraient, tout au contraire, leur sembler comme un singulier surcroît de travail peu valorisant.

7. La bibliométrie dans la veille en général

Nous venons de montrer l'attrait que pourrait avoir l'introduction de l'outil bibliométrique dans le processus de veille technologique. Il est évident que cet outil peut créer de nouvelles perspectives dans le traitement de l'information propice à l'innovation. Mais il n'est qu'un outil parmi les autres dans le système d'aide à la décision de la veille industrielle. La figure 8 positionne cet outil dans un système de surveillance globale.

La présentation que nous venons de faire de l'activité de la veille technologique aura pu paraître trop restrictive pour bon nombre de personnes. Les auteurs ont l'habitude d'affecter la veille technologique à une surveillance comprenant le secteur économique, juridique, environnement, normes, ainsi que tout ce qui touche les produits et les services.
La vision personnelle de la veille technologique qui vient d'être exposé à l'avantage de très bien segmenter les fonctions et donc d'obtenir probablement une plus grande performance par des compétences plus adaptées aux domaines de surveillance.

LA PLACE DE LA BIBLIOMETRIE DANS LA VEILLE TECHNOLOGIQUE

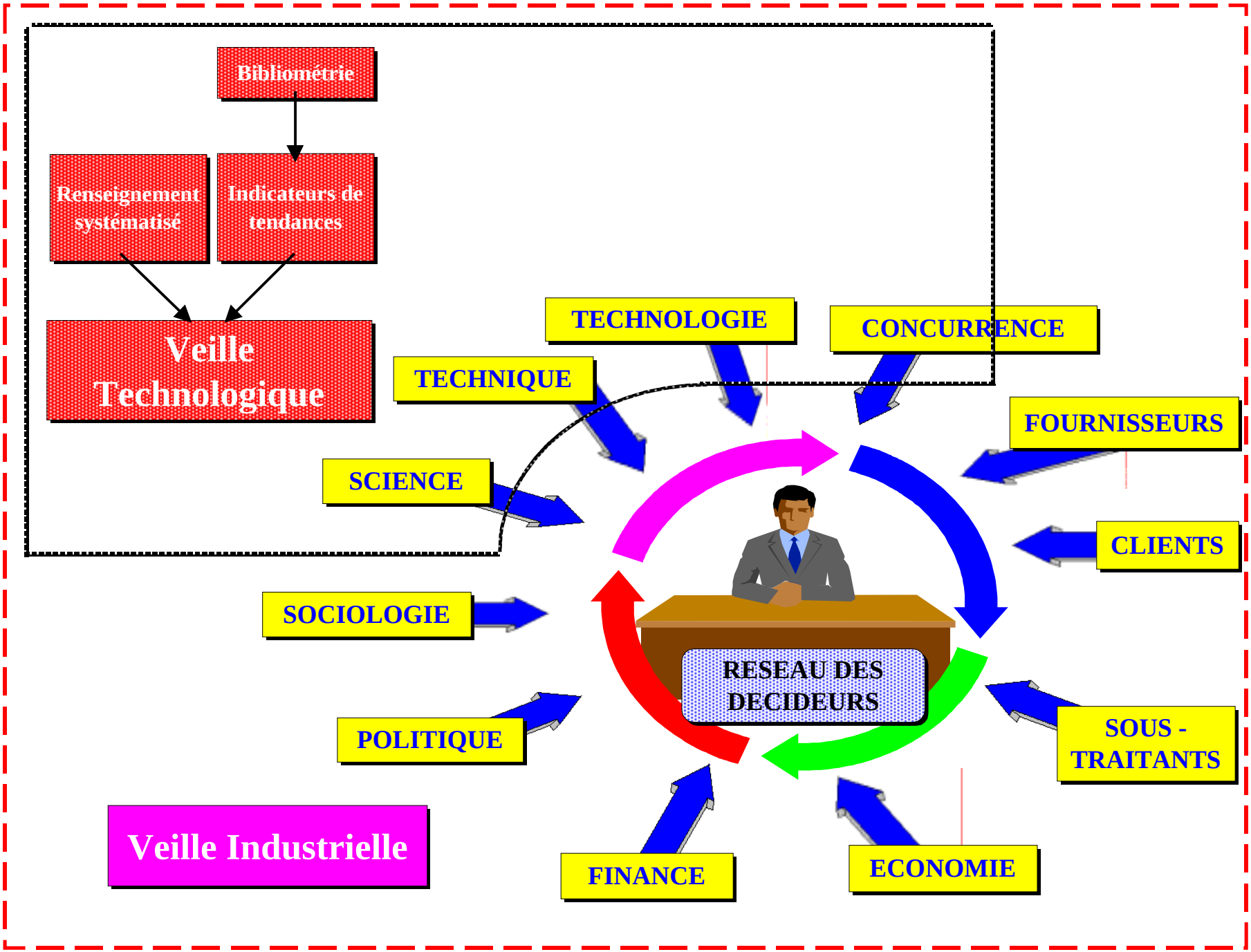


Figure 8

La source des informations exploitées en veille technologique et en bibliométrie: les bases de données

III. La source des informations exploitées en veille technologique et en bibliométrie: les bases de données

On définit les banques de données comme "**un ensemble de données relatif à un domaine défini des connaissances et organisé pour être offert aux consultations d'utilisateurs**" (Journal Officiel, 17 janv 82).

J-P Courtial dans [COUR90] introduit les bases de données en les nommant des bibliothèques informatisées. Cette dénomination est bien adaptée à leur fonction, mise à part le fait que les bases de données ne mettent pas à disposition les documents originaux mais permettent seulement de passer une commande auprès de l'organisme producteur. Effectivement, outre l'emprunt d'un ouvrage, les bibliothèques mettent, avant tout, à disposition de nombreux moyens pour trouver les documents qui semblent répondre correctement à votre recherche: répertoires classés par auteur, par thème, par type...

Les bases de données offrent exactement le même genre de service mais avec des moyens plus rapides, plus performants et sur des ensembles de documents dont l'étendue n'est aucunement comparable.

En fait, les grands producteurs de répertoires scientifiques sous l'accroissement de la production mondiale de documents et de la demande ne pouvaient plus se contenter des versions papiers de répertoires. Dans les années 60, 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 70, avec la mise au point de la communication des données par télématique, les versions informatisées des grands bulletins signalétiques sont devenues accessibles à tous. Plus récemment, avec l'évolution des progrès des technologies de l'information et de la communication deux autres modes de diffusion des bases de données sont venus compléter les deux précédentes: le vidéotex, le disque optique numérique.

Plusieurs intermédiaires sont concernés par ces modes de diffusion...

A. La distribution commerciale des bases de données

Les divers modes de diffusion des bases de données empruntent plusieurs filières:

⇨ la vente de bibliographies imprimées:
Elle reste encore probablement le principal mode de diffusion des informations.

⇨ la distribution par réseau vidéotex:
Surtout développée en France, elle favorise l'accès aux bases par les utilisateurs novices en permettant la connexion sans abonnement et en aidant la recherche par des menus guidés.

⇨ la distribution par disque optique compact (CD-Rom):
Elle permet à l'utilisateur de disposer d'une base de données sur place (sur son micro-ordinateur), de la consulter sans le stress de l'interrogation en ligne et autant de fois qu'il le veut (le coût étant couvert lors de l'achat).
L'utilisation du CD-Rom suscite quelques inconvénients:

- ❍ bien qu'il puisse contenir 550 méga-octets, il reste d'une capacité faible obligeant le découpage du fonds documentaire par année ou groupe d'années
- ❍ un temps de réponse relativement long
- ❍ le coût est généralement élevé car la plupart des utilisateurs ont un taux de consultation assez faible
- ❍ le délai de mise à jour reste la principale raison du faible succès des CD-Rom, l'utilisateur étant encore obligé de consulter les versions en ligne pour trouver les articles récents.

⇨ la distribution en ligne par les serveurs professionnels:
Les serveurs professionnels ont été conçus pour être manipulés par des utilisateurs avertis et non pas directement par l'utilisateur final de l'information. En prenant en considération ceci, on peut alors distinguer quatre intermédiaires dans la diffusion des informations en ligne:

- ❍ le producteur: collecte, indexe(1) et organise le fonds documentaire
- ❍ le serveur: se charge de la mise à disposition des données via le réseau des télécommunications. Il a aussi un rôle de formation des utilisateurs. La France a investi à ce niveau dans son propre serveur: *Questel* filiale de *France Télécom*
- ❍ le réseau de télécommunication: assure la fiabilité des liaisons entre les clients et le serveur

(1) Le terme *indexer* recouvre deux sens différents. Un sens informatique signifie la création d'un fichier informatique inversé pour permettre une recherche rapide dans un champ d'une base de données. Un sens documentaire correspond à l'analyse de documents pour en dégager les principaux concepts sous forme de mots-clés et sous forme de résumé. C'est sous ce dernier sens que les termes *indexation* ou *indexer* seront employés tout au long de ce chapitre. (voir aussi la partie III.C. de ce chapitre)

- l'utilisateur averti: bibliothécaire ou documentaliste, il doit être formé au langage d'interrogation des serveurs et à la transcription des questions des demandeurs sous forme de requêtes d'interrogations propres aux langages et à la base de données consultée.

Les systèmes documentaires en ligne ont l'avantage de posséder une très grande capacité de stockage, permettant de retrouver, presque immédiatement, une information précise parmi des millions de références (*Chemical Abstracts:* 8,5 millions; *Pascal:* 7 millions...).

B. Les différents types de bases de données

Les informations mises à disposition dans les bases de données ne sont pas exclusivement du type signalement d'article. Plusieurs autres types de documents y sont accessibles. Nous en dénombrons quatre:

⇨ la référence bibliographique:
Références de livres, articles, colloques, brevets ou documents non-publiés (littérature grise, thèse, documents administratifs...)

⇨ le texte intégral:
Il s'agit de la saisie informatique du texte complet de documents originaux. A bien distinguer avec une image numérique de ce texte (sans reconnaissance de caractères); l'image ne permettrait pas d'accéder à chacun des mots du texte pour la recherche.

⇨ les données factuelles:
L'élément de base de ce système d'information, souvent nommé *banque de données*, est constitué par des données que l'utilisateur recherche directement, en général sans faire appel à des descripteurs (annuaires, répertoires, catalogues, données scientifiques, économiques, statistiques ou financières, horaires de transports...).

⇨ les graphiques et images:
Là, le document n'appartient plus au monde textuel. La recherche s'effectue alors sur un ensemble de quelques descripteurs textuels affectés à chaque image.

Le traitement bibliométrique, comme son nom l'indique, a été développé pour examiner les bibliographies. **Les types de bases exploitées en bibliométrie sont donc ceux qui recèlent les documents de la première de ces quatre catégories, bien que les recherches actuelles se dirigent activement vers l'analyse de documents en texte intégral.**

Pour le traitement bibliométrique des références, une des caractéristiques importantes des documents escomptés est la structure sous laquelle ils sont diffusés. La durée des traitements bibliométriques dépend principalement de la rigidité de cette structure. Lorsque les documents collectés sont bien structurés, le temps de traitement devient presque négligeable devant le temps consacré à la validation et à l'analyse des résultats.
Le document structuré comporte des balises permettant de distinguer les diverses parties logiques du document. Ces parties sont nommées *champs*. Chaque champ contient un élément d'information particulier.

A titre d'exemple nous avons choisi une référence brevet de la base *WPIL* du producteur *Derwent* obtenue sur le serveur *Orbit*. Cette table 1 présente à gauche la référence lors de sa visualisation et à droite la signification du contenu de chaque champ.

FORMAT DE RECEPTION	SIGNIFICATION DES RUBRIQUES
-1-	<- Numéro d'édition de la référence
AN - 92-009829/02	<- ACCESSION NUMBER: n° d'entrée de la référence dans la base
TI - Patches for topical or transdermal drug delivery - with adhesive layer contg. polyacrylate adhesive and film former	<- TITLE: Titre du brevet référencé
TT - PATCH TOPICAL TRANSDERMAL DRUG DELIVER ADHESIVE LAYER CONTAIN POLYACRYLATE ADHESIVE FILM FORMER	<- TITLE TERMS: Termes du titre (normalisés par DERWENT)
PR - 90.06.25 90DE-020144	<- PRIORITY NUMBER: Date et numéro de dépôt du brevet
PN - EP-464573-A 92.01.08 (9202) DE4020144-A 92.01.09 (9203)	<- PATENT NUMBER: Numéros et dates des brevets délivrés
AP - 91.06.24 91EP-110409 90.06.25 90DE-020144	<- APPLICATIONS DETAILS: Dates et numéros de publications pendant les procédures
DS - AT BE CH DE DK ES FR GB GR IT LI LU NL SE	<- DESIGNATED STATES: états d'extensions ou de désignations
PA - (LOHM) LTS LOHMANN THERAPI	<- PATENT ASSIGNEE: Organisme déposant
IN - MULLER W,MINDEROP H,TEUBNER A	<- INVENTORS: Inventeurs du brevet
LA - G	<- LANGUAGE: Langue de rédaction du brevet
CT - (G)DE3843238 DE3843239 EP-305758 EP-379933	<- CITED PATENTS: Brevets cités lors de la procédure d'antériorité
IC - A61L-015/16 A61F-013/02 A61M-037/00	<- INTERNATIONAL PATENT CLASSIFICATION: Classif. internationale des brevets
DC - A96 B07 D22 G03 A14 P34 P32	<- DERWENT CLASSES: Classification DERWENT
MC - A04-F06E5 A08-P01 A12-V03A B04-C03B B12-M02F D09-C04B G03-B02D1 G03-B04	<- MANUEL CODES: Deuxième classification DERWENT
AB - (EP-464573) Topical or transdermal patches comprise a backing layer, an adhesive layer and a release liner. The adhesive layer comprises 100 pts.wt. of a polyacrylate adhesive (I), 5-150 pts.wt. of a polyacrylate-compatible film former (II), 0.250 pts.wt. of non-plasticising active agents and/or additives, and 10-250 pts.wt. of plasticising active agents and/or additives. ADVANTAGE - Inclusion of (II)...	<- ABSTRACT TERMS: Résumé du brevet

Table 1: Référence d'un brevet de la base WPIL sous le serveur Orbit

1. Les bases de données scientifiques

Juste pour donner une idée de la couverture thématique des bases de données, nous avons listé la centaine de bases tenue à jour par le serveur anglais *Orbit* (table 2).

Nous n'avons dans cette table qu'une partie des bases scientifiques, techniques et technologiques accessibles en ligne. Le nombre de serveurs de bases de données est considérable; les principaux consultés sont *BRS, Datastar, Dialog, ESA, Questel, Orbit, STN.* Le serveur qui a actuellement le plus grand nombre de bases est *Dialog* avec un nombre dépassant les 400.

Pour donner un ordre de grandeur sur le contenu de ces bases de données nous avons choisi la plus réputée dans le domaine de la chimie: *Chemical Abstracts.* Elle "épluche" un panel de 9000 revues spécialisées, les brevets de 27 pays, les nouveaux livres, les rapports de colloques scientifiques et les rapports de recherches gouvernementaux. L'ensemble des références se répartit de 1967 à nos jours et répertorie plus de 9 millions de références. On peut estimer le nombre de références supplémentaires à 500 000 par an, ce qui correspond à une activité de dépouillement d'environ 10 000 documents par semaine.

Ces quelques chiffres ne pourraient être plus rassurants sur la qualité de l'exhaustivité géographique et temporelle de l'information formalisée en chimie compulsée par *Chemical Abstracts.*

La table 3 donne un exemple de référence bibliographique provenant de la base de donnée *Pascal* sur le serveur français *Questel.*

Les études bibliométriques de la science ont l'habitude de prendre en considération non seulement l'existence d'un document mais aussi son impact sur la communauté scientifique en examinant les citations dont il fait l'objet. Cette information n'est présente que dans trois bases de données, toutes produites par un organisme américain, l'*ISI.* Ces bases sont probablement les plus exploitées en bibliométrie et particulièrement dans les pays anglo-saxons. Elles sont présentées un peu plus loin dans ce chapitre (*Les bases de données de l'ISI*).

Databases By Subject

Business
ABI/INFORM
ACCOUNTANTS
API Energy Business News Index (APIBIZ)
Chemical Economics Handbook
Chemical Industry **Noies**
CorpTech
ENERGYLINE
LABORDOC
Materials Business File
PIRA - Paper, Printing and Publishing, Packaging, and Nonwovens Abstracts
PNI (Pharmaceuiical News Index)
RAPRA Abstracts
World Surface Coatings Abstracts

Chemistry
Analytical Absiracts
Beilstein
Biotechnology Abstracis
Chemical Abstracis
Chemical Abstracts Service Source Index
Chemical Dictionary
Chemical Economics Handbook
Chemical Engineering and Biotechnology Abstracts
Chemical Industry Notes
Chemical Reactions Documentation Service
Chemical Safety NewsBase
ChemQuest
CORROSION
PESTDOC
RINGDOC
Sianaard Drug File
Standard Pesiicide File
VETDOC

Directories
American Men and Women of Science
Chemical Abstracis Service Source Index
ChemQuest
CorpTech
Cuadra Cirectory of Databases
Oirectory of American Research & Technology
National Union Caialcg Codes
Scientific and Technical Books & Serials in Print
Who's Who in Technoiogy

Energy & Earth Sciences
API Energy Business News Index (APIBIZ)
APILIT
APIPAT
Electric Power Industry Abstracts
Energy Bibliography
ENERGYLINE
Geobase
GeoMechanics Abstracts
GEOREF
IPABASE
POWER
Remote Sensing
TULSA (Petroleum Absrracts)
WPIA/WPILA

Engineering
Aqualine
Biotechnology Abstracis
Chemical Abstracts
Chemical Engineering and Biotechnology Abstracts
COLD
COMPENDEX'PLUS
CORROSION
Electronic Publishing Abstracts
GeoMechanics Abstracts
GEOREF
ICONDA
INSPEC
MICROSEARCH
SAE Global Mobility Oatabase
Supertech
Weldasearch

Health, Safety & the Environment
Aqualine
Chemical Safety NewsBase
ENVIROLINE
Food Science and Technology Abstracts
Health and Safety Executive
National Institute for Occupational Safety and Health Technical Information Center
PNI (Pharmaceutical News Index)
Remoie Sensing
Safety Science Abstracts
WasteInfc

Materials Science
Ceramic Abstracts
CORROSION
Engineered Materials Abstracts
Food Science and Technology Abstracts
Imaging Abstracts
Materials Business File
METAOEX
Metals Data File
PIRA - Paper, Printing and Publishing, Packaging, and Nonwovens Abstracts
RAPRA Abstracis
Weldasearch
World Ceramics Abstracts
World Surface Coatings Abstracts

Multidisciplinary
COLD
Electronic Publishing Abstracts
ISTP Search
Japan Technology
Library and Information Science Absiracts
MICROSEARCH
NTIS
SciSearch
Supertech
Tropical Agriculture

Patents & Trademarks
APIPAT
Chinapats
CLAIMS
CLAIMS Classification
CLAIMS Compound Registry
CLAIMS Reassignments
Current Patents - Evaluations
Current Patents - Fast Alerts
Orug Patents International
INPADOC/INPANEW
JAPIO
Legal Status
LitAlert
Patent Status File
RAPRA Trade Names
UK Trademarks
US Classification
US Patents

Table 2: Liste des bases de données accessibles sur le serveur Orbit

```
 4/20330 - (C) CNRS
 NO   : PASCAL 92-0420379 INIST
 FT   : Diagnostic en temps reel par systeme expert. Application a un systeme
        de diagnostic embarque sur automobile
 ET   : (Real-time monitoring using expert systems. Application to a
        diagnosis system on board a vehicle)
 AU   : GERLINGER Gilles; MORIZET MAHOUDEAUX Pierre (dir.)
 DT   : These; LM
 SO   : FRA; DA. 1991-06; 184 p.; ABS. fre/eng; BIBL. 124 ref.; Th. doct. :
        Syst. expert./Universite de Compiegne. FRA/1991/91COMP366S
 LA   : FRE
 FA   : Il existe un interet grandissant a l'heure actuelle pour les systemes
        experts (SEs) dits "temps reel".  A cela, on peut voir deux raisons.
        Tout d'abord un SE est capable, dans des cas ou il n'existe pas de
        solution algorithmique satisfaisante, de resoudre un probleme en un
        temps realiste (polynomial), notamment a l'aide d'heuristiques
        appropriees.  Ensuite, les applications temps reel peuvent beneficier
        de certaines caracteristiques des SEs, particulierement interessantes
        vis-a-vis des contraintes de temps, comme la possibilite de faire
        progresser le travail deja effectue sans le remettre en cause dans sa
        globalite ou la capacite a estimer le sous-espace de recherche le
        plus prometteur en fonction de l'etat du systeme.  Cependant,
        l'integration d'un SE dans une application temps reel souleve bon
        nombre de difficultes.  Globalement, le SE doit satisfaire trois
        exigences principales: 1) l'integration dans l'environnement
        exterieur, c'est-a-dire avec les autres composants ***logiciels*** de
        l'application temps reel; 2) le fonctionnement en temps reel, qui
        recouvre des concepts comme le fonctionnement en continu, la prise en
        compte de donnees asynchrones, la focalisation d'attention,
        l'activation, etc...; 3) la prise en compte du temps dans le
        raisonnement, afin que le SE puisse determiner la position relative
        d'evenements dans le temps.  Tous ces aspects ont ete developpes a
        partir d'un generateur de SEs existant, le systeme SUPER.  La
        demarche suivie est de faire du SE une des pieces d'un systeme temps
        reel conventionnel.  Les points forts concernant par consequent les
        problemes d'interruption en cours de raisonnement, de communication
        et d'interaction entre modules et enfin de temps de reponse.  Une
        application d'un tel systeme a ete realisee dans le cadre du projet
        europeen PROMETHEUS avec la mise au point d'un systeme de diagnostic
        embarque sur vehicule
 CC   : 001D02C; 430A06D
 FD   : Intelligence artificielle; Diagnostic; systeme expert; Temps reel;
        multitache; Multi-agent; Raisonnement temporel
 LO   : INIST-TD 80933.T91COMP366S   0000
```

Table 3: Référence d'un article dans la base Pascal sur le serveur Questel

2. Les bases de données brevets

L'information brevet à travers les bases de données n'est pas aussi simple à maîtriser que l'information en science académique.

Il faut d'abord faire remarquer que la plupart des bases brevets entendent par "document brevet" aussi bien les demandes de brevet que les brevets délivrés. Il faut aussi remarquer qu'il y a coexistence de fichiers dont la finalité est différente, les documents sont donc différents:

⇨ bases de type documentaire:
Les brevets y sont traités pour leurs aspects scientifiques, techniques et technologiques. Par exemple, pour *WPI/L* de *Derwent* la référence renseigne sur l'innovation, donc sur toute la famille que ce soit au moment du dépôt de la priorité, des extensions, des procédures d'examen, ou des accords. Ce producteur a aussi introduit plusieurs systèmes d'indexations par codes pour aider dans la recherche des documents (*Derwent Code, Manuel Code, Chemical Code*...).

⇨ bases de type répertoire:
Principalement gérées par des organismes officiels de dépôt (*INPI, OEB*...), les brevets y sont traités pour leur aspect juridique. Par exemple *FPAT* et *EPAT* vont renseigner sur les modifications de l'état juridique de brevets dans le cas de procédure d'opposition. Dans ces bases, il y a donc une référence pour chaque brevet d'une même famille. Elles comportent peu d'éléments d'indexation pour connaître le fond de l'innovation si ce n'est un ensemble de codes de classification, bien souvent la *Classification Internationale des Brevets*.

⇨ base de type texte intégral:
Met à disposition les textes intégraux des documents. Par exemple, *US Patents* contient les brevets américains depuis 1970 avec le texte intégral de leurs revendications.

⇨ bases de type IST:
Certaines bases qui sont consacrées à l'information scientifique et technique prennent aussi en compte depuis peu l'information brevet concernant leurs spécialités (par exemple *Chemical Abstracts*).

Des travaux en bibliométrie ont repris l'idée de l'impact des documents par la citation pour l'appliquer au document brevet. On peut estimer qu'il existe l'équivalent de la citation scientifique pour les brevets puisque les procédures juridiques, lors d'un dépôt, demandent d'indiquer la liste les documents antérieurs proches de l'innovation. Cette liste sera complétée

lors de la recherche d'antériorité de l'office de dépôt. Mais hélas, ces "citations" sont très peu souvent présentes dans les bases ou alors partiellement renseignées. Par exemple, *Derwent* ne les introduit que depuis 78 et *FPAT* ne fait pas références aux articles scientifiques. *EPAT* et *USPAT* ont ces renseignements mais ces bases n'ont pas une couverture géographique mondiale. Comme pour les bases scientifiques, il existe aux Etats-Unis une société qui a créé une base brevet sur les mêmes principes que ceux de l'*ISI* et qui permet de traiter bibliométriquement les "citations" brevets. Cette base est exploitée uniquement à titre privé. Nous présenterons tout de même cette société, le *CHI*, un peu plus loin (*Les bases de données du CHI*) car ses travaux en bibliométrie sont très intéressants du fait qu'elle travaille sur des documents brevets.

3. Les bases de données de l'ISI

L'institut pour l'information scientifique (*Institute for Scientific Information*), à Philadelphie aux Etats-Unis, doit sa création à Garfield. Il a eu l'idée de constituer un répertoire qui aurait une couverture interdisciplinaire et qui regrouperait uniquement les articles publiés par le coeur des périodiques scientifiques. Ce coeur de revues est déterminé par le taux de citations dont elles font l'objet [GARF79]. Le but de l'ISI est de couvrir les revues les plus prestigieuses à travers le monde dans toutes les branches de la science.

En 1955, il met au point le *Science Citation Index*. 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 la science "molle", sont venus se joindre au SCI: le *SSCI* (*Science Social Citation Index*), publié depuis 1973, suit 1 400 autres périodiques, et depuis 1978, et 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'auteur)
- le *Source Index* (répertoire des publications par noms d'auteur)
- 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 (citations d'articles publiés dans la revue) pendant une année
- le facteur d'impact (I_F) de chaque revue
- le classement des revues par thèmes (128) selon leurs valeurs du I_F
- les 15 revues les plus "citées" par une revue précise
- les 15 revues qui citent le plus une revue précise
- ....

Ce journal est très employé lors d'études bibliométriques de revues scientifiques.

Pratiquement toutes les versions imprimées de ces répertoires (*SCI*, du *SSCI*, et A*&HCI*) sont de nos jours consultables par les réseaux professionnels de télécommunication. Le *JCR* est la seule exception.
Il est à remarquer que la version informatisée du SCI ne couvre plus que 3300 revues ce qui équivaut à environ 10 millions de citations en plus par années. Un exemple d'une référence, provenant de la base SCI, est présenté à la table 4.

Au départ, le but de ces répertoires n'était pas de mesurer les "performances" des chercheurs, des équipes ou des laboratoires mais plutôt d'établir des relations pouvant exister entre les divers travaux de recherches 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 verrons qu'un grand nombre d'études bibliométriques se basent principalement sur les seules données offertes par l'ISI.

Nous verrons plus tard les critiques majeures qui ont été formulées à l'encontre des déficiences de ces bases de données (*SCI, SSCI, A&HCI*) en tant que source d'informations. Elles sont finalement des représentations de la science très tendancieuses car leur mode de sélection d'articles désavantage des pays et des domaines par rapport à leur réel statut.
Par exemple, une étude comparative entre le SCI et la British Library Lending Division a montré que l'ISI défavorisait le Japon et l'Union soviétique, avantageait les Etats-Unis et le Royaume-Uni et traitait convenablement la science française [CARP81].

Au vu des limites des banques ISI pour une étude bibliométrique, les chiffres, qu'elles fournissent, sont à considérer comme des témoignages et non des affirmations. L'analyse des citations ne doit pas être prise comme un but mais plutôt comme le point de départ d'une enquête.

```
-22-
AN  - 9148-001247
GA  - GP388
TI  - CONSTRUCTION OF THE PAULI POTENTIAL, PAULI ENERGY, AND
      EFFECTIVE POTENTIAL FROM THE ELECTRON-DENSITY
AU  - HOLAS A; MARCH NH
OS  - POLISH ACAD SCI,INST PHYS CHEM, PL-01224 WARSAW,  POLAND
      (REPRINT);
      UNIV OXFORD,DEPT THEORET CHEM, OXFORD OX1 3UB,  ENGLAND
SO  - PHYSICAL REVIEW A  V44 (N9) P5521-P5536; 1991
DT  - U (ARTICLE)
LA  - EN (ENGLISH)
FS  - CURRENT CONTENTS/P (PHYSICAL, CHEMICAL & EARTH SCIENCES)
CC  - UI  (PHYSICS); UI  (PHYSICS)
NC  - 31
KP  - GROUND-STATE; DIFFERENTIAL-EQUATION; FUNCTIONAL THEORY;
      KINETIC-ENERGY; SYSTEM; PHASE; ATOM
RF  - 89-0840-3 (DENSITY FUNCTIONAL METHOD; ABINITIO CALCULATION;
      CRYSTALLINE SILICON; QUASIPARTICLE ENERGIES FOR SEMICONDUCTORS)
AB  - The Kohn-Sham (KS) one-electron Schrodinger equations assume
      the existence of a one-body effective potential upsilon-eff(x),
      defined to generate the correct electron density rho(x) of the
      ground state.  This paper returns to the electron-density
      description of an N-fermion system.  It is best thought of as
      starting from a given rho(x), ideally to be obtained from
      diffraction experiments.  A method is then set up that focuses
      predominantly on the way the "correct" upsilon-eff(x) can be
      "recovered," if it exists, from such an experimental density.
      Certainly the method has associated with it one practical
      disadvantage in common with the KS procedure; an order of N
      Euler equations have to be solved, with input information
      rho(x), though the "unknown" potential upsilon-eff(x) does not
      now appear.  In this program, we have found it most helpful to
      work with the Pauli potential and energy, which enable the
      N-fermion problem to be converted to a boson problem for the
      density amplitude [rho(x)]1/2.  The way the above-mentioned
      Euler equations determine the Pauli potential and energy is
      worked out explicitly.  Examples that embrace the important
      area of atomic central-field calculations are presented to
      illustrate the method.  As a by-product, the theory developed
      can afford a direct test as to whether a given electron density
      is, in fact, representable via a one-body potential
      upsilon-eff(x).
CR  - ARYASETIAWAN F, 1988,  PHYS REV B (V38, P2974)
CR  - BALTIN R, 1985,  PHYS LETT A (V113, P121)
CR  - BARTOLOTTI LJ, 1982,  J CHEM PHYS (V77, P4576)
CR  - BROWN PJ, 1972,  PHILOS MAG (V26, P1377)
CR  - DAWSON KA, 1984,  J CHEM PHYS (V81, P5850)
CR  - DIRAC PAM, 1930,  P CAMBRIDGE PHIL SOC (V26, P376)
CR  - FERMI E, 1928,  Z PHYS (V48, P73)
CR  - FOCK V, 1930,  Z PHYS (V15, P126)
CR  - HARRIMAN J, 1984,  LOCAL DENSITY APPROX
CR  - HARTREE DR, 1927,  P CAMBRIDGE PHILOS S (V24, P111)
CR  - HERRING C, 1988,  PHYS REV A (V37, P31)
CR  - HOHENBERG P, 1964,  PHYS REV (V136, PB864)
CR  - HUNTER G, 1986,  INT J QUANTUM CHEM (V29, P197)
CR  - KOHN W, 1965,  J PHYS REV (V140, P1133)
CR  - KOZLOWSKI PM, 1989,  INT J QUANTUM CHEM (V36, P741)
CR  - LASSETTRE EN, 1985,  J CHEM PHYS (V83, P1709)
CR  - LAWES GP, 1979,  J CHEM PHYS (V71, P1007)
CR  - LEVY M, 1982,  PHYS REV A (V26, P1200)
CR  - LEVY M, 1984,  PHYS REV A (V30, P2745)
CR  - MARCH NH, 1959,  NUCL PHYS (V12, P237)
```

Table 4: Référence d'un article dans la base SCI sur le serveur Orbit

4. Les bases de données du CHI

Le SCI a été conçu pour la recherche d'information bibliographique et non pour le comptage rigoureux des publications et des citations. C'est pourquoi, au début des années 70, la *National Science Foundation* américaine a demandé à une petite firme de Philadelphie, la *Computer Horizon Incorporated* (*CHI*), dirigée par Francis Narin, de mettre son index sous une forme qui le rende utilisable pour la production d'indicateurs.

Ce "nettoyage" a nécessité un travail considérable: normalisations 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 la base, soit un peu plus de deux mille cents 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 également en mathématiques que ce décalage est le plus marqué. La représentation de la science, par cette base, 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 base CHI. Plus on s'intéresse à des disciplines étroitement définies et récentes, plus ces défauts deviennent rédhibitoires.

L'Advisory Board for the Research Councils, en Grande-Bretagne, a réalisé une comparaison systématique dans le domaine de la physique, entre la base CHI et la base des *Physics Abstracts* gérée par une grande société savante américaine. Pour l'ensemble du domaine, certains écarts, au total peu significatif, existent. Mais dans quelques spécialités, les discordances sont telles que les tendances s'inversent parfois [ADIS86].

Il faut donc considérer que ces bases peuvent fournir simplement des données pour des analyses de première approximation de l'état général de la science d'un pays.

Outre la restructuration et l'exploitation de la base de l'ISI, le CHI a constitué sa propre base bibliométrique de brevet. 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 ayant un aspect technologique.

Elle contient les données de 59 pays reparties selon plusieurs plans de classement:

- en 376 classes de brevets de l'office américain des brevets
- en 57 classes de produits nommées *US Standard Industrial Classes* (SIC). Ces classes SIC comportent aussi des sous-classes: environ 100 par classe soit plus de 700 000 sous classes au total. Le regroupement dans les classes et sous-classes SIC est réalisé par l'*Office of Technology Assessment of the US Patent Office* qui prévient que des erreurs d'affectation existent.
- en regroupant les brevets par similarité technologique. On retrouve par exemple les brevets dans des classes telles que la "pharmacie", la "chimie" ou la "technologie de l'information".

Il faut noter qu'il n'y a aucune correspondance entre les trois classifications.

Cette base brevet a surtout comme caractéristique, intéressante en bibliométrie, de saisir les informations concernant les "citations" de brevets ou d'articles scientifiques antérieurs.

Le CHI a aussi comme activité la publication du *Science Indicators Reports* depuis 1972. Les chercheurs du CHI ont développé de nombreux indicateurs bibliométriques à partir de leurs bases. Cette revue est un répertoire de ces indicateurs.

Ces bases qui présentent une certaine qualité au niveau des facilités de traitement bibliométrique ne sont exploitées qu'à titre privé. Par contre, le CHI propose de réaliser des études à la demande.

C. Les champs indexés des bases de données

Comme pour les bibliothèques, le nombre considérable de références dans une base de données a incité le développement des procédés pour la recherche. Ainsi, les références bibliographiques ne sont pas de simples signalements de documents originaux.

Une partie des champs d'une référence comporte les informations du contexte dans lequel le document original apparaît: le titre, les auteurs, la source, la date de publication, le nombre de pages, la langue de l'écrit... Ces champs ne produisent pas de nouvelles informations par rapport au document de départ.

Par contre, un autre ensemble de champs présente un supplément d'information. Cette création d'information se trouve répartie dans des champs comme le résumé, les mots-clés, les classifications documentaires... **Ce sont des informations qui ont pour objet de décrire analytiquement le contenu du document original. Ce traitement d'analyse et d'affectation de termes au document pour mieux le décrire se nomme en langage documentaire l'*indexation*.**

Il faut faire très attention à la finalité de celle-ci. Elle n'est pas réalisée dans le souci de condenser l'information mais elle a pour objectif d'améliorer la communication. **Les descripteurs associés aux documents par l'indexation ne sont là que pour permettre de le retrouver. Lors d'études bibliométriques exploitant les informations de ces champs, cette notion doit être constamment présente à l'esprit pour ne pas les interpréter incorrectement.**

On peut se référer à la référence présentée par la table 3 pour retrouver ces deux catégories de champs.

1. Les catégories d'indexations

Les descripteurs sont en général des formes nominales ou des codes:

⇨ <u>des mots-clés unitermes:</u>
ce système a l'inconvénient de mélanger des sens distincts (la requête "*droit* et *travail*" trouvera des articles sur le "*droit du travail*" et sur le "*droit au travail*")

⇨ <u>des descripteurs composés</u> (2 à 3 termes)

⇨ <u>des codes de classification:</u>
les descripteurs peuvent être codés pour représenter simplement des notions conceptuelles difficiles à réduire à quelques mots. Les classifications ou plans de classement permettent de situer un document dans la partie de la connaissance traitée par le système documentaire. On trouve des classifications hiérarchiques (ex: la *CIB*), des classifications non-hiérarchiques (ex: les *manuel-codes* de *Derwent*, les *section-codes* de *CA*, les *concept-codes* de *Biological Abstracts*).

Pour les indexations par termes (uni- ou composés) on distingue trois natures d'indexation manuelle:

- une indexation libre:
  l'indexeur n'est soumis à aucune contrainte et s'efforce de respecter des règles générales (nombre, organisation des informations...)

- une indexation contrôlée:
  les termes d'indexation sont obligatoirement choisis dans une liste préétablie. Cette liste peut être une simple énumération de termes classés alphabétiquement (listes d'autorité) ou bien elle peut être structurée selon une organisation définissant des relations sémantiques entre les termes (thesaurus).

Il est souvent reproché à l'indexation d'être une opération fondamentalement inconsistante et aléatoire. H Le Crosnier dans [LECR90] exprime ce phénomène en ces termes:

> *"Indexer, c'est attribuer à un document des descripteurs en fonction de critères implicites (aspect communication), avec une large dose d'incertitude (dans quelle mesure le terme choisi est discriminant?) et un coefficient de confiance dans l'opération fort limité (aurait-on adopté le même choix demain?)."*

Yves Courrier a étudié dans [COUR76] l'opération d'indexation, il en dit ceci:

> *"Deux indexeurs choisiront pour le même document très peu de descripteurs identiques parfois moins de 50%. Ce sera aussi le cas pour un même indexeur à deux périodes différentes."*

Cette seconde caractéristique devra aussi être bien connue lors de traitements bibliométriques sur des champs indexés. **La qualité d'indexation des bases de données peut être un critère décisif au moment du choix des sources pour la collecte des références.**

2. L'enrichissement de l'indexation

Certaines bases de données, toujours dans le souci de perfectionner la performance du système de recherche, usent de certains procédés.

⇨ l'autopostage:

Dans le cas d'une indexation par thesaurus ou par classification hiérarchique, le système informatique peut avoir été conçu pour compléter le travail de l'indexeur. Lorsque l'indexeur affecte un terme spécifique à un document, le système rajoutera automatiquement à ce document le ou les descripteurs génériques situés aux niveaux supérieurs de la hiérarchie. Ainsi le document pourra être extrait lors d'une interrogation plus large. Un tel processus améliore donc le taux de rappel du système documentaire (nombre de documents pertinents retrouvés par rapport au nombre de documents existants).

⇨ la pondération:

L'utilisation de la pondération dérive de l'indexation par des méthodes statistiques (voir dans le chapitre suivant *Le traitement du langage naturel*). Dans les bases de données en ligne il n'existe pas de réelles pondérations des descripteurs par une valeur statistique de leurs poids pour le document où ils sont affectés. Mais certaines proposent une pondération simplifiée qui partage les descripteurs en *descripteurs principaux* et *descripteurs secondaires.*

La pondération doit permettre d'augmenter le taux de pertinence du système documentaire (nombre de documents pertinents retrouvés par rapport au nombre de documents extraits)

ex: *Chemical Abstracts* pondère en deux niveaux de code son champ *section-codes*. Il positionne le code correspondant au thème principal du document en première position (*Principal-code*). Ensuite, les autres codes sont tous aux mêmes niveaux de pondération quels que soient leurs emplacements. Ceux-ci sont nommés les codes secondaires (*Cross-codes*).

```
CC  - SEC31-5; SEC1; SEC22
```

La base *INSPEC* utilise aussi une pondération à deux niveaux de poids pour sa classification (*Physics Abstracts*). Les codes des thèmes principaux sont marqués d'une étoile.

```
CC  - *B1265; B1130B; *C7410D; C6140D; C5210B
```

⇨ les indicateurs de rôles

Ces indicateurs sont destinés à préciser le type de relation qui unit les descripteurs. Les descripteurs ne sont donc pas tous au même niveau significatif puisque selon leur position ils seront soit descripteurs principaux soit descripteurs qualificatifs de ces descripteurs principaux (c'est-à-dire qu'ils qualifient le contexte dans lequel le descripteur principal associé est à considérer).

La technique des indicateurs de rôle permet de diminuer le taux de bruit du système documentaire (nombre de documents non pertinents extraits par rapport au nombre de documents extraits).

ex: le champ IT de WSCA avec 2 niveaux de qualification et une forme d'autopostage

```
IT  - Optical Fibres: coated, ultraviolet-curables, acrylates/stabilisers
      (heat); Ultraviolet-curable Coatings: optical fibres,
      acrylates/stabilisers (heat); Stabilisers, Heat: ultraviolet-
      curables, piperidine (hindered)/phenols (hindered); Piperidine:
      groups, hindered, stabilisers (heat), ultraviolet-curables;
      Acrylic/Urethanes: ultraviolet-curables, stabilisers (heat) &
```

Il faudra parfaitement savoir, lors d'une étude bibliométrique, par quel procédé d'indexation un champ est renseigné. La méconnaissance des conditions de création de chacun des champs de la base pourrait entraîner l'utilisation de traitements bibliométriques inadéquats. Les résultats des analyses seraient alors lourds de conséquences; l'information en sortie serait fortement biaisée.

La bibliométrie: méthode d'évaluation des sciences et des techniques

IV. La bibliométrie: méthode d'évaluation des sciences et des techniques

Nous allons, tout au long de ce chapitre, présenter les différentes méthodes de traitements bibliométriques qui ont été développées jusqu'à nos jours. Le résultat de cet exercice n'est certainement pas exhaustif mais les principales grandes écoles de pensées y sont décrites.

Ces descriptions de méthodes s'étendent plus particulièrement sur les successions de traitements impliqués par leurs mises en oeuvre. Les raisons théoriques qui ont permis de les étayer et les critiques de leur validité ne seront pas toujours abordées. Cette présentation a simplement pour objet de mieux faire comprendre les fondements communs à toutes ces méthodes.

A. Concept

On peut dire que la bibliométrie est fondée sur deux postulats:

- ☞ Le premier postulat serait celui-ci:
  Une publication 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 du chercheur 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 le chercheur doit dépeindre dans son argumentation.

- ☞ Un second postulat voudrait que:
  L'activité de recherche soit la confrontation de connaissances acquises par des travaux émanant d'autres auteurs avec les propres réflexions du chercheur. **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 l'acceptation de ces deux postulats, l'étude des publications permettrait d'appréhender les connaissances, leurs structures suivant les écoles de pensées et leurs évolutions.

Ces postulats, qui ont été définis initialement pour la recherche scientifique, ont ensuite été admis pour les formes de communication écrite des connaissances techniques ou technologiques.

S'appuyant sur ces deux postulats, la caractéristique de la bibliométrie est d'établir des études de publications sur des données quantitatives et non plus simplement subjectives (avis des pairs). Ces données quantitatives sont calculées à partir de comptages statistiques de publications ou d'éléments extraits de ces publications.

B. Définition

La bibliométrie est un terme générique qui décrit une série de techniques qui cherchent à quantifier les processus de la communication écrite.

Les auteurs attribuent l'invention du terme ***bibliométrie*** soit à A Pritchard [PRIT69] soit à P Otlet [ESTI69].
Pritchard suggérait de remplacer le terme ***bibliographie statistique*** (*statistical bibliography*) employé depuis 1923 lors de sa première utilisation par E. Wyndham Hulme [HULM23]. *Bibliographie statistique* pouvait prêter à confusion et être interprétée comme une bibliographie sur la statistique. De plus, le terme bibliométrie se rapproche du même coup de termes établis comme *biométrie*, *économétrie* ou *scientométrie*.
Il a défini 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 qu'en a donné Raising en 62 [RAIS62] alors qu'elle était toujours connue 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"*

Depuis la bibliométrie a fortement repoussé son application au-delà des frontières de la bibliothéconomie. Hawkins [HAWK77] plus récemment définit la bibliométrie comme *"les analyses quantitatives des caractéristiques bibliographiques d'un corps de littérature"*. Mais

par cette définition, une des activités de la bibliométrie n'est pas prise en compte: l'études de la circulation des données.

L'ancien concept définit l'unité fondamentale des analyses bibliométriques des comptages statistiques comme étant le document. Encore de nos jours, certains auteurs restent sur ces positions [BROA87]. Ils estiment que les analyses bibliométriques, qui traitent de données non plus déterminées sur cette unité mais sur un découpage du contenu de cette unité, s'éloignent de la bibliométrie pour se rapprocher de l'analyse linguistique statistique.

Un autre terme, la ***scientométrie***, est apparu parallèlement à celui de la *bibliométrie*. Il est originaire d'un terme Russe signifiant l'application de méthodes quantitatives pour l'histoire des sciences (Dobrov & Korennoi 1969 [DOBR69]). Les auteurs ont donc voulu distinguer les deux concepts précédents par ces deux termes. Récemment Brookes [BROO88] a précisé de nouveau cette distinction. Tandis que la bibliométrie aurait pour objet étudié les livres ou les revues scientifiques et pour objectif les activités de communication de l'information. La scientométrie aurait pour objet les 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.

Donc, la bibliométrie regrouperait les méthodes pour aider 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.

Un troisième terme l'***infométrie*** serait le terme générique embrassant *bibliométrie* et *scientométrie*. Sa définition est bien plus large: l'infométrie est l'application des modèles et des méthodes mathématiques et statistiques de façon à dégager des lois relatives à l'information scientifique et technique.

Pour notre part, nous aurons une approche plus pragmatique de la définition de la bibliométrie. Cette définition rejoint beaucoup celle donnée par White et Mac Cain dans [WHIT89]:
La bibliométrie est simplement l'application de méthodes statistiques ou mathématiques sur des ensembles de références bibliographiques.

Cette définition permet d'intégrer l'ensemble des traitements cités par les précédentes définitions. Ces traitements considèrent les objets étudiés comme étant les signalements des références ou comme étant des éléments extraits de ces références.

La scientométrie serait alors une conception englobant la bibliométrie comme un des outils pour permettre de dresser des bilans de santé d'un système de recherche. Les études scientométriques prennent en compte dans leurs analyses d'autres facteurs comme les ressources et la façon dont ces ressources sont transformées en publications [TURN90] ou comme des phénomènes sociologiques. La scientométrie s'introduit plus dans un contexte d'évaluation de la science pour renseigner les instances politiques de recherches nationales ou locales.

Les différentes spécialités que l'on trouve en bibliométrie peuvent être découpées selon cette liste:

- modélisation de distributions bibliométriques (lois Bradford, Lotka et Zipf et des notions sur l'avantage du cumul)
- indicateurs univariés (mesures purement quantitatives basées sur des calculs de ratio)
- indicateurs relationnels (analyses statistiques descriptives des relations entre les éléments étudiés; donnent des indications plus qualitatives)
- analyse bibliométrique des brevets (application des méthodes bibliométriques aux références brevets)
- modélisation mathématique de la circulation des livres (lois sur la diffusion et la communication des ouvrages)

Les quatre premiers points sont abordés dans ce mémoire. Le dernier ne sera pas évoqué car il ne s'applique essentiellement qu'à des activités bibliothéconomiques et par conséquent il ne peut être d'aucune utilité dans un système de veille technologique.

C. Bref historique

La première étude bibliométrique est supposée avoir été faite par Cole et Eales en 1917 [COLE17]. Ils avaient réalisé une analyse statistique de la littérature de l'anatomie publiée entre 1(50 et 1860 pour montrer les fluctuations d'intérêt pendant cette période.

Dix ans après, une étude (la troisième chronologiquement) fut la première analyse des citations. Gross & Gross [GROS27] comptabilisèrent les citations présentes dans les bibliographies d'articles de journaux en chimie, puis rangèrent les journaux dans l'ordre du nombre de citations reçues. Ils constituèrent ainsi une liste de journaux en chimie qu'ils considéraient comme indispensable.

La théorie, qui aura été à l'origine d'un grand nombre d'articles en bibliométrie, a été décrite un peu plus tard. Cette théorie est maintenant connue sous le nom de *loi de Bradford*. Elle a été proposée par l'anglais Bradford en 1934 [BRAD34] mais l'article n'a pas beaucoup fait d'émules à cette époque. Il l'a repris en 1948 dans son livre *Documentation*. Depuis c'est certainement l'article qui a fait couler le plus d'encre dans la discipline.

A la suite de cette théorie, la plupart des travaux étaient consacrés à la recherche de formules mathématiques qui s'ajustent le mieux possible aux données bibliographiques. Certains auteurs se sont aussi penchés sur la formulation de relations mathématiques permettant de relier différentes lois vérifiant un même concept de répartition (Lotka, Zipf, Pareto, Mandelbrot, Weber…). L'objectif ambitieux était de découvrir la relation unifiant ces lois sous une loi universelle.

Aux Etats Unis, sous l'impulsion de sociologues, d'autres courants de pensées se sont développés au cours des années 60 et 70. De Solla Price a été l'un des plus fervents artisans de l'explosion de la discipline pendant cette période.
La création, à Philadelphie au début des années 60 par E. Garfield, de l'*Institute for Scientific Information* (ISI) a permis aux travaux en bibliométrie de prendre un nouvel essor. Une nouvelle école de pensée, fondée sur l'étude des citations, se greffe autour de cet institut. Dans ce mouvement, les premières recherches sur les réseaux de citations se développent: méthode du couplage de Kessler, analyse des co-citations de Small…
Au début des années 70, toujours autour de l'ISI et à partir de volontés nationales, la bibliométrie étend son domaine d'application aux technologies grâce à la fondation du Computer Horizons Inc. (CHI).

Avec l'apparition des bases de données informatisées, la bibliométrie prend une nouvelle ampleur et l'on voit apparaître de nouveaux courants de pensées dans de nombreux pays.
Pour l'Europe de l'est, la Hongrie est le pays phare avec Braün, Schubert, Glänzel; Todorov étant lui de Bulgarie. La création de la revue Scientometrics en 1977 est pour beaucoup dans cette reconnaissance de l'activité bibliométrique dans ces pays.
En Europe de l'ouest, une prédominance anglo-saxonne persiste particulièrement en Hollande avec des auteurs comme Leydesdorff, Van Raan, Tijssen, Leeuw, Rip et Law. La Belgique, en la présence de Egghe et Rousseau, garde une recherche essentiellement axée sur les distributions bibliométriques.

Des pays comme le Japon et l'Inde ont aussi une recherche très active dans le domaine de la bibliométrie.

La France ne détient qu'une place assez faible à l'échelle internationale. Il y n'a pas de grand centre de recherche mais uniquement des petites cellules parfois constituées uniquement d'un seul individu:

- **Centre d'Etudes Techniques en Industrie Mécanique** (Devalan, Belot, Dumas):
  Le CETIM utilise la bibliométrie pour monter des dossiers technologie/marché pour des PME-PMI [DEVA91].

- **Centre de DOCumentation des ARmées** (Paoli):
  Outre l'activité de serveur des bases de données spécialisées en physique, le CEDOCAR est le promoteur du développement d'une station de travail bibliométrique [PAOL92].

- **Centre de Recherche Rétrospective de Marseille** (Dou, Hassanaly, Quoniam, La Tela):
  Formation universitaire DEA Veille Technologique et DESS d'IST, centre de recherche et de développement de méthodes et de logiciels bibliométriques, leurs insertions dans la veille technologique, conseils, et formations industrielles [DOU92].

- **Centre d'Etude et de REcherche en Science de l'Information/CNRS** (Turner):
  Le CERESI développe de manière générale des logiciels d'IST dont les applications vont de l'aide à la documentation jusqu'à la bibliométrie.

- **Centre Européen Scientifique de Mathématiques APpliquées** (Bédéccarax, Huot):
  Le CESMAP met en application l'Analyse Relationnelle dans des études bibliométriques pour des prestations de services auprès des industriels [BEDE92].

- **Centre de Sociologie de l'Innovation** (Callon, Courtial, Penan):
  Le CSI axe ses recherches sur l'aspect socio-cognitif de la science et développement des outils bibliométriques pouvant répondre à ces considérations [COUR90].

- **Institut National d'Information Scientifique et Technique/CNRS** (Polanco, Ducloy)
  Centre de documentation du CNRS, producteurs de bases de données (Pascal, Francis accessible sous Questel) et autres services. Dans le domaine de la bibliométrie, le DRPN de l'INIST conçoit une boîte à outils qu'ils envisagent d'installer sur une station de travail d'IST [DUCL92].

- **INSSIB** (Lafouge, Boucher)
  Formation: école nationale des conservateurs, DEA, DESS. T Lafouge est spécialisé dans la circulation des ouvrages [LAFO91] et le Professeur Boucher en lexicographie.

- **Laboratoire d'Evaluation et de Prospectives Internationales** (Miquel), **CERCOA/CNRS** (Gilbert), **Laboratoire d'information chimique et biologique du muséum national d'histoire naturelle** (Doré):
  Ces trois chercheurs mènent des travaux en commun et appliquent les méthodes bibliométriques pour des données en chimie, pharmacologie… [DORE87]
- **Observatoire des Sciences et des Technologies** (Barré)
  L'OST développe des indicateurs bibliométriques macro-économiques à l'échelle nationale à partir de données scientifiques, techniques, technologiques, économiques et sociologiques [BARR92].

Signalons aussi certaines unités de l'INRA et de l'INSERM qui utilisent la bibliométrie comme aide à l'analyse et à la programmation de la recherche.

D. La mesure de la science

Dans le terme bibliométrie, le suffixe "**métrie**" renvoie aussi bien à la ***mesure*** qu'à la ***métrique***.

La *mesure* est l'évaluation d'une grandeur par comparaison à une unité considérée comme étalon.

La *métrique* implique la création d'une convention qui permette de définir des "distances" entre l'ensemble des éléments étudiés.

La bibliométrie s'inscrit tout à fait dans ces deux concepts.
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. Le classement et la comparaison des éléments les uns par rapport aux autres, selon cette dimension, sont alors possibles.
Par contre, le concept de ***métrique*** est plus spécifique aux ***indicateurs relationnels***. 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 alors chercheront à disposer les éléments selon des calculs de "distances" qui devront estimer les degrés de ressemblance ou de dissemblance entre les éléments.

La bibliométrie est donc un outil de "mesure" auquel on fait appel pour aider à la comparaison et à la compréhension d'un ensemble d'éléments bibliographiques.

Comme White et Mac Cain l'ont si bien dit dans [WHIT89]: **la bibliométrie est aux publications ce que la démographie est aux populations**. Le *bibliométricien* exploite statistiquement des signalements bibliographiques comme le démographe étudie les populations: il n'est pas censé avoir lu les publications qu'il catégorise et comptabilise comme le démographe n'est pas sensé connaître les individus qu'il étudie. Heureusement pour lui, puisqu'il ne pourrait pas, de toute évidence, lire dans des temps raisonnables les ensembles de documents qu'il analyse!

1. Développement de la science

Price dans les années 60 suggéra que la courbe d'évolution des connaissances en science devait suivre une loi. Il étudia l'accumulation des écrits scientifiques. Il voulait montrer que la science se développe selon une ***courbe logistique***. Cette courbe est à l'origine le modèle appliqué au suivi du tonnage d'un minerai extrait d'un gisement naturel, de son démarrage à son épuisement. Elle s'applique aussi parfaitement à l'évolution au cours du temps de la production industrielle d'un produit résultant d'une innovation.

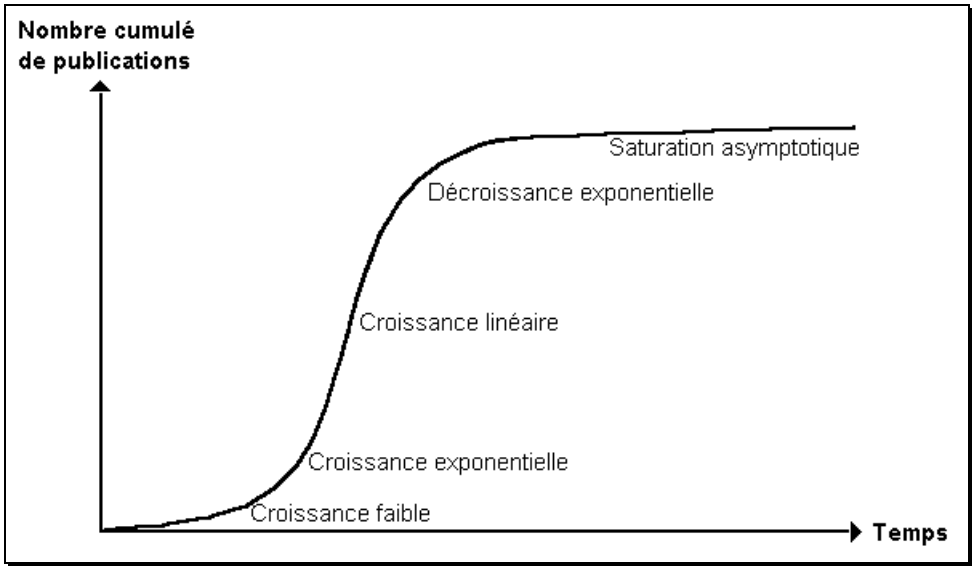


Figure 9: Courbe logistique de croissance de la science

Price comparait ainsi la connaissance scientifique à un élément naturel qui viendrait à s'épuiser avec le temps. Son étude avait prévu pour 1950 le point d'inflexion de la courbe vers le déclin. Les travaux de Tague en 1981 [TAGU81] montrent que nous serions bien dans une phase de croissance linéaire du nombre cumulé de publications scientifiques avec le temps, toutes sciences confondues. Mais rien ne nous a prouvés jusqu'à présent la validité d'une telle thèse au niveau de la connaissance mondiale.

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 dans [ASHO92]. Ils ont modélisé la courbe des publications dans le domaine du laser à l'échelle mondiale et à l'échelle de l'Inde. Ils sont partis de la formule suivante donnée par Sternman dans [STER85]:

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

où

. E(t) est la fonction qui représente l'effort de la communauté contribuant à l'évolution des connaissances dans un domaine.

. W est le nombre maximum de publications que cette communauté peut produire avant que la connaissance du domaine soit épuisée

. w(t) nombre de publications déjà parues

donc W - w(t) est le nombre de publications qu'il reste à publier 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 [BASS69]. 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.

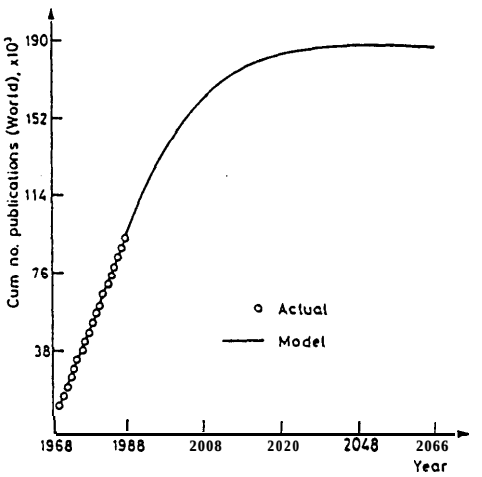
Dans l'article [ASHO92] la forme discrète de la formule:

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

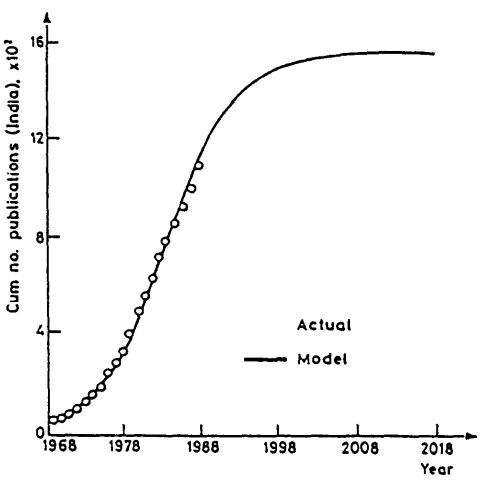
est ajustée aux données de la recherche mondiale et indienne. Les valeurs trouvées pour les paramètres sont reprises dans le tableau suivant:

	Mondiale	Indienne
p	$1{,}72.10^{-2} \pm 10^{-3}$	$5{,}79.10^{-3} \pm 10^{-3}$
q	$7{,}01.10^{-2} \pm 5.10^{-3}$	$18{,}2.10^{-2} \pm 1{,}9.10^{-3}$
W	175.383 ± 1864	1.698 ± 105

Les ajustements du modèle, réalisés à partir des points expérimentaux, sont présentés par les courbes de la figure 10



Model projections of world laser research output. Continuous line indicates model estimates circlcs denote the observed values. The model curve hrs a point of inflexion at the year 1983 reaches 90% of the saturation (maximum possible output) value around the year 2010.



Model projections of Indian laser rcscarch output. Continuous line indicates modelestimates and circlcs dcnote the observed values. The model curve hrs a point of inflexion at the year 198.5 and reaches 90% of the saturation (maximum possible output) value around the year 1990.

Figure 10 : Modélisation de l'évolution scientifique dans le secteur du laser

2. Le "coeur" et la "dispersion"

Comme on l'a défini plus haut, la bibliométrie est l'étude d'un ensemble de références bibliographiques par des méthodes mathématiques. **L'objectif d'une bibliographie est le recensement de tous les textes concernant un sujet donné. Cet ensemble de textes, par l'étude qui va en être faite en bibliométrie, est assimilable à un *corpus*.** Ce terme est employé en linguistique ou en lexicographie pour désigner l'ensemble des discours étudiés.

Dans ce sens, la bibliométrie va étudier non pas uniquement les signalements des références mais les renseignements contenus dans ces références. Ces renseignements sont évoqués, sous forme écrite, par divers types d'éléments: mots du titre, noms des auteurs, affiliation des auteurs, journal qui a publié… **Ces éléments suivent des modèles de *concentration* et de *dispersion* qui peuvent être présentés comme des *distributions statistiques***.

Ces modèles sont souvent symbolisés par "**un coeur et une dispersion**" (*core and scatter*). Cette appellation est due à l'apparence que les éléments prennent quand ils sont rangés selon un ordre de fréquence d'apparition:

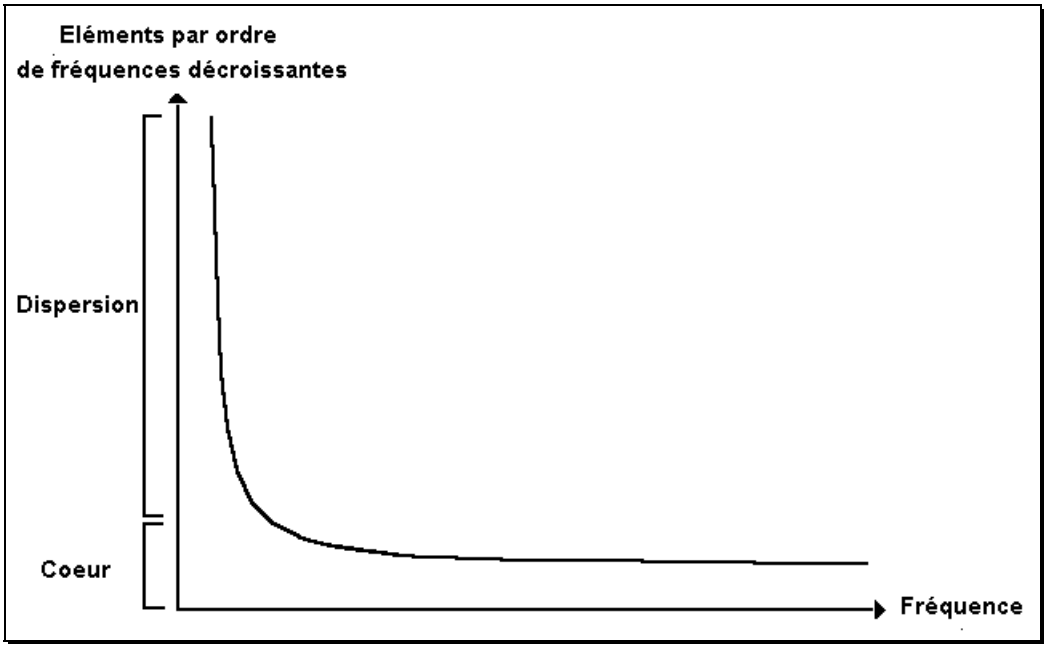


Figure 11: Coeur et de dispersion d'une distribution

☞ le "**coeur**" représente le groupe des éléments qui apparaissent le plus fréquemment dans le corpus. Dans l'absolu, c'est-à-dire pour la littérature dans son ensemble, ce sont les éléments qui co-apparaissent le plus avec le thème du sujet, c'est-à-dire avec les principaux termes décrivant le sujet.

☞ la "**dispersion**" représente les nombreux autres éléments à basse fréquence dans le corpus, donc ceux qui co-apparaissent très peu avec le sujet.

Ces modèles ont été largement étudiés pour certains éléments. Dans le paragraphe suivant nous exposerons en détail ces études ainsi que les modélisations que l'on en a fait: *loi de Bradford*, *loi de Lotka*. La loi de Bradford se focalise sur le coeur des journaux qui ont le plus d'articles sur un sujet donné. La loi de Lotka cherche à établir le coeur des auteurs qui publient le plus sur un sujet. Dans le premier cas les éléments considérés sont les revues scientifiques tandis que dans le second cas ce sont les auteurs.

Bien qu'elle n'ait pas 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 et de Lotka, la répartition d'un type d'élément est étudiée pour un sujet précis. Ces lois mettent en évidence la distribution des cooccurrences entre les éléments et le sujet. Les exemples qui vont suivre présentent des systèmes de concentrations qui ne sont pas établis par le choix d'un sujet. Ils montrent que les distributions présentant un coeur et une dispersion sont des phénomènes caractéristiques de la plupart des données bibliographiques.

⇨ Les journaux concentrent les termes des titres:

De façon générale, le coeur des termes d'un titre est induit par celui qui les écrits: pour un auteur le thème de prédilection, pour un journal les sujets principaux, pour un thème les descripteurs du thème ou les descripteurs des thèmes connexes.

Etude de Paisley [PAIS86] en 1986: pour 300 articles venant de 6 journaux, les 5 premiers termes sont pour:

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 termes du nom du journal et les termes des titres est remarquable.

⇨ Les descripteurs concentrent les termes du titre:

Etude de Lawson et al. [LAWS80] sur le thème de l'analyse énergétique.

Parmi 349 publications, les termes des titres les plus présents sont:

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

⇨ Les auteurs cités concentrent les journaux:

White dans [WHIT81], cherchant sur le sujet "Prehistoric great basin ecology", trouva 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 antropological 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 citation, Lawani [LAWA82] a trouvé que les citations des auteurs étaient le plus souvent les auto-citations, les citations de collaborateurs, les citations cachant une relation étudiant-enseignant et les citations des co-auteurs de l'article. 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 journaux concentrent les journaux cités:

Garfield dans [GARF79] a remarqué que les auteurs qui publient dans un journal ont tendance à citer des articles du même journal. 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 dans [WHIT87] ont comptabilisé les descripteurs des articles co-cités pour 18 concepts en science de la connaissance médicinale. Leur bibliographie étant faite sous 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 behavior	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* (ISI). Small et Griffith [SMAL74] donnent une illustration de cette concentration par l'étude des termes des titres pour les documents qui citent un livre de Lederer C M sur la physique du nucléaire et des particules. Les principaux termes sont:

decay	12
reactions	7
isotopes	5
neutronactivation	5

⇨ Les descripteurs concentrent les descripteurs:

Pour un descripteur qui apparaît au moins dans trois articles dans une base de données en ligne, 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 [MART83] a décrit une recherche sur la base INSPEC des articles sur le sujet *growth of crystals under weightlessness* en commençant uniquement sa requête par les descripteurs *gravity* et *crystal(s)*. 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 dans [WHIT89] a récapitulé les relations de concentrations par la table suivante et il a précisé qu'elle était probablement non exhaustive:

- ⇨ les auteurs prolifiques concentrent
  - ❍ leurs propres termes de titres
  - ❍ leurs propres auteurs cités
  - ❍ leurs propres journaux auquel 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 journaux
  - ❍ les descripteurs
- ⇨ les références fortement citées concentrent
  - ❍ les auteurs des citations
  - ❍ les titres des citations
  - ❍ les références co-citées
  - ❍ les descripteurs
- ⇨ les journaux 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 journaux
  - ❍ d'autres descripteurs

Quels sens donner aux distributions présentant un coeur et une dispersion:

Les exemples qui viennent d'être cités illustrent bien les significations que les auteurs ont données à ces deux zones:

- ☞ **le coeur entretient l'identité, la redondance**
- ☞ **la dispersion contient l'individualisation, la variété.**

Dans le premier exemple on a vu que les termes titres des articles d'un journal contiennent principalement les termes du nom du journal. Ceci indique bien la volonté d'entretenir l'identité du journal à travers les titres des articles qui y sont publiés que ce soit par les auteurs ou que ce soit par l'éditeur.
Un autre exemple caractéristique est celui des auteurs des références citées par un auteur. Un auteur prolifique cite principalement ses propres travaux, ce qui vérifie parfaitement que le coeur entretient l'identité. De plus, les autres auteurs qui sont fréquemment cités ont tous un lien social avec l'auteur. Ils entretiennent là encore une identité sociale. Moed et Van Raan ont même précisé dans [MOED86] que les auteurs cités qui font partie de la dispersion sont souvent des personnes 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, Nelson et Tague suivis de Price et de Weinberg. Ils ont donné deux explications opposées:

- **l'avantage du cumul** (*cumulative advantage*): plus un mot est à forte fréquence plus il sera réutilisé souvent
- **la spécialisation**: plus un mot 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.

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 encore pas bien compris.

Les collèges invisibles montrent 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:

- celui qui crée 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'un journal ou d'une certaine terminologie.
- celui qui crée l'effet de dispersion en balayant continuellement les nouveautés et les différentes personnes et idées ce qui crée dans ce cas 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.

Les études de CRANE étaient dirigées vers une approche sociale de la cognitivité scientifique avec des possibilités d'éliminer la trop grande empreinte cognitive par l'utilisation de la bibliométrie. **Il a retenu que le phénomène de distribution présentant un coeur et une dispersion est une propriété de la plupart des éléments ou des croisements d'éléments bibliométriques. La connaissance et la maîtrise de ces distributions d'occurrences ou de cooccurrences donnent déjà de grands renseignements sur le contenu des références ou les caractéristiques des divers acteurs et de leurs liens.**

Nous retrouverons ces notions importantes dans l'outil bibliométrique développé dans le cadre de cette thèse.

3. La modélisation des distributions bibliométriques

Les auteurs, 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.

En bibliométrie, ces lois s'intéressent principalement aux relations existant 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 auteurs, comme Brooks, 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 similarité avec les distributions statistiques standards.

Les bibliométriciens vont développer de nombreuses techniques pour traiter ces distributions très souvent mises sous forme de rangs: les statisticiens parlent de statistiques non paramétriques. Elles s'écartent en cela des statistiques classiques.

Dans la suite de cette partie, les descriptions des concepts seront présentées sans la rigueur mathématique et statistique à laquelle nous devrions faire preuve. Ces lois seront essentiellement discutées en termes intuitifs.

a) La loi de Bradford

⇨ L'étude de Bradford:

La description de la loi qui va suivre est celle proposée par Bradford dans son livre [BRAD48].

Il considère que l'activité de gestionnaire de bibliothèque est soumise à un "**chaos documentaire**" de la littérature. 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 il a pensé à sélectionner parmi tous ces 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 qui cernait 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 augmentera alors que la productivité diminuera.

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

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

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

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

comme $p < p_1 < \ldots < p_n$ et $r > r_1 > \ldots > r_n$

on peut imaginer de choisir les zones de façon à obtenir ($1^{ère}$ hypothèse)

$$p.r = p_1.r_1 = ... = p_n r_n$$

Par conséquent

$$p_1/p = r/r_1 = n_1$$
$$p_2/p_1 = r_1/r_2 = n_2 \ldots$$

où n_1, n_2, ... sont des constantes

Alors on peut écrire

$$p_1 = n_1.p$$
$$p_2 = n_2.p_1 = n_1.n_2.p$$
$$\ldots$$

On peut encore choisir d'établir les zones de façon à avoir (2^{nde} hypothèse)

$$n_1 = n_2 = \ldots = n_n = n$$

d'où

$$p_1 = n\ .p$$
$$p_2 = n^2.p\ldots$$
$$p_n = nn.p$$

Bradford ajoute à cette démonstration une formulation verbale:

> *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 de plus en plus faibles liens de parenté. 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 présente ensuite cette loi sous la forme d'un graphe à partir de données expérimentales:
Il constitua deux bibliographies en consultant la collection des résumés de périodiques du *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 la loi qu'il vient d'établir les zones ont, pour un nombre d'articles constant, un nombre de publications croissant avec l'exponentiel d'une constante *n* (nommée multiplicateur de Bradford). Cette relation suggère que la somme cumulée des articles soit proportionnelle au logarithme du nombre de publications correspondantes (voir table 5). Ainsi lorsque la courbe est tracée, les points sont presque alignés sur une droite (voir figure 12).

Bradford conclut que le cumul des articles pour un sujet donné, à part ceux produits par la première zone de production, est proportionnel au logarithme du nombre de producteurs concernés, quand ceux-ci sont classés par ordre de productivité.

Table 5: Table des données expérimentales collectées par Bradford

APPLIED GEOPHYSICS, 1928-1 93 1 INCLUSIVE

A	B	C	D	E
1	93	1	93	0'000
1	86	2	'79	0·301
1	56	3	235	0·477
1	48	4	283	0·602
1	35	5	329	0·699
1	28	6	364	0·778
1		7	392	0·845
1	20	8	412	0·903
1	17	9	429	0·954
4	16	13	493	1'11'4
1	15	14	508	1·146
5	14	19	578	1·279
1	12	20	590	1·301
2	11	22	612	1·342
5	10	27	662	1·431
3	9	30	689	1·477
8	8	38	753	1·580
7	7	56	802	1·653
11	6	68	868	1·748
12	5.		928	1·833
17	4	85	996	1·929
23	3	108	1,065	2·033
49	2	157	1,163	2·196
169	1	326	1,332	2·513

LUBRICATION, 193 1-JUNE 1933 (FEW 1933 REFERENCES)

A	B	C	D	E
1	22	1	22	0·000
1	18	2	40	0·301
1	15	3	55	0·477
2	'3	5	81	0'699
2	10	7	101	0·845
1	9	8	110	0·903
3	8	11	134	1·091
3	7	14	155	1·146
1	6	15	161	1·176
7	5	22	196	1·342
2	4	24	204	1·380
13	3	37	243	1·568
25	2	62	294	1·792
102	1	164	395	2·215

A : nombre de journaux produisant un nombre d'articles donnés ;
B : nombre d'articles correspondant pendant une période donnée ;
C : somme cumulée du nombre de journaux (colonne A) ;
D : somme cumulée du nombre d'articles (colonne B) multipliée par A ;
E : logarithme du nombre de la colonne C.

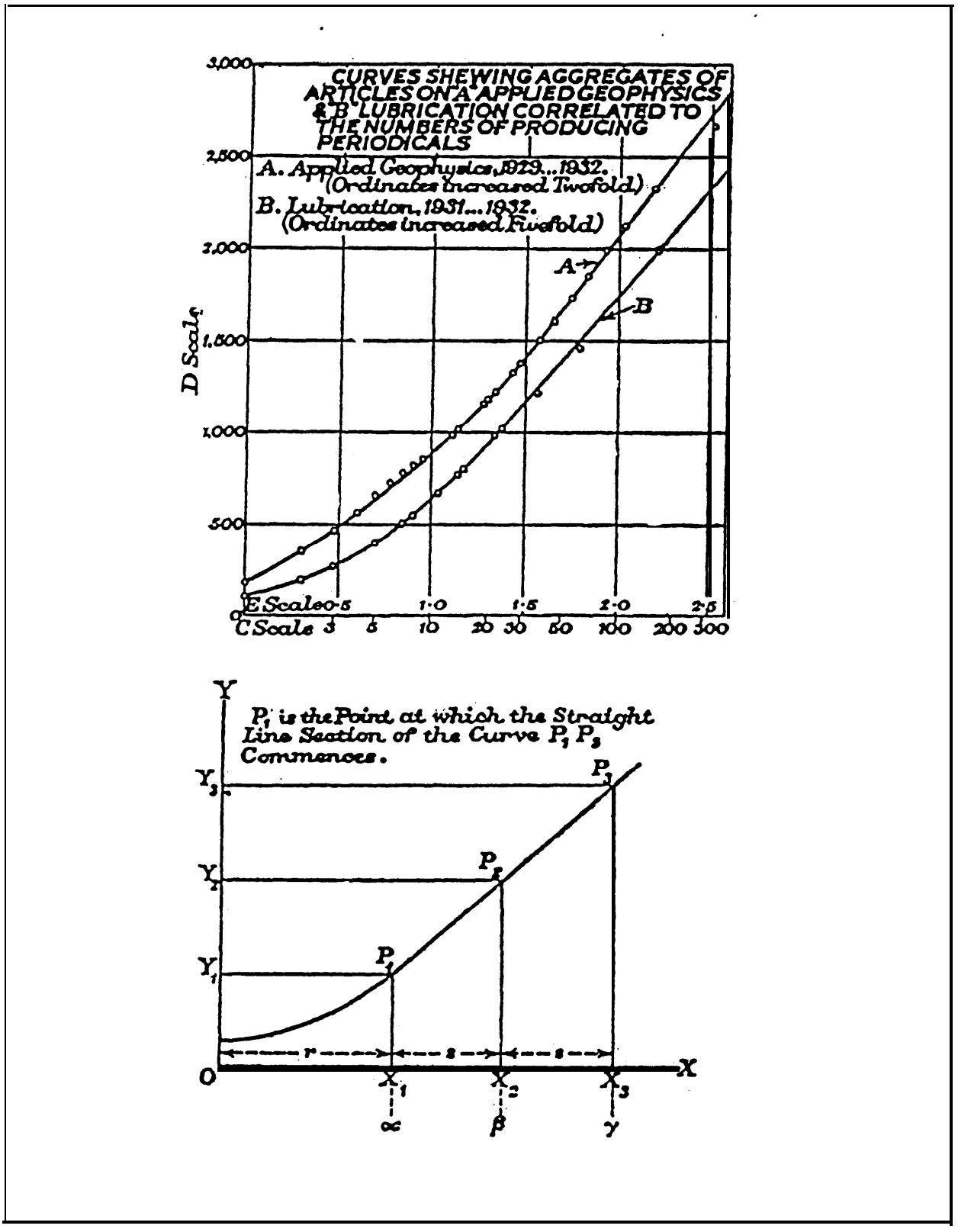


Figure 12: Courbes construites lors de l'étude de Bradford

Il a construit trois zones pour ces données. La première frontière est déterminée par le point de la courbe où commence la partie linéaire. Puis la seconde frontière a été fixée en reportant la distance 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.

Avec un oeil plus critique, Bradford a pu examiner qu'il y a en fait d'une part les périodiques qui ont une couverture très proche du sujet recherché, et d'autre part les périodiques qui ont des couvertures plus larges, **donc dans un certain sens plus productifs**.

⇨ 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 distribution statistique proposée par Bradford: Lawani [LAWA73], Alabi [ALAB79], Aleypeku [AIYE77]…

L'allure générale de la loi est correcte, mais certains détails animent encore de nos jours des débats passionnés:

- Le nombre de zones

  Des auteurs ont noté la difficulté de distinguer le coeur (noyau) des journaux de la dispersion dans une distribution de Bradford.

  Bradford découpa arbitrairement la courbe en trois zones équivalentes et trouva un multiplicateur de 2. Mais Mac Creery et Pao [CREE84] divisèrent la littérature en éthnomusicologie en 14 zones et ayant un multiplicateur moyen de 1,63. Wallace [WALL8] découpa la littérature au sujet du dessalement en 10 et trouva un multiplicateur de 1,8. Pontigo et Lancaster [PONT86] firent le découpage de la littérature sur la bactérie méthagonique en 4 parties avec un $n = 3,68$.

  De nombreuses méthodes moins empiriques pour déterminer ces zones ont été discutées par Egghe, Brooks et Rousseau mais toutes nécessitent forcément l'introduction d'une donnée arbitraire. **Seule la réflexion de Bradford, qui dit qu'une relative petite proportion de journaux peut satisfaire la requête d'une grande proportion d'articles sur un sujet, paraît sage**.

❍ La formulation mathématique

La volonté de trouver la formulation mathématique qui s'ajuste au mieux à ces données 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 [LEIM67] reprit la loi verbale faite par Bradford pour en établir une expression générale.

❑ L'année suivante Brooks [BROO68] montra son désaccord sur la formulation donnée par Leimkuhler et livra une formulation plus simple dressée à partir des représentations graphiques données par Bradford.

❑ En 1972 Wilkinson [WILK72] montra que leur discorde venait du fait qu'ils étaient partis des deux formulations données par Bradford. Or Bradford avait fait une erreur dans son analyse algébrique et sa formulation verbale étaient incorrecte (en 1948 Vickery avait déjà noté dans [VICK48] qu'il n'y avait pas concordance entre sa représentation graphique et sa formulation verbale). Wilkinson trouva que le la formulation graphique était plus proche des données empiriques que de l'expression verbale.

❑ Plus récemment, en 1984, Maia et Maia [MAIA84] ont su donner finalement une formulation mathématique de la loi de Bradford qui parait satisfaisante bien que relativement complexe. Elle est proche de celle de Brooks à la différence qu'elle ne décrit pas seulement la partie rectiligne de la courbe ($R(n) = k \text{ Log } n$) mais la totalité de la distribution.

Ils présentent leur formulation comme ceci:

Considérons que la collection des périodiques est organisée en classes de productivité décroissante désignées par l'indice $k = 1, 2, \ldots, n$

p_k = le nombre de périodiques dans la classe k

m_k = le nombre d'articles dans la classe k

on a alors r_k le nombre moyen d'articles par périodique dans la classe k

$$r_k = m_k / p_k$$

La première hypothèse de Bradford est de trouver les classes de façon qu'elles aient le même nombre d'articles

$$m_1 = m_2 = m_3 = \ldots = m_k = m$$

ce qui leur permet de trouver

$$p_k = \beta_{k-1}\, p_{k-1} \qquad \text{avec} \qquad \beta_{k-1} = m / P_{k-1}\, r_k$$

Selon la seconde hypothèse Bradford propose de considérer la série des β_k comme constante

$$\beta = \beta_1 + \beta_2 + \beta_3 + \ldots + \beta_k$$

d'où $\beta = r_k / r = p_k / p_{k-1}$ pour $k > 1$

soit R(k) la somme cumulée des articles jusqu'à la classe k

$$R(k) = m_1 + m_2 + m_3 + \ldots + m_k = k\ m$$

et N_k la somme cumulée des périodiques jusqu'à la classe k

$$N_k = p_1 + p_2 + p_3 + \cdots + p_k$$

la fonction de R en fonction de N_k est définie selon la relation:

$$R(N_k) = J \log (N_k / S_k)$$

avec

$$J = m / \log \beta$$
$$S_k = (p_1 / \beta).(\beta^k - 1) / (\beta - 1)$$

0 L'affaissement de Groos (Groos droop)

Une autre remarque a été faite en 1967 par Groos [GROO67] en ce qui concerne la fin de la distribution telle que l'avait présentée Bradford. Il nota que la courbe en fin de la distribution a une tendance à s'affaisser (voir la figure 13). Cette nouvelle partie est appelée l'affaissement de Groos (*Groos droop)* en honneur à son découvreur. Plusieurs interprétations ont été données mais celle qui est la plus communément acceptée estime que **cet affaissement reflète le caractère d'une bibliographie incomplète**.

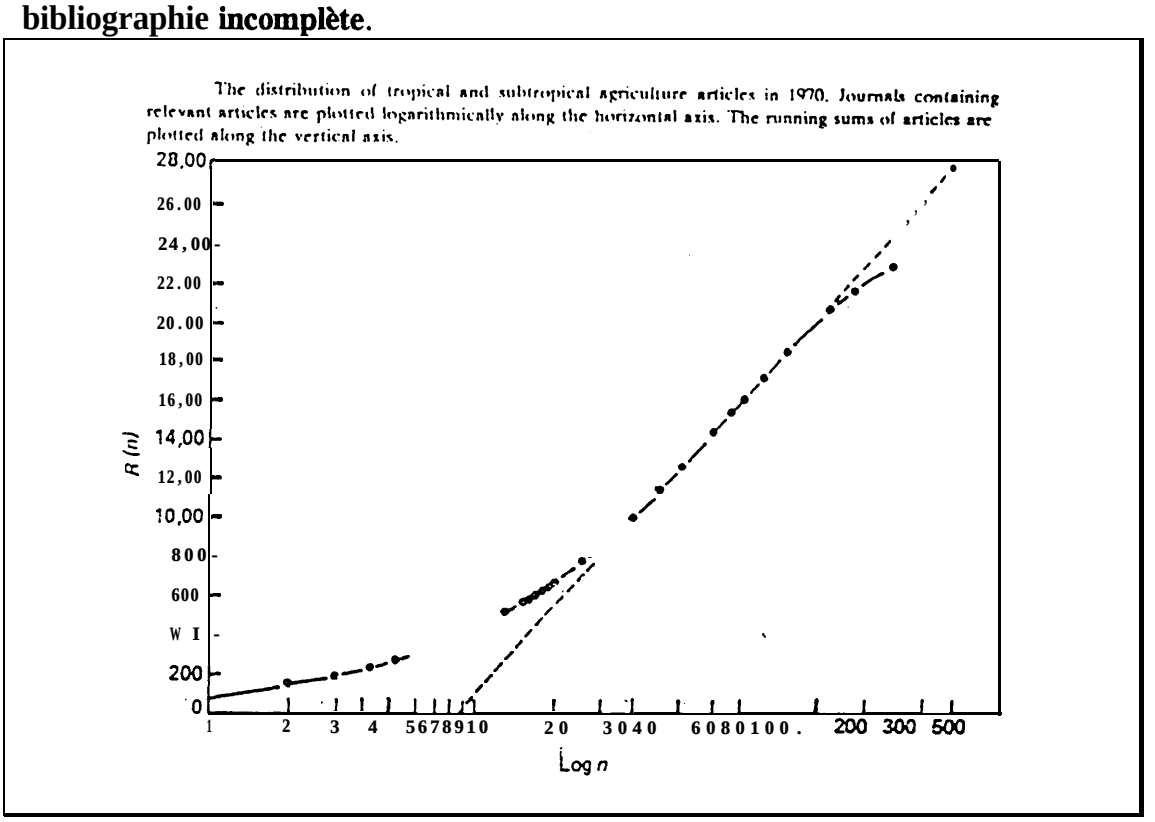


Figure 13: Affaissement de Groos étudié par Lawari S M [LAWARI73]

b) La loi de Lotka

⇨ L'étude de Lotka [LOTK26]:

Il lui semblait intéressant de déterminer la part de contribution de chaque chercheur au progrès de la science.

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 commençant par les premières lettres A et B. Puis il a cumulé tous les auteurs qui n'avaient qu'une entrée (un article dans le CA), puis ceux qui en avait 2, 3…
La même manipulation a été aussi faite avec l'index des auteurs du journal Auerbach's Geschichtstafeln der Physik (J.A. Barth, Leipzig, 1910)

Il a d'abord présenté ses résultats sous la forme d'un histogramme qui représentait le pourcentage d'auteurs en fonction du nombre d'articles qu'ils ont publiés (voir figure 14). Puis les mêmes résultats sont présentés selon des échelles logarithmiques pour les deux axes (voir figure 15).

Les points étant pratiquement alignés selon une droite, il était possible d'estimer que ces résultats suivaient la relation

$$\text{Log } y = n \text{ Log } x + b$$

$$y = x^n * C$$

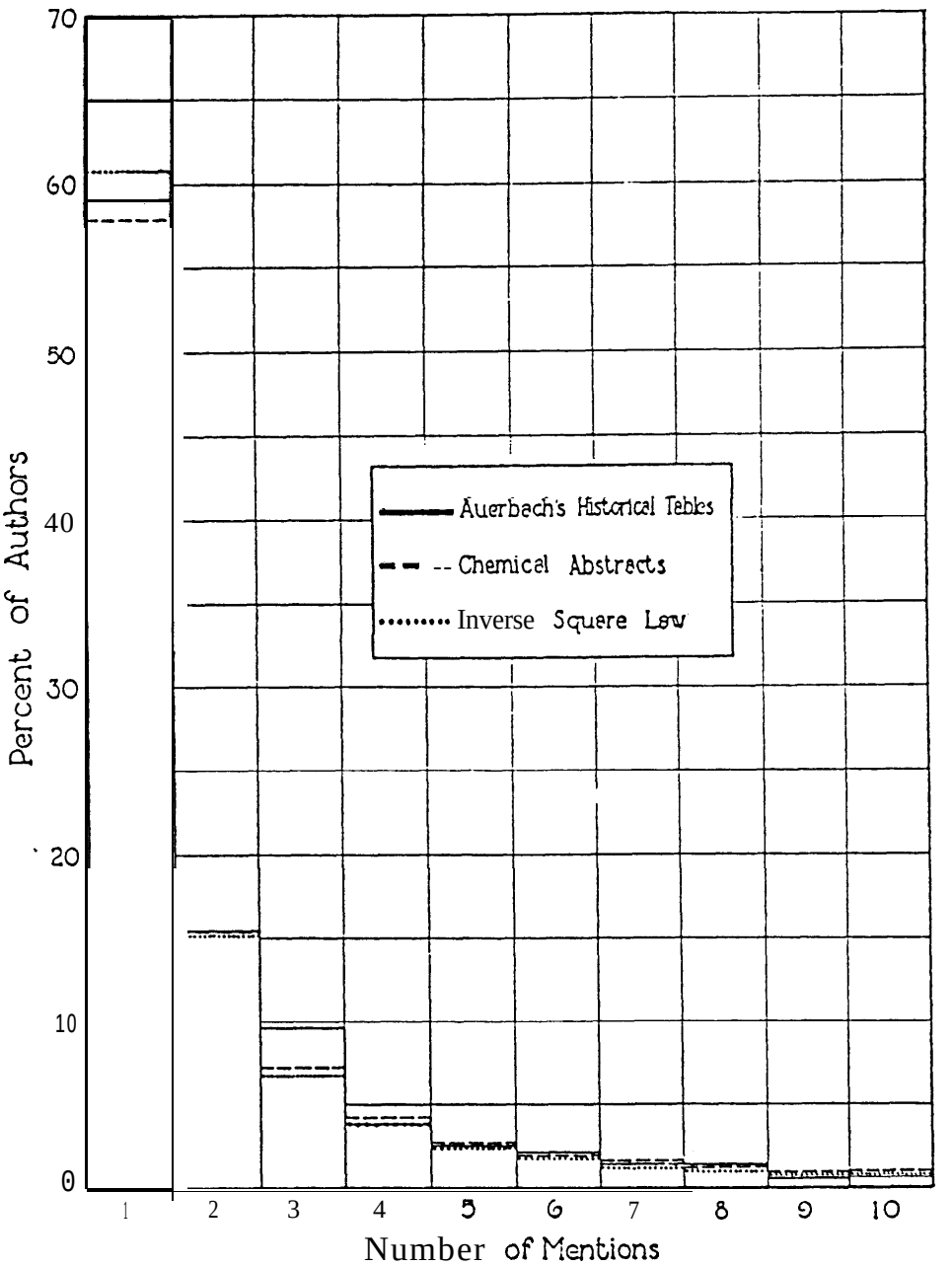
avec x = le nombre de publications
y = le nombre (ou pourcentage) d'auteurs ayant x publications
n = pente de la droite
C = constante dont la valeur est égale à l'intersection de la droite avec l'axe des y, c'est à dire le nombre d'auteurs n'ayant publié qu'une fois

La pente de la droite pour ces deux séries de valeurs proches de la valeur -2 (-2,021 ± 0,017 et -1,888 ± 0,007), il obtenait la relation empirique:

$$\mathbf{y = Constante / x^2}$$

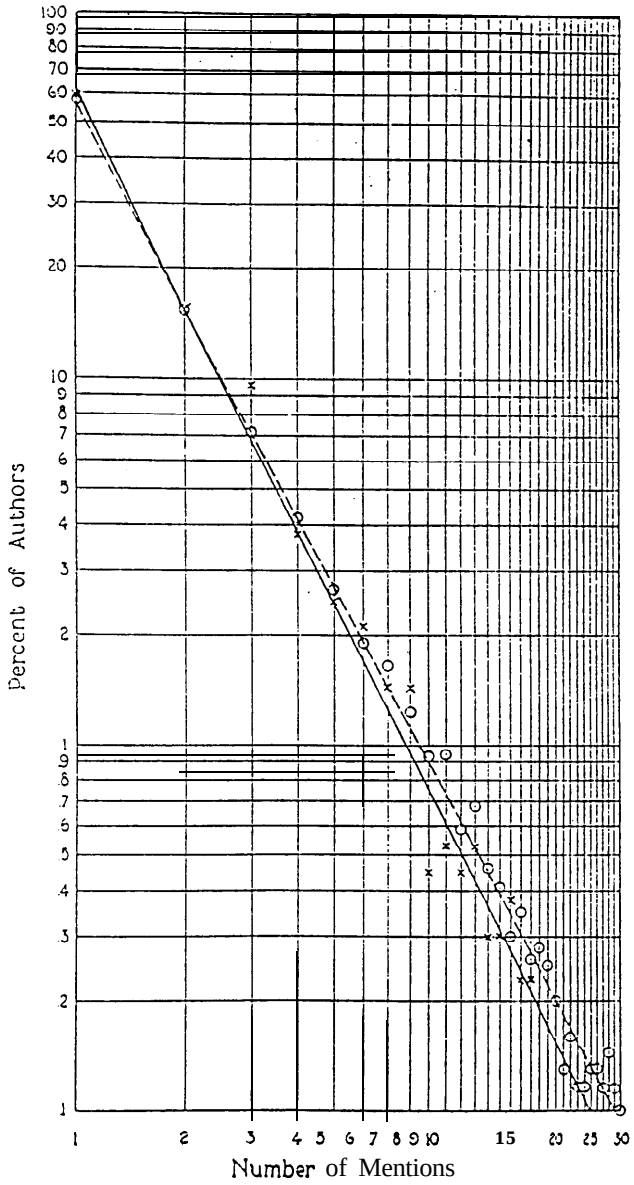
Il y a, par rapport au nombre d'auteurs qui publient 1 article, 4 ($1/2^2$) fois moins d'auteurs qui en publient 2, 9 ($1/3^2$) fois moins qui en publient trois, … n^2 fois moins qui en publient n. **On a donc une productivité scientifique qui diminue selon une loi de type carré inverse**.

Remarque: dans son cas, la part des auteurs qui avaient publié un seul article était proche de 60 %.



Frequency diagram showing per cent of authors mentioned once, twice, etc., in Auerbach's Geschichtstafcln der Physik, entire alphabet, and in the decennial index of Chemical Abstracts 1907–1916, letters A and B. The dotted line indicates frequencies computed according to thc inverse square law.

Figure 14: Présentation faite par Lotka de la distribution des auteurs sous forme d'histogtamme



Logarithmic frequency diagram showing number of nuthors mcntioned oncc, twice, etc., in Auerbach's tables (points indicated by crosses), and in Chemical Abstracts, letters A and B (points indicated by circles). The fully drawn line indicates points given by inverse square law, csponent = 2; thc line of dashes corresponds to csponent 1.89.

Figure 15: Présentation faite par Lotka de la distribution des auteurs sous forme logarithmique

Lotka n'a pas pu tester son modèle à d'autres journaux car ceux-ci n'avaient pas d'index d'entrée classé par auteur.

⇨ Controverse sur le modèle Lotka

Price précisa dans son livre [PRIC63] que la formule de Lotka tendait à surestimer le nombre d'auteurs à forte productivité. En fait, les données, que Price a collectées, montraient que le nombre de personnes à plus forte productivité avait une chute 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 dont [RADH79], [VOOH74]. Mais Subramanyam [SUBR79] attribue ces résultats à la mauvaise sélection de l'échantillon ou à une négligence en ce qui concerne les articles à auteurs multiples.

Murphy a même montré dans une étude [MURP73] que la littérature de l'humanité en général suivait une loi de type Lotka. Elle a été, elle aussi, contestée peu après par Coile qui a testé statistiquement les résultats de Murphy [COIL77].

En conclusion, l'universalité de la loi de Lotka est peut-être à remettre en cause.

c) La loi de Zipf

⇨ L'étude de Zipf

Zipf reprit une idée qui avait déjà été exposée en 1919 par Estoup [ESTO16] et étendit grandement sa portée. Parmi tous les sujets traités dans son livre [ZIPF49], Zipf se posait la question suivante: **à quelle fréquence les mots apparaissent-ils dans un texte littéraire?**

Zipf comptabilisa les occurrences des 29.899 mots différents qu'il trouva dans le l'oeuvre *Ulysses* de Joyce. Il les classa par ordre décroissant de fréquence et il affecta chaque mot du rang 1 (le plus fréquemment apparu) au rang 29.899 (le moins fréquemment apparu).

Il trouva qu'en multipliant la valeur du rang (r) par la valeur de la fréquence correspondante (f) il obtenait un produit (C) qui était constant pour l'ensemble de la liste de mots (voir table 6).

Table 6: Valeurs étudiées par Zipf pour le livre *Ulysse* de Joyce

I Rank (r)	II Frequency (f)	III Product of I and II (r * f = C)
10	2.653	26.530
20	1.311	26.220
30	926	27.780
40	717	27.780
50	556	27.800
100	265	26.500
200	133	26.600
300	84	25.200
400	62	24.800
500	50	25.000
1.000	26	26.000
2.000	12	24.000
3.000	8	24.000
4.000	6	24.000
5.000	5	25.000
10.000	2	20.000
20.000	1	20.000
29.899	1	29.899

D'où la formule de Zipf:

$$f \times r = C$$

Zipf a nommé cette loi, **le principe du moindre effort**, car elle suggère que les gens choisissent et utilisent plus facilement des mots familiers que des mots inhabituels par pure paresse. Donc la probabilité d'occurrence d'un mot familier est bien plus élevée que celle des autres dans le vocabulaire usuel.

⇨ Représentation sous forme de graphe

Zipf 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 peut être représentée suivant la figure 16.

Selon la loi de Zipf cet histogramme suivrait une courbe de la forme 1/x à ceci près que la courbe est continue là où l'histogramme est discontinu. Cette précision est applicable à toutes les lois bibliométriques auxquelles les auteurs appliquent abusivement des formules mathématiques continues alors que les données sont discrètes.

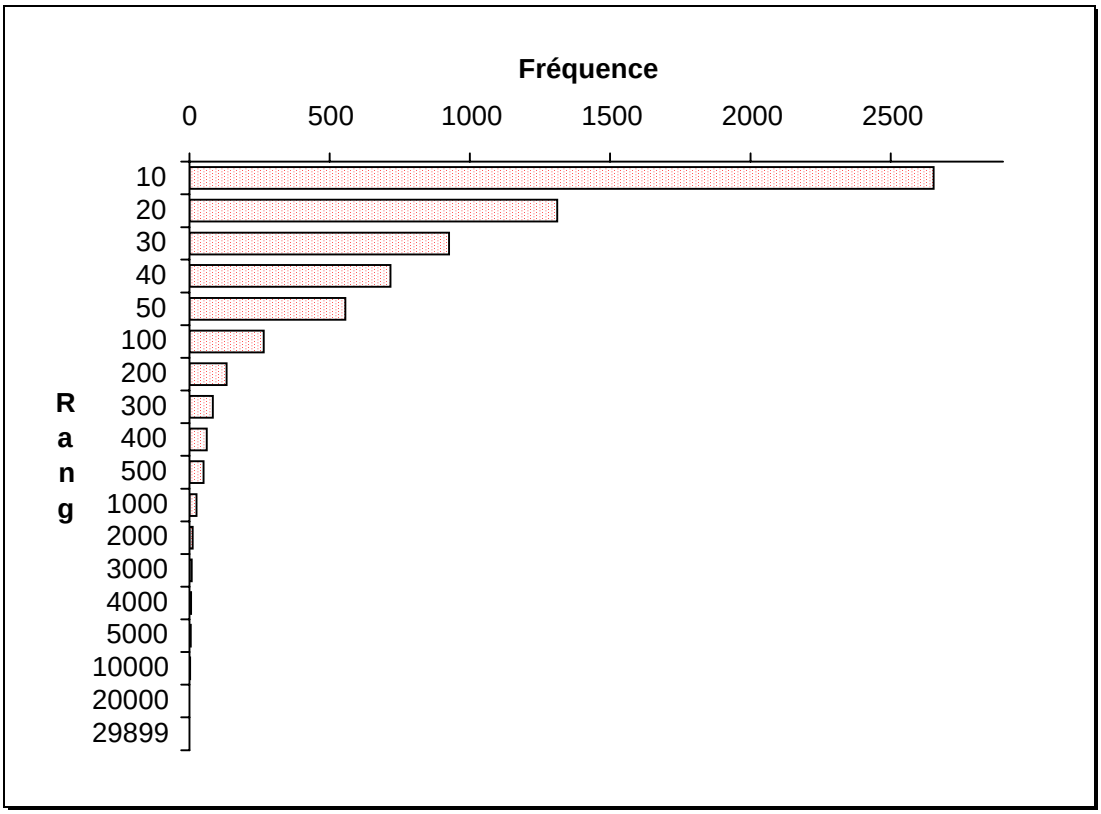


Figure 16: Présentation des données de Zipf sous forme graphique

⇨ Formulation mathématique

Là encore cette loi a donné l'impulsion à la formulation par de nombreux auteurs de divers modèles mathématiques. 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 [FAIR69]. 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 $1/n(n+1)$ donne la fraction du nombre total de mots différents qui apparaissent à la fréquence n.

❑ Fedorowicz dans [FEDO82] reconnaît quatre écoles de pensées qui se penchent sur une formulation théorique de la loi de Zipf:

- Hill, Woodstruffe, Sichel, Crowley et d'autres montrent que la loi de Zipf peut être dérivée par différents procédés stochastiques incluant le modèle Bose-Einstein de Hill, la binomiale négative de Bliss et Fisher et la distribution des séries logarithmiques de Fisher.
- Bon nombre de modèles classiques de dérivations peuvent être manipulés de façon à produire des distributions hyperboliques. C'est une approche caractérisée par des théories probabilistes. Les fonctions hyperboliques incluent la fonction Beta de Simon et la distribution de l'avantage cumulatif de Price.
- La troisième approche est due à Mandelbrot qui par son modèle de la théorie de l'information étudie les structures statistiques du langage.
- La dernière approche est basée sur les travaux de Herdan dans un nouveau domaine de la linguistique quantitative.

❑ Dans un second article [FEDO82], Fedorowicz a présenté puis appliqué une formulation de la loi de Zipf faite par Booth (inspirée de l'approche de Mandelbrot). Booth a proposé une loi qui améliore celle de Zipf pour les mots à très faibles fréquences d'occurrences. Vue l'abondance des mots qui apparaissent rarement, de nombreux mots 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

$$\mathbf{G_m = (kT)^{1/ß} [1/(2^{m-1})^{1/ß} - 1/(2^m)^{1/ß}]}$$

avec T = longueur du texte
ß = constante > 0

Fedorowicz a testé cette formulation sur la base de données MEDLINE pour les fichiers inversés des champs en langage libre (titre et résumé). Cette étude lui a permis de vérifier que la relation entre le fichier index (mots et leurs adresses dans le fichier *postings* associé) et le fichier *postings* (liste des références correspondant à chaque entrée c'est-à-dire un mot) est bien du type Zipf.

Il voudrait 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.

❑ L'approche de l'expression de la loi de Zipf par la théorie de l'information est principalement le fruit du travail de Mandelbrot [MEND53]. Il a proposé une formule plus générale et mieux ajustée que la loi de Zipf:

$$f(r) = k\,(r + c)^{-\mu}$$

Avec f(r) = fréquence du mot
et r = le rang du mot

La constante c améliore l'ajustement pour les mots communs, dont les rangs sont peu élevés L'exposant 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.

d) 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 et unique 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 qui permettraient de les unifier sous une seule et même loi.

⇨ Ressemblance des lois Zipf-Bradford-Mendelbrot:

Quand Brooks a publié sa dérivation simplifiée de la distribution de Bradford [BROO68], il découvrit qu'elle avait une forte ressemblance avec la loi de Zipf. La loi de Zipf ne décrit pas exactement la distribution de Bradford. Il est vrai que multiplier le rang d'un journal par son nombre d'articles, contribuant à un thème, aboutira à 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 $f \times r = C$.

Mais bien avant Brooks, en 1960, Kendall avait déjà montré que Zipf et Bradford étaient deux lois structurellement similaires [KEND60].
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). [EGGH91]

⇨ Ressemblance des lois Lotka-Pareto-Zipf:

Parker-Rhodes et Joyce [PARK56] en 1956 ont présenté la distribution de mots d'un texte comme Lotka l'avait fait pour les auteurs dans le passé.

$$n(u) = k\, u^{-2}$$

n(u) = nombre de mots d'un vocabulaire qui apparaissent, dans un texte suffisamment long, avec la fréquence u.

Price indiqua aussi que la formule de la loi de Lotka ressemblait à un autre paradigme économique, la loi de Pareto utilisée comme représentant la distribution des revenus (suit une loi en $1/n^{1,5}$).

⇨ L'unification:

L'existence d'une fonction commune à ces trois lois a été démontrée récemment par Hubert [HUBE78] et Chen [CHEN85].

Mais le premier auteur à avoir ressenti le besoin de réunir ces trois lois sous un même principe a été De Solla Price. Price proposa 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 l'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 comme ceci: supposons une population d'individus essayant d'atteindre un but: un nombre de publications (Lotka), une acquisition d'articles pertinents (Bradford)… Cette loi est régie par le fait que si un individu (le scientifique, le journal…) a du succès alors sa probabilité de succès augmentera pour une tentative ultérieure, alors qu'un échec ne réduit pas la probabilité de succès pour sa prochaine tentative.

La distribution de l'avantage cumulatif est présentée par Price selon la formulation de densité:

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

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
- $Beta(n,m+2)$ = 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 $Beta(a,b) = (b-1)!\ a^{-b}$.

Cette formulation ne contient donc qu'un seul paramètre en la présence de la variable m.

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 [EGGH86]
- Concernant l'application de la théorie de Price pour modéliser la fréquence de citation d'une publication, Budd et Hurt dans un récent article [BUDD91] l'ont expérimentée puis l'ont comparée avec des données réelles qu'ils avaient recueillies dans les bases de données de l'ISI. Les résultats montrent que pour leurs données 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, dans un article en 3 volets, 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 humaine, la loi quantitative fondamentale de l'activité humaine (HAITUN 82c)**.

Même si sa formule exacte n'est pas encore bien définie, on peut retenir que la loi de Zipf est la distribution de base en bibliométrie.

e) Mesures synthétiques des distributions

L'analyse d'une distribution par la statistique moderne est fondée sur le calcul de valeurs qui permettent de la caractériser. La connaissance de ces valeurs devrait déjà bien définir la nature de la distribution.

L'analyse d'une distribution se fait donc sur l'examen de la forme de la distribution accompagnée de ses mesures synthétiques. Ces mesures synthétiques sont de deux types:

- Les premières servent à étudier la tendance centrale de la distribution
- les secondes informent sur la dispersion des données autour de cette tendance centrale.

Les deux mesures synthétiques les plus connues sont la moyenne arithmétique et la variance. Elles sont toutes les deux construites sur la théorie des moments (la moyenne représente le moment d'ordre 1 par rapport à l'origine et la variance représente le moment d'ordre 2 par rapport à la moyenne). La simple connaissance de ces deux mesures définit suffisamment bien les distributions pour permettre de les comparer aisément et rapidement.

⇨ Les mesures synthétiques courantes s'appliquent-elles aux distributions bibliométriques?

Comme on vient de le voir, Haitun a précisé que les distributions, contrairement aux lois des sciences naturelles, ne sont pas Gaussiennes mais hyperboliques (Zipfiennes). Les distributions Gaussiennes (loi normale, de poisson, négative binomiale, etc) ont des moments stables (moyenne, variance, etc). **A l'inverse les distributions hyperboliques ont des moments 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.

⇨ Mesures de concentration:

❑ Conscient de cette absence de mesures synthétiques Pratt a proposé des mesures de concentration (dispersion) des distributions bibliométriques et des leurs éléments [PRAT77]. 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, il voulait pouvoir estimer la concentration de chaque élément pour la distribution. Appliqué à la distribution de Bradford cet index devait prétendre mesurer le degré de concentration des articles sur un sujet dans une collection de journaux.

❑ Juste après la publication de Pratt, Carpenter [CARP79] fit remarquer la ressemblance entre l'indice de Pratt et l'indice de Gini. 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.

❑ L'interprétation bibliométrique de l'indice de Pratt n'est pas reconnue par tous. Drott [DROT80] observa 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 Hustopeckÿ et Vlachÿ dans [HUST78].

❑ Récemment, Egghe a modifié l'indice de Pratt pour l'appliquer dans des cas particuliers aux distributions bibliométriques [EGGH87] [EGGH88].

On peut encore consulter [EGGH90] et [BURR91] pour des discussions générales sur des mesures de concentrations

⇨ Théorie de la communication:

De nombreux auteurs émettent l'idée que la communication scientifique suit des processus de transmission que l'on retrouve dans la nature.

❑ Goffman et Newill [GOFF64] ont appliqué un modèle "épidemique" à la propagation des idées scientifiques. **Ils estimaient que la communication des idées par les publications était 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.

❑ En 1948, Shannon avait donné naissance à la "**théorie de l'information**" qui fut renommée par la suite "**théorie de la communication**". **L'outil essentiel de cette théorie est la mesure de la variété moyenne ou complication des signaux par une équation mathématique**. Il a développé celle-ci pour l'étude de la transmission des signaux par voie téléphonique.

Elle a été largement utilisée par les écologistes (Cf Legendre & Legendre [LEGE84]). En écologie, le concept de diversité des espèces est prépondérant pour évaluer la richesse d'un milieu. Cette 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. Lafouge et Quoniam dans [LAFO91] ont rappelé comment cette mesure de diversité pouvait s'appliquer aux distributions bibliométriques.

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 d'ordre a de Rényi utilisée en biologie est transposable en science de l'information:

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

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 $H_1 = -\sum_{i=1}^{n} (p_i \text{Log} p_i) = H_{Shannon}$

$N_1 = \text{Exp } H$ 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 procure 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, $p_i = 1/n$.

Il faut bien être conscient que cette notion de diversité représente une toute autre information que celle induite par la théorie des moments.

L'applicabilité et la validité de cette théorie pour des données bibliométriques expérimentales n'ont pas encore été confirmées. Mais dans le cas où cette théorie se montre adaptée aux données expérimentales, alors la théorie de la communication offrirait des mesures synthétiques pour caractériser des distributions bibliométriques. D'après ce qui vient d'être exposé, **les deux premières mesures caractéristiques seraient la taille de l'échantillon étudié et l'entropie de Shannon**.

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 donc une entropie négative, une néguentropie, ce n'est que pour des raisons de simplicité qu'on la qualifie d'entropie.

4. Indicateurs univariés

Cette partie présente la bibliométrie univariée. **Le principe de cette dernière est de constituer des indicateurs qui vont permettre de comparer les éléments des corpus bibliographiques entre eux**. Le problème dans l'élaboration de ces indicateurs va se situer dans le choix du système de mesure pour comparer de façon équitable les éléments.

Ces indicateurs univariés sont généralement considérés comme des informations purement quantitatives.

a) Le dénombrement des publications: indicateur de productivité

Parmi tous les indicateurs univariés envisageables, le plus simple est le dénombrement des publications. **Ce simple comptage est généralement considéré comme la mesure même d'une productivité**, que ce soit au niveau d'individus, au niveau d'organismes, au niveau de disciplines ou au niveau de nations. 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 des auteurs ont cherché à 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 vont être présentées dans les lignes qui vont suivre.

Mais quelle que soit la mesure employée l'évaluation reste subjective. **Les évolutions temporelles de ces mesures sont toujours les indications les 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.

b) Le dénombrement des citations: Est-il un bon indicateur de qualité?

Juger de la productivité par une mesure de quantité (nombre de publications) est déjà bien, mais pas satisfaisant aux yeux des bibliométriciens. Ils ont immédiatement cherché à connaître quel indicateur leur permettrait d'**évaluer la qualité d'une publication**.

Plusieurs facteurs sont considérés comme rentrant en jeu dans l'évaluation de la qualité d'une publication. Les plus importants sont:

- le taux de citation de la publication, la renommée du périodique où elle est publiée
- le type de la publication (livre, périodique, rapport…)
- les publications dont l'origine est la collaboration entre chercheurs
- la nature du contenu (fondamental, méthodologique, expérimental, synthèse,…)

…etc…

Le facteur qui a été considéré pendant longtemps comme donnant une indication de qualité est le taux de citations. **On estimait 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, en voilà quelques unes:

- Les citations erronées qui renvoient à des sources secondaires plutôt qu'à l'auteur principal de la découverte [MCRO89]
- L'auto-citation qui est évaluée de 10 à 30% par article [MCRO89]
- La différence de nature entre toutes les citations: certaines citations seront faites dans un contexte de critique [NADE83] (confrontation entre la théorie de Kuhn et celle de Popper)
- L'inertie de la citation: le délai de temps important entre la publication et la citation [MCCA89]
- L'influence de la revue où a paru l'article: à 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 [VANR88]
- La variation de la pratique de la citation dans chaque discipline [MOED85] [GARF82]
- La taux de citation dépend du type de document (article ou synthèse): par exemple les articles de méthodes sont plus cités car les auteurs, pour éviter de décrire la méthode qu'ils utilisent, citent d'autres documents où elle est expliquée [PERI83]
- un auteur cite plus facilement la science nationale [STEV89]
- La variation du taux de citation selon la nationalité des auteurs [MCROB89]

- 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 [COUR90].

Pour ajouter à ces critiques de fond, des problèmes 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:

- mauvaise couverture thématique
- géographique et temporelle
- 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, les 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* ***Erreur! Source du renvoi introuvable.*** *(et, a fortiori, dans la communauté scientifique internationale telle 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"* [COUR90]

c) La mesure des journaux

Beaucoup d'investigations en bibliométrie s'intéressent à la revue scientifique en tant qu'unité d'analyse. Ceci n'est pas étonnant puisque le journal 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 périodiques et aux revues spécialisées, la création de moyens d'évaluation de ces derniers est fondamentale pour bien gérer les abonnements aux journaux scientifiques.

⇨ L'étude de Bradford:

L'étude de Bradford, exposé plus haut, a été l'une des premières méthodes mise au point pour aider cette gestion bibliothécaire de revues.

⇨ La citation:

Un peu plus tard la création du *Journal citation reports* (JCR) par l'ISI à permis d'offrir de nouvelles données pour juger de l'importance des journaux par l'intermédiaire des citations. Cette idée n'était pas tout neuve puisqu'elle avait déjà été exploitée par Gross et Gross en 1927 [GROS27].

L'ISI recueille dans ce journal le nombre de citations dont font l'objet 4200 revues scientifiques. **Cette citation 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 à d'articles et plus elle a de chance d'être citée. Pareillement plus l'existence de cette revue est importante et plus son taux de citation augmente. Pour nuancer ces facteurs influents de nombreux indices ont été proposés. Nous allons présenter les deux plus connus parmi ceux-ci. Ce sont les deux indices fournis par le JCR:

- Le facteur d'impact (I_F):

$$\mathbf{I_F = c(x) / (p(x\text{-}1) + p(x\text{-}2))}$$

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.
Le facteur d'impact a été proposé par Garfield en 1969 pour être introduit dans le JCR.

- l'indice d'immédiateté (I_I):

Il a été introduit en même temps que le facteur d'impact dans le JCR par Garfield.

$$\mathbf{I_I = c(x) / p\,(x)}$$

c'est à dire le nombre de citations, reçues l'année *x* 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**: "how rapidely a journal's material is picked up an used" [GARF79].

Des algorithmes plus élaborés pour déterminer l'importance d'une revue ont été développés par Bennion et Karschamroon [BENN84] ainsi que par He et Pao [HE86].

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 [BUFF91] dans son domaine.

❑ Critique de la couverture et de la qualité des données du JCR:

En plus de tous les biais introduits par la pratique de la citation (listés dans la partie précédente), des critiques propres au JCR entrent en jeu. Tous les chercheurs utilisant le JCR ont remarqué des inadéquations ou des problèmes: Vlachy [VLAC85], Carpenter & Narin [CARP81], Rice et Al [RICE89]. Notamment ces derniers ont rapporté la possibilité d'avoir **jusqu'à 25% d'erreurs de mesure en utilisant le JCR du SSCI 1977-85**, la plupart 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 de temps.

⇨ La typologie:

La comparaison des journaux peut être menée par des approches bibliométriques plus classiques, ne faisant pas intervenir la pratique des citations. L'article de Dou et al. [DOU90], par les solutions graphiques qui y sont proposées, est un bon exemple de ce genre d'étude.
Ces auteurs établissent des cartes typologiques de thèmes par revues qui permettent de mieux connaître les thèmes de prédilection et les spécialités qui les distinguent. Dans ce cas la comparaison se fait donc 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 publiés ont été collectées par consultation du *Chemical Abstracts* accessible en ligne de 82 à Juillet 87. Ces données sont 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 sections codes de CA.

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 (figure 17).

Un découpage des références d'une revue par année fournit aussi l'évolution thématique des articles de la revue par une succession de cartes typologiques.

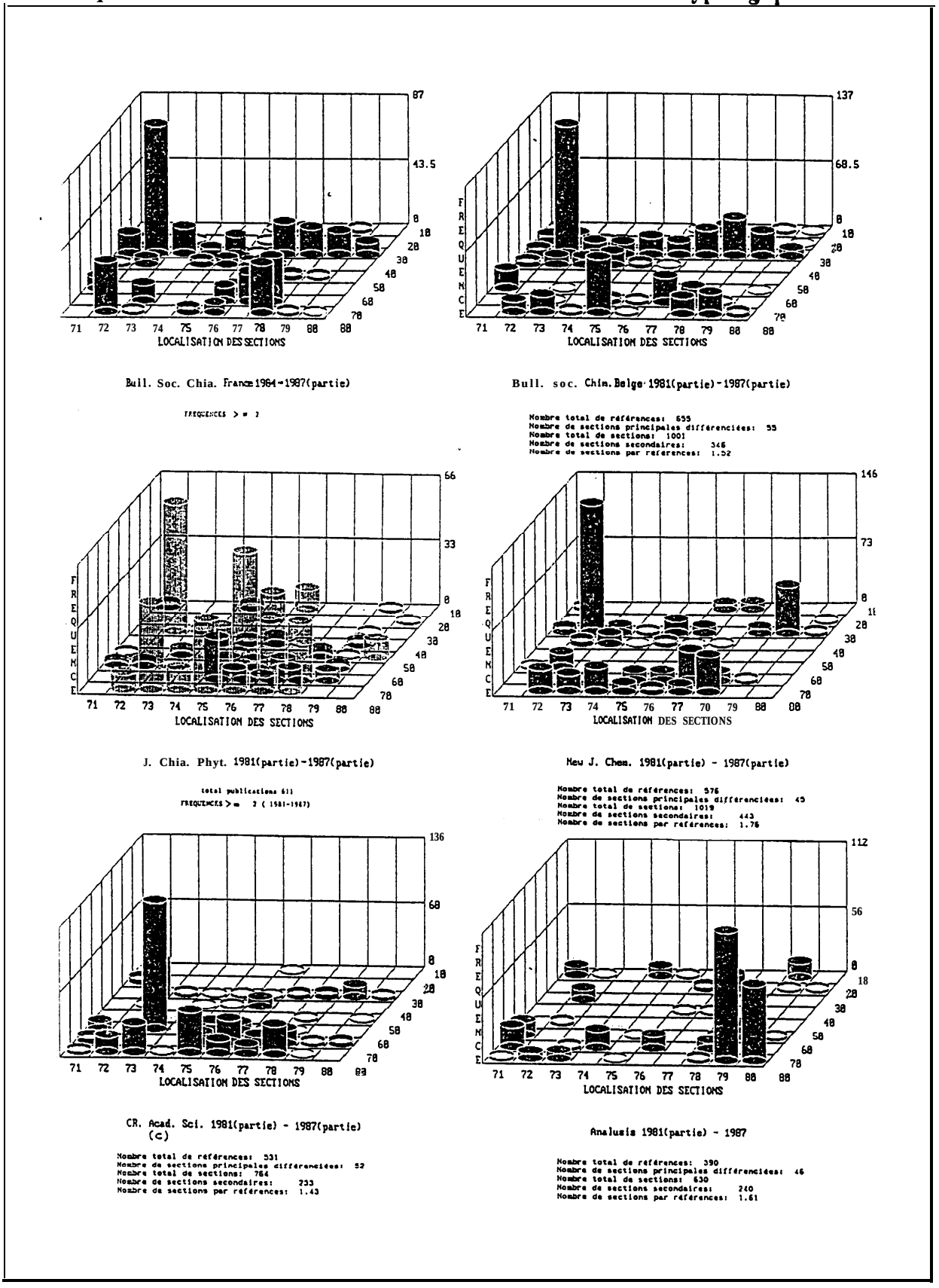


Figure 17: Typologie des thèmes abordés par des journaux en Chimie

d) La mesure des chercheurs

Il est bien connu que le nombre d'articles est encore le principal moyen d'évaluation des chercheurs pour les instituts de recherche. 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 [DEMA92]. Est-ce que cet indicateur est suffisant? Bien évidemment, la réponse est négative, mais peu de solutions 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 pour 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 caractéristique 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 bibliographique peut jouer le rôle d'étalon à la mesure?

⇨ La citation:

Comme pour les journaux de nombreuses études bibliométriques introduisent **la citation dont l'auteur fait l'objet**. Les données, comme pour toutes celles concernant la citation, sont uniquement produites par l'ISI soit sous forme papier soit par consultation en ligne [GARF81].
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 [MCRO89]:

- seul le premier auteur de l'article cité est considéré
- problème de l'homonymie
- une auto-citation encore plus accentuée

- erreurs d'autographe

⇨ Les co-signatures:

Dans ce genre d'évaluation des auteurs, **pour être totalement irréprochable il faudrait prendre en considération le cas des co-signatures**.

Une étude sur l'implication des différentes procédures de comptabilisation de la productivité des auteurs a été menée par Brandis et Oluic-Vukovic [PRAV91]. Quatre procédures de comptabilisation ont été menées simultanément sur le même ensemble d'auteurs:

- Comptage normal (*normal count*):
  donne un crédit équivalent à tous les auteurs d'une même publication; il y a donc comptabilisation pour un auteur de tous les articles où il a été signataire.
- Paternité fractionnée (*authership fractional*):
  la part de 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 les parts de contribution des ces publications.
- Comptage direct (*straight count*):
  seul le premier auteur reçoit la paternité de la publication.
- Comptage direct modifié (*modified straight count*):
  chaque publication est attribuée à un seul auteur. Ce n'est plus le premier auteur mais 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 procédures C et D ont, elles, en commun de réduire 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.

Mise à part le problème du comptage, **la co-signature d'articles 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 dans la partie traitant des *cartes relationnelles*).

e) La mesure des laboratoires

Les méthodes sont toujours les mêmes mais utilisées pour une unité d'analyse ciblée sur les centres de recherche. On se trouve dans le même cas d'études que lors de l'évaluation d'un chercheur; mis à part qu'il ne faut plus considérer un individu mais plusieurs.

⇨ Difficulté des comptages:

Effectivement, **il est pratiquement impossible pour le traitement bibliométrique de considérer comme unité d'analyse le centre de recherche, si on utilise les bases de données**. Le centre de recherche est, en toute logique, détectable par l'indication de l'affiliation des auteurs des publications dans les références. Or cette affiliation est très mal représentée dans les bases de données accessibles en ligne pour trois raisons:

- la plupart des bases n'indiquent que l'affiliation du premier auteur
- il n'existe aucune norme pour la saisie des noms et l'adresse des centres d'où une multiplicité de leurs écritures
- affiliation inexistante dans les citations des bases de l'ISI

La solution est donc de collecter les données pour chaque individu appartenant à l'organisme et de les compiler. Le traitement par les noms des auteurs crée des ambiguïtés pour certains d'entre eux mais réduit pratiquement à néant la diversité d'écriture.

Outre ces difficultés inhérentes aux traitements, les méthodes sont similaires à celles exposées précédemment. Des exemples d'étude utilisant à la fois le nombre de publications et le nombre de citations ont été montrés dans des articles comme [REMY91], [BAUI92], [PETE91].

f) La mesure des pays

C'est peut-être l'unité d'analyse la plus employée en bibliométrie. La comparaison des pays en fonction de leur contribution scientifique a toujours été importante dans les études bibliométriques car celle-ci s'intègre bien dans des investigations scientométriques à l'échelle nationale. Connaître la situation du pays dans les divers domaines scientifiques par rapport à celles des autres pays est une indication précieuse pour budgétiser les programmes

de recherche nationaux. Il est probable que c'est pour répondre à de tels besoins que la bibliométrie a vu son champ d'action s'élargir subitement.

⇨ Indice d'avantage:

Pour évaluer la part de contribution d'un pays à la science, les études découpent généralement la mesure suivant deux dimensions: les domaines et le temps. La science du pays est divisée en disciplines homogènes que l'on va mesurer. Puis on suit l'évolution de cette mesure au cours du temps.

Le choix des mesures pour sonder un domaine est toujours aussi restreint: soit la mesure de la productivité, par le nombre des 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 pondérer ces mesures selon la part réelle qu'elles représentent par rapport aux autres domaines et par rapport aux autres pays**. Le premier a avoir introduit ce calcul est Price dans [PRIC81a]. Il voulait montrer par ce travail comment rendre un tableau de données plus accessible et plus rapidement interprétable.

Dans le cas de l'évalutation 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 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 dans le cas du nombre de publications comme mesure:

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

et où le dénominateur symbolise la valeur escomptée

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

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

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

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 que le pays a dans un domaine est pondéré par le poids qu'a celui-ci 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.

Certaines études exploitent cet indice sous forme graphique en disposant les pays par domaine selon deux axes. L'un des deux 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 le poids du pays 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 selon, soit une contribution trop importante dans le domaine soit une trop faible (voir figure 18).

Cet indice a été utilisé pour différents objectifs d'étude:

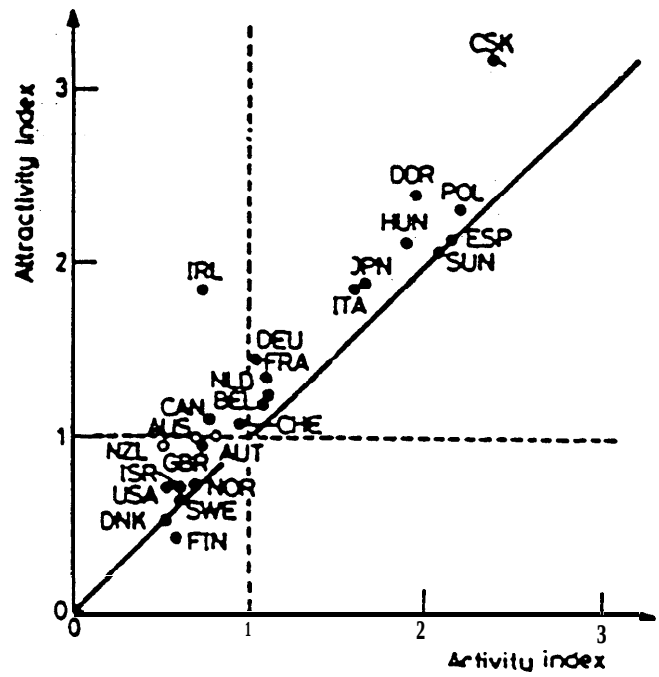
- ☞ Par Barré pour comparer les publications de 11 pays au travers de la base Pascal dans [BARR91] (voir *Analyses par croisements d'unité bibliographiques*).
- ☞ Schubert et Braun le mettent en oeuvre pour la citation dans [SCHU86] et évaluent 25 pays dans la recherche en chimie.
- ☞ Callon et Leydesdorff s'en servent pour estimer l'état de santé de la recherche française dans [CALL87] grâce aux deux mesures. Ils font remarquer que la citation comme indicateur n'est pas à considérer comme une valeur bien sûre puisque l'ISI privilégie plus particulièrement la couverture de certains pays par rapport à d'autres.

(Nous verrons que cet indice est aussi appliqué pour les études des dépôts de brevets par nation, Cf *La mesure des techniques et des technologies - Indicateurs univariés*)

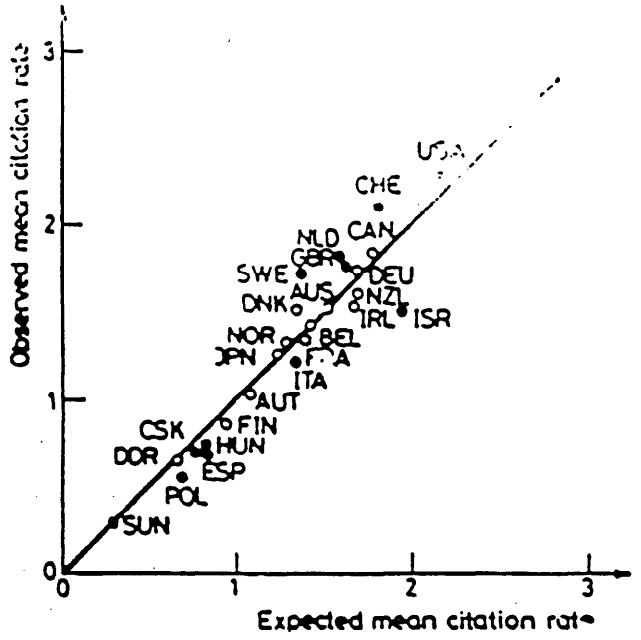
g) La mesure d'un domaine

Par l'emploi des différentes mesures présentées précédemment, de nombreux auteurs ont publié des études pour cerner les évolutions et les grands acteurs d'un domaine.

Un bon exemple est présenté dans l'article de Czerwon [CZER90]. Il montre l'emploi d'un panel de mesures pour analyser la dynamique de la spécialité *Monte Carlo methods in latice field theory* dans la physique théorique des hautes énergies: évolution des publications dans le temps, répartition des publications par pays, évolution des citations dans le temps, distribution des auteurs, distribution des revues spécialisées…



Relational chart displaying the attractivity *vs.* activity indices in chemistry (Papers published in 19784979; cited in 1980). Symbols: o – only *AI* differs significantly from 1, • – both *AI* and *AAI* differ significantly from 1 (at a 95% confidence level).



Relational chart displaying the observed *vs.* expected citation rates in chemistry (Papers published in 1978–1979, cited in 1980). Symbols: o – *RCR* does not differ significantly from 1, • – *RCR* differs significantly from 1 (at a 95% confidence level).

Figure 18: Représentation graphique du ratio de l'indice d'avantage faite dans [SCHU86]

h) Classement des indicateurs bibliométriques univariés

Dans [VINK88], Vinkler a recueilli un ensemble d'indicateurs bibliométriques univariés qu'il a classé selon plusieurs critères:

⇨ Nature du comptage:

Les indicateurs peuvent se diviser en deux groupes qui dépendent de la donnée de départ de la mesure:

- ❍ indicateurs de publications
- ❍ indicateurs de citations

⇨ Nature du calcul:

Ces deux types d'indicateurs peuvent eux-mêmes se partager en fonction des différents types des mesures qu'ils mettent en valeur

1 - type de mesures à caractéristique simple:
un comptage simple comme un nombre d'articles, un nombre de citations reçues

2 - type de mesures à caractéristique spécifique:
productivité en fonction d'un autre facteur comme un nombre d'articles par an en fonction du nombre de chercheurs ou du budget

3 - type de mesures à caractéristique balance:
comparaison entre une entrée et une sortie comme un nombre de citations données comparé au nombre de citations reçues

4 - type de mesures à caractéristique distribution:
mesure d'une donnée sous forme de part comme un nombre d'articles non cités par rapport au total d'articles publiés

5 - type de mesures à caractéristique relative:
mesure par rapport à une valeur étalon comme un nombre de citations par article en 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:

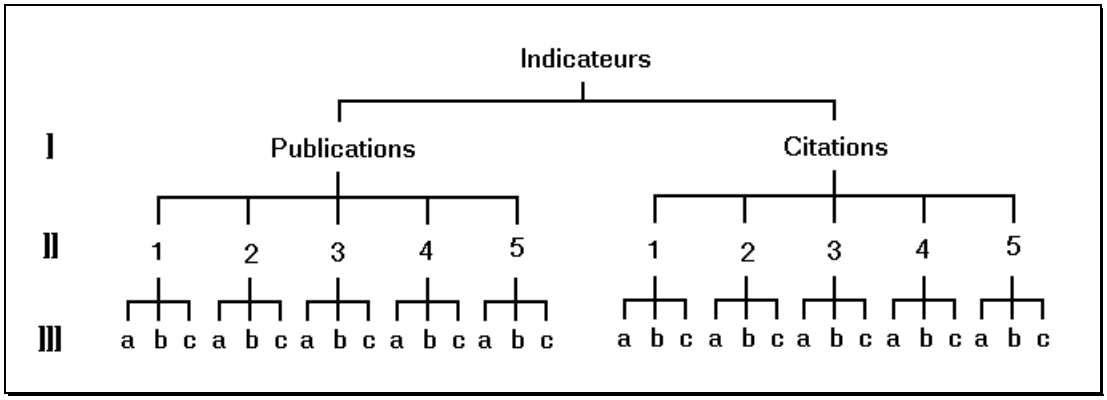
a - quantité

b - impact

c - quantité/impact

☞ Classification des indicateurs bibliométriques univariés:

On peut symboliser cette répartition par une arborescence, bien qu'elle ne soit pas de nature hiérarchique (les niveaux de l'arborescence pourraient être intervertis):



niveau I: type de donnée comptabilisée
niveau II: type de calcul appliqué à ces comptages
niveau III: nature de la mesure obtenue

☞ Niveau d'évaluation:

Ces indicateurs peuvent être mis en oeuvre pour différents niveaux d'évaluation:

	Niveau d'évaluation		
Type dévaluation	Micro	Meso	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 mesures sont utilisées comme des indicateurs en comparant les différentes valeurs des mesures entre les pays, les organisations, les thèmes, les types de publications…

Le problème majeur est de savoir à quel niveau et quel type d'évaluation il faut élaborer pour être sûr que la mesure soit totalement adaptée à l'unité bibliographique étudiée.

5. Les cartes relationnelles

Rapidement, les bibliométriciens ont 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 unique ordonnancement. Ils cherchèrent à disposer les éléments sur des cartes en deux dimensions plus adaptées à résumer le phénomène de coeur et de dispersion des littératures.

Cette idée de carte implique de pouvoir positionner les éléments les uns par rapport aux autres grâce à une métrique ou à 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 donc un degré de ressemblance ou de dissemblance entre les éléments.

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

Une fois que l'adaptation de ces méthodes aux études bibliométriques a été 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. Les exploitations de ces méthodes se sont alors diversifiées 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 tendance alors à parler d'indices "qualitatif".

a) Les méthodes des co-citations

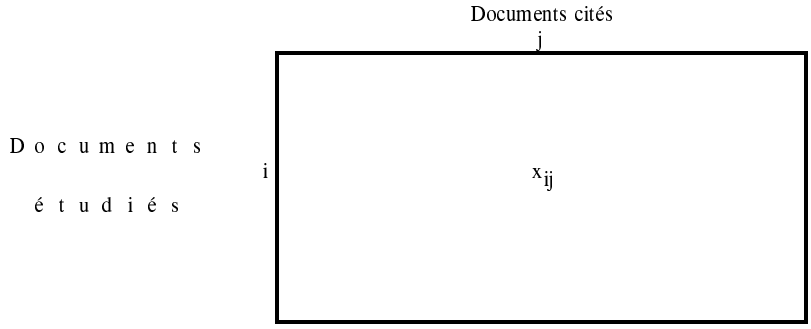
Les premières constructions de ces cartes ont été fondées à partir de réflexions se servant de la pratique de la citation comme élément relationnel entre les documents. Ces méthodes ont été développées par l'école de pensée américaine initiée par Garfield. Elles sont le fruit de collaborations entre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

(1) L'association bibliographique (bibliographic coupling)

C'est Kessler qui le premier a enrichi 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 [FANO56], Kessler élabora la méthode d'analyse bibliométrique par l'**association bibliographique** (*bibliographic coupling*) [KESS63]. **Kessler postula 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 l'association.

Il constituait un tableau comportant d'un côté les documents étudiés et de l'autre l'ensemble des citations qu'ils effectuent.



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

Ce tableau en entrée de classification automatique permet de construire des agrégats de documents selon la ressemblance de la pratique de citation qu'ils font aux travaux antérieurs.

Kessler a comparé les résultats de sa méthode à ceux d'une analyse sur les thèmes indexés et a conclu qu'il y avait une très forte corrélation des groupes formés par ces deux méthodes [KESS65].

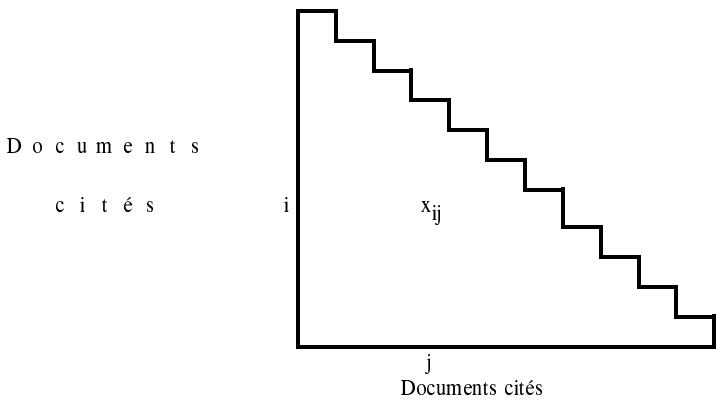
Cette méthode est malheureusement tombée en désuétude, probablement parce qu'à l'époque elle nécessitait une masse de données trop importante pour les systèmes informatiques de l'époque: la méthode impose la présence d'une entrée de tableau (ligne) par document à étudier.

(2) L'analyse de co-citation de documents

Small s'est inspiré des premiers résultats de la méthode de Kessler pour développer une méthode de cartographie qui est certainement la plus employée en bibliométrie: la ***co-citation analysis***.

Pour pallier au 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.



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

Ce tableau perd la trace des documents à l'origine des citations.

Il est d'adord traité pour normaliser la mesure de distance entre les documents cités, et ensuite injecté 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, ont été disposés par une méthode de cadrage multidimensionnelle non métrique.

Une fois cette structure établie, on cherche à connaître quels sont 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 une force de lien entre les différentes spécialités que représentent les souches**

Cette méthodologie développée par Small en 1973 [SMAL73], a été utilisée pour de multiples études, par exemple:

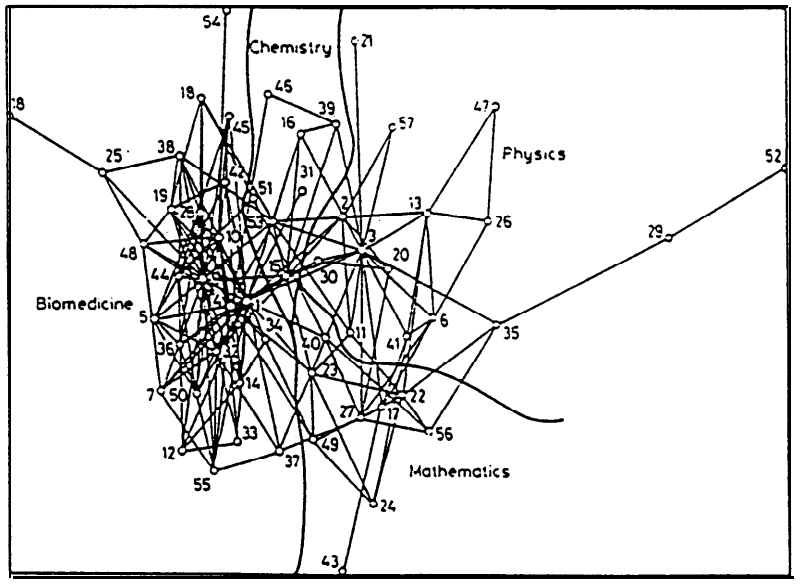
- ☞ Small et Griffith [SMAL74] et Griffith et al. [GRIF79] dans le domaine de la science naturelle
- ☞ Griffith et Small [GRIF83] en science sociale et en science de la connaissance
- ☞ Salton et Bergmark [SALT79] pour la science de l'informatique
- ☞ Small [SMAL81], Saito [SAIT84], Marshakova [MARS81] pour la science de l'information
- ☞ Mullins et al. [MULL84] et Hargens et al. [HARG80] dans le domaine de la sociométrie

…etc…

Mais, la plus importante est celle que Small et Garfield ont dirigé 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 que l'on vient d'exposer. 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 considérés 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 85, ils construisent un nouveau modèle [SMAL85]. La mesure des co-citations est calculée à partir d'une pondération de chaque citation par le nombre de citations présentes dans le document citant. De plus, le modèle consisté à créer des clusters 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 84, à partir de 13931 paires de liens entre documents cités, on obtient 3932 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-exécuté pour ces nouvelles relations, 502 agrégats sont générés en deuxième génération. Puis par une troisième itération ces 502 passent à 57 agrégats finaux. Ce résultat nous suggère une décomposition de la science en 57 "secteurs". Ces 57 secteurs sont liés entre eux. Leurs liens sont représentés sur une carte que l'ISI aime nommer ***Atlas*** de ***la*** *science* (figure 19).



1979 C4 map of science

Figure 19 : Atlas de la science obtenu à la suite de l'étude de 1978 sur les fronts de recherche de l'ISI

(3) L'analyse de co-citation d'auteurs

L'étude des co-citations d'auteurs déplace l'unité de l'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 contexte de co-citations, mais il focalise l'attention sur une durable et intéressante unité à mi-chemin entre les documents et les journaux: les auteurs eux-mêmes.

Les données sont donc obtenues par le comptage du nombre de documents qui citent deux auteurs simultanément. La matrice d'entrée pour l'analyse 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. Selon White [WHIT89] une interprétation des axes de la carte finale peut aboutir soit:

- ⇨ aux différents thèmes de prédilection des auteurs
- ⇨ aux différences de styles de travaux (tel que le degré de formulation mathématique).

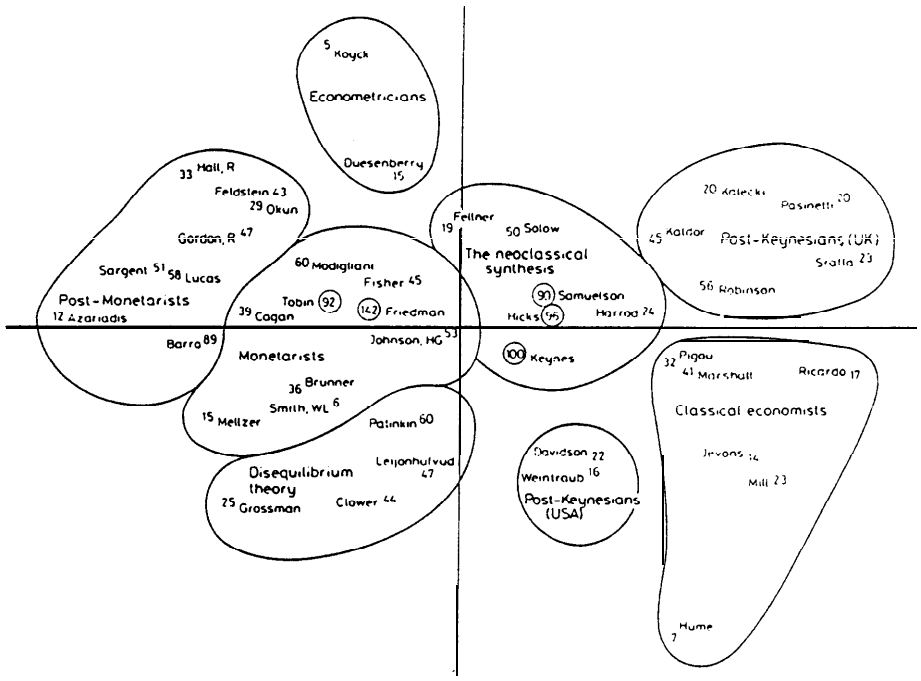
Il argumente aussi que les auteurs qui sont proches sur la carte n'ont pas seulement le sujet et la méthode en commun, mais aussi un lien de type collaboration (d'autres possibilités de liens sont le langage, la période, le pays ou l'idéologie). **La cartographie est ainsi le symbole du jeu social et de la structure intellectuelle**.

La technique a été introduite par White et Griffith [WHIT81] pour cartographier la science de l'information. D'autres auteurs l'ont appliqué pour diverses études:

- ☞ en ethnologie et en sociologie par Hopkins [HOPK84]
- ☞ en macro-économie et en génétique de la Drosophile par Mac Cain [MCCA83], [MCCA86]
- ☞ en théorie micro-économique par Penan [PENA92]

Ce dernier 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 d'eux pratiquement en automatique alors que jusqu'à présent les cartes imposées l'intervention d'un expert pour donner des noms aux groupes.

Un exemple de carte d'analyse de co-citation d'auteurs provenant de l'étude [MCCA83] est donné par cette figure 20:



A map of macroeconomics

Figure 20: Carte de co-citation d'auteurs dans le domaine de la macro-économie

(4) L'analyse contextuelle des co-citations

Imaginée par Small, ces premiers travaux ne sont restés qu'à l'état de projets. (résumé de ces premiers travaux SMALL 82). **Il a essayé de mettre au point des techniques pour isoler certains passages de textes de document dans une analyse deco-citation**.

Il estime que les documents cités symbolisent les concepts de ceux qui les citent et donc leur analyse permettrait de reconstituer la connaissance des principaux thèmes. Généralement les auteurs qui citent de précédents travaux le font en liaison avec une phrase qui reprend un concept exposé dans le document cité. L'analyse de cette phrase indiquerait à quelle unité d'information la citation fait référence. Connaître dans quel contexte la citation est faite dans le texte du document citant apporterait des informations que la seule analyse de titres ne pourrait pas livrer.

Une seconde idée est l'analyse de contexte de co-citations, c'est-à-dire l'analyse des phrases dans lesquels deux documents seraient cités en tant que support à l'argumentation. Pour que ce soit réalisable, il ne faut considérer que les co-citations assez proches (faites dans le même paragraphe par exemple) pour que leurs relations soient raisonnablement claires.

Mais les synoptiques d'algorithmes imaginés par Small imposent l'emploi d'un ahurissant matériel et de techniques trop complexes pour qu'il puisse les automatiser.

(5) 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 à la validité de cette mesure lors de la présentation de la citation comme indicateur univarié. 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 [JAGO90] qui présente un agrégat-coeur construit par l'analyse de co-citation de la base SCI en 1978. 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 citer une étude approfondie de validation de l'analyse des co-citations (ainsi que de l'analyse des mots associés) qui a été commanditée par le *UK advisory board for research councils* [HEAL86]. Les résultats livrés par le CSI ont été confrontés par des interviews de spécialistes scientifiques. Les principaux points de critiques formulés en conclusion sont ceux-ci:

- Les méthodes des co-citations réduit les risques de mauvaise évaluation en se basant sur une évaluation des articles par le "plébiscite" d'autres scientifiques du domaine en question. Le prix de 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 révélera jamais l'émergence des nouvelles spécialités en temps réel**

- Ce qui intéresse le plus les décideurs ne sont pas les finesses de descriptions des domaines mais plutôt les interfaces entre les domaines car elles sont sources de renseignements précieux. Hors le **déséquilibre de la pratique de citation entre chaque discipline de la science la méthode des co-citations déforme ces zones interdisciplinaires**.

Des critiques ont été aussi formulées sur les méthodes mathématiques que ces analyses de co-citation 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 [LEYD87].

- Comme pour l'évaluation anglaise, le *Advisory council for science policy (RAWB)* hollandais a financé une étude sur la méthode. Il n'a pas reposé 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 ces études, ils ont établi un projet de conception d'une simulation informatique par génération aléatoire, sous des contraintes, d'un tissu de co-citation dans un corpus virtuel (répartition Zipfienne des citations, nombre de citations par référence, la valeur de la citation maximale…) [OBER88]. En conclusion, l'instabilité statistique des résultats des structures de groupes de l'analyse de co-citation ne semble pas avoir été correctement appréciée par l'ISI. **Et l'étude révèle de sérieux problèmes. Ces problèmes suggérent que les résultats de l'analyse de co-citation ne peuvent pas**

être pris sérieusement comme une preuve pertinente de formulation de la politique 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.

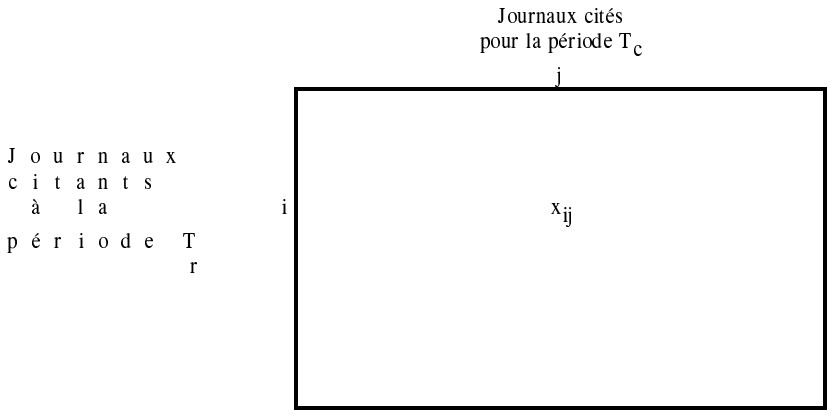
En conclusion, on peut dire que la co-citation reflète le monde de la science comme les scientifiques la perçoivent et plus sous un aspect historique et épistémologique que de prospective.

(6) L'analyse des citations-croisées de journaux (cross-citation)

Pour décrire un domaine de recherche plus étendu qu'un front de recherche et au moyen des citations, les études bibliométriques se reportent sur une unité d'analyse encore plus large: les journaux. 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 journaux, leur réseau de communication pour identifier les journaux centraux et périphériques dans la spécialité, l'existence de sous-spécialités, et pour modéliser le flot d'information transitant entre eux.

Différentes approches pour identifier les groupes de journaux sont proposées, mais toutes traitent une **matrice de citation-croisée** (*cross-citation*) comportant en ligne les journaux citants et en colonnes les journaux cités:



x_{ij} = nombre d'articles dans le journal *i* qui citent des articles du journal *j*

Le choix de la période T_c de temps 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".

Il est évident qu'une forte activité de publications pour un journal influencera son taux de citations Ci. Pour réduire cette influence les auteurs ont envisagé de nombreuses possibilités de normalisation. Dans l'article [TODO87] Todorov et Braun ont récapitulé les principaux calculs proposés dans les travaux antérieurs. Ils traitent aussi des propositions faites pour inhiber les fortes valeurs de la diagonale (les articles citent souvent des articles du même journal) caractéristiques de ce type de matrices que Price a intitulé ***matrices de transaction*** [PRIC81b]

Différentes méthodes mathématiques peuvent être ensuite appliquées à cette matrice normalisée:

- Leydesdorff dans [LEYD86] a appliqué des méthodes d'analyse factorielle ou de positionnement multidimensionnel qu'il a ensuite comparé avec des méthodes de classifications automatiques (simple lien et Ward) dans [LEYD87].
- Doreian a mis au point par Doreian d'une technique basée sur la méthode du block-modelling [DORE85]
- Narin et Carpenter [CARP73], Niyamoto et Nakayama [NIYA83] ont eux exploité les méthodes de regroupement traditionnelles eu analyses bibliométriques.
- Agirre et al. appliquèrent l'analyse des correspondances dans [AGIR91].

La principale critique, rencontrée dans la littérature, 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, contient un pourcentage d'erreurs non négligeable et une couverture restreinte et hétérogène.

b) Les méthodes des cooccurrences de mots (co-word)

Cette méthode établit l'analyse de la rhétorique des publications scientifiques en considérant que les mots clés, affectés par les auteurs (titres) ou les indexeurs (mots-clés) des bases de données scientifiques, reflètent les étapes de l'argumentation scientifique des auteurs.

Quand une paire de mots (-clés) est utilisée pour indexer un grand nombre d'articles alors ces deux mots représentent une forte association entre les problèmes ou les concepts auxquels ils se réfèrent. Donc cette méthode est basée sur l'étude des cooccurrences de mots par des méthodes statistiques pour **découvrir les agrégats de mots, symbolisant les thèmes de problématiques, et leurs situations les uns par rapport aux autres**

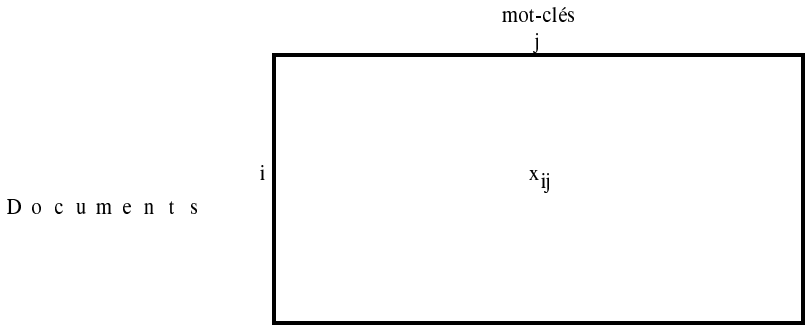
La méthode des cooccurrences de mots a été presque exclusivement conçue par une école de pensée française. **Elle est en fait la conséquence d'une conjonction entre le principe de modélisation de la science (bibliométrie-scientométrie) et l'approche sociologique de la science (représentation sociale de la connaissance)**. La collaboration entre le CSI et l'INIST (ex CDST) 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 base scientifique multidisciplinaire Pascal produite par l'INIST. Elle a été ensuite reprise par divers auteurs pour la plupart appartenant aux écoles de pensée Hollandaise et Anglaise.

(1) Les réseaux "socio-techniques" (mots associés)

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'Ecole des Mines de Paris et du Centre de Documentation Scientifique et Technique (actuel INIST). La dernière version de cette méthode a été mise au point au cours de la thèse de Michelet [MICH88] et a abouti à la création du logiciel **Leximappe**. C'est ce logiciel qui est exploité actuellement dans de nombreuses études par les deux centres de recherches [LAW88], [CALL91] et qui sera exposé dans cette partie. Il fait aussi l'objet de nouveaux projets en tant que moteur d'analyse auquel on adjoint soit des pré-traitements [POLA91] soit des post-traitements [CARD92].

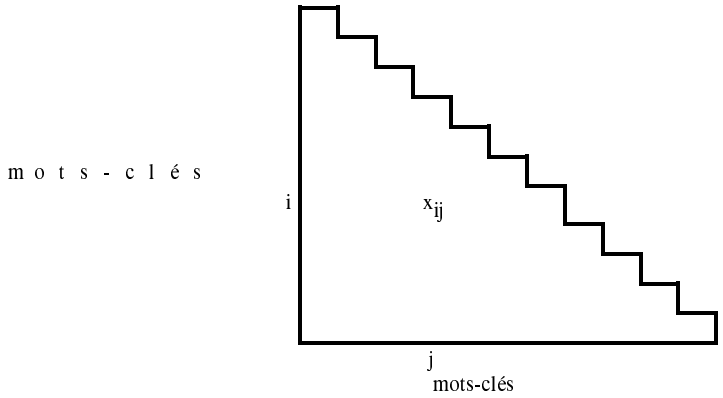
⇨ La méthode Leximappe:

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 présents dans l'ensemble des références étudiées (jusqu'à concurrence de 1500 mots-clés).



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

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).



x_{ij} = valeur du coefficient d'équivalence pour le mot-clé *i* avec le mot-clé *j*

Ensuite, l'objectif de la méthode est de réaliser des agrégats de mots-clés sur la base des mesures de "distance" entre mots-clés. Mesures qui sont récapitulées dans la matrice carrée d'association. Pour ceci, 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 **taille maximale de 10 mots-clés par agrégat**. Les classes résultantes sont donc très **hétérogènes**: la première classe obtenue sera constituée des mots-clés les plus fortement liés alors que la dernière sera constituée de tous les mots clés "rebuts" puisque étant très faiblement liés à tous les autres mots-clés.

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 (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 par le nom de "diagramme **stratégique"** (figure 21).

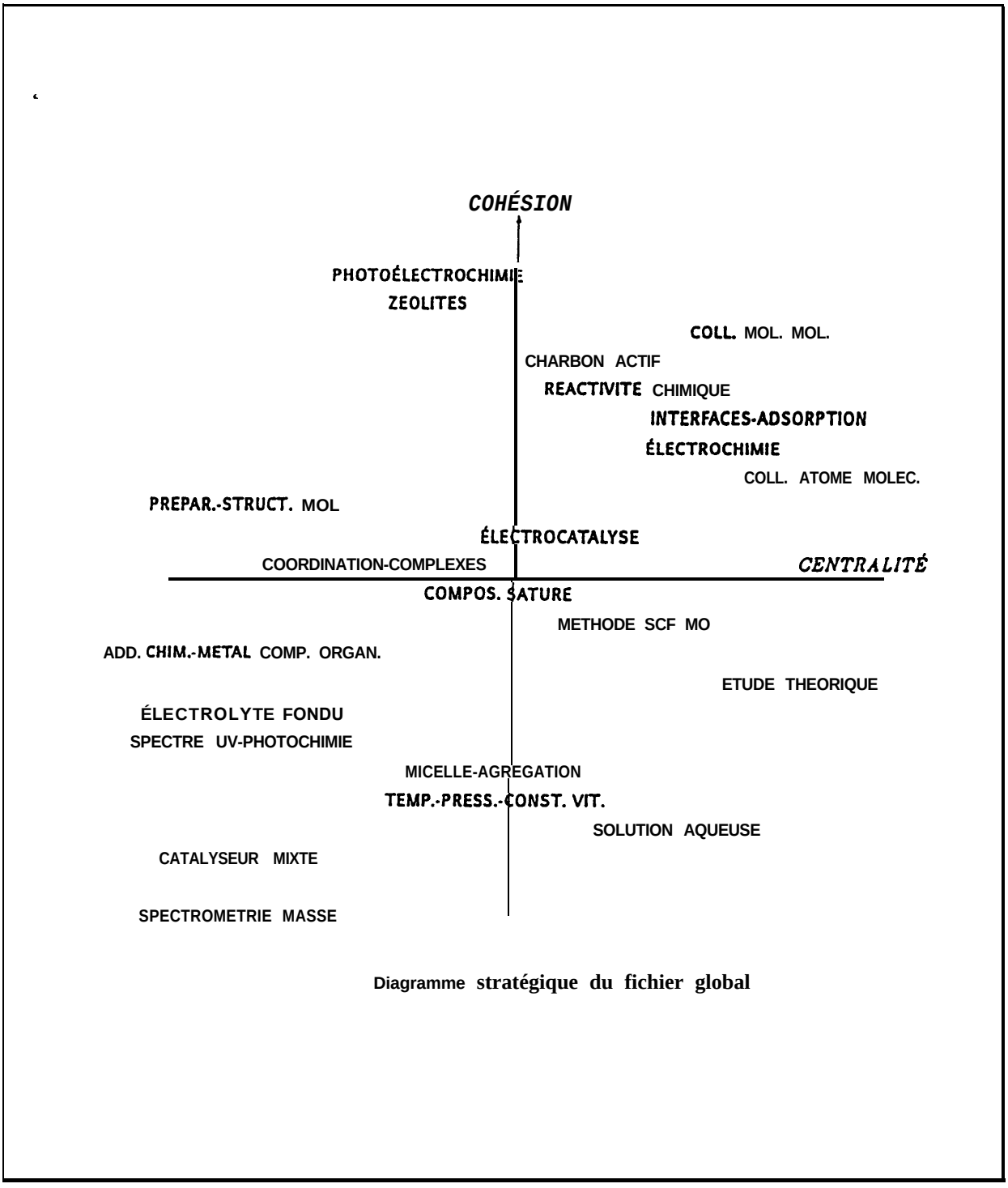


Diagramme **stratégique du fichier global**

Figure 21: Diagramme stratégique des mots-clés [MICH88]

⇨ La critique:

Ici, ce n'est pas le principe 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, que l'on peut faire au développement informatique qui en a été fait, est de ne pas avoir donné suffisamment de moyens de jouer sur le traitement en fonction des données à étudier: seuil de troncature automatique, agrégat de même taille, mesure des "distances" entre mots fixée une fois pour toutes et principe d'agrégation irrévocable.

Le logiciel 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 son mémoire de thèse comme une "boite noire"). **Mais en contre partie, l'utilisateur du logiciel n'a pas la possibilité de nuancer son étude en fonction de la spécificité des données de départ.**

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 étant les représentants synthétiques des ces problématiques. Ainsi LEXIMPAPPE a été mis au point pour traiter les descripteurs de la base scientifique multidisciplinaire Pascal produite par l'INIST. Lors d'études ultérieures, cette méthode a été employée avec d'autres champs où les mots présents ont été jugés comme étant de bons représentants de ces problématiques: le champ titre pour les publications scientifiques ou le champ titre normalisé pour les références brevets de la base Derwent. **Cette méthode est donc confinée à l'analyse du contenu d'un seul champ et pas n'importe lequel**

(2) Les autres méthodes d'analyse des cooccurrences de mots

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

⇨ Analyse des mots du titre par Leydesdorff (Hollande) [LEYD87]:

Pour éliminer un possible "effet d'indexation" pendant son étude, Leydesdorff a utilisé les **mots originaux du titre et du résumé des documents**.

L'échantillon des références est constitué des documents originaux rédigés par un laboratoire en Biochimie de l'université d'Amsterdam durant la période 1979-1982: il n'a sélectionné que 47 textes complets sur 57 documents. Après élimination des mots en dessous d'un seuil de fréquence (titre < 2, résumé < 4) et des mots triviaux, il construit la matrice carrée et symétrique du croisement des mots avec une diagonale vide qu'il transforme en matrice de **corrélation de Pearson**. Il traite cette matrice par la **méthode de Ward** pour constituer les agrégats des mots.

Il explique qu'il a choisi cette procédure mathématique car la méthode des liens simples lui semble totalement inadaptée. Comme la matrice est presque "vide" elle génère un effet de chaînage sur le premier groupe.

Ses conclusions concernant son analyse des co-words sur les titres et les résumés sont:

- La cooccurrence des mots semble très bien traduire la spécificité des axes de recherche. C'est un très bon indicateur des structures internes pour un groupe d'auteurs restreints.
- Les mots des résumés sont moins spécifiques que ceux des titres et concordent moins bien avec le sujet

⇨ Analyse des mots associés modifiée par Law et Whittaker (Grande Bretagne) [LAW92]

Law et Whittaker ont conduit une étude sur les mots-clés d'un échantillon de références collecté sur la base française Pascal. Ils ont ensuite découpé l'ensemble des références en 5 périodes de temps pour les analyser par une analyse des co-words.

La procédure est là strictement la même que celle développée par le CSI/INIST mise à part quelques modifications:

- le regroupement les mots dans les groupes:
  en cours de regroupement il n'y a pas de calcul de distances entre les agrégats formés et les points restants mais simplement une répartition des mots en descendant la liste décroissante des mesures d'association des paires de mots
- la taille maximale des agrégats:
  les mots sont réunis dans un agrégat jusqu'à une taille limite de 15 mots

- le calcul de deux nouveaux indices pour mesurer le chevauchement entre les thèmes de sujets similaires qui surviennent au cours des périodes successives de temps.

- les graphes fournis pour chacune de ces périodes. Ils sont construits de façon à:
  - disposer les agrégats concernant les mêmes concepts à peu près aux mêmes positions sur les 5 graphes
  - symboliser les intensités de relations entre et à l'intérieur des agrégats par des nuances d'épaisseur de traits
  - rendre compte du nombre de documents contribuant à la création de ces agrégats par des surfaces de carrés différents
  - le calcul des deux nouveaux indices permet de construire des graphes qui retracent la "génération" des thèmes.

(3) 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.

En fait, les mots-clés appartiennent à plusieurs catégories de sens selon qu'ils représentent un aspect technique, technologique, composant chimique, caractéristique physique, traitement, condition expérimentale, matériel utilisé, 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: 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 d'analyse de mots-clés:

- Dans une étude menée pour l'entreprise Burton-Corbin (fabricant de compresseurs) [DEVA90], 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

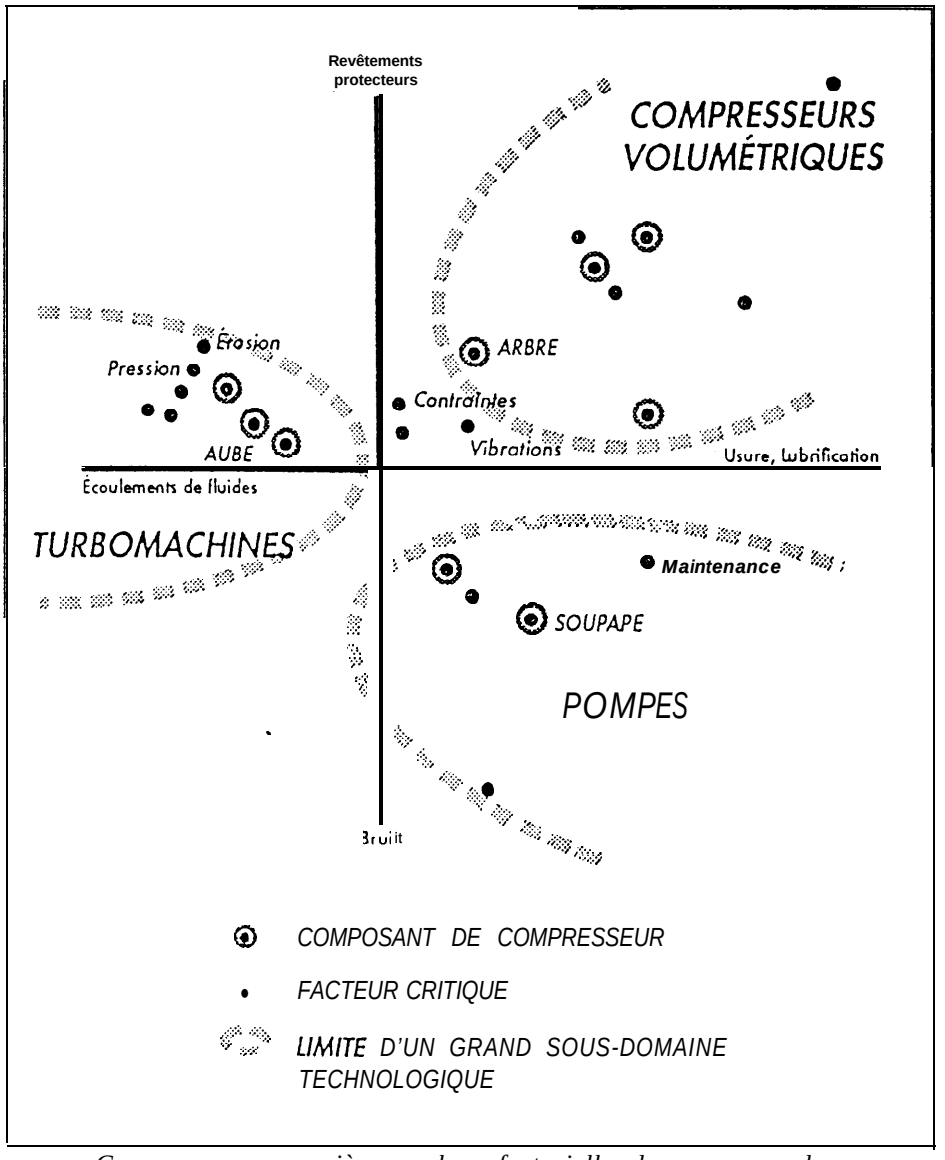
de sollicitation (pour 30000 références provenant d'une quinzaine de bases). Le graphe factoriel (figure 22), obtenu par une analyse des correspondances de cette matrice, a donné une image assez générale qui a permis à la société Burton-Corbin de recentrer l'étude sur les compresseurs volumétriques. Une seconde matrice, croisant les mêmes catégories de mots-clés mais ceux-ci étant choisis à un niveau plus fin (plus que 1000 documents concernés), a fourni une seconde image plus détaillée (figure 23), après une analyse des correspondances et une classification automatique.

- La même technique a servi pour d'autres études bibliométriques du CETIM:
  - tableau technologies - marchés dans le domaine de la productique [BELO]
  - tableau équipements - composants pour les références de trois années de la base du CETIM [DEVA89]
  - tableau technologies - marchés dans le domaine des revêtements de surfaces [BELO92]

(4) Qu'apportent ces nouveaux travaux?

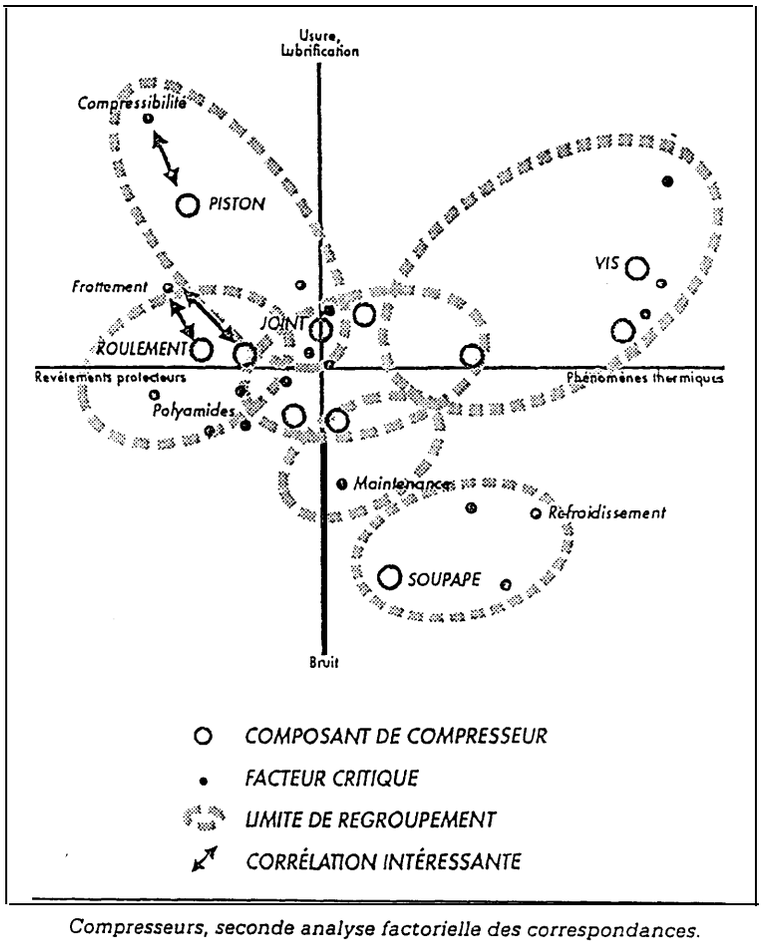
Les travaux que nous venons de citer ont cherché à améliorer soit la procédure d'agrégation par des méthodes statistiques plus adaptées soit la qualité des représentations graphiques finales soit la spécificité de l'étude. **Mais Leximappe conserve le considérable avantage d'automatiser pratiquement toute la chaîne des traitements dans l'analyse alors que celle-ci reste principalement manuelle**.

Remarque: ceci n'est plus vrai pour le CETIM puisqu'il vient récemment d'acquérir l'outil bibliométrique DATAVIEW conçu pendant cette thèse. Il leur offre maintenant un traitement totalement automatisé des corpus bibliographiques.

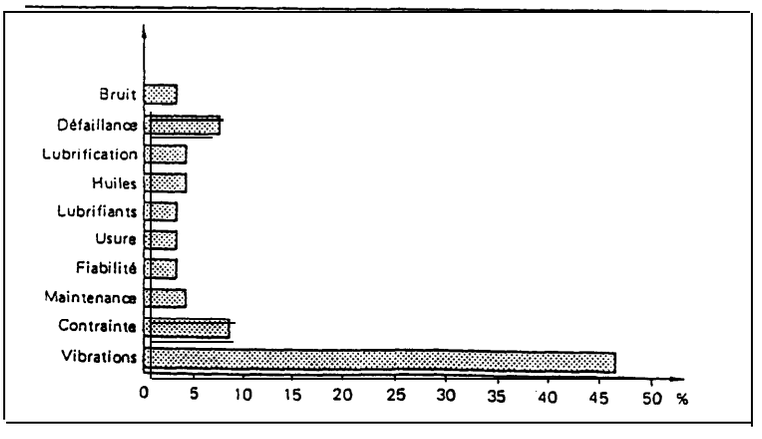


Compresseurs, première analyse factorielle des correspondances.

Figure 22: Première analyse factorielle [DEVA90]



Compresseurs, seconde analyse factorielle des correspondances.



Pourcentages de co-occurrences entre le composant « arbre » et les facteurs critiques (nombre **total** *de co-occurrences = 92).*

Figure 23: Seconde analyse factorielle [DEVA90]

c) Les autres analyses de relations

Plus récemment les auteurs en bibliométrie ont élargi les domaines d'application des méthodes de cartographie pour représenter l'information contenue dans d'autres champs que les citations, les mots-clés ou les titres.

Les premières méthodes de cartes relationnelles, que nous venons de présenter, étaient fondées sur des principes sociologiques. L'analyse de co-citation est bâtie sur des conceptions épistémologiques de la science (Popper, Kuhn…) et l'analyse par les mots sur des conceptions de la représentation sociale en science (Moscovici, Jodelet…). **Les méthodes mathématiques employées ne sont dans ces conditions que des outils pour essayer de représenter ces concepts par l'intermédiaire de la littérature scientifique**

Les méthodes qui sont exposées par la suite s'éloignent des premières par leur **approche moins conceptuelle mais plus pragmatique**. Plus aucun phénomène sociologique n'est à mettre en évidence, mais simplement une **volonté de mieux comprendre de vastes et complexes masses d'information**. **La statistique descriptive est exploitée dans ce cas exactement pour ce qu'elle a été créée: un outil de description de données ou d'information et non pas un outil de modélisation de phénomènes sociologiques**

Dans ce contexte, tout type d'information est traitable par ces méthodes mathématiques tout au moins si elle peut se mettre sous une forme quantifiable. Pour l'information scientifique contenue dans les références bibliographiques, toute unité d'information peut faire l'objet d'une étude. Des analyses sur des unités bibliographiques comme les auteurs, les codes documentaires, les pays… sont alors apparues.

Certaines techniques mathématiques sont bien adaptées pour traiter des tableaux autres que des tableaux symétriques d'association (encore nommés *tableaux de transactions bibliométriques*). Des tableaux croisant des unités d'informations différentes seront donc analysables (cas d'analyse déjà rencontrée pour l'*association bibliographique*, l'*analyse des citations-croisées* des journaux et les *tableaux de contingences de mots-clés*). **Les renseignements apportés par les analyses de croisements sont bien souvent d'une très grande qualité et d'une très grande pertinence**.

Voici quelques exemples de ce genre d'étude…

(1) L'analyse des codes documentaires

Un des systèmes d'indexation des références dans les bases des données est l'affectation de codes appartenant à une classification documentaire. La classification d'une base de données découpe le domaine couvert par la base en secteurs d'activité (Cf la partie *Les champs indexés des bases de données*). Ces secteurs sont symbolisés par des codes (succession de caractères alphanumériques).

L'utilisation de ces codes s'est avérée propice à la réalisation d'études bibliométriques. Ils permettent l'analyse des thèmes de recherches 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:

- ☞ condensation de concepts en un seul code donc un sens plus synthétique
- ☞ absence de synonymie donc aucune ambiguïté sémantique
- ☞ pérennité dans le temps donc pas de déviances du langage
- ☞ diversité des termes plus faible donc meilleure qualité statistique
- ☞ traitement plus facile donc gain de temps

Nous avons recueilli ici quelques travaux bibliométriques exploitant l'information des codes documentaires:

⇨ Réseau de paires de codes par Dou et al.:

Dans les analyses bibliométriques relationnelles, le traitement le plus simple est la construction de **réseaux de cooccurrence**. Cette construction est simple parce qu'elle ne fait appel à aucun calcul mathématique. Mais elle offre l'avantage de livrer des cartes dont l'interprétation est immédiate car elle ne fait appel qu'à la notion de fréquence de co-apparition d'unité bibliographique. En fait, on peut considérer que le principe est le même que celui d'une analyse statistique. **La mesure de la "distance" entre les éléments n'est pas calculée par une "normalisation" des données mais elle correspond aux données brutes elles-mêmes: les fréquences de co-apparitions**. Cette méthode privilégie donc les éléments à très fortes fréquences en négligeant les relations entre les éléments rares (Cf partie *Tableaux d'indice d'association*).

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 (figure 24) [DOU88], [DOU90], [DOU91]. Elle s'est particulièrement penchée sur la **réalisation de logiciels spécifiques pour traiter en automatique les références bibliographiques collectées sur les serveurs professionnels.**

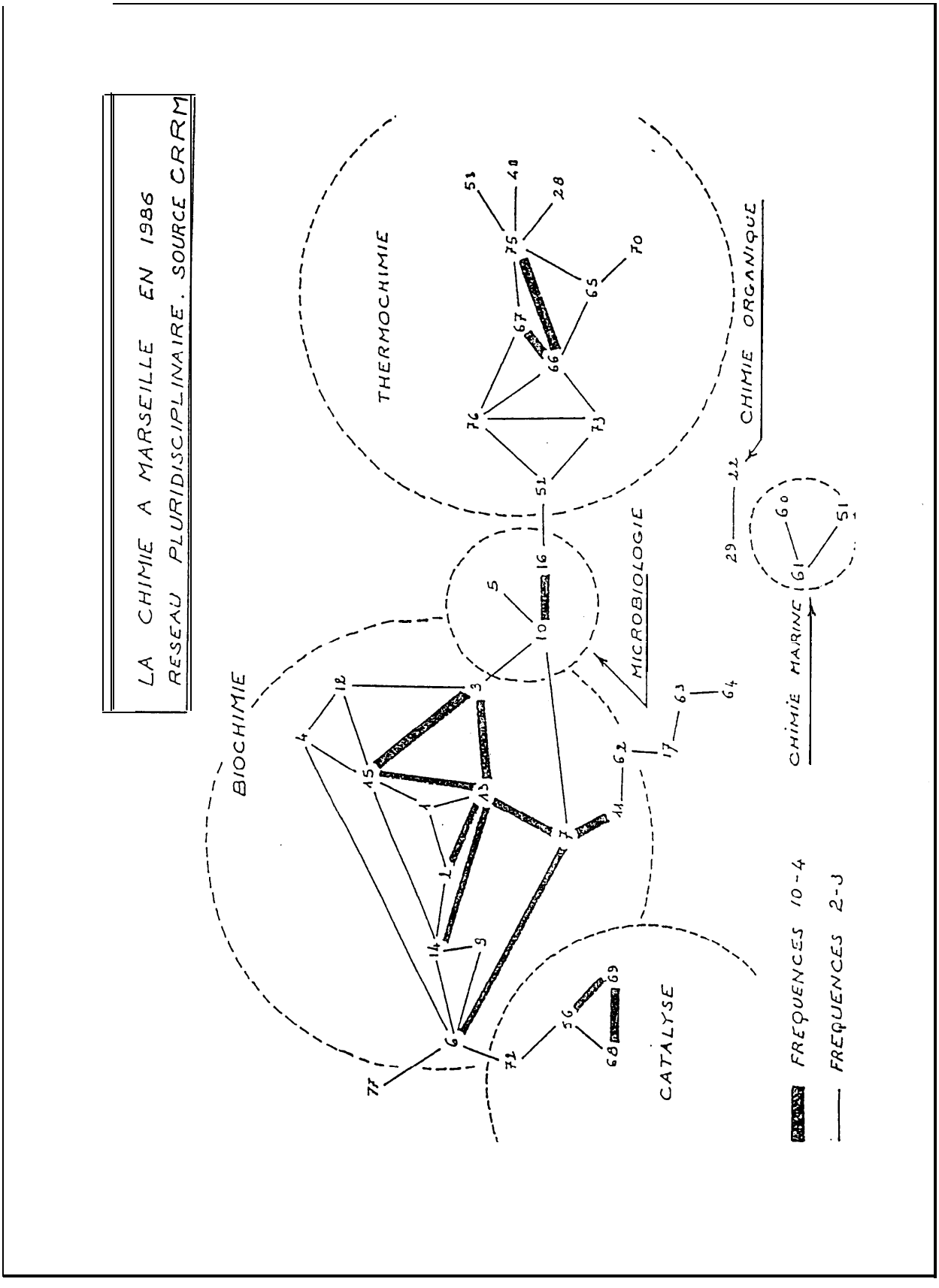


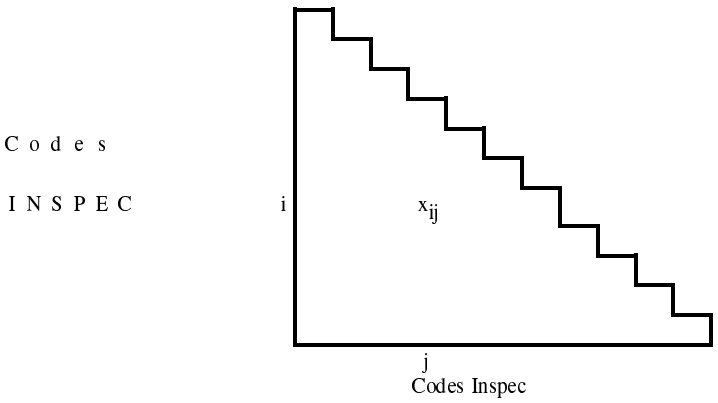
Figure 24: Réseau de paires de codes

⇨ Analyse de co-heading par Todorov et Winterhager:

Todorov et Winterhager ont imaginé en 1990 un nouveau traitement bibliométrique pour analyser les codes de classification documentaire [TODO90].

Le principe de la méthode est le suivant:

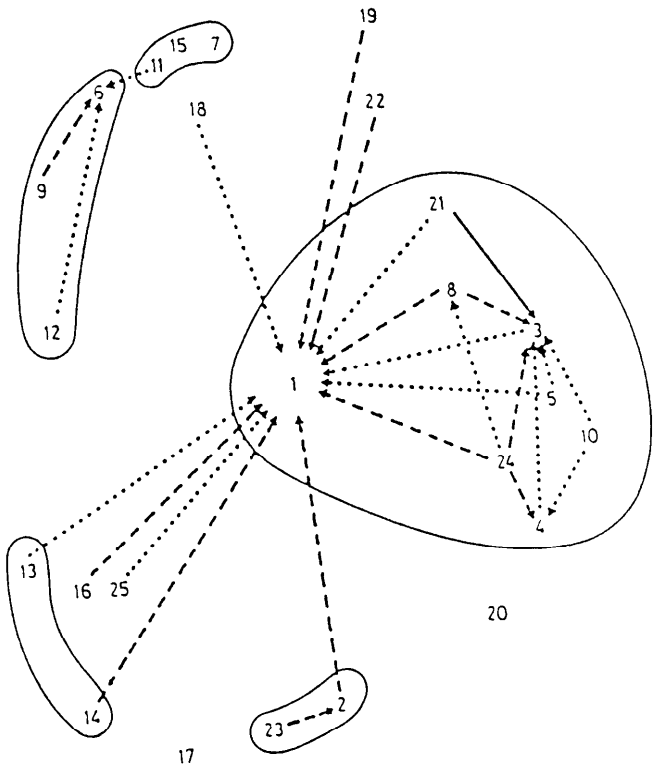
- comptage des occurrences et des cooccurrences de codes, puis construction de la matrice triangulaire des codes apparaissant le plus souvent (**un logiciel spécifique a été développé**).



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

- calcul de la matrice des "distances" des relations entre les codes (**mesure par l'indice d'inclusion**)
- **cadrage multidimensionnel** de cette matrice des relations pour positionner les points sur un plan (programme ALSCAL)
- définition des agrégats de codes par une **classification hiérarchique ascendante** sur la base de la matrice des relations
- ajout sur le graphe des forces de relations les plus élevées par des traits reliant les points ainsi que la symbolisation des agrégats.

Les auteurs ont appliqué leur méthode dans des études à partir de la classification documentaire présente sur la base INSPEC (*Physics Abstracts*) [TODO91], [TODO92]. Un des graphes construits est illustré par la figure 25.



Inclusion map for 1988 Australian gcophysics. (Plot symbols are given in Table 6)

——— Inclusion index values ≥ 0.8
— — — — Inclusion index values ≥ 0.6
. . . . . Inclusion index values ≥ 0.5

Figure 25: Graphe de relation entre codes INSPEC [TODO90]

Pour les auteurs, 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 dont les suivants:

- ☞ 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 qui se limite aux articles des journaux les plus cités
- ☞ être plus objectifs que les mots employés par l'auteur car basés sur un jugement extérieur d'indexeurs.

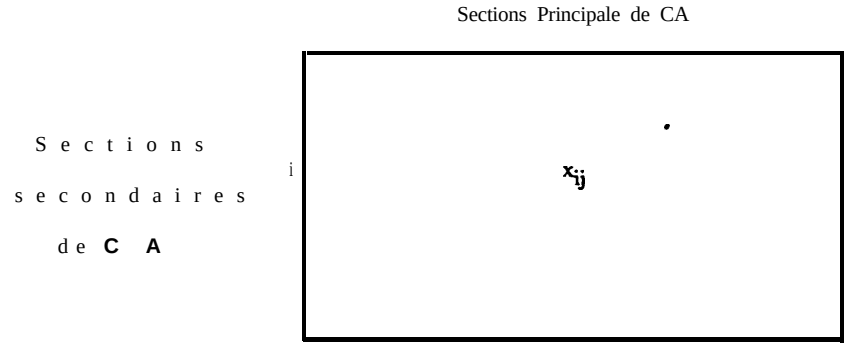
Dans leurs études, ils ont parallèlement étudié l'information contenue par les termes contrôlés d'INSPEC (champ CT: *Controlled Terms*). Les résultats sont identiques avec des renseignements plus détaillés pour les termes contrôlés.

⇨ Analyse de co-subfield par Van Raan et Peters:

Après avoir employé dans une étude précédente trois méthodes d'analyses bibliométriques pour caractériser la dynamique de l'ingénierie en chimie (une *analyse de citations* entre journaux, une *analyse des relations entre les journaux et les thèmes* par l'intermédiaire de codes de classification et une *analyse des co-words* sur les mots-clés), ils présentent dans [VANR89] une méthode basée sur l'étude des relations entre les codes de classification pour le même domaine. Le but est de décrire le transfert de connaissance entre différents thèmes et, si possible, tracer des processus de synthèse de connaissance.

La méthode d'analyse qu'ils mettent en oeuvre exécute les étapes suivantes:

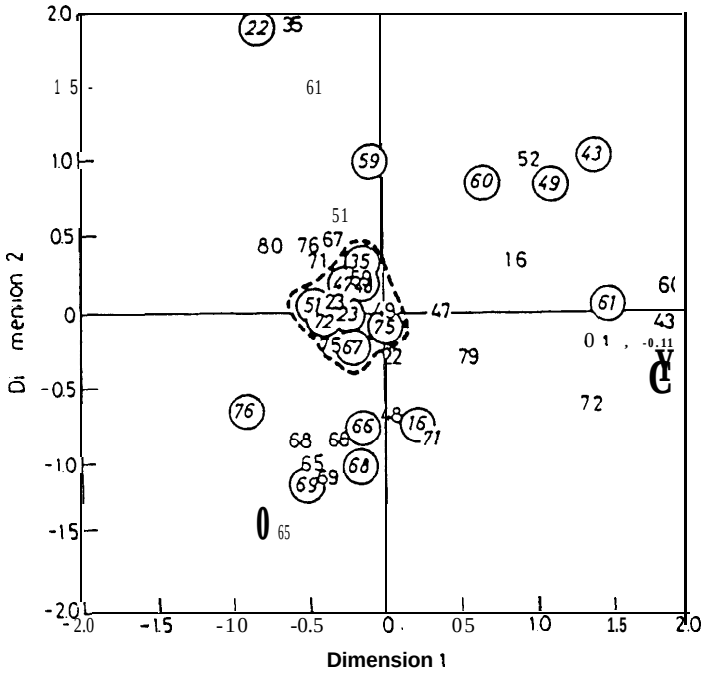
- ❍ Exploitation des 80 codes de classifications de Chemical Abstracts pour connaître les paires entre les codes principaux (champ SC: Main Sections) et les codes secondaires (champ SX: Cross Sections) de façon à construire les structures en réseaux des thèmes pour des périodes de temps successives (77-79, 80-82, 83-85)



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

Pour cette étude, le dénombrement des croisements de codes est réalisé par interrogation en ligne des combinaisons SC *x* SX *x* période.

0 Application la méthode **d'analyse par quasi-correspondance** (quasi-correspondence analysis QCA) qu'ils ont mis au point pour disposer les codes sur un espace à deux dimensions **(figure 26).**



QCA-Display of subfield-to-subfield relations for rhe pcriod 1983–1985. Circled italics: main CA-cections (main subfields). Bold: CA cross-sections (secondary subfields). Variance dim. 1: 18%, dim. 2:16%

Figure 26: Cartographie des codes CA par la méthode QCA [VANR89]

En comparant ces résultats à ceux obtenus par les précédentes études réalisées sur les données, les auteurs concluent que:

- ☞ l'analyse des cooccurrences de mots-clés produit des résultats plus fins et, contrairement à une classification focalisée sur le découpage d'une seule discipline, elle décrit bien l'émergence d'applications de techniques appartenant à une autre discipline. Dans le cas étudié, elle a montré un récent développement des modèles mathématiques, tandis que cette technique est noyée dans une section générale (48: unit operations & processes) dans l'analyse de "co-thèmes". Mais par contre l'analyse des cooccurrences 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.
- ☞ les inconditionnels de l'analyse des co-citations clament l'importance 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.

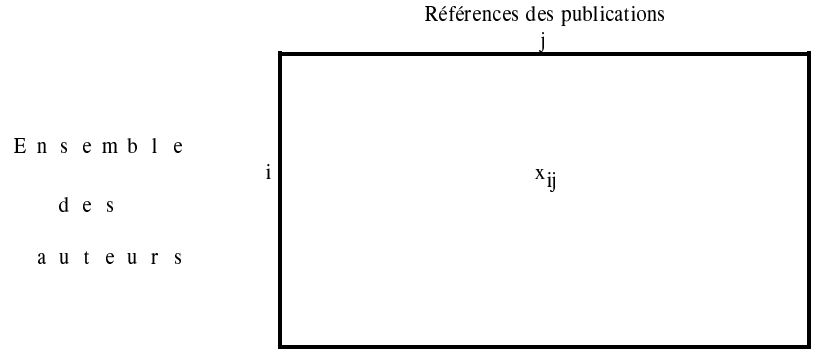
Ils considèrent donc que cette technique est à prendre en considération au même titre que les autres, chacune apportant une vision différente.

(2) L'analyse des co-signatures

Price et Beaver sont 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** [PRIC66]. Ils ont recherché à leur époque, manuellement, les "**collèges invisibles**" (*Invisible colleges*). La reconstitution des groupes de collaboration autour d'un auteur commençait par une recherche de tous auteurs qui avaient travaillé avec lui, puis les nouveaux auteurs qui avaient publié avec ces derniers et ainsi de suite.

Peters et Van Raan ont repris le concept en 1991 [PETE91] et l'ont adapté à une méthode de regroupement automatique. 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:

- construction d'une matrice de co-auteurs

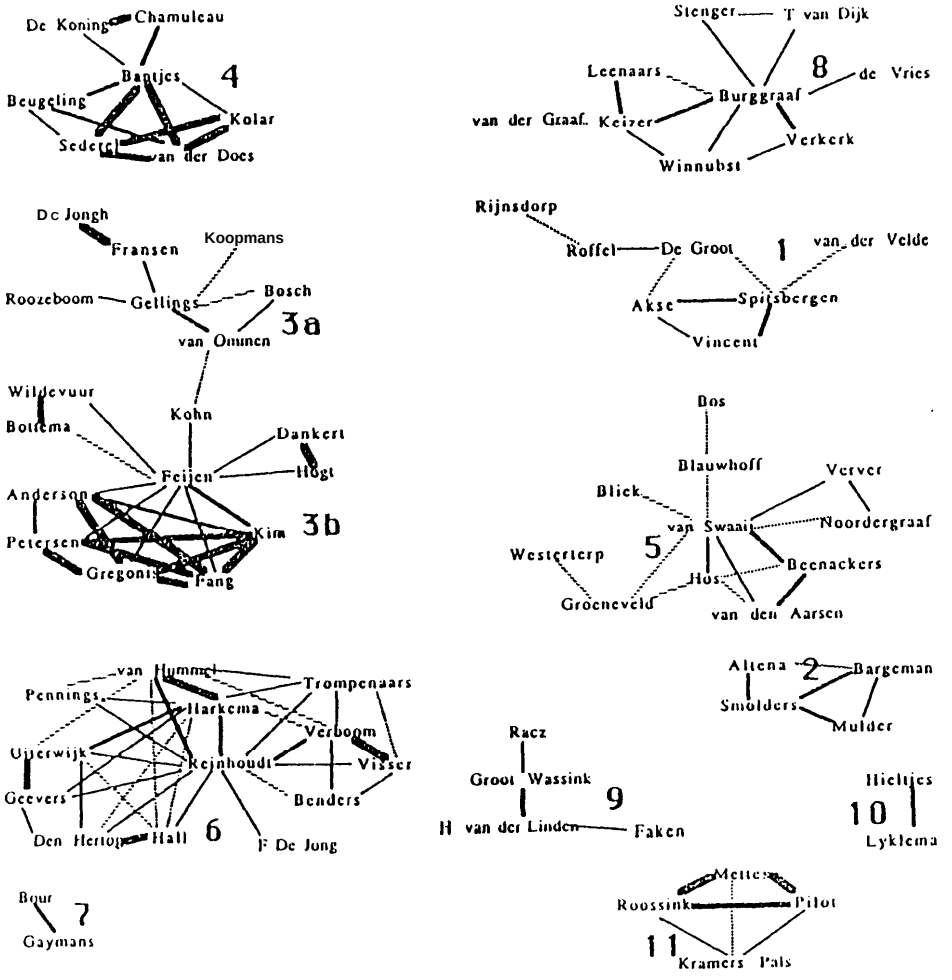


x_{ij} = 1 si l'auteur i a signé ou co-signé pour l'article j

- regroupement des auteurs par la méthode de classification à lien simple sur la mesure d'association du cosinus entre auteurs.
- représentation plane des groupes par construction manuelle, les nuances des traits entre les auteurs symbolisant des intervalles de valeurs des mesures d'associations du cosinus (figure 27).

Ils finissent par conclure que l'analyse des co-auteurs permet:

- d'**identifier 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.
- de **connaître les évolutions de ces groupes, par des comparaisons dans le temps**
- d'**identifier les spécialités phares ou pivots**
- d'**identifier les leaders dans les spécialités**



Co-author clusters based on authors of Twente University of Technology at Enschede between 1980-1983 (number of publications per author > 3; co-author strength > 0.20).

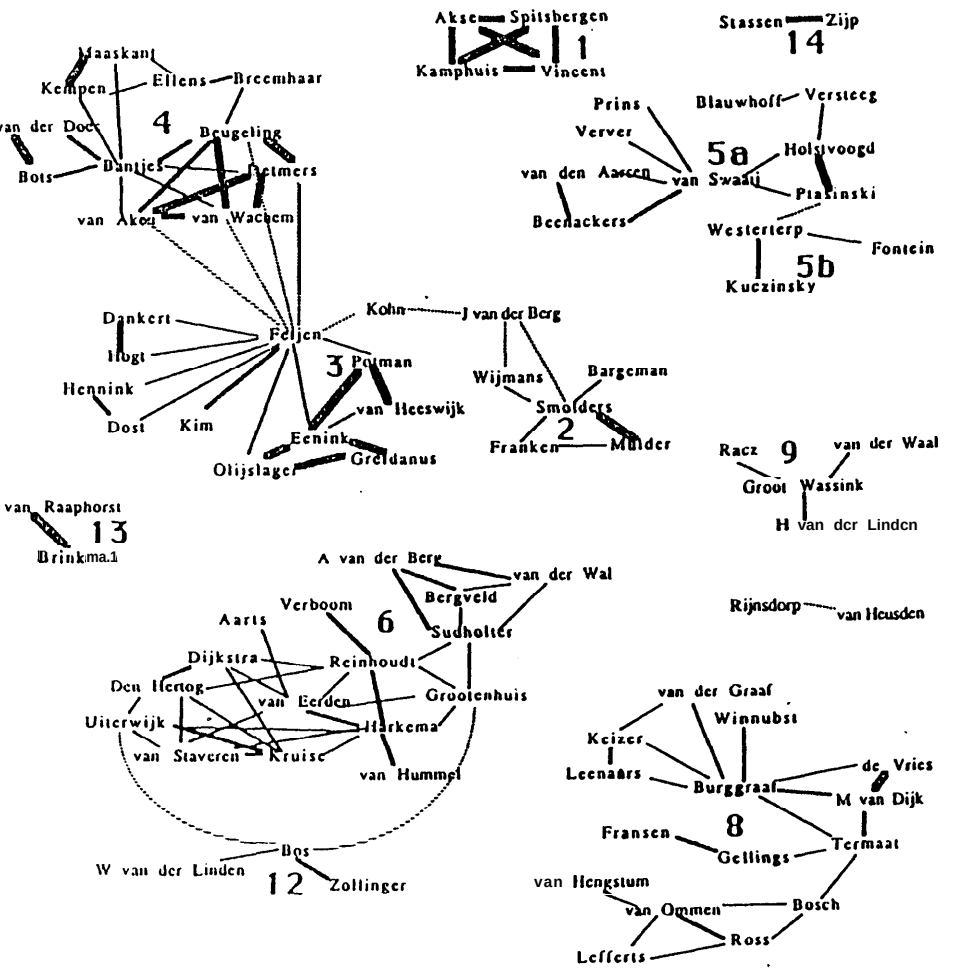
Average co-author strength:

0.20-0.30 ·····

031-045 ——

0.46-0.60 ——

> ——



Co-author clusters based on authors of Twente University of Technology at Enschede between 1384-1387 (number of publications per author > 4; co-author strength shown within clusters > 0.30, between clusters > 0.15).

Average co-author strength:

0.15-0.30 ·····

0.31-0.45 ——

0.46-0.60 ——

> 0.60 ——

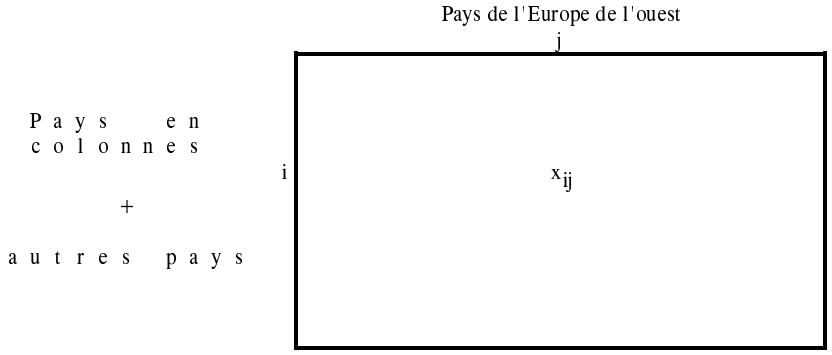
Figure 27: Réseau d'auteurs construit par une analyse des co-publications [PETE91]

(3) L'analyse des co-opérations internationales

Au cours d'une étude des co-opérations internationales des scientifiques des pays de la communauté européenne, Moed et al. ont voulu réduire ces collaborations sur une carte [MOED91], c'est-à-dire représenter graphiquement toutes les publications qui sont l'aboutissement d'un travail entre deux équipes de pays différents mais dont l'un des deux appartient à la communauté européenne.

Il était donc indispensable qu'ils utilisent une source qui ait les données des différentes affiliations des auteurs de publications. Ils ont donc collecté leurs données sur les bases de l'ISI (SCI, SSCI et A&HCI) puisque que ce sont les seuls qui saisissent toutes les affiliations des co-signataires (voir l'exemple de référence table 4 dans *Les bases de données de l'ISI*).

Deux ensembles de pays sont donc croisés pour constituer une matrice de fréquences de co-publications internationales:

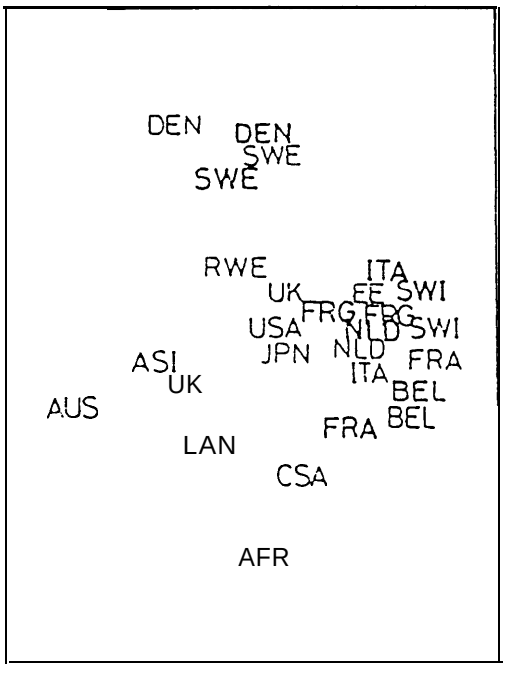


x_{ij} = nombre de publications co-signées par un centre du pays *i* avec un centre du pays *j*

Dans l'étude il y a:

- en colonne, 9 pays de l'Europe de l'ouest: Pays Bas, Belgique, Danemark, RFA, Grande Bretagne, France, Italie, Suisse et Suède
- en ligne, 18 pays comportant les 9 précédents plus les pays suivants: Afrique, Asie, Australie, Canada, Amérique centrale et Amérique du sud, Europe de l'est, Japon, reste de l'Europe de l'ouest, Etats Unis.

La carte est obtenue par injection de cette matrice "brute" dans une **analyse des correspondances** (figure 28).



Correspondence Analysis map of the ISC structure.
Bold label = selected Western European countries;
plain label = all countries/geographical regions

Legend:

BEL - Belgium
DEN - Denmark
FRA - France
FRG - Fed. Rep. Germany
ITA - Italy
NLD - Netherlands
SWE - Sweden
SWI - Switzerland
UK - United Kingdom

AFR - Africa
ASI - Asia, Japan excluded
AUS - Australia and Pacific
CAN - Canada
CSA - Central and South America
EE - Eastern Europe
JPN - Japan
RWE - Rest of Western Europe
USA - United States

Figure 28 : Carte des collaborations des pays européens [MOED91]

On a donc bien évidemment les pays de l'Europe de l'ouest qui sont représentés deux fois sur le graphe: une première fois (libellé appuyé) en tant que pays qui entretiennent des liens avec l'ensemble des pays du monde et une seconde fois (libellé clair) en tant que pays qui collaborent avec les pays de l'Europe de l'ouest. Donc la place du premier est dépendante de ses relations mondiales et celle du second de ses relations inter-européennes.
Les auteurs ont fait remarquer que les positions des pays sont similaires de leurs rapprochements géographiques. Ceci dénote que les chercheurs ont tendances à plus facilement entretenir des collaborations avec les pays qui partagent des frontières communes avec le leur.

(4) Analyses par croisement de deux unités bibliographiques

Nous venons de voir que certaines méthodes d'analyses bibliométriques ne sont pas basées sur la construction d'un tableau carré et symétrique. Toutes les techniques mathématiques n'imposent pas ce genre de matrice comme format des données de base. Les bibliométriciens ont su fouiller parmi la multitude des techniques mathématiques pour y découvrir de nouvelles possibilités. S'étant aperçus que certaines d'entre elles étaient très bien adaptées aux traitements de matrices asymétriques, ils ont vu là de nouvelles ouvertures d'analyses bibliométriques. **Au lieu de combiner uniquement les données du même champ, pourquoi ne pas chercher des corrélations entre les données de champs différents?**

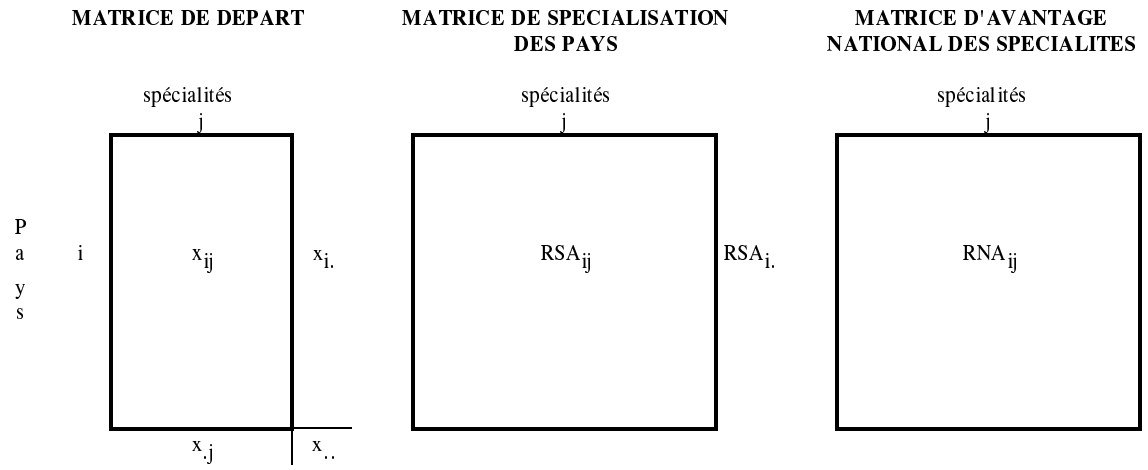
Dès lors que l'on se permet de croiser des données de différentes natures, la première analyse que l'on cherche à résoudre est de nature à répondre à l'interrogation du type "**qui fait quoi?**".

⇨ Analyse Pays - Codes avec pondération par l'indice d'avantage:

Pour le premier exemple, le "**qui**" se situe à une échelle internationale et le "**quoi**" au niveau des grands domaines d'activités de recherche. Cet exemple a été mené par Barré [BARR91] dans le but de caractériser 11 pays selon leurs activités de recherche et aussi de rapprocher ces activités en fonction de leurs répartitions dans ces 11 pays.

Lors d'une analyse des forces et des faiblesses des pays pour un ensemble de spécialités, les données doivent être normalisées. La possibilité offerte dans cet article est de constituer des indices nommés ***Revealed Scientific Advantages RSA*** (Terme et notion que Barré a emprunté à Patel et Pavitt, qui eux l'appliquent en technologie sous la dénomination de "Revealed Technology Advantage").

Cet indice est en fait le même que celui qu'avait déjà présenté Price dans [PRIC81a] (voir *Indicateurs univariés Les pays*). Dans le cas présent, il va servir à normaliser la matrice récapitulant les contributions des pays aux spécialités:



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

et que l'indice **RSA** se calcule ainsi: $RSA_{ij} = \dfrac{x_{ij} / x_{.j}}{x_{i.} / x_{..}}$

Il représente le ratio de la performance du pays *i* dans le domaine *j* sur la performance de ce même pays dans tous les domaines

On peut aussi établir le vecteur qui va caractériser pour ce pays *i* les points forts et points faibles parmi ses spécialités. C'est le vecteur de cet indice pour toutes les spécialités:

$$RSA_i = (RSA_{ij})_{j = 1,n}$$

Si en suite on veut étudier les domaines pour les comparer à travers leur développement pour ces pays, on calcule un nouvel indice que l'on peut nommer:
L'indice de ***Revealed National Advantages***:

$$RNA_j = \left(\frac{RSAij}{RSAi.}\right)_{i = 1, n}$$

Barré a employé ces indices à partir de données de la base PASCAL. Il a étudié 11 pays (France, RFA, UK, USA, Canada, Japan, Pays Bas, Suède, Italie, Inde, Australie) pour 107 spécialités. Ces spécialités ne sont pas celles de PASCAL, elles ont été définies par un groupe d'experts.

❑ Il construit la matrice, la normalise et en déduit les RNA_j des 107 spécialités. Ces vecteurs constituent une matrice qu'il va traiter par une classification ascendante hiérarchique pour regrouper les spécialités.

❑ Les experts, après interprétation des 9 groupes de spécialités fournies par la classification automatique, établissent en réalité 13 domaines majeurs.

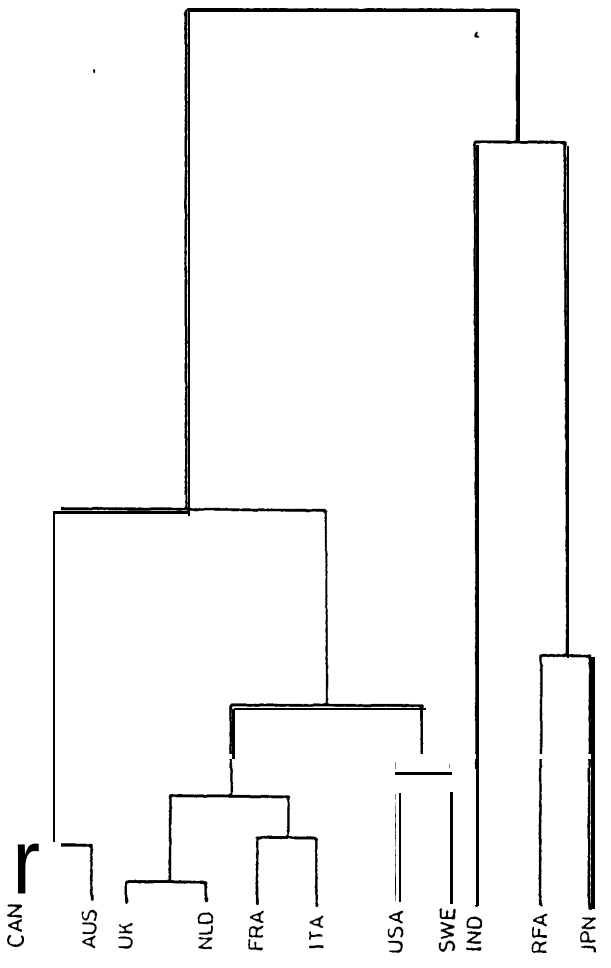
❑ Il y a reconstitution d'une matrice selon les regroupements en 13 domaines et normalisation par les indices RSA et RNA.

❑ La matrice RSA est classifiée pour **regrouper les pays ayant le même comportement d'activité dans ces domaines** (figure 29).

❑ La matrice RNA est elle classifiée pour **regrouper les domaines qui sont étudiés de manière similaire par les 11 pays** (figure 30).

⇨ Analyse Pays - Mot-clés:

Dans le même ordre d'idée, Billard et al. dans [BILL89] voulant répondre à la question "**Qui fait quoi? Et où?**" dans le domaine de l'état du système cardio-vasculaire de l'homme soumis à l'effort, ont étudié les corrélations entre les pays et les mots-clés pour un ensemble de références sélectionnées sur la base Medline. La matrice croisant ces deux unités bibliographiques a été analysée par la technique de l'analyse des correspondances suivie d'une classification automatique.



Hierarchical classification of the 11 countries according to the similarity of their specialization profile over the 13 fields. Two countrics have similar scientific profiles to the extent that thcy are linked togcther at a low level in thc order of hierarchical classification (for example, United Kingdom and Netherlands have the most similar scientific profiles)

Figure 29: Regroupement des pays selon leurs comportements d'activités scientifiques

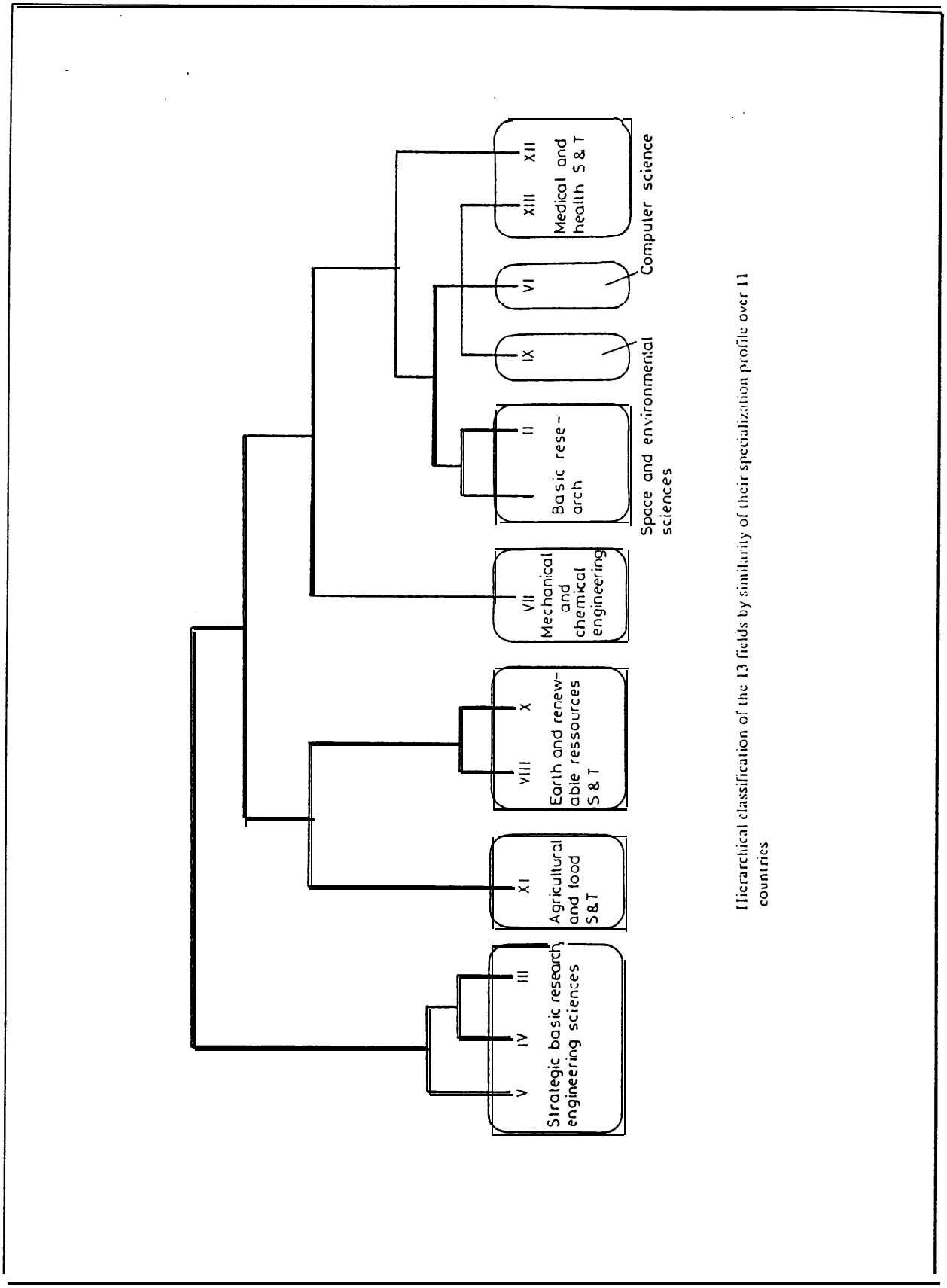


Figure 30: Regroupement des domaines qui sont étudiés de manière similaire dans les pays

E. La mesure des techniques et des technologies

Le besoin de maîtriser la connaissance technique ou technologique par des méthodes bibliométriques n'est apparu que très récemment. Alors que les premières études bibliométriques sur les publications scientifiques peuvent être datées du début de ce siècle, il n'en est pas de même pour les publications relevant des techniques et des technologies.

Initialement la bibliométrie a été imaginée 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. Les premiers à vouloir élaborer des indicateurs bibliométriques ont été les gestionnaires des politiques scientifiques nationales. Les Etats Unis, la Grande Bretagne et la Hollande ont tout particulièrement investi dans cette voie en instituant des centres spécialisés (*ISI* et *CHI* aux US, *SPRU* et *ABRC* en UK, *RAWB* et *groupes universitaires de Leiden* en Hollande). Par conséquent, les renseignements attendus touchent directement la recherche fondamentale et peu l'innovation industrielle.

Ce n'est qu'au début des années 80 que les premières études sur l'information technique et technologique ont surgi. On peut vraisemblablement attribuer cette éclosion aux travaux de Narin. Actuellement, avec les nouvelles exigences de surveillance de leur environnement, les entreprises et les instances gouvernementales industrielles préconisent, elles aussi, l'emploi d'indicateurs de tendances en sciences, techniques et technologies. Bien que ce nouvel axe soit devenu une spécialité pour certains centres de recherche en bibliométrie, le peu de travaux et surtout le peu d'originalité des traitements appliqués à cette catégorie information s'explique par le fait que nous ne sommes qu'à la naissance d'un nouveau champ d'activité de recherche. **Bien souvent il y a eu un simple glissement des techniques employées pour l'information scientifique à celle du monde industriel** Les chercheurs n'ont pas toujours apprécié la réelle adaptation des anciens traitements à la nature des nouvelles données, mais plus grave ils n'ont pas su réévaluer les besoins des industriels.

Cette partie sur l'application de la bibliométrie à des données techniques et technologiques paraît donc bien dépouillée en comparaison de celle de la mesure de la science, mais elle constitue sans doute un prolongement fructueux.

1. Remise en cause des postulats bibliométriques initiaux

Comme fondement de la bibliométrie appliquée à la science, nous avons émis le postulat suivant: la publication scientifique est une représentation objective de l'activité de recherche de son auteur car c'est le passage obligé de sa reconnaissance dans la collectivité scientifique. Un second postulat suppose qu'il existe des liens intellectuels entre les publications qui définissent des structures consensuelles.

On est en droit de se poser la question "**En est-il de même pour l'information technique et technologique?**"

Pour essayer de répondre, nous allons tout d'abord rappeler brièvement qu'elle est la nature de cette information:

⇨ L'information brevet:

La principale source information technique et technologique est le document brevet.

L'acte de dépôt du brevet est soumis à un principe de réciprocité: l'entreprise ou l'organisme qui dépose un brevet 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 contre-facteur).

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 peuvent vouloir 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 (ceci a très bien été décrit par Sommier dans [SOMM92]). **Le dépôt de brevets est la seule pratique garantissant la sauvegarde du patrimoine technologique de l'entreprise**. Comme nous l'avons indiqué dans la première partie de ce mémoire (Chapitre *La veille technologique*), l'innovation technologique est trop importante pour qu'elle ne soit pas considérée comme stratégique pour 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 dépôts n'ont pas la même fermeté concernant la présence de ces deux conditions pour que le brevet soit accordé. Ainsi il est bien connu que l'examen de la demande par un office français est bien moins strict que pour un dépôt dans un office américain. Mais après la publication officielle de la demande de brevet une période est laissée libre à toutes oppositions. Ainsi les concurrents peuvent venir demander certaines révisions ou même le rejet du brevet. Donc, comme il existe le terme de "libre concurrence" en commerce on pourrait donner le terme de "libre droit" pour le brevet. **En laissant jouer cette liberté les caractères innovants sont donc valides dès lors que le brevet fait partie d'une stratégie industrielle et dès lors qu'il n'est pas remis en cause** (faire attention que certaines bases de données brevets introduisent les références brevets au moment des publications des demandes sans indiquer si le brevet a été accordé ou non en fin de procédure !).

❑ Un brevet après son dépôt national (ou régional) peut être étendu (ou désigné) 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 sur 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'informations concernant les stratégies que ces entreprises vont mener pour leurs futurs produits** Là encore le brevet semble fournir des renseignements 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) est, elle, la **source d'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 lors du dépôt font référence aux brevets antérieurs plus ou moins proches de leur demande de façon à préciser en quoi leur brevet en diffère.
- ❍ les déposants font référence aux éventuels articles scientifiques justifiant l'invention et sa priorité.
- ❍ l'examinateur donne, après recherche d'antériorité, ses propres références aux brevets ou toutes publications s'approchant du brevet déposé.

Donc, ces citations portent moins sur des points conceptuels mais plus sur des références techniques. Mais plus important, e**lles ne sont pas régies par des phénomènes sociologiques comme l'auto-citation, les collèges invisibles, l'effet S^t Mathieu… Ces citations sont donc probablement plus légitimes que celles données par les chercheurs.**
De plus on sait qu'un brevet très cité a souvent une importance stratégique particulière, plus importante 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.

Nous pouvons donc 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 décrira pleinement l'activité de recherche et développement du déposant sur l'invention mais apportera en plus des renseignements sur les stratégies industrielles visées.
- ☞ Les documents 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 vecteurs de l'information:

- ❍ production centralisée:
  Toutes les demandes de dépôts passent par des organismes officiels. Ces offices, étant producteur 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 (offices nationaux, office européen, office PCT). Pour obtenir des données chevauchant ces diverses voies de dépôts, des producteurs privés collectent et saisissent aussi de leurs côtés les données des brevets (ex: Derwent).
- ❍ présentation formalisée:
  Une demande de dépôt doit satisfaire à des préconisations, 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.
- ❍ classification internationale:
  Pour caractériser le contenu du brevet, les offices de dépôts intègrent au 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 pour la constitution des rapports de recherche d'antériorité. Le fait que cette indexation soit identique pour tous les brevets est un avantage important par rapport aux bases scientifiques qui ont toutes leurs propres indexations. Au cours d'une étude bibliométrique, cette classification facilite la recherche et élimine les problèmes d'homogénéisation des données provenant de sources multiples (problèmes qui ne sont pas encore complètement automatisables pour les bases scientifiques).

Le brevet a donc une homogénéité que ne peut revendiquer aucun autre type de document. C'est pourquoi ils constituent un fonds documentaire qui se prête particulièrement bien à des opérations bibliométriques.

2. Indicateurs univariés

⇨ Les simples comptages statistiques:

Comme pour les études bibliométriques en science, le simple dénombrement statistique est l'indicateur de base. La plupart des études bibliométriques en technique/technologie est malheureusement restée à ce stade de la technique bibliométrique. Nous ne voulons pas dire par là que ces données n'aient 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 de renseignements.

Une partie de ces renseignements est récapitulée par le tableau suivant:

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

Les études menées par l'Institut Français des Pétroles peuvent être prises comme exemple de ces comptages statistiques [MOUR87a], [MOUR87b].

⇨ Les indicateurs de Battelle:

La société américaine Battelle, consultant en stratégie pour la gestion de la technique/technologie, étaye ses dossiers par ce genre d'indicateurs bibliométriques [ASHT83], [STAC92]. Battelle utilise particulièrement trois indicateurs qu'il nomme: *activity, immediacy, dominance*

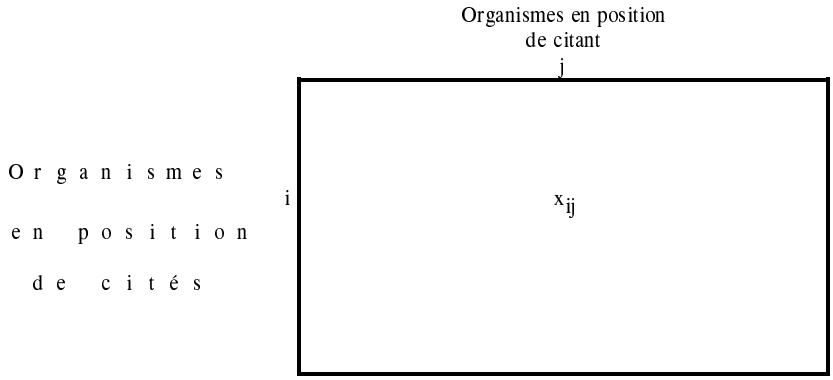
- l'indicateur d'activité (*activity*):
  correspond au comptage simple de brevets par période, par société, par inventeur..
- l'indicateur d'immédiateté (*immediacy*):
  mesure l'age de la technique/technologie en regardant l'écart de temps entre les brevets citants et les brevets cités. Si les brevets citent le plus souvent des brevets récents alors le turnover de la technique/technologie du domaine est très rapide.

Il propose deux types de représentation pour illustrer l'age de la technique/technologie:

- histogramme en 2 dimensions (figure 31)
- histogramme en 3 dimensions (figure 32)

- l'indicateur de dominance (*dominance*)

  mesure la pratique de 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 tant que cités. C'est une matrice analogue à celle des citations-croisées dans les études bibliométriques des journaux.



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 archétypes d'activité brevet. Ces cinq catégories se définissent ainsi:

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

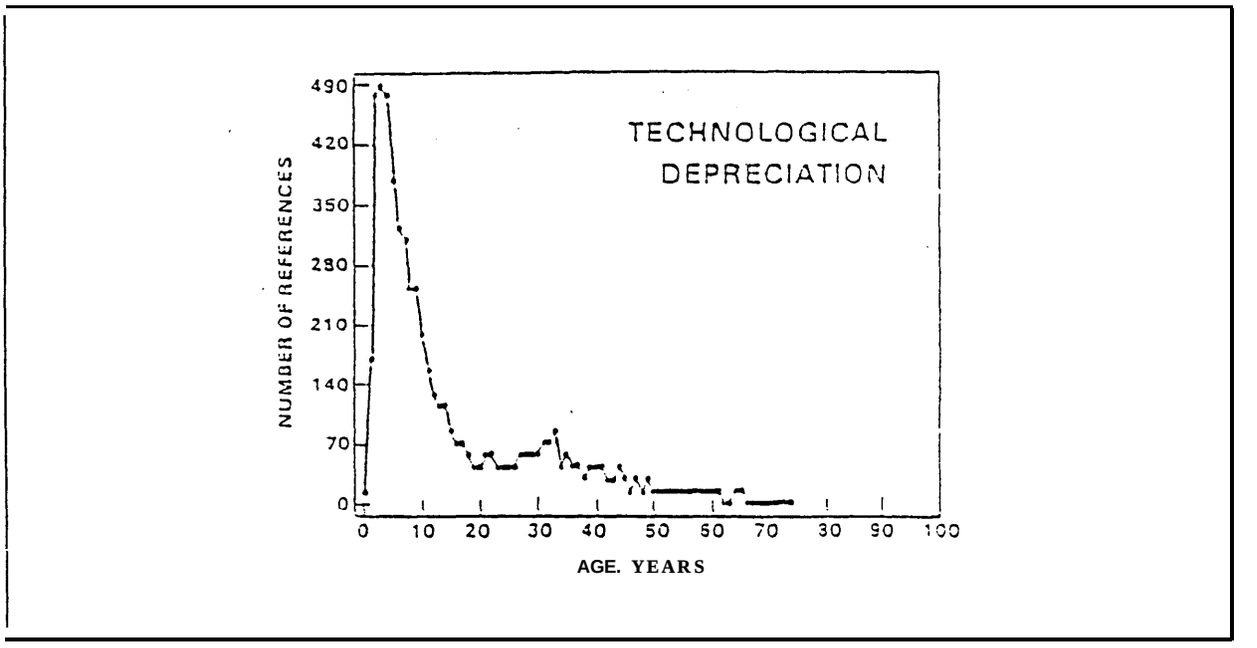


Figure 31: Représentation en 2D de l'âge des technologies

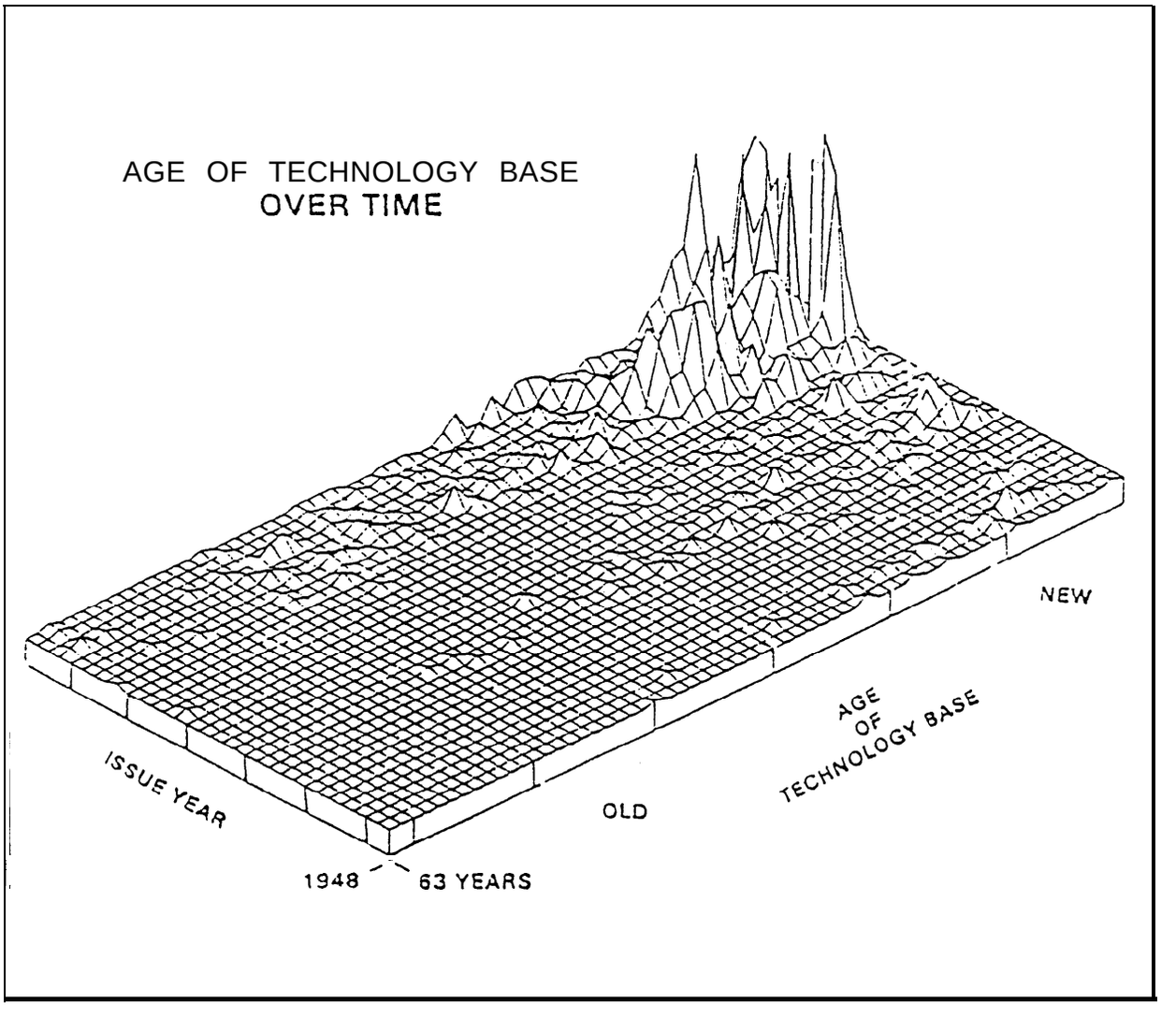


Figure 32: Représentation 3D de l'âge des technologies

⇨ Les indicateurs du CHI:

Narin avec ses collaborateurs du CHI sont probablement les premiers à avoir réalisé des travaux bibliométriques sur les brevets.

Le CHI a développé un panel considérable d'indicateurs bibliométriques, que ce soit pour évaluer la science ou pour évaluer la technologie. La recherche du CHI 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 technologies. Cette spécialisation a été fortement incitée par les instituts américains. Depuis bon nombre d'années, les instituts gouvernementaux américains mettent en place des indicateurs non seulement pour surveiller la science académique mais aussi l'évolution technologique des industries des nations industrialisées. **La création d'une base brevet, dédiée à ces études et structurée dans l'intention d'être exploitée pour des comptages bibliométriques** (Cf *Les banques de données du CHI*), leur a permis d'instaurer un grand nombre d'indicateurs concernant les brevets déposés aux Etats-Unis.

Le portefeuille des indicateurs du CHI que Narin exposa en 1989 au colloque des *Systèmes d'informations élaborées* de la *SFBA* était le suivant:

☞ 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 - indice d'activité
- A5 - avantage concurrentiel
- A6 - évolution technologique
- A7 - distribution géographique
- A8 - indice de dispersion
- A9 - Pourchasse de l'inventeur

☞ Indicateurs d'impacts:
- I1 - fréquence de citation de brevet
- I2 - ratios de performance de citation
- I3 - indice 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 "précurseuses" (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

Tous ces indicateurs sont des ratios établis à partir de comptages bibliographiques. Leur utilisation a été reprise par un grand nombre d'individus: certains pour les appliquer comme indicateur pour la science fondamentale (ce sont essentiellement des laboratoires de recherche), et d'autres dans leur application d'indicateurs technologiques (ce sont bien souvent des personnes du monde industriel ou des personnes travaillant étroitement pour les entreprises). En voici quelques exemples:

- Indicateur d'activité pour les publications scientifiques et pour les brevets

  Le premier exemple est l'application de l'indicateur d'activité à la fois pour les publications scientifiques et les brevets. Cet indicateur sera calculé pour une série de pays et pour chaque domaine soit de la science soit de la technique/technologie. Ces mesures permettent donc de comparer la part d'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 [DIMO90] à partir des indicateurs normalisés mis au point par le CHI.

  La normalisation va pondérer les données selon:

  - la capacité scientifique et technologique générale du pays
  - la capacité scientifique et technologique du pays dans le domaine considéré

  Soit

  N_{sp} = Nombre de documents d'un secteur pour le pays et pour une année
  N_{s} = Nombre de documents d'un secteur pour l'ensemble des pays pour l'année
  N_{p} = Nombre de documents tous secteurs pour le pays pour l'année
  N = Totalité des documents pour une année

  Remarque: Selon le cas, le terme "document" représente soit une publication scientifique soit un brevet.

On peut définir:

Un ***indice d'activité "sectorielle"*** du pays

$$I_{as} = N_{sp} / N_s$$

Un ***Indice d'activité "générale"*** du pays (moyenne pondérée)

$$I_{ag} = N_p / N$$

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

$$E_a = \frac{(I_{as} - I_{ag})}{I_{ag}} = \frac{I_{as}}{I_{ag}} - 1$$

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

Nous retrouvons donc, à une soustraction près, l'*indice d'avantage* qui a déjà été présenté dans les indicateurs univariés de mesure scientifique pour évaluer les pays. Dans ce cas il est aussi appliqué aux brevets.

Pour chaque section, un graphe permet de comparer les pays selon l'évolution de leur indice E_a en fonction du temps (figure 33).

Pour replacer les écarts d'activité de chaque pays selon leurs indices d'activités, une seconde représentation est réalisable (figure 34). Pour chaque section, on dispose la série des années pour chaque pays selon les axes I_{ag} en fonction de I_{as} pour (la série pour un pays est reliée par une ligne continue).

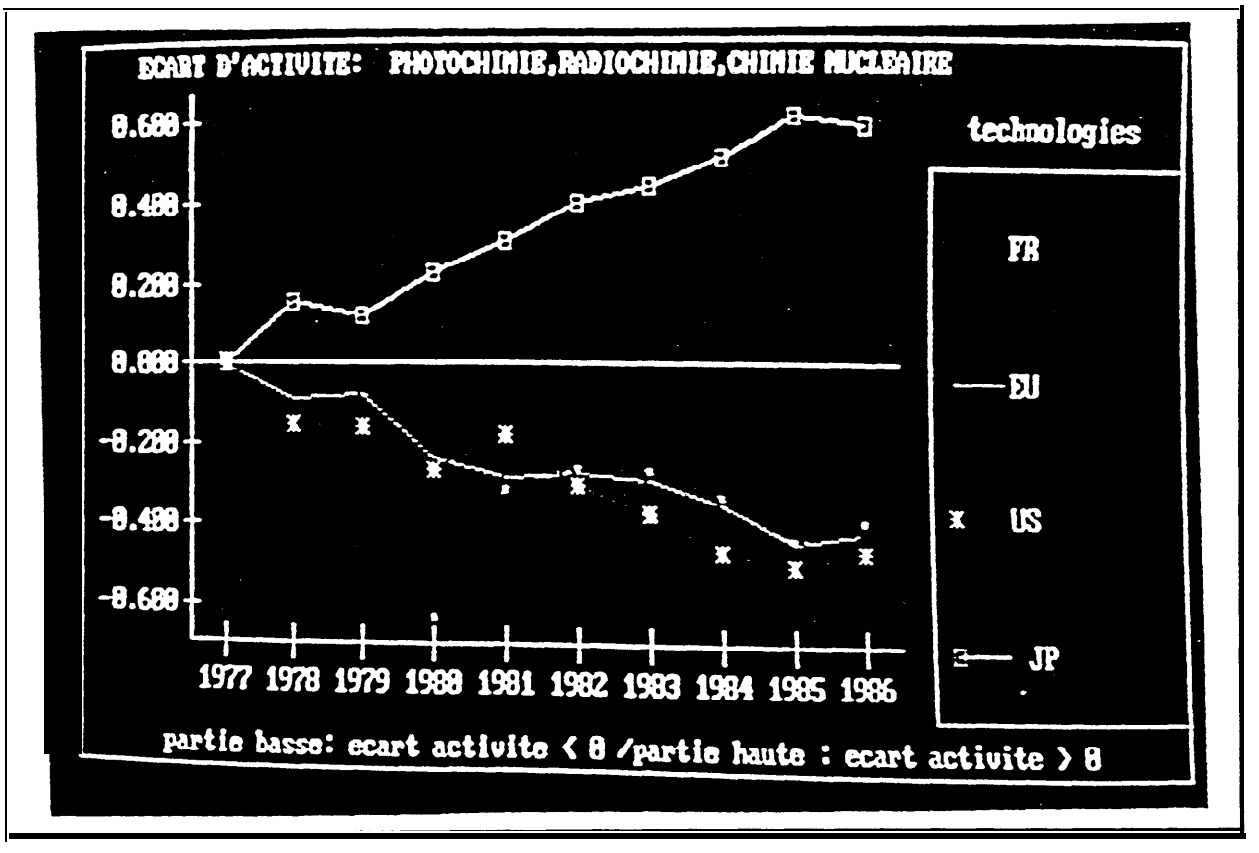


Figure 33: Graphe de l'indice d'avantage calculé pour les publications dans un secteur technologique

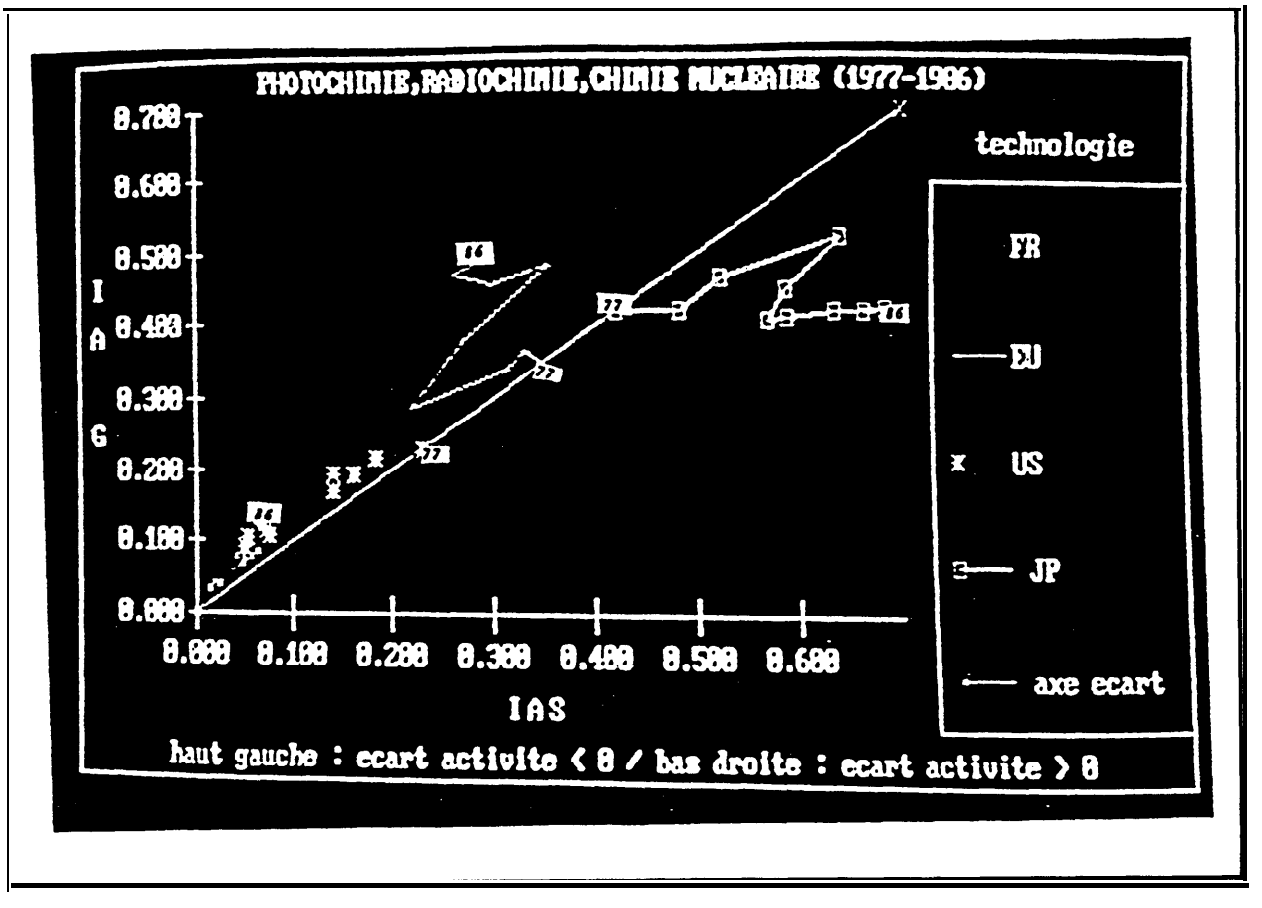


Figure 34: Graphe de la balance activité sectorielle - activité générale pour le même secteur

- La balance scientifique (relations science - technologie):

  La seconde étude donnée en exemple a été faite par l'Observatoire des Sciences et des Techniques [BARR92]. Elle analyse 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 sa 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 principes de la mesure mis en place sont les suivants:

  - la production scientifique d'un pays: c'est la part mondiale de ce pays dans la publication scientifique répertoriée par l'ISI
  - la production technologique d'un pays: c'est le pois mondial de ce pays dans les brevets accordés aux Etats Unis
  - la citation par un brevet d'une publication scientifique est la marque de l'utilisation par ce brevet de la connaissance scientifique.
  - 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 est exposée. Leur application à cinq pays (Grande Bretagne, Allemagne, Japon, Etats unis et France) s'appuient sur des données fournies par le CHI.

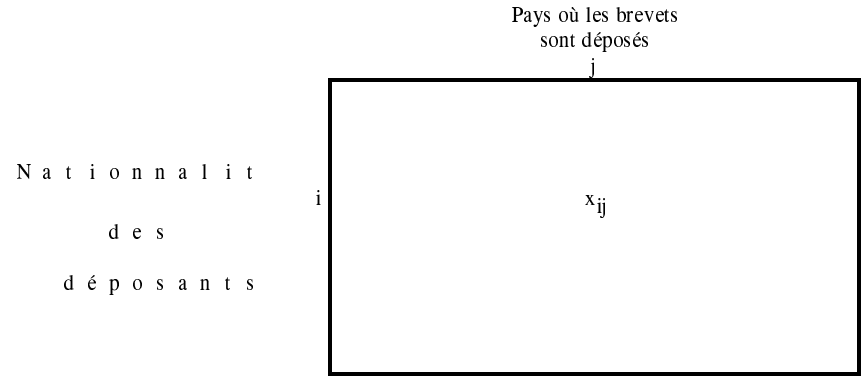
3. Les cartes relationnelles

Les études mettant en oeuvre des techniques statistiques pour présenter les structures sous-jacentes à des tableaux de relations sont pratiquement inexistantes. Quelques équipes françaises travaillent dans cette voie. Nous présentons ici quelques exemples d'analyses possibles à partir de l'information brevet.

⇨ Etude nationalité de dépôt - pays de dépôt:

Doré et al. 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**. Ces fréquences analysées par des traitements statistiques leur ont permis de dégager les corrélations significatives entre ces deux ensembles [DORE87].

Ils ont considéré comme source des 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 CA, soit 71770 brevets. La construction de la matrice de contingence analysée croisait les 11 premiers pays déposants (plus la réunion de tous les autres) avec les 10 principaux pays de dépôts (plus une classe de type divers et deux classes correspondants aux procédures régionales EP et PCT).



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

Les traitements statistiques exécutés sur ce tableau ont fait appel à une **analyse des correspondances** (figure 35), une **classification automatique** (figure 36) et une analyse typologique par construction d'un **arbre de longueur minimale** (figure 37).

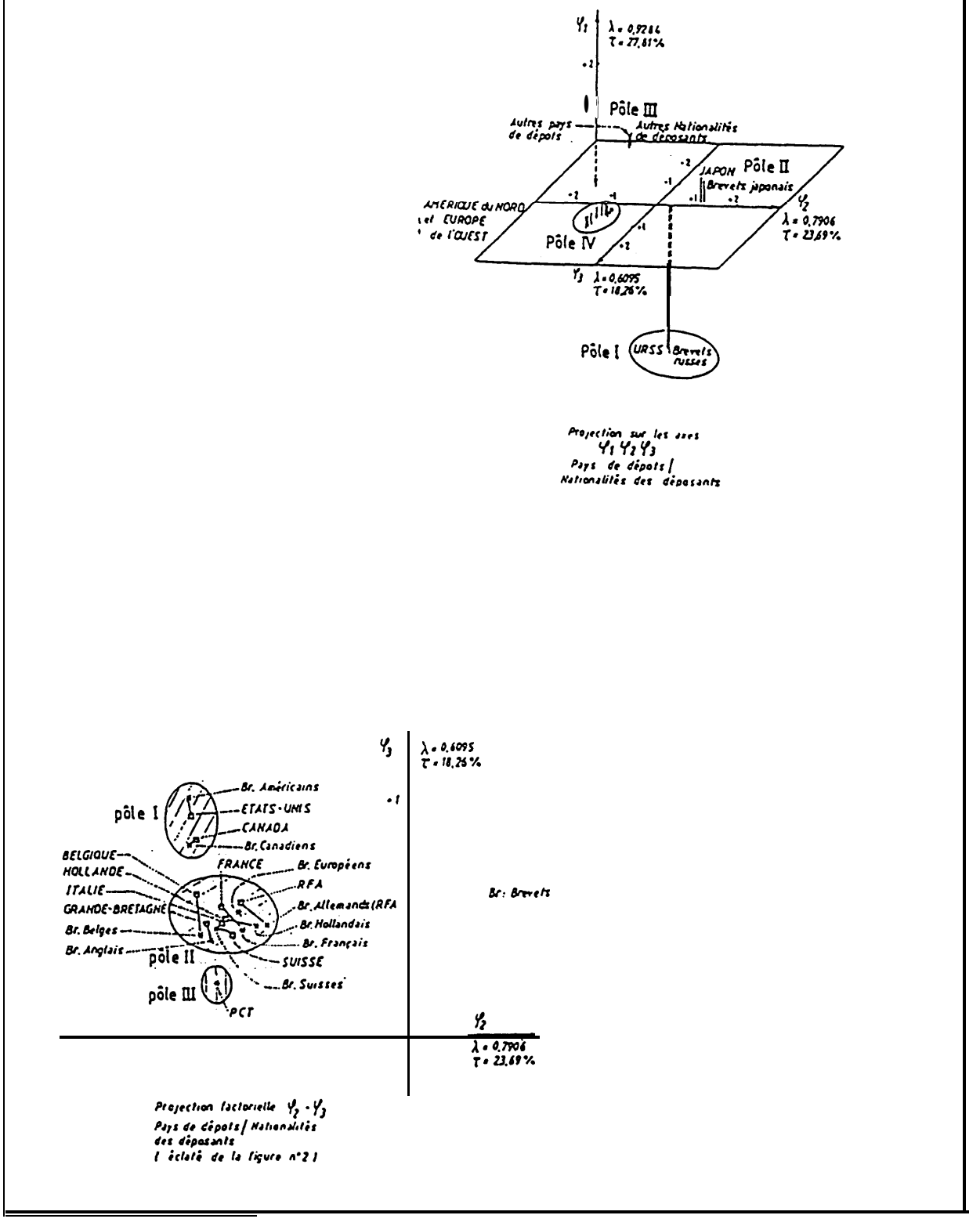
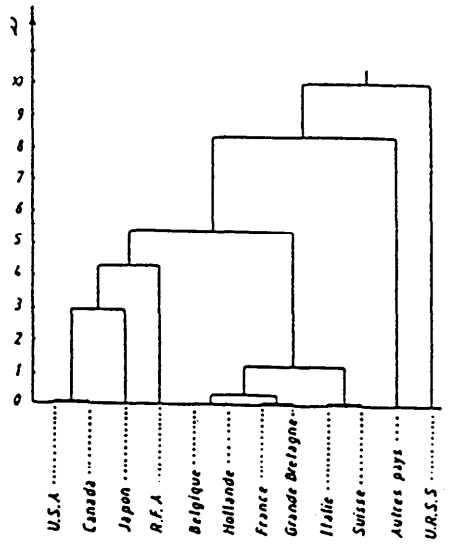
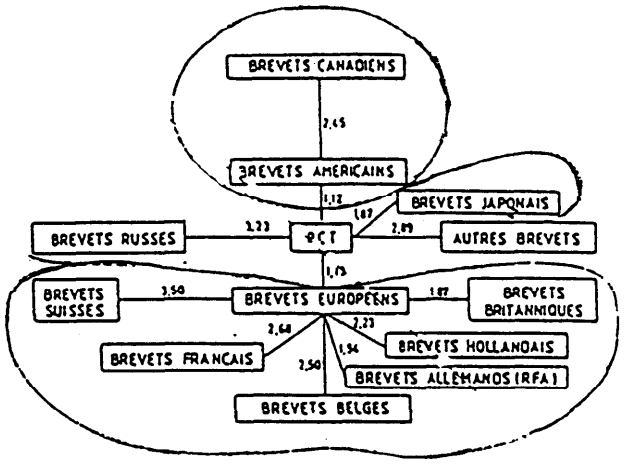


Figure 35: Analyse factorielle des correspondances appliquée au tableau



Classification ascendante hiérarchique des déposants des différents pays.

Figure 36: Regroupement par classification hiérarchique ascendante



Classification non hiérarchique des différents pays de dépot. (Arbre de longueur minimale)

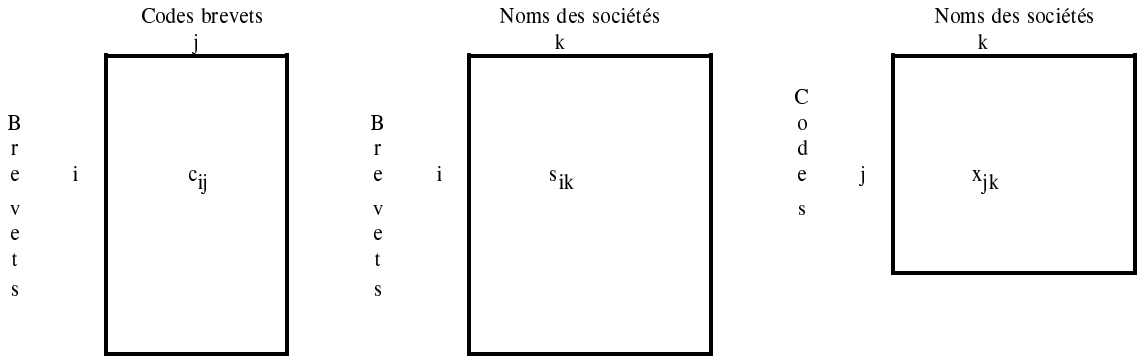
Figure 36: Construction par la méthode de l'arbre mininum

⇨ L'analyse relationnelle:

Récemment, le centre de mathématiques appliquées d'IBM France (CESMAP) a trouvé dans la bibliométrie un bon domaine d'application des méthodes statistiques qu'ils ont conçues. 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**.

L'article de Bédécarrax et Huot [BEDE91], présente la méthodologie mise en oeuvre dans l'application de l'analyse relationnelle pour le traitement des corpus provenant de la base de données brevets *Derwent*. L'intérêt porte particulièrement sur l'exploitation des relations entre les champs noms de sociétés, numéros de brevets et codes CIB décrivant le brevet.

Leurs méthodes statistiques exploitent des matrices de ce type:



c_{ij} = 1 si le code *j* est présent dans le brevet *i* sinon 0

s_{ij} = 1 si la société *k* a déposé le brevet *i* sinon 0

x_{ij} = nombre de brevets déposés par la société *k* et comportant le code *j*

Plusieurs traitements sont possibles soit à partir de ces matrices de départ soit à partir de matrices de similarité relationnelle bâties à partir de ces dernières. Les méthodes de sériation ou de classification que le CESMAP propose vont **réorganiser les données contenues dans ces matrices sans élimination ou déformation des données** (permutation des lignes et 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).

ces matrices étant assurée par l'outil bibliométrique présenté dans ce mémoire). Puis, le CESMAP exploite statistiquement les matrices pour dégager les indicateurs de tendances.

Un article de Huot et al. [HUOT92] présente de façon didactique une étude employant l'analyse relationnelle suivie de sa représentation factorielle des codes (figure 38). Une autre étude employant l'analyse relationnelle est présentée dans ce mémoire pour montrer de quelle manière insérer l'analyse bibliométrique dans le processus de veille technologique (*Exemples d 'études bibliométriques*) .

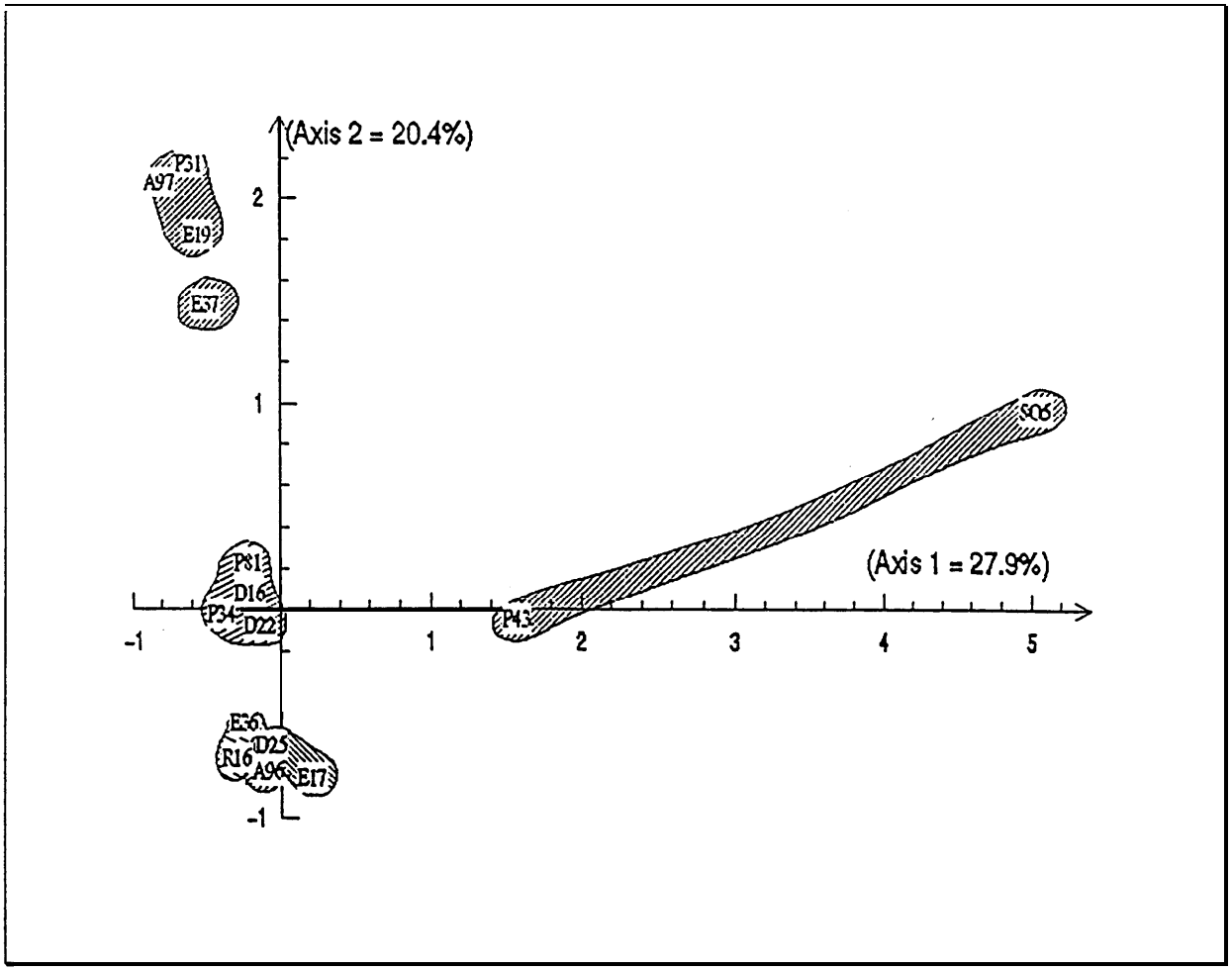


Figure 38: Graphe d'analyse factorielle relationnelle

Le logiciel bibliométrique DATAVIEW: outil d'aide à l'élaboration d'indicateurs de tendances

V. Le logiciel bibliométrique DATAVIEW: outil d'aide à l'élaboration d'indicateurs de tendances

Ce chapitre va présenter le logiciel bibliométrique qui a été conçu dans le cadre de ce doctorat: le logiciel DATAVIEW.

La présentation, dans le chapitre précédent, de toutes ces méthodes bibliométriques avait pour objet de bien faire comprendre pour quelles raisons la réalisation d'un tel logiciel bibliométrique est profitable. Elle va nous permettre de mieux apprécier pourquoi ces méthodes ne pourraient pas convenir dans un processus de veille technologique.

A la suite de ces réflexions, nous exposerons quelles fonctions l'outil bibliométrique doit disposer pour permettre de répondre aux besoins de la veille technologique. Nous verrons ensuite par quels moyens nous envisageons de satisfaire ces fonctions et quelles solutions informatiques ont été développées dans DATAVIEW.

A. Où est l'outil bibliométrique?

Les auteurs des articles bibliométriques ne parlent pratiquement jamais de l'**automatisation des manipulations de données** qu'ils effectuent. Les interventions de logiciels informatiques se limitent à deux étapes dans une étude bibliométrique:

- ⇨ au moment de la collecte:
  les bases de données étant l'une des principales sources des données bibliométriques, l'utilisation de matériels informatiques et logiciels de communication est indispensable pour consulter les serveurs.

- ⇨ au moment de l'examen des données:
  on a vu que pour aider l'interprétation des données bibliométriques les bibliométriciens font souvent appel à des outils informatiques et mathématiques pour les représenter sous une forme graphique: courbes, histogrammes, nuages de points...

Deux remarques sont à retenir sur ces deux étapes:

- ⇨ Les deux interventions informatiques ne sont pas exclusives aux études bibliométriques:
  La communication des données par les télécommunications est devenue courante dans de multiples activités qui ne concernent pas que le domaine de l'information scientifique et technique. D'autre part, les logiciels pour élaborer des représentations graphiques des données n'ont jamais été développés pour le traitement bibliométrique. On ne peut donc pas parler d'outils informatiques bibliométriques.

- ⇨ Les données traitées par ces deux étapes ne sont pas du même type:
  En effet, au moment de la collecte, les données sont textuelles (références bibliographiques) tandis qu'à la seconde étape les données sont numériques. La fonction même de la bibliométrie est l'intermédiaire entre ces deux étapes: **transformer des informations qualitatives en valeurs quantitatives**. Il serait donc légitime de nommer ***logiciel bibliométrique*** l'outil informatique qui réaliserait cette tâche. Il jouerait le rôle d'une **passerelle entre le format texte et le format tabulé**, entre le corpus de références et les traitements graphiques et statistiques.

Les auteurs de la discipline ne font pratiquement jamais référence à cette phase de traitement, tout bonnement parce qu'ils ne l'ont pas automatisée. Ceci est un inconvénient dans un processus de veille technologique. **Un des premiers principes en veille technologique est de fournir des informations élaborées dans un laps de temps le plus réduit possible**. La collecte est devenue immédiate grâce aux serveurs. Le traitement des données numériques est depuis longtemps le domaine d'excellence des informaticiens. Le maillon intermédiaire dans la chaîne des traitements bibliométriques manque dangereusement pour le processus veille technologique.

Jusqu'à présent, cette exigence de temps n'était pas prioritaire dans les études bibliométriques. Mais pour son introduction dans un système de surveillance industrielle, la bibliométrie ne doit souffrir d'aucun "poids mort" dans sa chaîne de traitements.

☞ Comment cette phase intermédiaire est-elle réalisée actuellement?:

❍ traitement manuel:

Cette première solution est le traitement manuel. Rédhibitoire, nous ne l'envisagerons pas.

❍ interrogation par SGBD:

Bien souvent, en l'absence d'autres solutions, les comptages de fréquence d'éléments bibliométriques sont réalisés par le langage d'interrogation Système de Gestion de Base de Données (SGBD): soit directement lors de l'interrogation en ligne grâce au langage de commande du serveur, soit en local grâce à un SGBD commercialisé. Pour ce deuxième cas l'ensemble de références peut être un fonds documentaire interne ou il peut avoir été collecté sur un serveur et mis au format d'import du SGBD.

Le principe du SGBD est de livrer des réponses uniquement après une formulation de questions. Cette étape de formulation est basée sur une idée préconçue du résultat attendu et elle prédétermine donc, dans une certaine mesure, la réponse. Or, lors d'une analyse bibliométrique, on recherche tout simplement l'inverse. Le but est de faire un examen systématique et complet de l'ensemble des éléments pour qu'en les structurants ils livrent d'eux-mêmes les signes qu'ils contiennent. Cette solution est donc impropre.

❍ des outils spécifiques:

Quelques outils ont été développés spécifiquement pour cette phase de transformation du texte au numérique:

❑ Le premier que l'on peut citer est le **Toolbox de Brooks**. Brooks a développé un logiciel pour estimer les caractéristiques des lois bibliométriques (Bradford, Lotka, Zipf). C'est donc un outil d'utilité très restreinte surtout pour la veille technologique.

❑ On peut aussi évoquer des **programmes créés par des auteurs** en bibliométrie pour leurs études. Mais ces programmes restent limités à un seul type de traitement et ne font que vaguement les évoquer (ex: Todorov [TODO90], son programme comptabilise les occurrences et les cooccurrences des codes INPEC).

❑ Il y a encore les sociétés comme l'**ISI** ou le **CHI** qui ont réalisé leurs propres traitements informatiques. Ils s'en servent dans les études qu'ils mènent dans le cadre de prestations de service. Ces traitements ne sont jamais évoqués et restent de toute manière à usage privé.

❑ Il existe par contre trois logiciels commercialisés qui peuvent prétendre au titre de logiciel bibliométrique: **PATSTAT+** (développé par Derwent Publications LTD [DERWEN]), **BATTELLE** (développé par Battelle Development Corporation [BATELL]) et **LEXIMAPPE** (développé par le CSI et le CDST).
Les deux premiers ne permettent d'analyser que des fichiers comportant des références brevets provenant des serveurs. Patstat+ reconnaît seulement les références de la base Derwent et Battelle les bases Derwent et US Patents. En plus de ces limites de formats, les traitements se restreignent à quelques champs et croisements de champs.
Quant à Leximappe, il n'offre qu'une méthode d'analyse et ne permet de traiter qu'un champ à la fois. De plus, cette méthode a été spécialement mise au point pour le champ mots-clés de la base Pascal, sa validité sur tout autre type de champ peut être remise en cause (Cf *Bibliométrie - Indicateurs relationnels - méthodes des cooccurrences de mots*).

❑ Le CRRM développe depuis de longues années des applications informatiques à finalités bibliométriques. La gamme de ces outils est variée.
Les applications ayant des objectifs bien spécifiques:

- **DATACODE**: analyse des codes des classifications CAS, INSPEC, WPI
- **DATAGET**: comparaison de fichiers obtenus par des GETs
- **DATAGE**: analyse des ages des technologies
- **DATASTRA**: analyse des adresses d'individus

En ce qui concerne les applications réalisant des traitements à façons:

- **DATALINK** [DOU87]
- **DATRANS** [QUON88]

B. Caractéristiques des traitements bibliométriques

Pour réaliser cet outil informatique, il faut tout d'abord déterminer ses fonctions. Connaissant les contraintes imposées par l'activité de veille technologique et en se référant aux exemples de méthodes bibliométriques que l'on vient d'exposer, nous déterminerons quels services nous attendons de cette automatisation informatique.

1. Diversité des sources

Dans le passé, la plupart des études bibliométriques constituaient leur bibliographie à partir du coeur de la littérature reproduite par les bases de l'ISI. **Actuellement, les objectifs des études bibliométriques ne sont plus axés sur une évaluation globale de la science mais dérivent vers une connaissance pointue de spécialités**. Ceci se vérifie d'autant plus lorsqu'elles viennent soutenir des activités de veille technologique.

Les besoins en information scientifique sont à la fois variés et exigeants en ce qui concerne l'exhaustivité. La qualité de couverture des thèmes de recherches par spécialité est importante, mais plus encore, la couverture des divers acteurs dans cette spécialité ne doit souffrir d'aucun favoritisme national ou critère d'excellence. L'invention et l'innovation suivent bien souvent des chemins non-conventionnels. Leurs émergences ne se feront pas forcément selon les canons de la science. **Il est donc indispensable ne pas se contenter d'un système de sélection élitiste mais tout au contraire de pratiquer un "ratissage" le plus large et le plus objectif possible**.

Par exemple, Balmer et Martin [BALM91], voulant connaître ce qui se faisait dans une spécialité pointue en biologie, ont dans un premier temps consulté la base SCI de l'ISI. Après avoir constaté de grosses lacunes au moment de la validation de la recherche par les experts, ils se sont finalement réorientés sur la base spécialisée MEDLINE. Mais ils précisèrent qu'il ne faut pas oublier que les bases de l'ISI offrent tout de même deux avantages non négligeables:

- permettent l'étude des collaborations puisque l'affiliation de chaque auteur est présente.
- permettent l'évaluation mondiale d'impacts de travaux et de chercheurs grâce aux citations

Une entreprise peut encore vouloir analyser des bases internes quelles soient de type bibliographique, de type "matière grise" (rapports internes...), de type fichiers clients, de type sondages...

L'outil informatique doit être suffisamment flexible et maniable pour permettre de manipuler des données textuelles sous des formes et des structures diverses.

2. Diversité des éléments bibliométriques

Comme on a pu le voir, les traitements bibliométrique sont variés. **Ils prennent en compte sous des modalités différentes la diversité des informations contenues dans les références bibliographiques**.

Les études bibliométriques n'examinent pas les références bibliographiques d'un seul tenant. Selon les objectifs de l'étude, **un ou plusieurs éléments bibliographiques sont choisis comme étant les seules données textuelles à considérer**. Ces éléments sont bien souvent soutirés de champs bibliographiques bien distincts.

Par exemple, pour les *analyses de co-citations*, un seul type élément est considéré par le traitement. Cet élément, pour les différentes variantes de ces analyses, est présent dans un champ unique: le champ citation. Mais pour chacune d'elles l'élément considéré n'est pas le même:

- pour la co-citation de documents, l'élément étudié est la citation au complet
- pour la co-citation d'auteurs, l'élément étudié est uniquement l'auteur présent dans le champ citation

Par contre, pour une *analyse des citations-croisées de journaux* deux catégories d'éléments sont prises en compte. Le titre de la revue où est publié l'article citant et le titre de la revue de l'article cité. Le premier élément est contenu dans le champ source (OS pour l'ISI) et le second dans le champ citation (CR pour l'ISI).

Les éléments considérés dans les analyses bibliométriques sont donc des sous-structures des formats bibliographiques proposés par les serveurs (découpage des données en champs).

Donc, l'outil informatique devra permettre l'étude de toute partie de texte d'une référence bibliographique.

3. Diversité des traitements bibliométriques

La bibliométrie est une discipline qui fait appel à la mesure. Cette mesure est forcément établie sur des données numériques.

Toutes les évaluations bibliométriques, comme les exemples l'ont montré, sont calculées à partir d'une mesure unique: l'occurrence. Pour un élément bibliographique, on peut présenter ses occurrences selon deux états:

- son état primaire:
  c'est-à-dire la simple localisation des occurrences. C'est le cas de toutes les méthodes bibliométriques qui calculent les mesures à partir d'un inventaire des présences et des absences des éléments dans l'ensemble des références (exemple: analyse des mots-associés).
- son état condensé:
  c'est-à-dire le dénombrement de ses occurrences, la fréquence.

Une troisième mesure est courante en bibliométrie: la cooccurrence. Les cooccurrences pour deux éléments ne sont en fait que la combinaison de l'état primaire des occurrences de chacun d'eux. Des traitements bibliométriques n'utilisent que l'état condensé de cette mesure, la fréquence des cooccurrences.

Sur la base de ces trois mesures, **les études bibliométriques exploitent les données sous différentes apparences**:

- Les listes:
  fréquences d'occurrences, fréquences de cooccurrences
- Les tableaux:
  matrices présences-absences d'occurrences, matrices de fréquences de cooccurrences.

L'outil informatique devra fournir en sortie toutes ces catégories de données numériques sous leurs diverses apparences d'édition.

C. Solutions de la conception informatique adoptée

Maintenant que les fonctions de l'application sont connues dans leurs grandes lignes, nous allons exposer quelles sont les solutions adoptées pour la conception informatique.

1. Solution à la diversité des sources

La diversité des formats proposés par les serveurs impose de passer par une première étape de traitement que l'on a l'habitude d'appeler ***reformatage***.

⇨ Structure reconnaissable des références:

L'ordinateur est une machine douée d'une phénoménale puissance de calcul mais pour cela il faut lui décrire la succession des tâches qu'elle a à réaliser de façon très précise.

On retrouve cette caractéristique pour tous les traitements qu'on a à lui faire exécuter. Le traitement des données textuelles ne faillit pas à cette règle. Pour l'ordinateur un texte n'est qu'une succession de caractères les uns à la suite des autres. **Il faut donc lui donner des repères pour pouvoir traiter différemment les diverses parties de ce texte**. Dans le cas de signalement bibliographique on peut différencier trois parties dans le texte:

- les intitulés des champs
- les contenus des champs associés
- les marques de séparations entre chaque référence

Ces trois repères constituent la structure des références bibliographiques. Comme cette structure varie d'un serveur à l'autre, il est donc nécessaire de passer par une structure établie qui sera reconnue par l'application informatique. Pour réduire au minimum le temps pour réaliser cette étape de transformation de la structure des données textuelles, **nous avons choisi de prendre comme format de référence celui qui est le plus souvent rencontré en standard sur les serveurs professionnels**. Il n'est pas totalement rigide pour admettre le plus grand nombre de variantes de ces formats.

La table 7 présente de façon schématique les divers formats de références textuels acceptés en entrée du logiciel DATAVIEW:

```
☞ Début de ligne
|
|-9-
|XXXXXXyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|      yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|      yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|
|XXXXXXXXyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|        yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|XXXXXXXXXyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|         yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|         yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|
|XXyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|      yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|      yyyyyyyyyyyyyyyyy
|XXXXXX   yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|      yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
|
|-10-
|
|
|
| ....
|
|
|
|-99999-
|XXXXXXyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy ...
|      yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy ...
|
|
|
```

X = symbolise un caractère de l'intitulé de champ
y = symbolise un caractère du contenu du champ
9 = symbolise un caractère numérique

Table 7: Exemples de structures d'entrée reconnues par DATAVIEW

Donc DATAVIEW reconnaît comme:

- repère d'intitulé d'un champ:
  une chaîne de caractère en début de ligne ne dépassant pas 10 caractères (cette chaîne sera à préciser pour indiquer que le traitement concernera ce champ).
- repère du contenu d'un champ:
  soit décalé de six blancs en début de ligne si l'intitulé du champ est inférieur à 6 caractères soit aligné sur l'intitulé dans le cas contraire.
- repère de séparation entre les références:
  une valeur numérique encadrée d'un tiret de chaque côté, l'ensemble en début de ligne.

⇨ Recours à plusieurs sources:

Comme il est très rare que les systèmes documentaires informatisés présentent une exhaustivité supérieure à 75%, l'utilisation d'une seule base de données pour la collecte des références peut introduire des lacunes si le domaine est à cheval sur plusieurs bases. Pour éviter cela, **les bibliométriciens ont pris l'habitude d'avoir recours à plusieurs bases**. **Par conséquent, l'ensemble des références n'est pas homogène car elles sont rédigées sous différents formats.**

L'étape de reformatage des diverses structures de références vers un format unique va permettre de les exploiter simultanément par différents traitements bibliométriques. Deux problèmes majeurs sont à résoudre lors de ce reformatage:

- Eliminer les références doublons:

  La même stratégie d'interrogation appliquée à plusieurs bases de données fournit souvent un ensemble de références communes à plusieurs sources. Lors des comptages d'occurrences, **pour ne pas donner un poids plus fort aux informations présentes dans ces références**, il est indispensable de réduire le superflu d'information à une seule référence. Cette action bien souvent réalisée au cours du reformatage des références est le ***dédoublonnage***. Ce dernier est la comparaison d'une séquence de caractères extraite de chaque référence et choisie de façon à être le plus univoque possible. Par exemple: nom du premier auteur + date de publication + début du titre de la publication...

- Homogénéiser les données des différentes sources:

  Malgré l'existence de normes internationales (ISO, AFNOR, ISSN, ISBN, CODEN, format UNISIST) certains fonds documentaires, dont l'origine est antérieure aux standardisations, ont préféré conserver leurs conventions (langue, pays, coden), leurs abréviations (source, affiliation des auteurs), et leur syntaxe (dates, noms d'auteur).

  Il faut donc homogénéiser le contenu de champs pour **obtenir un "langage" commun à toutes les références**.

  Ainsi, mise à part la structure extérieure des références (répartition des champs, intitulés des champs...), le reformatage va aussi **modifier le contenu des champs**. Cette normalisation des champs a pour objectif de construire un format bibliométrique et non un format documentaire, c'est-à-dire:

  - éliminer les champs inutiles pour l'étude bibliométrique
  - redistribuer les informations entre les champs ou à l'intérieur des champs en fonction du format commun établi. Selon les bases, les données ne sont pas toujours

distribuées dans les mêmes champs ou ne sont pas présentes selon la même séquence à l'intérieur des champs.

- ajouter des champs pour individualiser les informations qui constituent des éléments habituels d'analyses bibliométriques, quitte à introduire des redondances entre les champs. Par exemple, les dates sont des données souvent examinées lors des études bibliométriques. Les formats documentaires livrés par les serveurs n'individualisent pas celles-ci dans des champs spécifiques.
- standardiser les champs rédigés différemment (pays, langues, auteurs, dates...)
- préciser clairement la présence ou l'absence d'information dans un champ.
  La fréquence d'absence de renseignement dans un champ peut être une source d'information tout aussi importante que les données qui y sont introduites. Les formats serveurs ne présentent pas les champs non renseignés. Ainsi, il n'est pas rare de collecter des références où le champ renseignant sur l'affiliation des auteurs ne soit pas présent. Lors du reformatage ces champs absents dans certaines références devront être insérés avec une mention indiquant l'absence d'information à l'intérieur, par exemple: CHAMPxVIDE.
- harmoniser les champs descripteurs (indexés):
  Pour les champs descripteurs contenant une **indexation enrichie** (Cf *L'enrichissement de l'indexation*) ou une indexation hiérarchique, il faut se poser quelques question:

  Veut-on donner la même signification à tous les descripteurs (génériques, spécifiques, indicateurs de rôles, pondérés, non pondérés)?

  ☞ Si la réponse est affirmative, le reformatage doit niveler toutes les différences entre ces descripteurs.

  ☞ Sinon, faut-il tous les conserver mais en les marquants pour les identifier dans les traitements ultérieurs ou faut-il en éliminer?
  Là, le reformatage devient plus complexe car tous les cas particuliers sont à envisager.

Tout ceci est encore automatisable par des procédures informatiques. Le seul inconvénient est l'**homogénéisation des descripteurs provenant de plusieurs indexations**. Chaque producteur de base de données a son propre système d'indexation. Pour décrire un même concept, ces systèmes d'indexation n'utiliseront pas forcément les mêmes descripteurs. Une simple réunion de ces ensembles de descripteurs crée une diversité de termes pour symboliser un même concept. **Le poids de ce concept est réparti sur plusieurs éléments et donc son importance**

statistique minimisée. La solution est la ré-indexation de l'ensemble des descripteurs sur un thesaurus unique? Actuellement, ceci n'est réalisable que manuellement, mais les recherches en cours pour adapter les méthodes d'analyses sémantiques à l'information scientifique et technique laissent présager une possibilité d'automatisation.

⇨ Exemple de mise au format bibliométrique:

Nous présentons ici un exemple simple de reformatage qui transforme une référence bibliographique de la base WSCA (*World Surface Coatings Abstracts*) soutirée du serveur Orbit en une référence au format bibliométrique (Table 8 et 9).

Le reformatage ne conserve que les éléments utiles aux traitements bibliométriques à mettre en oeuvre pour l'étude. Il va donc:

- éliminer les champs dont le contenu est superflu: les numéros d'introduction dans la base (ABN et AN)
- découper le champ SO renseignant sur la source en deux champs distincts: le nom du journal (SOJ) et la date de la publication (SOD). Les autres éléments sont éliminés car il n'apporte pas d'information intéressante mais par contre ils peuvent perturber les comptages
- normaliser les champs langue (LA), type de document (DT) et code documentaire (CC) en éliminant les données inutiles et pouvant être source d'erreurs de saisie
- indiquer l'absence des champs non renseignés dans cette référence: les champs spécifiques aux documents de type brevet (PA, PT, PN), et les champs nom chimique (CN), nom commercial (TN), localisation géographique (LO), nom de la compagnie (CO).
- créer un champ mots-clés qui réunis les descripteurs du champ IT et ST après nivellement des niveaux de signification du champ IT, élimination des descripteurs redondants, normalisation des séparations entre les descripteurs.

Cette mise au format bibliométrique a été réalisée en automatique par des applications informatiques. **Les champs obtenus sont "propres" et sont directement exploitables par des analyses bibliométriques. Chaque champ ne comporte que des éléments bibliométriques pertinents**.

```
-3-
ABN - 90-09535
AN  - 309535
AU  - LE TOULLEC M
IS  - 9012
TI  - Painting and colouring of motor vehicle plastics.
SO  - Plast. Mod. Elast. 1990, Vol 41 No 7, 134-7
LA  - French (FR; XE)
DT  - J  (Journal, etc.)
CC  - 49  Pretreatment and Application
IT  - Plastics: vehicles (motor) &, pretreatment/painting/bulk colouration;
      Vehicles, Motor: plastics &; Painting: plastics, vehicles (motor);
      Pretreatment: plastics
AB  - Painting and colouring are surveyed, with descriptions of methods of
      improving adhesion, e.g. in-mould coating, in-mould release (of
      polyurethane reaction injection moulded parts), solvent vapour cleaning,
      and flame treatment (of polyolefins). Details are provided of in-line
      painting, this being especially suitable for sheet moulding compounds
      (SMC). Bulk colouring is the preferred method for internal components,
      e.g. dashboards. Developments in drying tend to favour the use of UV and
      microwaves, and in some cases vapour injection curing is used. (In
      French)
ST  - reaction injection moulding; RIM; plastics substrate; vapour curing;
      ultraviolet curing; radiation curing
```

Table 8: Référence WSCA au format documentaire

```
-3-
AU  - LE TOULLEC M
TI  - Painting and colouring of motor vehicle plastics.
IS  - 9012
PA  - CHAMPxVIDE
PT  - CHAMPxVIDE
PN  - CHAMPxVIDE
SOJ - Plast. Mod. Elast.
SOD - 1990
LA  - French
DT  - J
CC  - 49
CN  - CHAMPxVIDE
TN  - CHAMPxVIDE
LO  - CHAMPxVIDE
CO  - CHAMPxVIDE
ST  - RIM;bulk colouration;motor;painting;plastic;plastics substrate;
      pretreatment;radiation curing;reaction injection moulding;
      ultraviolet curing;vapour curing;vehicle;
AB  - Painting and colouring are surveyed, with descriptions of methods of
      improving adhesion, e.g. in-mould coating, in-mould release (of
      polyurethane reaction injection moulded parts), solvent vapour cleaning,
      and flame treatment (of polyolefins). Details are provided of in-line
      painting, this being especially suitable for sheet moulding compounds
      (SMC). Bulk colouring is the preferred method for internal components,
      e.g. dashboards. Developments in drying tend to favour the use of UV and
      microwaves, and in some cases vapour injection curing is used. (In
      French)
```

Table 9: Référence WSCA au format bibliométrique

⇨ Solution informatique:

Cette étape de reformatage n'étant pas une préoccupation spécifique à la bibliométrie, des logiciels commerciaux réalisent déjà ce genre de traitement de données textuelles. **La finalité de l'outil bibliométrique conçu au cours de cette thèse n'étant pas de redévelopper des applications déjà existantes sur le marché, nous avons choisi de nous décharger de cette étape sur un produit commercialisé**. Au même titre, nous n'allions certainement pas redévelopper des programmes d'analyses statistiques existant depuis une dizaine d'années.

Le logiciel de reformatage sélectionné pour ce pré-traitement bibliométrique est INFOTRANS, produit développé par I+K [INFOTR]. **Ce logiciel sera directement relié à DATAVIEW pour permettre à l'utilisateur de faire appel à ses fonctions de reformatage à tout moment pour tous les pré-traitements de mise au format bibliométrique**.

2. Solution à la diversité des éléments bibliométriques

Après l'étape de reformatage, les données sont distribuées dans les champs par catégorie d'information bibliométrique. **Il va falloir indiquer comment découper ces champs pour obtenir les éléments bibliométriques à considérer pour les comptages d'occurrences**.

⇨ Notion de forme graphique:

En fonction du serveur, de la base et du champ étudié les caractères qui séparent ces éléments bibliographiques sont différents. Reprenons l'exemple précédant (format d'une référence WSCA, table 8).

❑ Si on considère que les éléments bibliométriques à traiter sont les mots du titre alors le séparateur est l'espace et le point. Si le point n'est pas défini comme étant un séparateur le dernier mot du titre sera *"plastics."* et non *"plastics"*.

❑ Si on considère que le champ à traiter est le champ ST, dans ce cas deux traitements sont possibles

- ❍ les éléments bibliométriques sont les mots: uni-termes
- ❍ les éléments bibliométriques sont les descripteurs composés: multi-termes

Dans le premier cas, les séparateurs sont le point-virgule et l'espace. Tandis que dans le second cas, le séparateur à définir est simplement le point-virgule. **On voit ici que l'élément bibliométrique n'est pas obligatoirement un uni-terme**.

❑ Un autre exemple est lors du traitement des noms d'auteurs; il vaut mieux considérer comme élément bibliométrique l'ensemble nom-prémons plutôt que le simple nom pour réduire les ambiguïtés d'homonymie.

Par conséquent, nous avons décidé de nommer ces éléments bibliographiques sous l'appellation de *forme graphique*. Nous avons repris l'appellation que Lebart et Salem donnent aux unités statistiques de comptages dans leurs traitements statistiques de réponses aux questions libres d'enquête [LEBA88]. Par commodité, l'appellation s'est réduite à ***forme***. Donc, on nomme *forme* l'unité bibliométrique considérée lors des traitements dans DATAVIEW. **La forme est une suite de caractères, encadrée de part et d'autre par un caractère séparateur-de-forme, et qui symbolise une entité bibliométrique**.

⇨ Traitement des formes par DATAVIEW:

DATAVIEW permet à l'utilisateur, en début de session de traitement, de proposer les caractères qui seront reconnus comme séparateurs de forme. L'utilisateur peut désigner plusieurs séparateurs (10 au maximum) qui seront considérés en même temps lors des traitements ultérieurs.

Cette liberté de liste de séparateurs permet d'être moins exigeant sur l'étape de reformatage.

Ainsi, le champ IT de l'exemple précédent (table 8) peut directement être traité par DATAVIEW sans phase de reformatage. Il suffit de préciser que les séparateurs de formes pour le champ IT sont le deux-points ":", le point-virgule ";", la parenthèse-ouverte "(", la parenthèse-fermée ")", le et-commercial "&" et la barre-oblique "/".

Les formes pour le champ de cette référence seront alors:

```
bulk colouration
motor
Motor
painting
Painting
Plastics
plastics
pretraitment
vehicules
Vehicules
```

Nous voyons qu'un caractère en majuscule n'est pas reconnu comme identique au même caractère en minuscule. Donc pour que le concept *"VEHICULES"* regroupe les deux formes "vehicules" et "Vehicules" il faut lors du reformatage transformer tous les caractères soit en majuscule soit en minuscule.

3. Solution à la diversité des traitements ultérieurs

L'unité de mesure dans les études bibliométriques est l'occurrence des formes.

En fait, pour la plupart de ces formes cette unité équivaut au recensement du nombre de références où elles sont présentes. Ceci est vrai pour les formes qui font partie d'un des ***champs contrôlés***. Cette appellation signifie que le contenu du champ a été retraité par une personne (indexeur). Ces champs ne se limitent pas uniquement à ceux dont l'objectif est d'extraire des descripteurs du texte de l'article (mots-clés, codes). On peut aussi considérer que les champs auteur, source, affiliation... sont des champs de type contrôlé puisque les données subissent une analyse et une normalisation avant d'y être introduites.

Tous ces champs traités par une intervention humaine sont généralement épurés de toute information redondante, si bien que le nombre d'occurrences des formes qu'ils contiennent est réduit à l'unité dans chaque référence.

Par contre, les autres champs dont le contenu est rempli par un ***langage libre***, ont souvent des redondances de formes (titres, résumés). Le nombre d'occurrences dans une référence varie pour chaque forme.

Pour DATAVIEW l'unité bibliométrique choisie est la référence. Ne connaissant pas le biais que pourrait introduire la prise en compte, trop systématique, de toutes les occurrences d'une forme, nous avons préféré uniquement comptabiliser la présence ou l'absence de la forme dans la référence, c'est-à-dire la présence ou l'absence du concept qu'elle symbolise.

⇨ La notion de fréquence d'une forme et la notion de l'occurrence d'une forme:

DATAVIEW est conçu pour fournir ces deux types de dénombrements pour chaque forme: le nombre des références comportant la forme et le nombre d'occurrences de cette même forme.

Pour différencier ces deux comptages nous leur avons donné deux noms:

- ***fréquence***: abréviation de ***fréquence de la forme***, c'est-à-dire selon notre unité le nombre de références
- ***occurrence*****:** abréviation de ***fréquence des occurrences de la forme***, c'est-à-dire le nombre de fois que la forme apparaît dans l'ensemble des références

Pour la simple référence de l'exemple précédent on obtiendrait:

FORME	FREQUENCE	OCCURRENCE
bulk colouration	1	1
motor	1	2
Motor	1	1
painting	1	1
Painting	1	1
Plastics	1	1
plastics	1	2
pretreatment	1	1
vehicules	1	2
Vehicules	1	1

⇨ La notion de paire:

Pour les formes, les deux comptages de fréquence et d'occurrence sont maintenus en parallèle. Par contre, **la notion de cooccurrence n'est plus retenue lors du comptage des paires de formes**.

Si l'occurrence était l'unité des comptages la cooccurrence pour l'exemple précédant la fréquence des cooccurrences entre la forme "Plastics" et la forme "vehicules" serait égale à 2. Pour les formes "vehicules" et "motor" elle s'élèverait à 4 (c'est la combinaison entre les 4 occurrences de ces 2 formes).

Maintenant, imaginons deux formes, dont les deux fréquences des occurrences soient de 10 dans une même référence (ce pourrait être le cas dans un résumé), la fréquence de leurs cooccurrences serait de 100. N'est-ce pas donner une importance surestimée à la relation entre ces deux formes? Imaginons deux autres formes qui apparaissent simultanément dans 100

références et n'ayant qu'une occurrence dans chaque référence; la fréquence de leurs cooccurrences serait à elles aussi de 100. Mais leur relation n'est-elle pas plus importante? **N'est-il pas plus remarquable d'avoir 100 références ayant toujours deux formes en commun plutôt qu'une seule comportant toutes ces occurrences?**

DATAVIEW ne considère donc pas les cooccurrences mais, comme pour la fréquence de forme, le nombre de références comportant la co-apparition de deux formes (deux concepts). Pour différencier cette notion de celle de la cooccurrence nous l'avons nommée la ***paire***. Le terme de *paire* est l'abréviation de ***paire de formes***. **La *fréquence d'une paire* comptabilise donc la présence ou l'absence de la paire des deux formes dans les références**.

Pour un champ comportant n formes, le nombre de paires générées suit donc une règle de dénombrement de type combinaison:

$$C_n^p = \frac{n!}{p!\ (n-p)!}$$

comme p = 2 on a $$C_p^2 = \frac{n!}{2!\ (n-2)!} = \frac{n\ (n-1)}{2}$$

⇨ La notion d'indice d'association:

Tous les comptages effectués par DATAVIEW se basent sur ces deux notions de fréquence: la fréquence de forme et la fréquence de paire. Une nouvelle catégorie de mesures pour estimer la force de relation entre deux formes est calculée à partir de ces deux types de fréquence. Ces mesures sont nommées des ***indices d'association***.

La plus simple mesure de relation entre deux formes est le comptage de la fréquence de la paire. Mais elle est en fait biaisée puisqu'elle ne prend pas en compte l'importance des fréquences relatives aux deux formes.

Prenons l'exemple suivant où nous allons comparer la fréquence d'association de deux paires de formes:

		Fréquence
Forme	A	100
	B	50
	X	10
	Y	8
Paire	A=B	15
	X=Y	8

A la simple vue des fréquences de paires, la paire A=B paraît avoir un lien deux fois plus important que la paire X=Y. **Mais est-ce une bonne estimation des associations?**
Si on relativise la fréquence des deux paires aux fréquences des formes constitutives, il est clair que la paire A=B ne constitue qu'une part minime des relations que les formes A et B doivent avoir avec l'ensemble des formes. Tandis que la paire X=Y représente une association totale des deux formes X et Y. Chaque fois que la forme Y est utilisée la forme X l'est à ses cotés. La simple fréquence d'une paire est donc une donnée qui ne permet pas de mesurer l'intensité d'association entre les formes. Il est alors difficile de reclasser les liaisons des paires selon leur importance par cette seule mesure.

Pour relativiser le poids des formes pour chaque association il est possible d'obtenir par DATAVIEW des indices statistiques de similitude ou de dissimilitude: *indices d'associations*.

Ces indices se calculent à partir d'un tableau qui résume les données des associations entre deux formes:

		Forme X	
		Présence	Absence
Forme Y	Présence	N_A	N_B
	Absence	N_C	N_D

où, si F_X = Fréquence de la forme X dans le corpus
F_Y = Fréquence de la forme Y dans le corpus
F_{XY} = Fréquence de la paire X=Y dans le corpus
M = Nombre de références du corpus

on a
N_A = Nombre de co-présences de X et Y = F_{XY}
N_B = Nombre de présences de Y en l'absence de X = $F_Y - F_{XY}$
N_C = Nombre de présences de X en l'absence de Y = $F_X - F_{XY}$
N_D = Nombre de co-absences de X et Y = $M - F_X - F_Y + F_{XY}$

plus la relation $N_A + N_B + N_C + N_D = M$

Les indices combinent ces quatre valeurs de façon à donner plus ou moins de poids à chacune d'elles dans la mesure de l'association entre X et Y. L'utilisateur dispose d'une liste d'indices statistiques binaires dans laquelle il peut choisir celui qui lui paraît le mieux retracer la mesure d'association entre les formes (voir liste en Annexe 1).

⇨ Edition des résultats:

Comme pour le reformatage, **les analyses statistiques ou représentation graphique des données bibliométriques ne sont pas effectuées sous DATAVIEW**. Ces outils existant déjà sur le marché avec souvent des traitements difficilement égalables (ex: Excel pour les tableurs et SAS pour les traitements statistiques), il paraît inutile d'y perdre son énergie.

Par contre, DATAVIEW est là pour dégager les caractéristiques bibliométriques d'un corpus de références et les éditer sous une disposition qui permettent des traitements ultérieurs. Ces dispositions d'édition sont de deux types: les listes de fréquences et les tableaux.

Les différentes listes de fréquences sont:

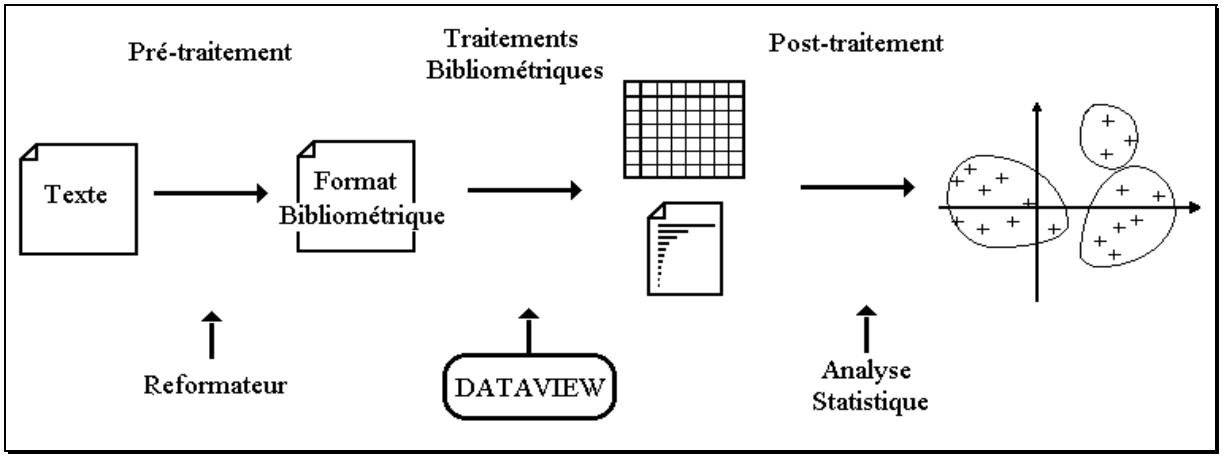
- liste des fréquences de formes
- liste des fréquences de paires
- distribution des fréquences par rang de formes
- distribution des fréquences par rang de paires
- répartition du nombre de formes par champ

Les différents Tableaux sont:

- matrice de présence-absence
- matrice symétrique
- matrice de contingence
- tableau de bord généralisé (tableaux de Burt)

Ces éditions sont bien évidemment fournies dans un format qui puisse être directement importé dans les applications informatiques exploitées pour les post-traitements statistiques (logiciels statistiques et tableurs).

Pour conclure cette partie, on peut préciser de nouveau la position centrale que joue DATAVIEW dans le traitement bibliométrique:



D. Description de la chaîne de traitement de DATAVIEW

Nous allons présenter dans cette partie la diversité des traitements réalisables avec l'outil informatique DATAVIEW. La succession des manipulations que l'utilisateur doit effectuer pour dégager les caractéristiques bibliographiques d'un corpus sera exposée.

L'illustration de cet outil bibliométrique sera bâtie autour d'exemples de traitements pour un échantillon de dix références. Cet échantillon n'a aucun sens en tant que tel, sa taille est réduite au strict minimum pour une meilleure présentation didactique. Ce sont dix références brevets provenant de la base Derwent. Elles sont reproduites dans leur intégralité en ANNEXE 2 (le nom du fichier informatique correspondant est "10refs."). La signification de chaque champ a déjà été donnée à la table 1.

1. Présentation générale des modules

La session de travail des traitements bibliographiques appliqués à un fichier s'effectue le passage par une succession de menus représentés sur la figure 39. Ces menus sont disposés dans le menu principal selon l'ordre dans lequel l'utilisateur va les employer.

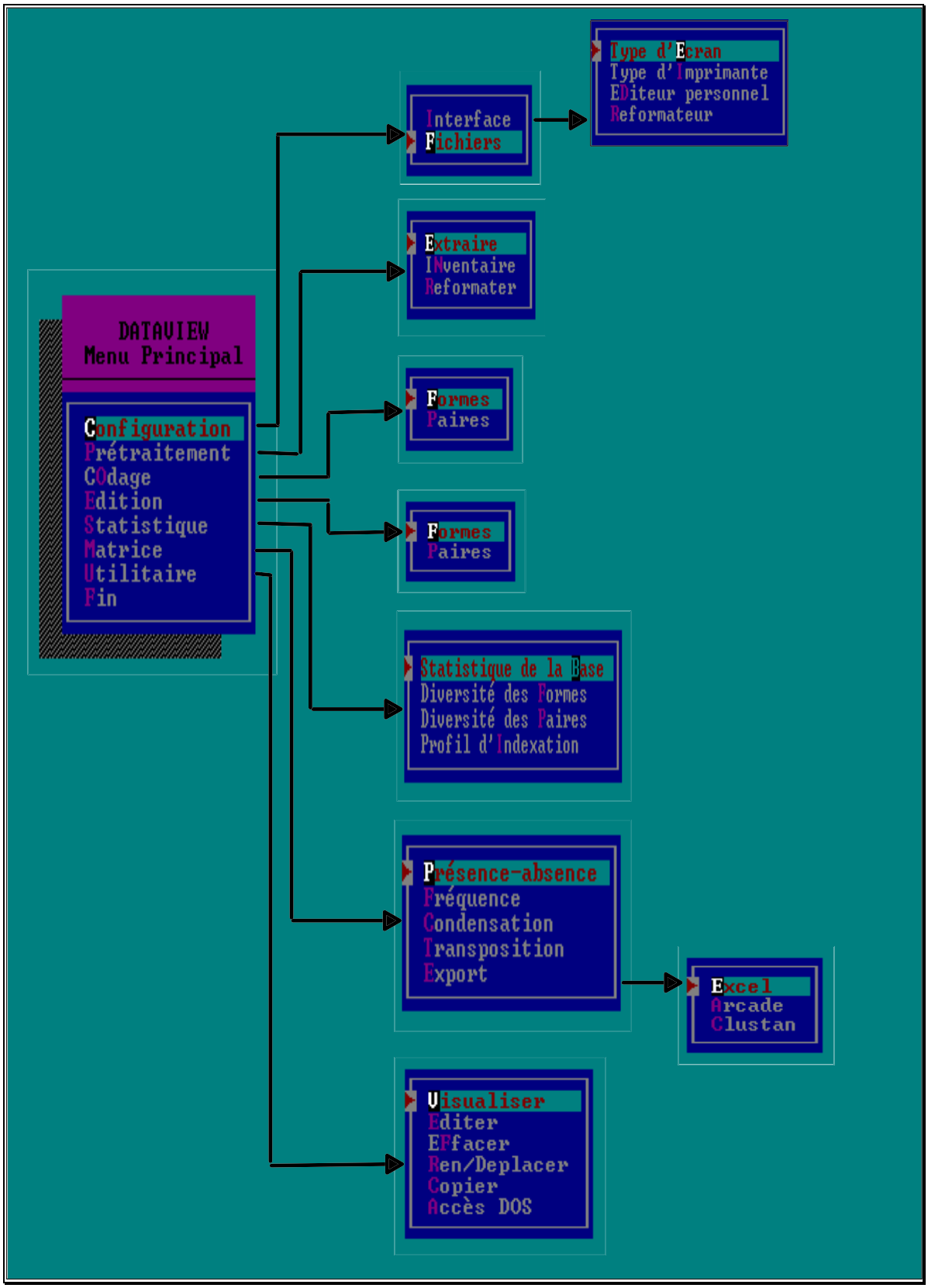


Figure 39: Synoptique des menus de DATAVIEW

2. Configuration de la session de travail

Pour débuter une étude bibliométrique d'un fichier contenant l'ensemble des références à étudier, il faut tout d'abord configurer la session pour les traitements ultérieurs.

- Cette configuration demande à l'utilisateur d'introduire dans une première fenêtre de dialogue (exemple figure 40):

  - ☞ le nom du fichier d'origine qui contient les références (extension fixe: "*****."):
    Pour notre exemple le fichier se nomme "10refs." et il se situe sur le répertoire C:\EXEMPLE. On peut remarquer que la recherche de ce fichier peut se réaliser grâce à une fenêtre de type ascenseur qui permet de se déplacer dans l'arborescence des répertoires du disque dur. Ainsi, l'utilisateur ne se souvenant plus du nom et du chemin d'un fichier peut le retrouver aisément. Cette option de gestion de fichier et de répertoire est indispensable car DATAVIEW va créer très rapidement un grand nombre de fichiers de travail ou d'éditions de résultats. Cette option est donc accessible à chaque introduction d'un nom de fichier dans les fenêtres de dialogue de DATAVIEW.

  - ☞ un nom pour le fichier de travail pour DATAVIEW (extension fixe: "***.job**"):
    Les champs bibliographiques ne sont pas tous utiles lors d'une étude bibliométrique. Pour gagner du temps en traitement informatique, on constitue un fichier intermédiaire qui ne contient que les champs intéressant l'étude. Le champ qui est traité dans l'exemple présent est le champ des Codes Derwent (DC), on a donc choisi de nommer ce fichier "DC.JOB"

  - ☞ le nombre de champs que l'on veut traiter en même temps:
    DATAVIEW a été conçu pour pouvoir traiter prochainement plusieurs champs sans étape de reformatage. Dans la version actuelle, il ne traite qu'un champ à la fois par session de travail. Cette saisie du nombre de champs est donc pour l'instant bloquée à 1.
    Pour traiter les corrélations entre plusieurs champs, il faudra alors passer par une étape de reformatage pour introduire les données de ces différents champs dans un champ unique (Cf *Extraction et homogénéisation des champs étudiés*).

  - ☞ la longueur maximale des formes:
    C'est-à-dire le plus grand nombre de caractères que peut avoir une forme. Les raisons de cette saisie sont purement informatiques. La constitution des données bibliométriques pour chaque forme se fait par l'intermédiaire d'une gestion informatique de fichiers à accès directs qui imposent une longueur d'enregistrement fixe. Cette longueur est donc fonction de celle des formes. Si la longueur introduite

ici est inférieure à la longueur maximale des formes du corpus alors certaines formes seront tronquées et créeront des ambiguïtés lors des comptages informatiques. Si c'est le cas, lors de l'étape des constitutions des caractéristiques bibliométriques (la phase du "codage") un panneau d'alerte le fera remarquer à l'utilisateur et lui proposera de recommencer les comptages en choisissant automatiquement la longueur maximale que DATAVIEW aura rencontrée. Cette saisie n'est donc pas très importante car elle sera réajustée automatiquement si besoin est.

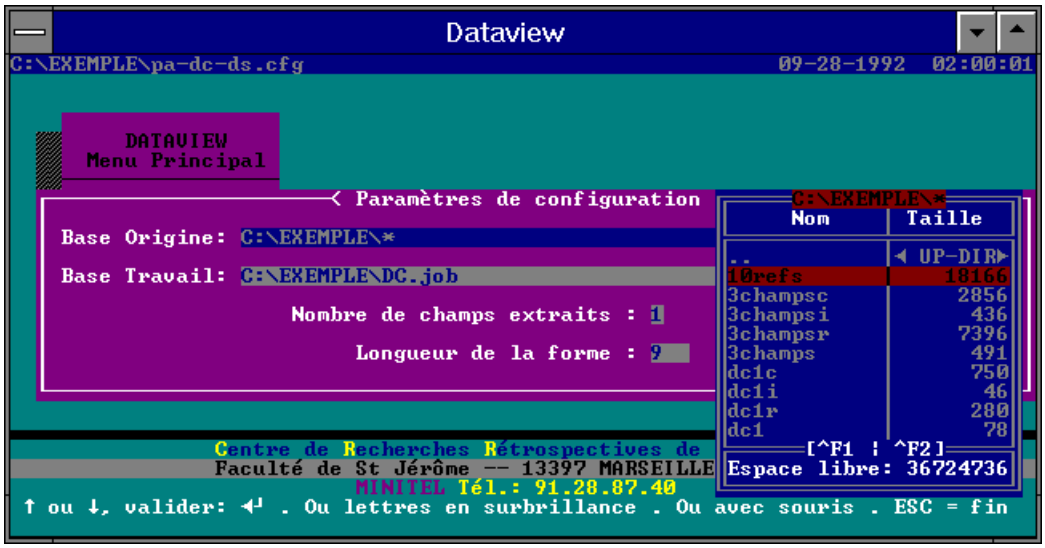


Figure 40: Première fenêtre du paramétrage de la configuration

- Les saisies de cette fenêtre étant validées, une seconde apparaît pour demander de préciser (figure 41):

  ☞ <u>L'intitulé du champ qui sera traité:</u>

  Dans l'exemple, l'intitulé du champ est "DC - " ("DC" + deux espaces + "-" + un espace). C'est une suite univoque de caractères qui permet de reconnaître le champ qui sera étudié pendant la session de travail.

  ☞ <u>Les séparateurs à considérer pour ce champ:</u>

  Ce sont les caractères qui seront reconnus comme séparateurs de formes lors du codage-comptage. Dans le cas du champ DC le caractère qui délimite les codes Derwent les uns des autres est l'espace (espace juste avant le curseur jaune de la figure 41).

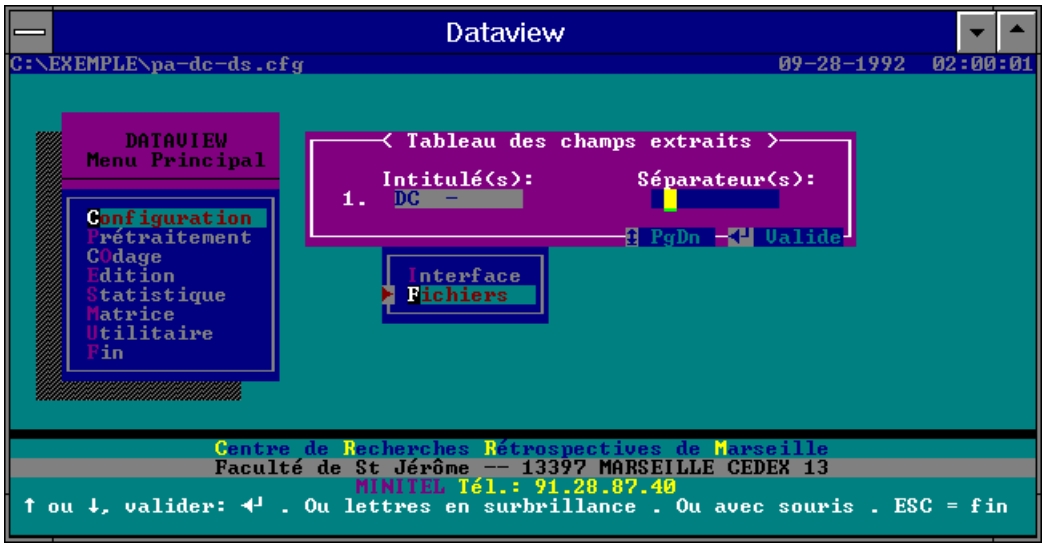


Figure 41: Paramétrage du champ traité

- Lorsque ces deux fenêtres sont validées, ces paramètres de travail sont conservés dans un fichier dont le nom est introduit par l'utilisateur (figure 42). Ce fichier de sauvegarde (extension fixe: "***.cfg**") permettra de réutiliser plus tard cette configuration de travail sans la redéfinir.

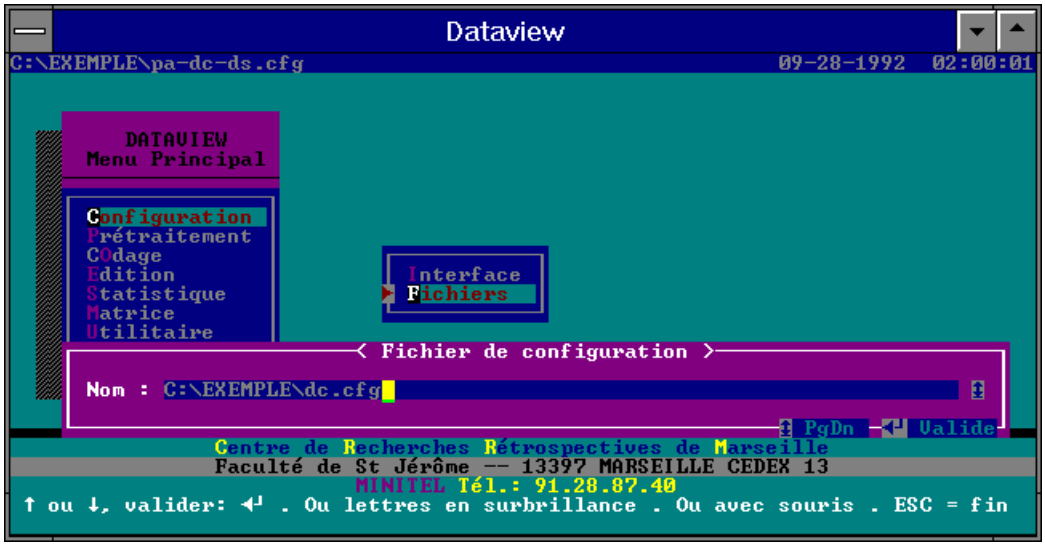


Figure 42: Sauvegarde des paramètres de configuration

3. Extraction et homogénéisation des champs étudiés

Cette seconde étape va constituer "**la base de travail**" exploitée par DATAVIEW.

Deux cas se présentent:

- ☞ le champ que l'on souhaite traiter est déjà sous un format bibliométrique (champ "propre") et alors un module d'extraction de DATAVIEW va permettre de constituer cette base de travail
- ☞ le champ nécessite une phase de pré-traitement pour la "nettoyer" et alors on aura recours à un reformateur soit trouvé dans le commerce soit développé en interne.

⇨ Extraction de champ propre:

Le champ DC, que l'on vient de configurer en tant que champ de travail, vérifie d'origine cette condition de propreté bibliométrique. Il est donc simplement extrait du fichier d'origine pour être placé dans le fichier de travail défini dans la configuration. Ceci est réalisé automatiquement par l'option *Extraction* du sous-menu *Prétraitement.*

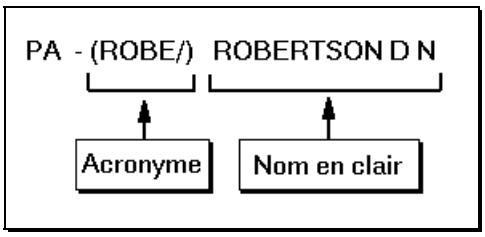
Le résultat de l'extraction pour notre exemple est la table suivante:

```
-1-
DC  - A96 B07 D22 G03 A14 P34 P32
-2-
DC  - B07 C07 D22 P34
-3-
DC  - A96 B07 D22 B01 P32
-4-
DC  - B01 D22 B07 P32
-5-
DC  - B02
-6-
DC  - A96 B07
-7-
DC  - B07
-8-
DC  - A96 B07 D22 P34
-9-
DC  - B07 S05 P34
-10-
DC  - B05 D21 B07 P32
```

Table 10: Fichier de travail: DC.JOB

⇨ Extraction et homogénéisation de champs:

Pour étudier le champ des organismes déposant des brevets (*Patent Assigny*: PA), il est recommandé d'éliminer le nom en clair des organismes pour ne conserver que l'acronyme. Les variantes du nom d'une même société sont grandes (filiales, rachats, fusions...). Pour limiter cela, Derwent réindexe ces noms pour les réunir sous une forme unique de 5 caractères nommés acronymes. Les deux informations sont présentes dans le champ PA:



Le reformateur va donc faire double emploi, extraire le champ PA et éliminer les noms de sociétés en clair. Pour un reformatage aussi simple, les résultats de comptages bibliométriques sont nettement améliorés. La table 11 montre le fichier de travail résultant:

```
-1-
PA  - (LOHM )
-2-
PA  - (DERM-)
-3-
PA  - (ROBE/)
-4-
PA  - (ROBE/)
-5-
PA  - (SQUI )
-6-
PA  - (KIMY/)
-7-
PA  - (SMIK )
-8-
PA  - (SEKI )
-9-
PA  - (MEDT )
-10-
PA  - (PHAR-)
```

Table 11: fichier de travail PA.JOB

⇨ Extraction de plusieurs champs:

DATAVIEW est actuellement conçu de telle façon qu'il ne peut comptabiliser que les paires entre des formes du même champ. Or, pour connaître les corrélations entre les formes présentes dans deux champs différents, le comptage des paires constituées entre ces deux champs est indispensable. Pour réaliser ces comptages nous mettrons alors bout à bout le contenu de ces champs. Puis nous comptabiliserons les paires comme si les formes appartenaient au même champ.

Si on veut connaître les domaines où chaque société dépose ses brevets et à quelles fréquences, il faut former les paires entre un champ faisant référence à la société déposante (champ PA) et un champ faisant référence aux thèmes abordés par l'innovation (on a choisi ici le DC).

Le reformatage entrepris pour réunir les deux ensembles de formes doit aussi homogénéiser les séparateurs des formes. Dans le cas présent, l'espace ne peut pas être défini comme séparateur puisque certains acronymes ont un espace comme dernier caractère (voir table 11). Donc, le reformatage doit injecter, à la place des espaces du champ DC, un autre caractère qui servira de séparateur: le point-virgule par exemple. Un point-virgule est aussi inséré avant la première forme du second champ, ici DC, pour délimiter la dernière forme du premier champ avec la première forme de ce second champ lors de leur réunion.

Le fichier de travail constitué est donc de cette forme:

```
-1-
PADC- (LOHM );A96;B07;D22;G03;A14;P34;P32
-2-
PADC- (DERM-);B07;C07;D22;P34
-3-
PADC- (ROBE/);A96;B07;D22;B01;P32
-4-
PADC- (ROBE/);B01;D22;B07;P32
-5-
PADC- (SQUI );B02
-6-
PADC- (KIMY/);A96;B07
-7-
PADC- (SMIK );B07
-8-
PADC- (SEKI );A96;B07;D22;P34
-9-
PADC- (MEDT );B07;S05;P34
-10-
PADC- (PHAR-);B05;D21;B07;P32
```

Table 12: Fichier de travail PA-DC.JOB

La même procédure peut être réalisée pour toutes les combinaisons de champs fécondes en renseignements. On peut donc vouloir conjointement étudier les déposants, les thèmes qu'ils abordent dans leurs brevets et les pays où ils étendent ces brevets. Le fichier de travail à construire est alors de ce type:

```
-1-
CHP - AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;LI;LU;NL;SE;
      (LOHM );A96;B07;D22;G03;A14;P34;P32
-2-
CHP - AT;BE;CH;DE;DK;ES;FR;GB;IT;LI;LU;NL;SE;
      (DERM-);B07;C07;D22;P34
-3-
CHP - CA;JP;AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;(ROBE/);A96;B07;
      D22;B01;P32
-4-
CHP - CA;JP;US;AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;LU;NL;SE;(ROBE/);
      B01;D22;B07;P32
-5-
CHP - ;(SQUI );B02
-6-
CHP - ;(KIMY/);A96;B07
-7-
CHP - AU;CA;JP;KR;US;AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;LU;NL;SE;
      (SMIK );B07
-8-
CHP - ;(SEKI );A96;B07;D22;P34
-9-
CHP - AU;CA;FI;JP;NO;AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;KR;LU;NL;SE;
      (MEDT );B07;S05;P34
-10-
CHP - AT;JP;AT;BE;CH;DE;DK;ES;FR;GB;GR;IT;LU;NL;SE;
      (PHAR-);B05;D21;B07;P32
```

Table 13: Fichier de travail PA-DC-DS.JOB

4. Détermination des caractéristiques bibliométriques: le "codage"

Une fois la base de travail constituée, vient alors l'**étape principale des traitements de DATAVIEW**. Elle génère les fichiers contenant les **caractéristiques bibliométriques du fichier de travail**. Ceux-ci ne sont que l'expression du comptage de la diversité des formes du texte ainsi que de leurs localisations les unes parmi les autres.

Ces fichiers prendront le même nom que le fichier de travail mais suivis de nouvelles extensions. Ils sont au nombre de six:

- le fichier "***.sta**" contient les renseignements sur les statistiques des fréquences de la base de travail.

- ☞ les fichiers "***.lex**" et "***.ocf**" contiennent toutes les données des fréquences et des positions des formes dans la base
- ☞ les fichiers "***.prt**" et "***.ocp**" contiennent toutes les données des fréquences et de position des paires dans la base
- ☞ un dernier fichier "***.faf**" ne sert que pour de la création des cinq précédents.

Le premier est rédigé en texte lisible (Cf *Editions des résultats - Les statistiques des fréquences de la base*) mais **sa mise en forme est directement lisible par le tableur Excel**. Ce logiciel a été choisi comme le standard DATAVIEW pour les exportations de données bibliométriques vers un tableur. L'exploitation de ce fichier est présentée dans la partie *Les traitements infographiques en sortie de DATAVIEW.*

Les autres fichiers sont créés dans un format codé illisible. La lecture de leurs données ne s'effectue que par l'intermédiaire de DATAVIEW. Leurs formats structurés (fichiers à accès direct) permettent des **recherches bien plus rapides** que pour les fichiers en texte lisible (fichiers séquentiels). Les recherches des données concernant un ensemble de formes, leurs tris et la reconstitution de leur réseau de paires sont fortement accélérés par l'emploi de tels fichiers. **Donc, les traitements effectués sur la base de ces fichiers sont bien plus rapides que ceux réalisés directement à partir du fichier de travail. Ces fichiers sont comparables aux fichiers permutés créés par les gestionnaires de bases de données (SGBD), mis à part qu'ils contiennent en plus des données purement bibliométriques**.

En revanche, la constitution de ces fichiers fait appel à un traitement assez long. **Pour améliorer le temps de création de ces fichiers un algorithme de *H-coding* des formes a été mis au point au CRRM [LATE87]**. C'est pour cette raison que cette étape d'inventaire et de dénombrement de la base de travail est nommée dans DATAVIEW sous l'appellation de "**Codage**". C'est cet algorithme qui donne toute sa puissance à cet outil. Il livre des résultats dans des temps raisonnables, ce qui permet de réitérer certains traitements dans le seul but d'affiner les résultats.

5. Editions des résultats

Quelles sont finalement les caractéristiques bibliométriques proposées par DATAVIEW. Elles se partagent en quatre grandes catégories:

a) Les statistiques des fréquences de la base

Ces données sont toutes regroupés dans le fichier "statistique" ("*.sta") généré par DATAVIEW au cours de l'étape du codage-comptage. Ces données ne renseignent pas sur les entités bibliométriques (formes, paires) elles-mêmes mais sur leurs regroupements par rangs de fréquence.

Pour l'exemple du fichier de travail DC.JOB, le fichier DC.STA est le suivant:

```
Nom de la base traitée    : C:\EXEMPLE\DC.JOB
Intitulé du champ traité  : "DC  - "
Nombre de références      :  10
Longueur forme configurée :  9
Longueur maxi. des formes :  3
Nbre formes maxi.par référ:  7   (Ref n° 1)

                         --- Codage des Formes ---
Nombre total de formes    :  13
Nombre total de réf.-hapax:  2
Fréquence mini. des formes:  1
Fréquence maxi. des formes:  9
Nombre total d'occurrences:  35
Occurrence maxi.des formes:  9

Nombre potentiel paires   :  59
Nombre potentiel triplets :  62

                       --- Répartition des Formes ---
Valeur du X                Nb Formes à Freq X                  Nb Formes à Occ X
      9                           1                                   1
      5                           1                                   1
      4                           3                                   3
      2                           1                                   1
      1                           7                                   7

                         --- Profil d'Indexation ---
Valeur du X                Nb Réf. ayant X Formes
      7                           1
      5                           1
      4                           4
      3                           1
      2                           1
      1                           2

                          --- Codage des Paires ---
Nombre réel de paires     :  35
Fréquence mini.des paires :  1
Fréquence maxi.des paires :  5
Occurence totale de paires:  59

                       --- Répartition des Paires ---
Valeur du X                Nb Paires à Fréq X
      5                           1
      4                           3
      3                           3
      2                           5
      1                          23
```

Table 14: fichier DC.STA

❑ **L'entête** de ce fichier donne des renseignements généraux sur la base de travail traitée.

❑ **La partie du *codage des formes*** informe sur:

- ☞ le nombre de formes rencontrées
- ☞ le nombre de références ne contenant qu'une forme (*référence-hapax*)
- ☞ la plus faible et la plus forte valeur dans les fréquences de forme
- ☞ le nombre d'occurrences et la valeur maximale dans les fréquences des occurrences.
- ☞ une estimation du nombre de paires et de triplets qui seraient créés à condition que toutes les formes soit différentes dans chaque champ (formule de la combinaison).

La répartition des formes mise sous forme d'histogramme donne la figure 43:

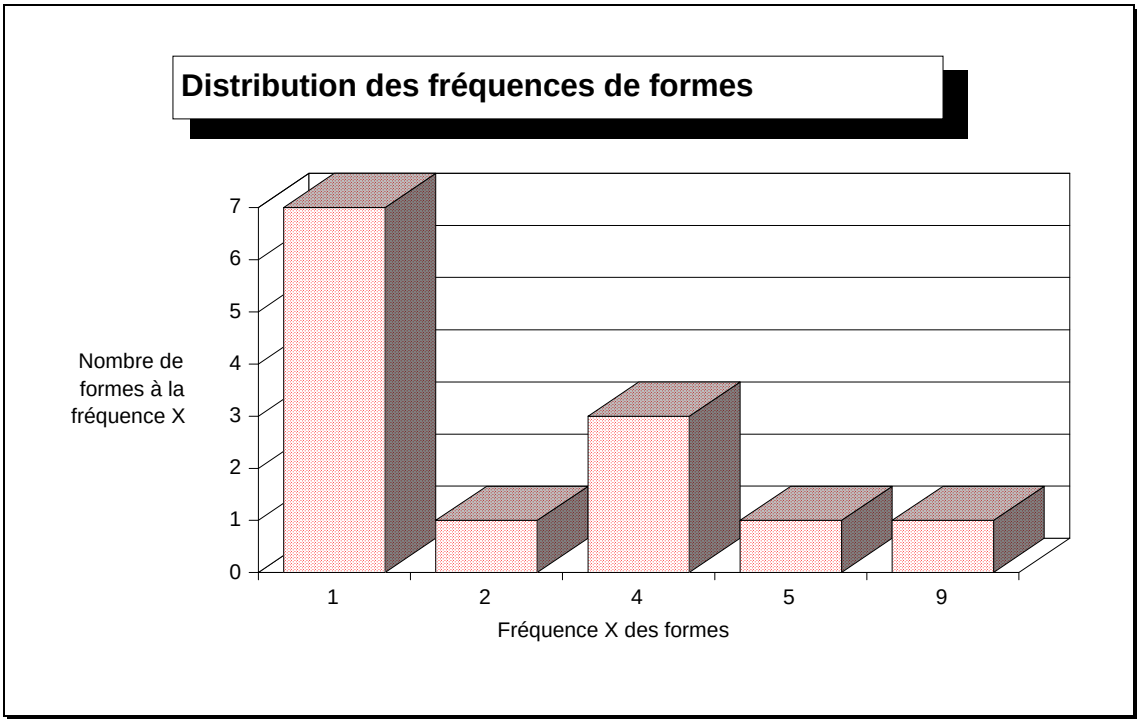


Figure 43: Distribution des fréquences de formes

On peut remarquer que pour un aussi faible corpus de références, la distribution ne suit pas parfaitement la courbe hyperbolique attendu (voir *Le "coeur" et la "dispersion"* du chapitre II).

❑ **La dernière partie** donne les mêmes informations mais pour les **paires de formes**: nombre de paires, fréquence maximale et minimale, occurrence des paires. L'appellation un peu trompeuse d'occurrence des paires ne correspond pas à la cooccurrence comme nous l'avons exposé dans le paragraphe *Solution à la diversité des éléments bibliométriques.* Cette valeur est la somme des combinaisons (voir formule dans le paragraphe cité ci-dessus) des paires de formes de chaque champ après avoir éliminé les occurrences de formes redondantes. Dans le cas du champ DC, le nombre d'occurrences de paires est égal à celui estimé lors du comptage des formes puisqu'il n'y a jamais deux occurrences d'une même forme dans un même champ (caractéristique d'un champ indexé).

La distribution des fréquences des paires, quant à elle, suit un peu mieux une loi hyperbolique:

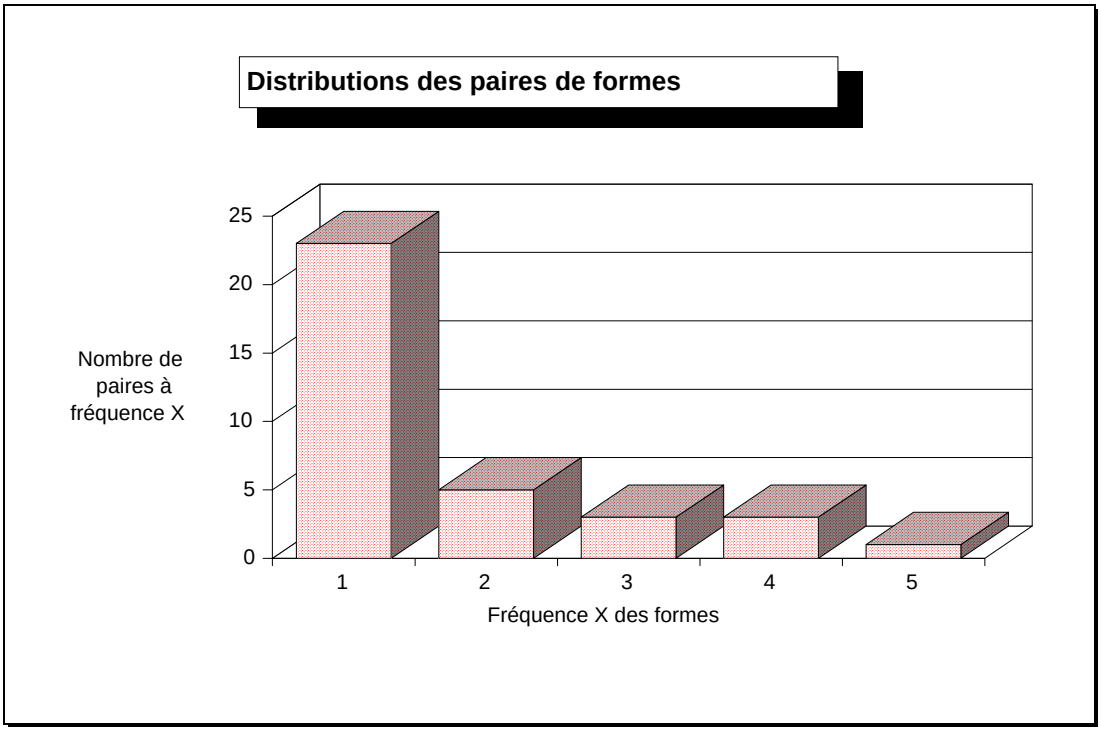


Figure 44: Distribution des fréquences de paires

❑ Avant cette distribution des fréquences des paires, une tout autre distribution est fournie. Elle indique la **répartition du nombre de formes par champ**. Cette distribution a été nommée ***Profil d'indexation*** mais elle est plus connue par les documentalistes sous le nom de ***Profondeur d'indexation*** [CHAU88]. **Contrairement aux deux autres, cette distribution s'apparente aux distributions gaussiennes**. Dans le cas présent, la moyenne du nombre de formes par champ avoisine la valeur 4.

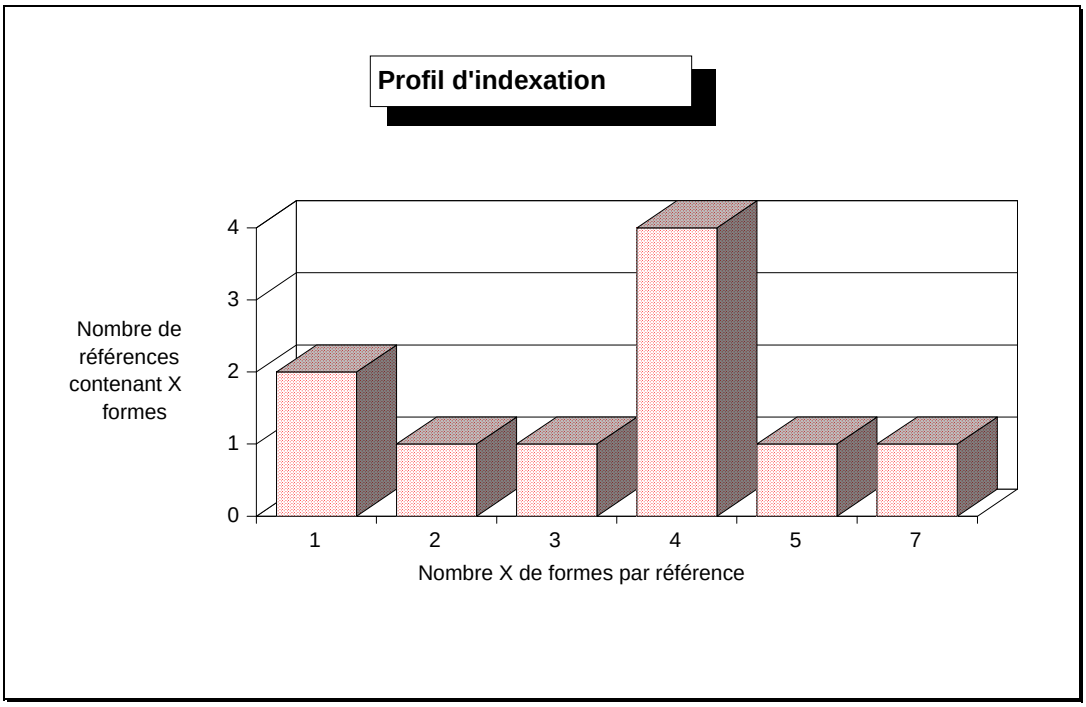


Figure 45: Profil d'indexation

Ce fichier "statistique" fournit les données de base pour construire les lois bibliométriques classiques (Bradford et Lotka). En l'absence de mesures synthétiques, les distributions elles-mêmes font office de référence lors des études bibliométriques.

Il sera profitable de se référer à ces données d'ensemble dès lors que l'on souhaite négliger des parties d'information dans les analyses. Ainsi, leur consultation sera bénéfique pour choisir les seuils de fréquences des formes pour les analyses de données statistiques. Elles permettront d'estimer la quantité d'information perdue et donc le bon sens des résultats à venir.

b) Les listes de formes

De nombreuses options d'édition sont possibles pour constituer ces listes de formes. L'utilisateur qui cherche à connaître la diversité des formes constituant le contenu d'un champ pourra le faire en activant la fenêtre de dialogue d'***édition des formes***:

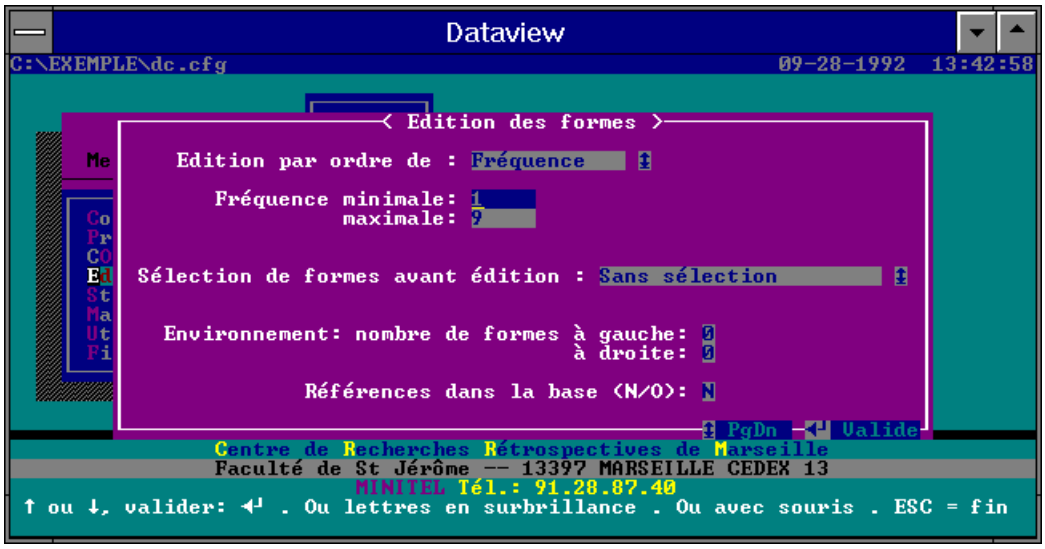


Figure 46: paramètres d'édition des formes

⇨ Intervalle de fréquence:

L'utilisateur peut choisir de n'éditer qu'une partie de l'ensemble de formes en déterminant un intervalle de fréquence. **Il est donc très facile d'éditer les formes qui constituent le coeur de la distribution après l'avoir déterminé à partir des données du fichier des statistiques des fréquences.**

⇨ Tri des formes:

Le classement des formes va permettre une meilleure lecture de la liste. DATAVIEW propose un classement par ordre des fréquences décroissantes (table 15) ou par ordre alphabétique (table 16).

EDITION DES FORMES DANS LES FREQUENCES : 1 - 9

FICHIER TRAITE: c:\exemple\dc.job

FREQUENCE	OCCURRENCE	HAPAX	FORME
9	9	1	B07
5	5	0	D22
4	4	0	P34
4	4	0	P32
4	4	0	A96
2	2	0	B01
1	1	0	S05
1	1	0	G03
1	1	0	D21
1	1	0	C07
1	1	0	B05
1	1	1	B02
1	1	0	A14

Table 15: édition par ordre des fréquences

Cette édition est assimilable au tri à plat des études statistiques employées en analyse d'enquête.

Il faut remarquer que pour notre exemple la valeur de la fréquence d'une forme est identique à celle de son occurrence. Ceci est tout à fait normal puisque le champ pris en exemple est celui de la classification contrôlée par Derwent. Un code de cette classification ne peut être attribué qu'une seule fois par document. Il ne peut donc pas être présent plusieurs fois pour une référence.

La colonne "**HAPAX**" fait apparaître les formes qui sont présentes seules dans un champ pour le renseigner. Ainsi dans l'exemple, deux champs DC sont renseignés par un seul code. Les deux formes employées dans ces références ont leurs colonnes HAPAX cochées d'une valeur égale à 1. Le fichier "DC.STA" nous avait informés sur le fait que deux références n'étaient renseignées que par 1 forme, l'édition des formes nous indique quelles sont ces formes-hapax.

⇨ Références d'origine:

On peut aussi vouloir connaître les références où apparait chaque forme (Table 16).

```
                    EDITION DES FORMES DANS LES FREQUENCES : 1 - 9


FICHIER TRAITE: c:\exemple\dc.job

FREQUENCE OCCURRENCE HAPAX  FORME
~~~~~~~~~ ~~~~~~~~~~ ~~~~~  ~~~~~
       1          1     0    A14        Ref.: 1
       4          4     0    A96        Ref.: 1, 3, 6, 8
       2          2     0    B01        Ref.: 3, 4
       1          1     1    B02        Ref.: 5
       1          1     0    B05        Ref.: 10
       9          9     1    B07        Ref.: 1, 2, 3, 4, 6, 7, 8, 9, 10
       1          1     0    C07        Ref.: 2
       1          1     0    D21        Ref.: 10
       5          5     0    D22        Ref.: 1, 2, 3, 4, 8
       1          1     0    G03        Ref.: 1
       4          4     0    P32        Ref.: 1, 3, 4, 10
       4          4     0    P34        Ref.: 1, 2, 8, 9
       1          1     0    S05        Ref.: 9
```

Table 17: Edition par ordre alphabétique et indication des références d'origine

⇨ Sélection d'un ensemble de formes:

Une option propose aussi à l'utilisateur de faire une sélection des formes qu'il a l'intention d'éditer en final (figure 47).

DATAVIEW présente alors les formes, dont les fréquences sont comprises dans l'intervalle choisi, dans une **fenêtre de type ascenseur**. C'est-à-dire que toutes les formes disposées dans cette fenêtre ne sont pas visibles en même temps mais elles sont accessibles en déplaçant le curseur à l'écran vers le bas (ou le haut) de la liste. Lors de ce déplacement la partie des formes apparentes à l'écran donne l'impression de "glisser" lançant apparaître les nouvelles formes (figure 48a).

L'utilisateur peut sélectionner les formes à éditer soit une à une, en se plaçant dessus et en activant la touche *Insertion*, soit en automatique par des sélections (et désélections) de formes répondant à un masque de reconnaissance (figure 48b et figure 48c). Les caractères de jokers sont les mêmes que ceux du MS-DOS: l'astérisque ("*" = remplace une suite de caractères quelconques) et le point-d'interrogation ("?" = remplace un caractère quelconque). L'astérisque de DATAVIEW couvre une plus large fonctionnalité que celle du DOS puisqu'elle peut servir de troncature à gauche ou peut être borné par un caractère précis à droite.

Ce mode de sélection est proposé dans DATAVIEW pour toutes les autres interventions de choix: sélection des paires à éditer, sélection des formes à distribuer en ligne ou en colonne d'un tableau.

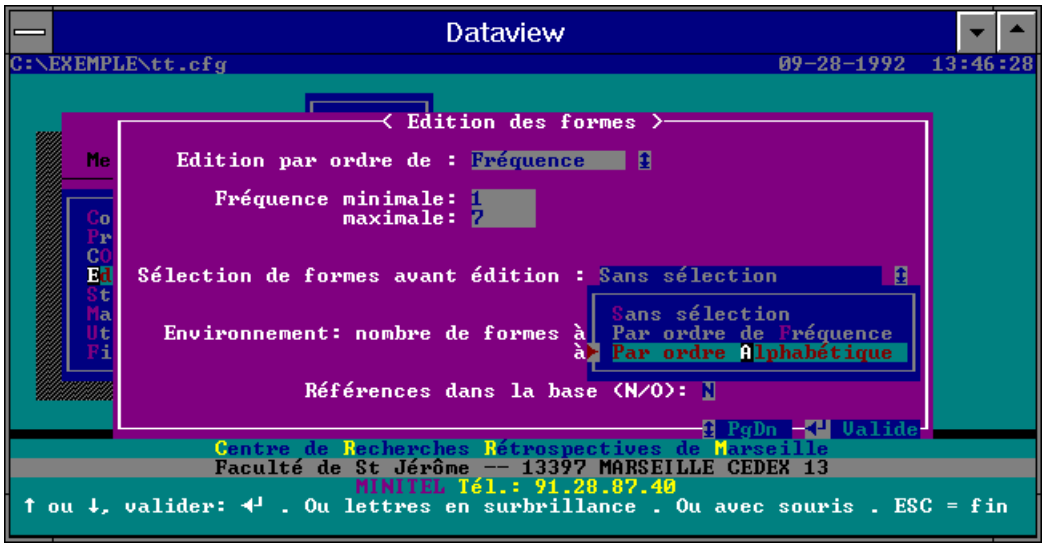


Figure 47: Option de sélection des formes à éditer

Figure 48a

Figure 48b

Figure 48c

⇨ Environnement des formes:

Une dernière option peut s'avérer utile lors d'étude de champ dont le contenu est en langage libre. Le champ n'étant pas contrôlé suivant des règles, un terme peut être utilisé dans plusieurs sens différents. Pour s'assurer que certains termes clés de l'étude expriment toujours le même concept, on peut prendre connaissance du contexte dans lequel il est utilisé. Ceci est possible par l'option ***Environnement*** (figure 46). L'utilisateur précise le nombre de formes qu'il veut connaître à droite et à gauche de chaque forme éditée.

L'exemple ci-dessous (table 18) présente l'environnement de trois formes sélectionnées (*drug, patch, transdermal*) parmi celles qui constituent les résumés des références (champ AB). Dans cette édition, le symbole "«»" remplace la forme sélectionnée dans son contexte. Devant chaque ligne d' "environnement", il est précisé la référence d'où il est soutiré.

```
                EDITION DES FORMES SELECTIONNEES DANS LES FREQUENCES : 5 - 10


FICHIER TRAITE: c:\exemple\ab.job

FREQUENCE OCCURRENCE HAPAX  FORME
~~~~~~~~~ ~~~~~~~~~~ ~~~~~  ~~~~~
       8         10     0    patch
- 8- (J03227919) Hydrophilic transdermal «» is composed of plastic
- 6-           (KR9006832) A transdermal «» for transdermal admin. of
- 8-                     - The transdermal «» provides good water absorption,
- 4-                         a band, disc, «» or bracelet. The ST1435
- 9-                           appts. is a «» system used for the
-10-                          as only one «» is required in 24
- 7-                disposal. The present «» fits comfortably against the
- 8-                simple structure. The «» causes no skin irritation.
- 3-                             such as a «» which adhere to the
- 5-                    via a transdermal «» or nasal inhalation solns.

       6         10     0    transdermal
- 1-  (EP-464573) Topical or «» patches comprise a backing
- 8- (J03227919) Hydrophilic «» patch is composed of
- 6-           (KR9006832) A «» patch for transdermal admin.
-10- (WO9115176) Compsn. for «» application of a cholinergic
- 8-     USE/ADVANTAGE - The «» patch provides good water
- 2-            a matrix for «» admin. of the compsn.,
- 5-    administration via a «» patch or nasal inhalation
- 2-  medicament adapted for «» admin.; (ii) a carrier,
- 6-   transdermal patch for «» admin. of a drug
- 2-               use in the «» admin. of a medicament,

       5         11     0    drug
-10-           activity comprises the «» and the ester of
- 6-                      admin. of a «» comprising; (a) drug reservoir
-10-   cholinergic or anticholinergic «» of high intrinsic specific
- 6-                 contact with the «» reservoir layer, adhesive-coated sheet,
- 6-              drug comprising; (a) «» reservoir layer contg. pharmaceutically
- 3-                  drug through the «» releasing surface to the
- 5- endocrine hypersecretory states, «» induced tardive dyskenesia, allergies,
-10-                fatty acid. Active «» is physostigmine, naloxone, nicotine,
- 3-               matrix releases the «» through the drug releasing
- 9-         other macromolecules. The «» delivery is dependent upon
- 3-    transdermally administering a «» has a drug-releasing surface
```

Figure 18: Environnement des formes lors de leur édition

c) Les listes de paires

Pour l'édition des paires nous retrouvons les mêmes options plus quelques nouvelles (figure 49). Nous n'expliquerons que ces dernières.

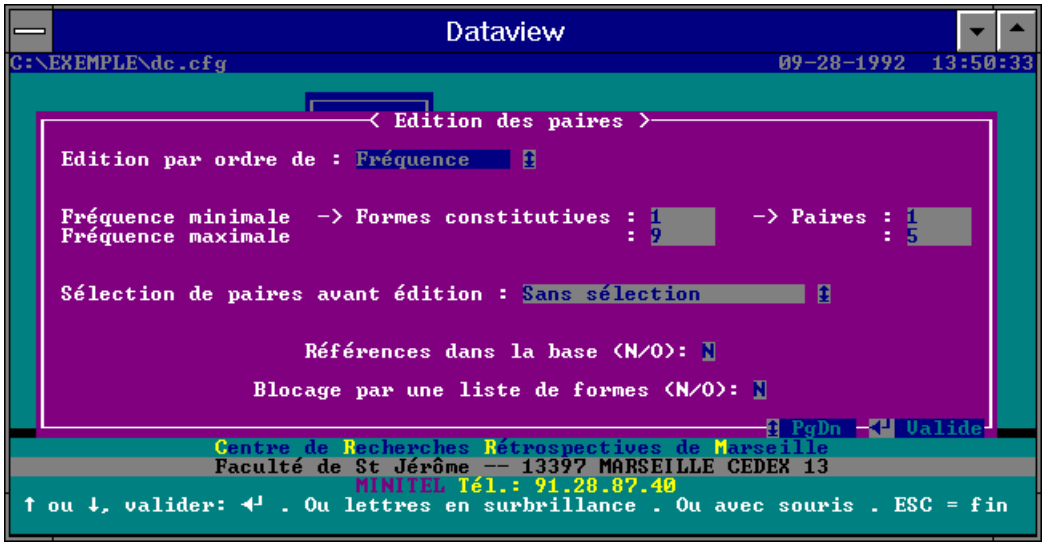


Figure 49: Paramètres d'éditions des paires

L'édition finale des paires comporte aussi de nouvelles données par rapport à celles fournies par l'édition des formes. Pour le cas du traitement du champ DC l'édition des paires fournit les renseignements suivants:

```
                          EDITION DES PAIRES DANS LES FREQUENCES : 1 - 5
Fichier traité: c:\exemple\dc.job
Intervalle de fréquences des formes constitutives: 1 - 9

Fréquence  Corrélation  Paire
~~~~~~~~~  ~~~~~~~~~~~  ~~~~~
        5        0.333  B07 = D22
        4        0.272  B07 = P34
        4        0.272  B07 = P32
        4        0.272  A96 = B07
        3        0.408  D22 = P34
        3        0.408  D22 = P32
        3        0.408  A96 = D22
        2        0.612  P32 = B01
        2        0.5    D22 = B01
        2        0.166  B07 = B01
        2        0.166  A96 = P34
        2        0.166  A96 = P32
        1        1      B05 = D21
        1        0.408  P32 = D21
        1        0.408  P32 = B05
        1        0.111  B07 = D21
        1        0.111  B07 = B05
        1        0.408  P34 = S05
        1        0.111  B07 = S05
        1        0.102  A96 = B01
        1        0.408  P34 = C07
        1        0.333  D22 = C07
        1        0.111  B07 = C07
        1       -0.25   P32 = P34
        1        0.408  G03 = P34
        1        0.408  G03 = P32
        1        0.333  D22 = G03
        1        0.111  B07 = G03
        1        0.408  A96 = G03
        1        0.408  A14 = P34
        1        0.408  A14 = P32
        1        1      A14 = G03
        1        0.333  A14 = D22
        1        0.111  A14 = B07
        1        0.408  A14 = A96
```

Table 19: Editions des paires de fichier de travail DC.JOB

On peut tout d'abord remarquer que la paire A96=B07 est la même que la paire B07=A96 et que la notion d'occurrence de paires est occultée. Par contre, la notion d'indice d'association est introduite. La mesure d'association proposée par défaut est celle calculée selon la formule de l'indice de corrélation (Cf ANNEXE 1).

⇨ Tri des paires:

Si l'utilisateur le désire il peut choisir de classer les paires non plus seulement par fréquences décroissantes mais aussi par valeurs d'indice d'association décroissantes. Une trentaine d'indices statistiques binaires est à sa disposition (ANNEXE 1). Il en choisit une grâce à l'ouverture d'une fenêtre de sélection de type ascenseur (figure 50)

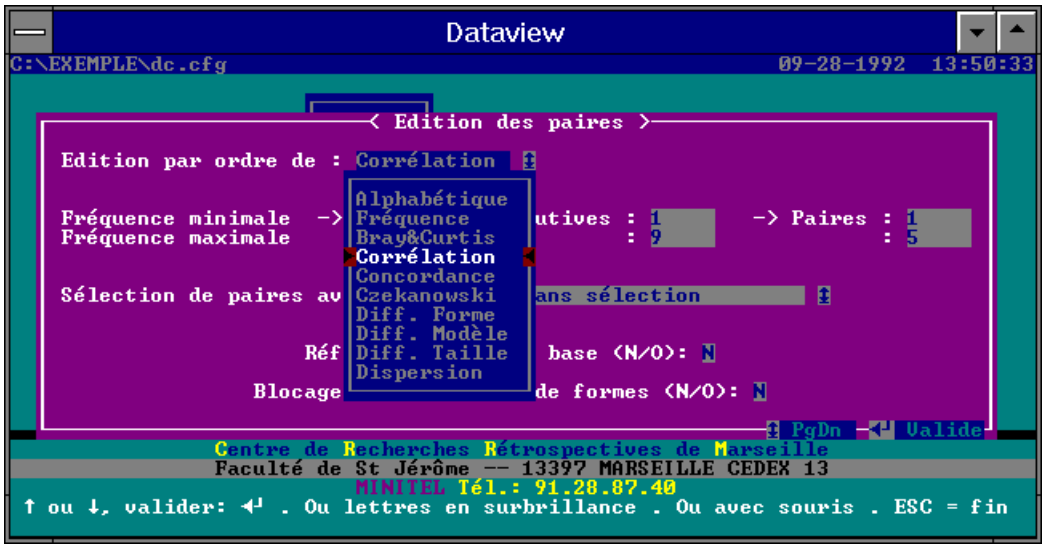


Figure 50: Liste d'indices d'association

Le classement par ordre alphabétique est toujours possible. Il y a deux formes à classer pour une paire, si bien que l'édition présentera deux fois la paire, la seconde fois les deux ayant permuté de place (table 20). Les listes sont donc deux fois plus longue que par un autre tri.

EDITION DES PAIRES DANS LES FREQUENCES : 3 - 5

Fichier traité: c:\exemple\dc.job
Intervalle de fréquences des formes constitutives: 1 - 9

Fréquence	Corrélation	Paire
4	0.272	A96 = B07
3	0.408	A96 = D22
4	0.272	B07 = A96
5	0.333	B07 = D22
4	0.272	B07 = P32
4	0.272	B07 = P34
3	0.408	D22 = A96
5	0.333	D22 = B07
3	0.408	D22 = P32
3	0.408	D22 = P34
4	0.272	P32 = B07
3	0.408	P32 = D22
4	0.272	P34 = B07
3	0.408	P34 = D22

Table 20: Edition de paires par ordre alphabétique

⇨ Intervalle des fréquences des formes constitutives:

L'indice de corrélation privilégie les paires dont les formes sont à très faible fréquence. Sur la table 19 les paires B05=D21 et A14=G03 ont la valeur de corrélation maximale (les valeurs de l'indice de corrélation varient de -1 à 1). Or ces deux paires n'apparaissent qu'une seule fois. Ceci s'explique par le fait que les formes constituant ces paires ont elles-mêmes une fréquence de un. Aucune des formes ne constitue d'autres paires avec d'autres formes de la base. Pour une édition des paires classées selon l'indice de corrélation, on peut vouloir éliminer ces formes rares qui vont s'introduire en tête de liste car elles n'apportent pas une information fiable. Ceci est réalisable en indiquant que l'édition ne prendra en considération que les formes ayant une fréquence supérieure à un seuil (figure 49: *fréquences minimale et maximale des formes constitutives*).

Cette réflexion peut être tenue également pour les fréquences de paires. L'utilisateur peut vouloir conserver les fréquences de paires très faibles mais uniquement pour les formes non rares.

⇨ Blocage de formes:

Cette notion de "**blocage**" correspond à la recherche de n-uplets à la demande. L'inventaire de l'ensemble des paires de formes est constitué automatiquement par DATAVIEW. Par contre, la connaissance des **triplets et des n-uplets de formes** n'est réalisable que sur demande de l'utilisateur. En fait, l'utilisateur demande l'édition des paires en précisant une liste de formes qui jouera le rôle de ***liste de formes de blocage*** (option figure 49).

Par exemple, pour connaître les triplets présents dans le fichier de travail "DC.JOB" dont l'une des formes constitutives est le code B01, l'utilisateur demandera l'édition des paires en précisant comme forme de blocage le code B01. DATAVIEW recherchera toutes les paires existantes mais seulement dans les champs comportant la forme bloquée. Cette forme bloque la recherche des paires pour une partie des références. C'est pour cette raison qu'elle est nommée *forme de blocage*. Le résultat de cette édition serait:

```
                    EDITION DES PAIRES DANS LES FREQUENCES : 1 - 5


Fichier traité: c:\exemple\dc.job
Intervalle de fréquences des formes constitutives: 1 - 9
Liste des formes du blocage: B01

Fréquence  Corrélation  Paire
~~~~~~~~~  ~~~~~~~~~~~  ~~~~~
       2         0.612  P32 = B01
       2         0.5    D22 = B01
       2         0.166  B07 = B01
       2         0      D22 = P32
       2        -1.088  B07 = P32
       2        -1.666  B07 = D22
       1         0.102  A96 = B01
       1        -0.25   A96 = P32
       1        -0.408  A96 = D22
       1        -1.769  A96 = B07
```

Table 20: Recherche des triplets formés avec le code BO1

Une partie des paires trouvées comporte bien évidemment la forme de blocage. On peut donc aussi trouver des formes qui constituent une paire avec B01 sans former de triplet. Ceci signifierait que les deux formes seraient seules dans une référence. Pour l'exemple présenté ce n'est pas le cas puisque toutes les formes qui constituent une paire avec B01 forment aussi une paire avec d'autres formes.
Les valeurs des corrélations des paires ne contenant pas B01 comme forme constitutive ont toutes chuté puisque des références où elles étaient présentes n'ont pas été considérées du fait qu'elles ne contenaient pas B01. Ces valeurs ne correspondent ni à la réelle mesure de l'association entre les deux formes, ni à une mesure de corrélations entre trois formes (les indices d'association ne peuvent mesurer que la force liant deux éléments). Cette mesure est plutôt une corrélation sous condition, la condition étant la présence d'une troisième forme B01.

Cet exemple présente un blocage de référence en fonction d'une forme choisie. DATAVIEW permet aussi de bloquer les références en fonction d'une liste de formes. Les paires sont alors recherchées seulement dans les références comportant toutes les formes de cette liste de blocage. **Ce traitement est donc assimilable à la recherche des n-uplets** (n étant le nombre de formes de blocage + 2) **par un choix déterministe de formes les constituant** (les n-2 formes de blocage).

⇨ Le cas des paires multichamps:

Prenons l'exemple du fichier de travail "PA-DC.JOB". L'édition des paires constituées lors de l'étude conjointe du champ PA et du champ DC permettrait de déceler rapidement: soit les principaux thèmes de recherche des sociétés en triant par ordre de fréquence, soit les thèmes spécifiques à chaque société en triant suivant une mesure d'association.

d) Recherche par chaînage de paires

Dans les systèmes documentaires classiques la recherche d'information est construite sur la base d'une logique booléenne (*et, ou, sauf*). Cette logique ne permet que de retrouver des éléments bibliographiques qui sont en lien direct les uns avec les autres. La méthode de recherche par chaînage de paires va par contre mettre en relation des éléments qui n'ont pas de lien direct mais qui ont en commun des données communes.

Prenons par exemple la base de travail "PA-DC-DS.JOB" (Cf *Extraction et homogénéisation des champs*). Admettons que l'on veuille connaître les sociétés qui travaillent dans le domaine précis A96 et qui ont une stratégie d'extension aux Etats Unis. Ces deux renseignements n'ont

pas forcément de relation directe puisque les brevets que la société étend aux Etats Unis ne sont pas obligatoirement sur un sujet couvert par le code A96. Donc l'emploi de la logique booléenne dans ce cas fournira directement le renseignement recherché. Une recherche combinant le code et le pays d'extension par l'opérateur "ET" est trop restrictive puisqu'elle impose la présence des deux éléments dans les mêmes brevets. L'emploi de l'opérateur "OU" nécessiterait lui une recherche manuelle des sociétés.

La méthode de chaînage de paires est développée dans DATAVIEW comme une itération d'éditions de paires constituées à partir des sélections de formes.

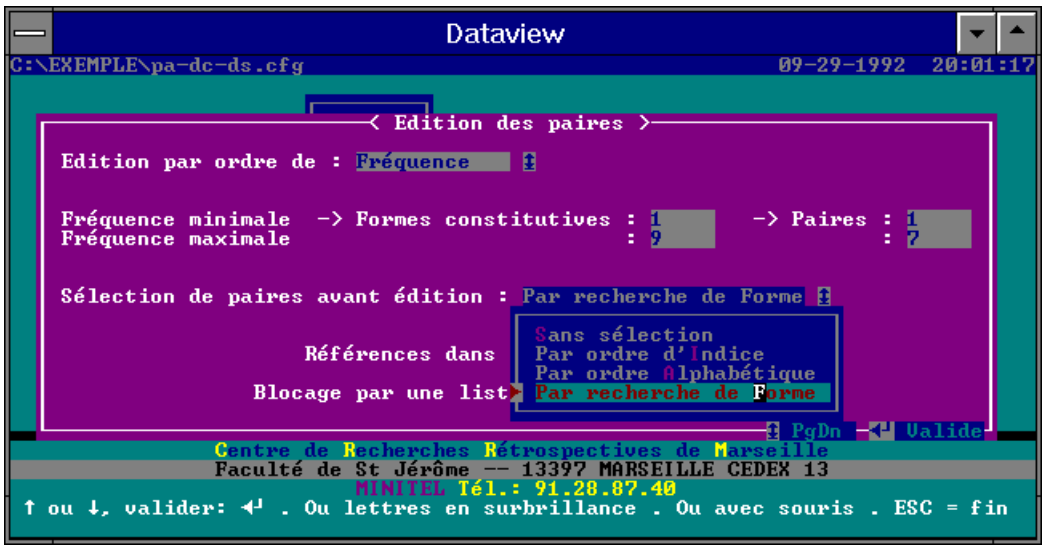


Figure 51: Recherche par chaînage de paires

Pour notre problème exposé plus haut, on choisit dans un premier temps de rechercher toutes les paires formées avec le code A96. DATAVIEW propose la liste des paires comportant cette forme dans une fenêtre de type ascenseur (figure 52a). Les nouvelles formes liées au code A96 sont réparties dans la première colonne. Dans cette liste de paires on sélectionne celles qui comportent les acronymes des sociétés. DATAVIEW recherche alors automatiquement toutes les nouvelles paires comportant les formes de la première colonne donc les paires avec les acronymes des sociétés. Le résultat de cette recherche est de nouveau présenté dans une fenêtre (figure 52b). Un seul acronyme forme une paire avec la forme US donc seule dans ces dix références la société ROBE/ a à la fois une activité d'extension aux Etats Unis et une activité de recherche dans le domaine A96. En remontant aux références brevets (ANNEXE 2) on peut s'apercevoir que ces deux formes ("A96" et "US") ne sont pas dans la même référence donc une recherche booléenne n'aurait pas retrouvé cette société.

Paire		Fréquence
A14	\ A96	1
B01	\ A96	1
B07	\ A96	4
D22	\ A96	3
G03	\ A96	1
P32	\ A96	2
P34	\ A96	2
(KIMY/)	\ A96	1
(LOHM)	\ A96	1
(ROBE/)	\ A96	1
(SEKI)	\ A96	1
AT	\ A96	2
BE	\ A96	2
CA	\ A96	1
CH	\ A96	2

[↵ Valide]
Sélection: 4

Figure 52a

Paire		Fréquence
GB	\(LOHM)	1
GB	\(ROBE/)	2
GR	\(LOHM)	1
GR	\(ROBE/)	2
IT	\(LOHM)	1
IT	\(ROBE/)	2
JP	\(ROBE/)	2
LI	\(LOHM)	1
LU	\(LOHM)	1
LU	\(ROBE/)	1
NL	\(LOHM)	1
NL	\(ROBE/)	1
SE	\(LOHM)	1
SE	\(ROBE/)	1
US	\(ROBE/)	1

[↵ Valide]
Sélection: 1

Figure 52b

Cette pratique de recherche non booléenne n'est pas une nouveauté puisque Price et Beaver utilisèrent la même logique pour constituer leurs collèges invisibles [PRIC66]. Pour créer des groupes de collaboration autour d'un auteur, ils commençaient par rechercher tous les auteurs qui avaient travaillé avec lui, puis cherchaient à connaître tous les auteurs qui avaient publié avec ces derniers et ainsi de suite.
Par contre ce qui est nouveau, c'est le fait d'avoir la possibilité de réaliser ces recherches en semi-automatique sur n'importe quel type d'éléments bibliographiques, ceci grâce à un outil informatique. Ce principe de recherche non booléenne est déjà à l'origine de travaux au CRRM [DOU90b].

e) Les tableaux

Dès lors que les listes de formes ou de paires deviennent trop conséquentes, le besoin d'une présentation plus condensée se fait ressentir. **La disposition des éléments selon une structure en tableau est l'un des moyens pour condenser l'information. Lorsque le tableau devient lui-même trop important pour que sa lecture reste aisée, alors on aura recours à des méthodes statistiques pour réorganiser l'information qu'il contient et faciliter son interprétation**. Ces méthodes seront présentées dans le paragraphe suivant *Les traitements infographiques en sortie de DATAVIEW.*

Price est l'un des premiers auteurs qui ait montré l'utilité d'exploiter les données présentées sous la forme de tableaux dans [PRIC81a] et [PRIC81b]. Depuis, l'utilisation de ce mode de présentation est devenue le principal moyen pour exposer les données bibliométriques. **DATAVIEW est tout particulièrement développé dans cet axe**. Il permet la construction de toutes les catégories de tableaux qu'ils soient l'ultime étape de l'étude ou bien un simple intermédiaire pour une analyse statistique plus complexe (dans ce cas on a l'habitude de nommer le tableau sous l'appellation de *matrice*). Nous avons classé les tableaux créés par DATAVIEW en 3 catégories:

(1) Tableaux des fréquences de paires

Un tableau des fréquences n'est que la forme tabulée d'une liste de fréquences de paires. **Selon le choix des formes distribuées en lignes et en colonnes l'appellation du tableau peut changer** mais leur construction sous DATAVIEW se réalise selon la même manipulation.

Une fenêtre de paramétrage demande à l'utilisateur d'indiquer ses options pour la construction de la matrice des fréquences de paires (figure 53).

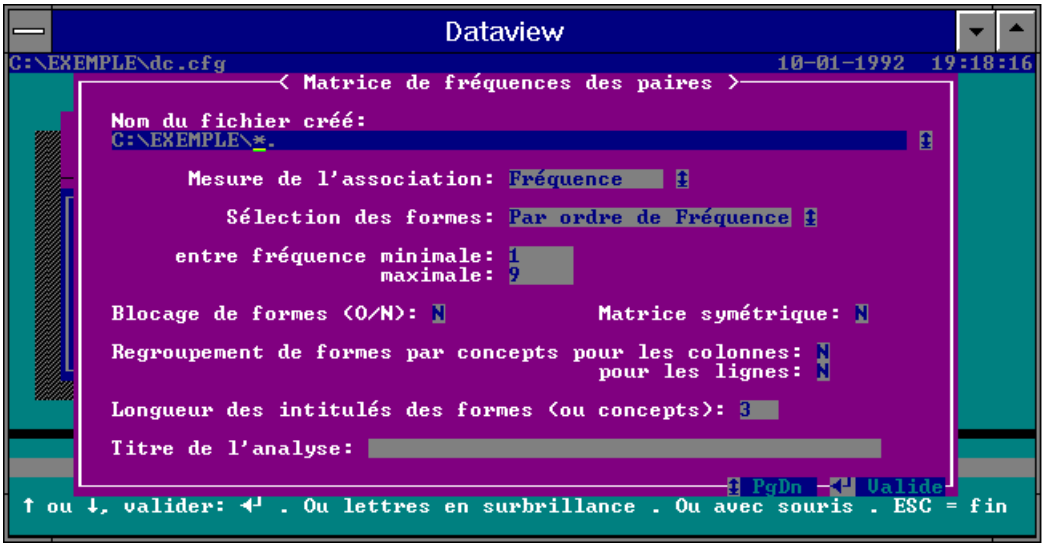


Figure 53: Options de construction des matrices de fréquences

Après validation des options de cette fenêtre de dialogue, DATAVIEW propose dans une fenêtre de sélection la liste des formes répertoriées dans la base de travail selon la sélection de l'intervalle des fréquences et selon l'option de tri des formes choisie (même présentation que celle de la figure 48). Cette liste est proposée deux fois. Une première fois pour sélectionner l'ensemble des formes à placer en colonne et la seconde fois pour sélectionner celles à placer en ligne.

Le tableau établi est conservé dans un fichier dont la structure ne permet pas la lecture de son contenu. Cette structure de fichier est celle employée par le logiciel statistique STAT-ITCF [STATIT]. Ce format de sortie a été choisi pour deux raisons principales:

- ☞ La gestion de ces matrices est facilitée par la création de fichiers différents pour les intitulés des lignes et des colonnes et pour les données de la matrice. Les intitulés étant de type texte les fichiers sont à accès séquentiel tandis que les données numériques de la matrice sont dans des fichiers à accès direct.
- ☞ Le laboratoire utilise ce logiciel statistique pour ses analyses bibliométriques

L'utilisateur qui veut consulter le tableau alors qu'il ne possède pas le logiciel STAT-ITCF pourra le faire en utilisant une option d'exportation de matrice proposée par DATAVIEW. Actuellement trois formats d'exportation sont offerts à l'utilisateur:

- ☞ exportation vers le tableur Excel de Microsoft (format texte avec séparation des colonnes par un caractère de tabulation)
- ☞ exportation vers le logiciel de statistique Clustan (format texte avec séparation des colonnes par des espaces). Ce format est aussi accepté par des tableurs comme Lotus123 et Multiplan...
- ☞ exportation vers le logiciel d'Analyse Relationnelle Arcade (format personnalisé du CESMAP d'IBM)

Les options de paramétrage de cette fenêtre (figure 53) seront explicitées au cours des présentations d'exemples dans les lignes qui vont suivre.

⇨ Tableau symétrique:

Un tableau est nommé symétrique si les mêmes formes sont distribuées en ligne comme en colonne. **Ce type de tableaux est construit pour représenter les relations existant entre les différentes formes de même nature**. Price dans [PRIC81b] les nomme ***matrices de transaction*** car elles symbolisent pour lui l'ensemble des transactions existant entre tous les membres d'un même groupe. Ces matrices sont très courantes en bibliométrie spécialement dans les analyses par cartographie (co-citation, co-word, co-signature, coopération).
Une option spécifique à ce type de matrice (figure 53) n'offrira à l'utilisateur qu'une fois la liste de sélection des formes, l'ensemble choisi étant automatiquement reporté pour la deuxième entrée du tableau.

Prenons comme premier exemple, la base de travail "DC.JOB". La construction de la matrice présentant la totalité des relations entre les différents Codes Derwent est donc de cette forme:

	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05	
a14	1	1				1			1	1	1	1		**a14**
a96	1	4	1			4			3	1	2	2		**a96**
b01		1	2			2			2		2			**b01**
b02				1										**b02**
b05					1	1		1			1			**b05**
b07	1	4	2		1	9	1	1	5	1	4	4	1	**b07**
c07						1	1		1			1		**c07**
d21					1	1		1			1			**d21**
d22	1	3	2			5	1		5	1	3	3		**d22**
g03	1	1				1			1	1	1	1		**g03**
p32	1	2	2		1	4		1	3	1	4	1		**p32**
p34	1	2				4	1		3	1	1	4	1	**p34**
s05						1						1	1	**s05**
	A14	**A96**	**B01**	**B02**	**B05**	**B07**	**C07**	**D21**	**D22**	**G03**	**P32**	**P34**	**S05**	

Tableau 1: Matrice de transaction

Remarque: les valeurs nulles ont été éliminées pour une plus grande clarté.

Ce tableau condense en une seule représentation les fréquences des formes et les fréquences des paires:

- la diagonale présente les valeurs des fréquences de formes
- hors de la diagonale sont disposées les valeurs des fréquences de paires.

❑ Les logiciels statistiques qui exploitent ces matrices acceptent des intitulés de colonnes et des lignes de taille limitée. Par exemple, STAT-ITCF limite le nom d'une ligne ou d'une colonne à 3 caractères, Clustan à 8 caractères et Arcade à 11 caractères. Les formes bibliométriques sont de longueurs très variables pouvant aller de 2 caractères pour les dates à plus d'une cinquantaine pour les noms de revues scientifiques. Couper les formes à une longueur fixe peut créer des ambiguïtés pour les intitulés de lignes ou de colonnes. Ces ambiguïtés se répercuteraient sur les représentations graphiques rendant l'interprétation laborieuse puisque des éléments graphiques portent des noms identiques.
Pour éviter cela, une option de paramétrage (figure 53) permet à l'utilisateur de préciser la ***longueur des intitulés*** à affecter aux lignes et aux colonnes. **Lors de la création de la matrice DATAVIEW reconnaîtra automatiquement les intitulés ambigus qui pourraient être générés et les transformera en remplaçant le dernier caractère par un caractère univoque**. Cette opération est appelée dans DATAVIEW le ***transcodage***. De plus, pour faciliter encore la tâche la phase de transcodage crée un fichier où une liste fait correspondre les intitulés des lignes et des colonnes de la matrice construite avec la forme réelle présente dans la base de travail.

Nous prendrons comme exemple un nouveau fichier de travail où seul le champ TT (Titre normalisé de Derwent) est extrait. Ce champ étant retraité par Derwent un séparateur unique délimite chaque mot, l'espace. En le choisissant comme séparateur de formes, nous pouvons construire une matrice de fréquences de paires qui décrit le réseau de relation qui relie tous ces mots. Cette matrice est assimilable à celle de l'analyse des co-word. Pour des raisons de format (matrice 100*x*100) nous n'avons conservé que la partie correspondant aux formes commençant par la lettre "C" car elle présente des traitements de transcodage. Si cette matrice doit être traitée ensuite par un logiciel limitant la longueur des intitulés à 3 caractères la matrice construite serait:

	CAL	CAN	CAR	CAV	CHA	CHO	CNS	COA	COM	COα	CON	COβ	COτ	COπ	COΣ	COσ	
cal	1	1			1		1				1						**cal**
can	1	1			1		1				1						**can**
car			1														**car**
cav				1					1					1			**cav**
cha	1	1			1		1				1						**cha**
cho						1											**cho**
cns	1	1			1		1				1						**cns**
coa								1				1					**coa**
com				1					1					1			**com**
coα										1			1				**coα**
con	1	1			1		1				1						**con**
coβ								1				1					**coβ**
coτ										1			2				**coτ**
coπ				1					1					1			**coπ**
coΣ															1		**coΣ**
coσ																1	**coσ**
	CAL	**CAN**	**CAR**	**CAV**	**CHA**	**CHO**	**CNS**	**COA**	**COM**	**COα**	**CON**	**COβ**	**COτ**	**COπ**	**COΣ**	**COσ**	

Tableau 2: matrice de transaction avec des formes transcodées

Le fichier de transcodage associé à cette matrice donne le sens des intitulés selon cette présentation:

```
                              TABLE DE TRANSCODAGE
                              ~~~~~~~~~~~~~~~~~~~~
     Nom du fichier STAT-ITCF   : C:\EXEMPLE\TT-TT.
     Titre de l'analyse         : matrice symétrique TTxTT
     Mesure d'intensité de lien : Fréquence

Variables-Colonnes:
~~~~~~~~~~~~~~~~~~~
  1°) CALCIUM        -> CAL
  2°) CANCER         -> CAN
  3°) CARRY          -> CAR
  4°) CAVITY         -> CAV
  5°) CHANNEL        -> CHA
  6°) CHOLINERGIC    -> CHO
  7°) CNS            -> CNS
  8°) COATING        -> COA
  9°) COMPOSITION    -> COM
 10°) COMPRISE       -> COα
 11°) CONGESTED      -> CON
 12°) CONSIST        -> COβ
 13°) CONTAIN        -> COτ
 14°) CONTOUR        -> COπ
 15°) CONTRACEPTIVE  -> COΣ
 16°) CONTROL        -> COσ
```

Table 22: Table de transcodage

⇨ Tableaux de contingence:

Un tableau est nommé tableau de contingence si en ligne et en colonne on ne distribue plus le même ensemble de formes mais au contraire deux ensembles dont les informations sous-tendues ne sont pas de même portée significative. Les statisticiens définissent le tableau de contingence comme étant le croisement de deux variables définissant chacune une partition sur la population étudiée.

Par exemple, pour la base de travail "PA-DC.JOB" les deux ensembles sont constitués des formes des deux champs: les sociétés déposantes en ligne et les codes de la classification Derwent en colonne.

	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05	
(derm-)						1	1		1			1		(derm-)
(kimy/)		1				1								(kimy/)
(lohm)	1	1				1			1	1	1	1		(lohm)
(medt)						1						1	1	(medt)
(phar-)					1	1		1			1			(phar-)
(robe/)		1	2			2			2		2			(robe/)
(seki)		1				1			1			1		(seki)
(smik)						1								(smik)
(squi)				1										(squi)
	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05	

Tableau 3: tableau de contingence

Nous sommes bien en présence d'un tableau de contingence puisque les deux ensembles de formes définissent chacune une partition sur la population étudiée, c'est à dire sur l'ensemble des références.

❑ Nous allons présenter maintenant une matrice de contingence construite en utilisant l'option ***Regroupement de formes par concept*** (voir figure 43). **Cette option va permettre de regrouper dans une même colonne plusieurs formes qui peuvent être considérées comme représentantes d'un concept donné**. Dans ce cas, en conservant comme principe que la fréquence d'un élément bibliographique est le dénombrement de références contenant cet élément, DATAVIEW considérera pour cette colonne toutes les références comportant au moins une de ces formes, c'est-à-dire comportant la présence du concept (que ce soit une ou plusieurs fois). Ce regroupement de formes par concept est aussi bien réalisable sur les lignes que sur les colonnes.

Comme exemple nous avons choisi de construire une matrice de contingence qui établit les liens entre les codes du champ DC et des regroupements de pays d'extension par régions géographiques. Donc en nous servant de la base "PA-DC-DS.JOB", nous avons créé cette matrice:

	A14	A96	B01	B05	B07	C07	D21	D22	G03	P32	P34	S05	
anglo	1	2	2	1	7	1	1	4	1	4	3	1	anglo
eur_s	1	2	2	1	7	1	1	4	1	4	3	1	eur_s
eur_n	1	2	2	1	7	1	1	4	1	4	3	1	eur_n
asie		1	2	1	5		1	2		3	1	1	asie
	A14	A96	B01	B05	B07	C07	D21	D22	G03	P32	P34	S05	

Tableau 4: tableau de contingence avec regroupement par concept en ligne

Le fichier de transcodage associé à la matrice donne toujours accès à l'utilisateur au sens réel des intitulés des lignes et des colonnes. Dans le cas d'un regroupement par concept il fournit la liste des formes réunies sous une même colonne ou une même ligne (table 23).

```
                              TABLE DE TRANSCODAGE
                              ~~~~~~~~~~~~~~~~~~~~
 Nom du fichier STAT-ITCF: C:\EXEMPLE\GEO.
 Titre de l'analyse      : Matrice Régions d'extension-Codes DC
 Mesure d'intensité de lien : Fréquence

Variables-Colonnes:
~~~~~~~~~~~~~~~~~~~
  1°) A14      -> A14
  2°) A96      -> A96
  3°) B01      -> B01
  4°) B05      -> B05
  5°) B07      -> B07
  6°) C07      -> C07
  7°) D21      -> D21
  8°) D22      -> D22
  9°) G03      -> G03
 10°) P32      -> P32
 11°) P34      -> P34
 12°) S05      -> S05

Variables-Lignes:
~~~~~~~~~~~~~~~~~
 1°) AU        -> anglo
     CA
     GB
     US
 2°) BE        -> eur_s
     CH
     ES
     FR
     IT
 3°) DK        -> eur_n
     FI
     NL
     NO
     SE
 4°) JP        -> asie
     KR
```

Table 23: Table de transcodage avec regroupement de formes en ligne

⇨ Les tableaux de bord (tableau de Burt):

Ce dernier type de tableaux constitue une récapitulation généralisée de toutes les relations qui existent entre plusieurs ensembles de formes. On ne se limite plus à connaître les liens entre un groupe d'éléments (tableau symétrique) ou entre deux groupes d'éléments (tableau de contingence) mais l'ensemble des liens qu'il soit intra-groupe ou qu'il soit inter-groupe. I**l correspond donc au tableau de Burt rencontré chez les statisticiens.**

Si on reprend l'étude des croisements entre les 3 champs DC, PA et DS, le tableau qui généralise ces relations est représenté par le tableau 5.

	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05
a14	1	1				1			1	1	1	1	
a96	1	4	1			4			3	1	2	2	
b01		1	2			2			2		2		
b02				1									
b05					1	1		1			1		
b07	1	4	2		1	9	1	1	5	1	4	4	1
c07						1	1		1			1	
d21					1	1		1			1		
d22	1	3	2			5	1		5	1	3	3	
g03	1	1				1			1	1	1	1	
p32	1	2	2		1	4		1	3	1	4	1	
p34	1	2				4	1		3	1	1	4	1
s05						1						1	1

	(derm-)	(kimy/)	(lohm)	(medt)	(phar-)	(robe/)	(seki)	(smik)	(squi)
(derm-)									
(kimy/)		1							
(lohm)	1	1							
(medt)				1					
(phar-)					1				
(robe/)		1	2			2			
(seki)		1					1		
(smik)								1	
(squi)									1

	at	au	be	ca	ch	de	dk	es	fi	fr	gb	gr	it	jp	kr	li	lu	nl	no	se	us
at	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
au	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2		2	2	1	2	1
be	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
ca	4	2	4	4	4	4	4	4	1	4	4	4	4	4	2		3	3	1	3	2
ch	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
de	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
dk	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
es	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
fi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	
fr	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
gb	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
gr	6	2	6	4	6	6	6	6	1	6	6	6	6	5	2	1	5	5	1	5	2
it	7	2	7	4	7	7	7	7	1	7	7	6	7	5	2	2	6	6	1	6	2
jp	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2		4	4	1	4	2
kr	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2		2	2	1	2	1
li	2		2		2	2	2	2		2	2	1	2			2	2	2		2	
lu	6	2	6	3	6	6	6	6	1	6	6	5	6	4	2	2	6	6	1	6	2
nl	6	2	6	3	6	6	6	6	1	6	6	5	6	4	2	2	6	6	1	6	2
no	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	
se	6	2	6	3	6	6	6	6	1	6	6	5	6	4	2	2	6	6	1	6	2
us	2	1	2	2	2	2	2	2		2	2	2	2	2	1		2	2		2	2

Tableau 5: Tableau de bord généralisé pour 3 champs

❑ Une dernière option de paramétrage est possible lors de la création d'une matrice: le ***blocage de formes***. Cette notion a déjà été présentée pour les éditions des fréquences de paires. Ici le principe est le même, lorsqu'une forme ou une série de formes est choisie comme forme de blocage, **le tableau créé ne considère que les formes présentes dans les références contenant les formes de blocage**.

Ainsi, pour le cas exposé précédemment, l'utilisateur de DATAVIEW peut vouloir établir son tableau de l'ensemble des relations entre les trois catégories de formes pour les références de brevets ayant été étendues au Japon. Il suffit donc de préciser lors de la construction du tableau de bord que la forme "JP" est choisie comme forme de blocage. Le résultat est le suivant:

	A96	B01	B05	B07	D21	D22	P32	P34	S05	(MEDT	(PHAR-	(ROBE/	(SMIK)	AT	AU	BE	CA	CH	DE	DK	ES	FI	FR	GB	GR	IT	JP	KR	LU	NL	NO	SE	US	
a96	1	1		1		1	1																											
b01	1	2		2		2	2																											
b05			1	1	1		1																											
b07	1	2	1	5	1	2	3	1	1																									
d21			1	1	1		1																											
d22	1	2		2		2	2																											
p32	1	2	1	3	1	2	3																											
p34				1				1	1																									
s05				1				1	1																									
(medt)				1				1	1	1																								
(phar-)			1	1	1		1				1																							
(robe/	1	2		2		2	2					2																						
(smik)				1									1																					
at	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	at
au				2				1	1	1			1	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	1	au
be	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	be
ca	1	2		4		2	2	1	1	1		2	1	4	2	4	4	4	4	4	4	1	4	4	4	4	4	2	3	3	1	3	2	ca
ch	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	ch
de	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	de
dk	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	dk
es	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	es
fi				1				1	1	1				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		fi
fr	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	fr
gb	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	gb
gr	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	gr
it	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	it
jp	1	2	1	5	1	2	3	1	1	1	1	2	1	5	2	5	4	5	5	5	5	1	5	5	5	5	5	2	4	4	1	4	2	jp
kr				2				1	1	1			1	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	1	kr
lu		1	1	4	1	1	2	1	1	1	1	1	1	4	2	4	3	4	4	4	4	1	4	4	4	4	4	2	4	4	1	4	2	lu
nl		1	1	4	1	1	2	1	1	1	1	1	1	4	2	4	3	4	4	4	4	1	4	4	4	4	4	2	4	4	1	4	2	nl
no				1				1	1	1				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		no
se		1	1	4	1	1	2	1	1	1	1	1	1	4	2	4	3	4	4	4	4	1	4	4	4	4	4	2	4	4	1	4	2	se
us		1		2		1	1					1	1	2	1	2	2	2	2	2	2		2	2	2	2	2	1	2	2		2	2	us
	A96	B01	B05	B07	D21	D22	P32	P34	S05	(MEDT	(PHAR-	(ROBE/	(SMIK)	AT	AU	BE	CA	CH	DE	DK	ES	FI	FR	GB	GR	IT	JP	KR	LU	NL	NO	SE	US	

Tableau 6: Tableau de bord généralisé avec blocage de forme

(2) Tableaux binaires

Les tableaux de ce type ne paraissent pas à première vue aussi intéressants que les tableaux de fréquences de paires. **Ils sont en fait la description la plus primitive des données caractéristiques de la population étudiée**. **Tous les autres types de tableaux peuvent être construits à partir de tableaux binaires**.

DATAVIEW a été conçu pour pouvoir construire des tableaux binaires car de nombreuses méthodes d'analyse statistique basent leurs calculs sur ces tableaux. **Pour un groupe défini de formes, ces matrices établissent la présence ou l'absence de chacune d'elles dans l'ensemble des références étudiées**. L'application de ces tableaux binaires au domaine de la bibliométrie va donc répartir l'ensemble des références sur l'une des deux entrées du tableau et l'ensemble des formes sur l'autre. Il a été choisi arbitrairement de donner la valeur 1 si une forme est bien présente dans une référence, sinon la valeur 0 est attribuée.
Cette simple notation de la présence ou de l'absence d'une forme et non son nombre par référence reste en droite ligne de notre conception du dénombrement entrepris jusque là. Par contre, une matrice d'occurrence contiendrait des valeurs autres que binaires. C'est par exemple le cas des matrices initiales dans la méthode des mots associés [MICH88].

DATAVIEW propose des constructions de matrices binaires avec les mêmes genres d'options que pour les tableaux de fréquences (figure 54).

Figure 54: Options de création de matrices binaires

Pour une étude ne concernant que les Codes Derwent pour notre petit échantillon le tableau créé est le suivant:

	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05	
Réf 1	1	1				1			1	1	1	1		**Réf 1**
Réf 2						1	1		1			1		**Réf 2**
Réf 3		1	1			1			1		1			**Réf 3**
Réf 4			1			1			1		1			**Réf 4**
Réf 5				1										**Réf 5**
Réf 6		1				1								**Réf 6**
Réf 7						1								**Réf 7**
Réf 8		1				1			1			1		**Réf 8**
Réf 9						1						1	1	**Réf 9**
Réf 10					1	1		1			1			**Réf 10**
	A14	**A96**	**B01**	**B02**	**B05**	**B07**	**C07**	**D21**	**D22**	**G03**	**P32**	**P34**	**S05**	

Tableau 7: Matrice binaire de présence-absence

Comme on peut le voir, si toutes les formes sont sélectionnées le tableau binaire est la description exacte du contenu de la base de travail, rien n'y est omis et rien n'y est ajouté. C'est pour cela que ces tableaux prennent une grande part dans les méthodes d'analyse statistique.

Un tableau de fréquences de paires est déjà une condensation d'information puisque les individus de la population (les références) sont rassemblés selon des croisements. Un tableau de fréquence perd la trace de chaque forme.

La matrice de présence-absence a l'avantage de permettre non seulement l'étude des relations entre formes mais aussi l'étude des relations entre références.

⇨ Tableaux disjonctifs complets:

Pour étudier des groupes formes qui ne sont pas de même nature (portées significatives différentes), certaines méthodes statistiques permettent d'utiliser des tableaux binaires uniquement si ceux-ci vérifient une propriété: **les éléments dans les groupes de formes sont mutuellement exclusifs**. **Ce qui signifie qu'on ne peut trouver par référence qu'une seule forme d'un groupe à la fois**. Les tableaux vérifiant cette caractéristique sont nommés ***disjonctifs complets***.

Comme cette propriété est très rare pour les données bibliométriques, nous proposons une astuce pour passer outre cette contrainte. L'**ajout de références "virtuelles"** permet d'éclater les formes d'un même groupe, présentes simultanément dans une même référence, dans des références séparées. Il y a introduction de plusieurs entrées par référence mais les fréquences des formes restent les mêmes (sommes colonnes constantes). Cette méthode met en jeu une très forte combinatoire. Un module a été développé pour mener à bien cette réalisation. Il n'a pas encore été fusionné aux autres modules de DATAVIEW mais il offre déjà cette construction de matrices disjonctives complètes de manière automatisée.

Les techniques statistiques utilisant ce type de matrice sont les seules qui permettent d'étudier conjointement plus de deux groupes de formes de catégories différentes. Ces techniques n'ont pratiquement pas été employées en bibliométrie, par contre elles sont courantes pour les analyses d'enquêtes (AFCM).

Pour notre exemple, nous reprendrons le cas de l'étude des interdépendances entre les trois champs PA, DC et DS. Le tableau présenté ci-dessous considère uniquement les deux premières variables pour des raisons de mise en page. Mais nous pouvons facilement imaginer que le nombre de références "virtuelles" seraient démultipliées en ajoutant le groupe des formes du champ DS:

	CHAMP PA									CHAMP DC												
	DERM-	KIMY/	LOHM	MEDT	PHAR-	ROBE/	SEKI	SMIK	SQUI	A14	A96	B01	B02	B05	B07	C07	D21	D22	G03	P32	P34	S05
			1							1												
			1								1											
			1												1							
Réf 1			1															1				
			1																1			
			1																	1		
			1																		1	
	1														1							
Réf 2	1															1						
	1																	1				
	1																				1	
						1					1											
						1						1										
Réf 3						1									1							
						1												1				
						1														1		
						1						1										
Réf 4						1									1							
						1												1				
						1														1		
Réf 5									1				1									
Réf 6		1									1											
		1													1							
Réf 7								1							1							
							1				1											
Réf 8							1								1							
							1											1				
							1														1	
				1											1							
Réf 9				1																	1	
				1																		1
					1									1								
Réf 10					1										1							
					1												1					
					1															1		

Tableau 7: Tableau binaire disjonctif complet

Ce tableau est la forme primitive du tableau de bord présenté précédemment.

(3) Tableaux d'indice d'association

Un dernier type de tableau peut être construit grâce à DATAVIEW: **les tableaux d'association**.

Le tableau de contingence donne déjà une idée de ce qu'on entend par un indice d'association. **Le but de l'emploi d'un tel indice est de mesurer une intensité de lien entre deux formes**. Dans le tableau de contingence la mesure mise en exergue est simplement un "indice" de fréquence de paires. Ceci permet d'estimer l'importance du lien entre ces deux formes. Mais nous avons vu que cette fréquence est une mesure qui ne prend pas en considération le poids statistique de chacune des deux formes constituant une paire (Cf *Solution à la diversité des traitements ultérieurs*).

Les statisticiens ont donc mis au point une série d'indices statistiques pour donner plus ou moins de poids aux différentes fréquences (fréquence des deux formes, fréquence de la paire et fréquence d'absence des deux formes). Une grande partie de ces mesures est calculée par DATAVIEW (liste en ANNEXE 2). Nous avons vu que l'utilisateur peut choisir une de ces mesures comme critère de classement des paires dans les éditions de paires (Cf *Les listes des paires*). De la même façon il peut en sélectionner une lors d'une construction de matrice de paires. Il peut accéder à la liste des indices statistiques par l'ouverture de la fenêtre associée à l'option *Mesure d'association* (figure 43). Les manipulations sont ensuite rigoureusement les mêmes.

La matrice symétrique des Codes Derwent, présentée plus haut, peut être reformulée grâce à la mesure d'association de l'indice de corrélation. Le tableau obtenu est le suivant:

	A14	**A96**	**B01**	**B02**	**B05**	**B07**	**C07**	**D21**	**D22**	**G03**	**P32**	**P34**	**S05**	
a14	1	0.40	-0.16	-0.11	-0.11	0.11	-0.11	-0.11	0.333	1	0.40	0.40	-0.11	**a14**
a96	0.40	1	0.10	-0.27	-0.27	0.27	-0.27	-0.27	0.40	0.40	0.16	0.16	-0.27	**a96**
b01	-0.16	0.10	1	-0.16	-0.16	0.16	-0.16	-0.16	0.5	-0.16	0.61	-0.40	-0.16	**b01**
b02	-0.11	-0.27	-0.16	1	-0.11	-1	-0.11	-0.11	-0.33	-0.11	-0.27	-0.27	-0.11	**b02**
b05	-0.11	-0.27	-0.16	-0.11	1	0.11	-0.11	1	-0.33	-0.11	0.40	-0.27	-0.11	**b05**
b07	0.11	0.27	0.16	-1	0.11	1	0.11	0.11	0.33	0.11	0.27	0.27	0.11	**b07**
c07	-0.11	-0.27	-0.16	-0.11	-0.11	0.11	1	-0.11	0.33	-0.11	-0.27	0.40	-0.11	**c07**
d21	-0.11	-0.27	-0.16	-0.11	1	0.11	-0.11	1	-0.33	-0.11	0.40	-0.27	-0.11	**d21**
d22	0.33	0.40	0.5	-0.33	-0.33	0.33	0.33	-0.33	1	0.33	0.40	0.40	-0.33	**d22**
g03	1	0.40	-0.16	-0.11	-0.11	0.11	-0.11	-0.11	0.33	1	0.40	0.40	-0.11	**g03**
p32	0.40	0.16	0.61	-0.27	0.40	0.27	-0.27	0.40	0.40	0.40	1	-0.25	-0.27	**p32**
p34	0.40	0.16	-0.40	-0.27	-0.27	0.27	0.40	-0.27	0.40	0.40	-0.25	1	0.40	**p34**
s05	-0.11	-0.27	-0.16	-0.11	-0.11	0.11	-0.11	-0.11	-0.33	-0.11	-0.27	0.40	1	**s05**
	A14	**A96**	**B01**	**B02**	**B05**	**B07**	**C07**	**D21**	**D22**	**G03**	**P32**	**P34**	**S05**	

Tableau 8: Tableau d'indice d'association

L'indice de corrélation est une mesure de similitude qui varie de -1 à 1. Le terme "similitude" indique que la valeur de l'indice augmente lorsque les relations entre les deux formes sont plus intenses. Ainsi, une forme ne pouvant pas entretenir des relations plus fortes avec d'autres formes qu'elle-même, la mesure de corrélation dans la diagonale est maximale et égale à 1.

La paire G03=A14 a elle aussi une mesure maximale alors que la fréquence de cette paire n'était que de 1. Nous nous trouvons devant le phénomène exposé dans le paragraphe *Les listes des paires.* Les deux formes n'apparaissent qu'une seule fois dans les références et de plus ensemble. La corrélation mesure ceci comme étant un état de relation maximale. Ces deux formes ont aussi de fortes valeurs avec les autres formes de la référence où elles sont. Ces valeurs sont plus faibles dès lors que ces autres formes sont présentes dans d'autres références (P32, P34, D22, B07, A96).
Par contre, les deux formes de la paire D22=B07, qui a la plus forte fréquence (= 5), ont leur relation revue à la baisse par l'indice de corrélation. Ces deux formes étant présentes dans de nombreuses références, elles entretiennent beaucoup d'autres relations et par conséquents la paire D22=B07 a une importance minimisée.
On peut aussi remarquer que pour les deux formes G03 et A14, les mesures entre elles et les formes qui ne sont pas présentes dans leur référence ne sont pas égales à -1 alors qu'elles ne sont jamais liées. Ces paires ont des valeurs proches de 0 du côté des valeurs négatives. Une valeur nulle indique que les deux formes ne sont pas en relation et plus la valeur s'approche de -1 plus les deux formes ont des comportements qui s'opposent. C'est le cas des formes B02 et B07. B07 est un code présent dans toutes les références sauf une seule, justement celle où B02 est présent pour son unique fois. La corrélation mesure cette relation comme étant totalement anti-liée.

Par cette petite démonstration, on peut voir tout l'intérêt qu'il faut porter à ces indices statistiques. **Mais pour bien maîtriser la logique qu'elles mettent en oeuvre, une bonne interprétation de la formule et une solide expérience empirique sont recommandées.**

Les logiciels statistiques du marché proposent aussi de tels indices d'association. Pour pouvoir calculer ces indices par ces logiciels, on doit y injecter la matrice binaire des éléments à mesurer. Or la taille de ces matrices est bien souvent limitée. L'utilisateur est donc dans l'obligation de fournir une matrice après lui avoir éliminé des éléments inférieurs à un seuil. Si bien que les mesures calculées à partir de cette nouvelle matrice ne sont plus exactement les mêmes. **DATAVIEW présente donc l'avantage d'assurer les mesures de ces indices à partir de la totalité des données de la base de travail**.

E. Les traitements infographiques en sortie de DATAVIEW

Dans une étude bibliométrique, avant toute chose, il faut évidemment connaître les questions que l'on cherche à résoudre. Dans un processus de veille technologique, elles auront été définies par les décideurs sous forme de Facteurs Critiques.

Pour les résoudre, il faut poser le problème par des questions de plus en plus précises. Cette dichotomie du problème doit permettre d'établir une liste de renseignements qu'il serait indispensable de maîtriser. Ces besoins en renseignements vont induire l'établissement d'ensembles d'éléments qu'il va falloir comparer par des systèmes de mesures.

Les systèmes de mesures les plus simples se contentent de simples classements des éléments suivant la mesure d'une seule réponse. Ce sont des indicateurs univariés. Les représentations graphiques vont donner une meilleure image de la position de chaque élément dans les classements. Elles présentent une vision globale et continue des données alors que les chiffres en eux-mêmes donnent des informations fragmentées et discontinues.

Mais ceci n'est pas suffisant. La connaissance des relations qu'entretiennent les éléments entre eux est bien plus fertile en renseignements. Ce qu'il est nécessaire de voir, ce sont les relations que l'ensemble des données construit. L'information utile à la décision est faite des relations d'ensembles. Il faut alors faire acte d'abstraction pour imaginer quels éléments mis en relation par des valeurs chiffrées pourraient représenter le phénomène à maîtriser. La disposition de ces relations sous la forme de tableaux est la présentation la plus commune à l'individu. La construction de tableaux va donc permettre d'expliciter clairement le phénomène à cerner.

Bien souvent, **ces tableaux sont trop complexes pour être interprétés tels quels**. En bibliométrie, non seulement leurs tailles est bien souvent démesurée mais la variété des profils des données pour chaque élément ne permet pas de retrouver les principales tendances des relations.

Pour faciliter leur synthèse on a alors recours à des représentations graphiques. Ces représentations ne veulent surtout pas se substituer aux données. Elles présentent le phénomène de manière globale tout en cherchant à faire apparaître les similitudes de relations entre les éléments. **Les méthodes graphiques regroupent donc les éléments des tableaux selon leurs ressemblances sans chercher à expliquer pourquoi**. Les raisons de ces rapprochements ne peuvent s'expliquer qu'en remontant aux données brutes des tableaux. **Il faut donc prendre ces représentations comme des aides à l'interprétation des tableaux**.

En bibliométrie, le retour aux données initiales s'arrête à ce stade (retour à la lecture de la matrice) tandis qu'en veille technologique un second pas en arrière dans les données brutes peut être indispensable: **remonter aux références et mêmes aux documents primaires est souvent indispensable** (figure 55). Cette étape se fera ressentir soit pour des questions de validations de certains éléments dont l'état paraît non justifié aux yeux des experts, soit pour recueillir des informations plus précises sur des documents source d'un certain étonnement chez ces mêmes experts.

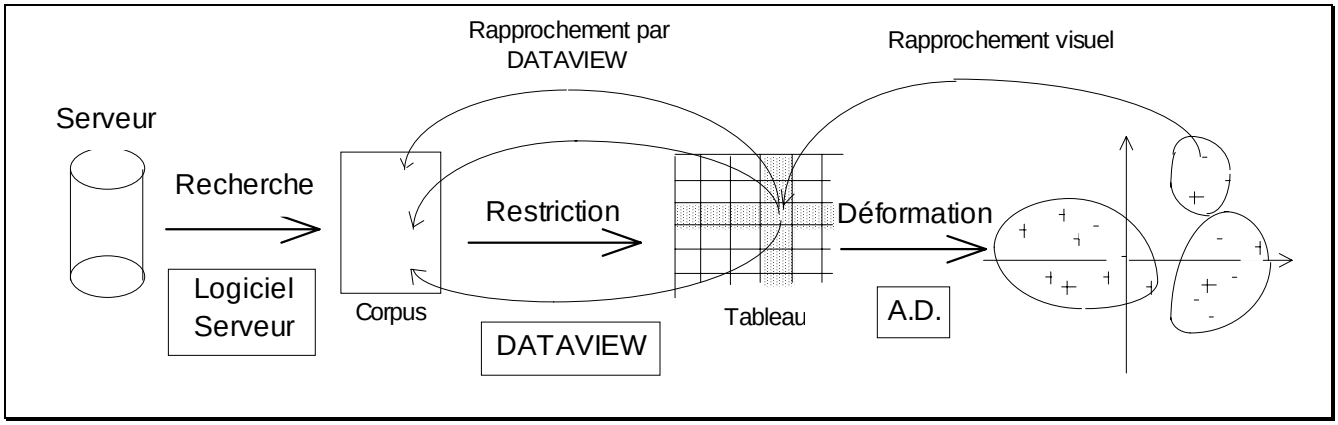


Figure 55: Retour aux données initiales pertinentes ou étonnantes

Il faut donc considérer ces méthodes de représentation graphique comme des "grilles de lecture" de documents primaires.

La qualité de l'analyse dépend du choix de la représentation. **Chaque méthode de représentation a son originalité et ne montre qu'une partie de l'ensemble des aspects des données**. Nous exposons les principales méthodes qui s'appliquent sur les données en sortie de DATAVIEW.

1. Traitement des distributions:

a) Les lois bibliométriques

Nous n'aborderons pas ici les études ciblées sur la modélisation des distributions bibliométriques quoiqu'elles puissent être aisément réalisées à partir des résultats fournis par les fichiers de DATAVIEW. **Nous nous attacherons uniquement à présenter leurs faisabilités et leurs validités apparentes pour les données bibliométriques**.

Toutes les données exploitées dans ces exemples sont tirées du fichier ("*.STA") des statistiques des fréquences de la base (Cf *Détermination des caractéristiques bibliométriques: le codage*). Les traitements graphiques sont réalisés grâce au logiciel *Excel*, tableur développé par la société *Microsoft* sous environnement *Windows* et comportant de nombreuses fonctions graphiques.

⇨ Loi de Lotka:

La construction de la représentation logarithmique de la distribution des contributions des auteurs à leur discipline est immédiate dès lors que la base de travail a été constituée et traitée par DATAVIEW.

La répartition par rang de fréquence de forme fournit DATAVIEW est précisément celle que Lotka a utilisée pour son étude.

	--- Répartitions des Formes ---	
Valeur du X	Nb Formes à Freq X	Nb Formes à Occ X
12	1	1
10	2	2
8	1	1
7	4	4
6	5	5
5	6	6
4	9	9
3	29	29
2	117	117
1	827	827

Table 24: Partie du fichier des statistiques des fréquences

Cette distribution est celle des auteurs d'un ensemble de 507 références d'articles abordant le sujet des marqueurs génétiques. On peut voir sur les figures 56, 57 et 58 que la loi de Lotka s'ajuste très bien à ces données. La qualité de la régression paraît satisfaisante.

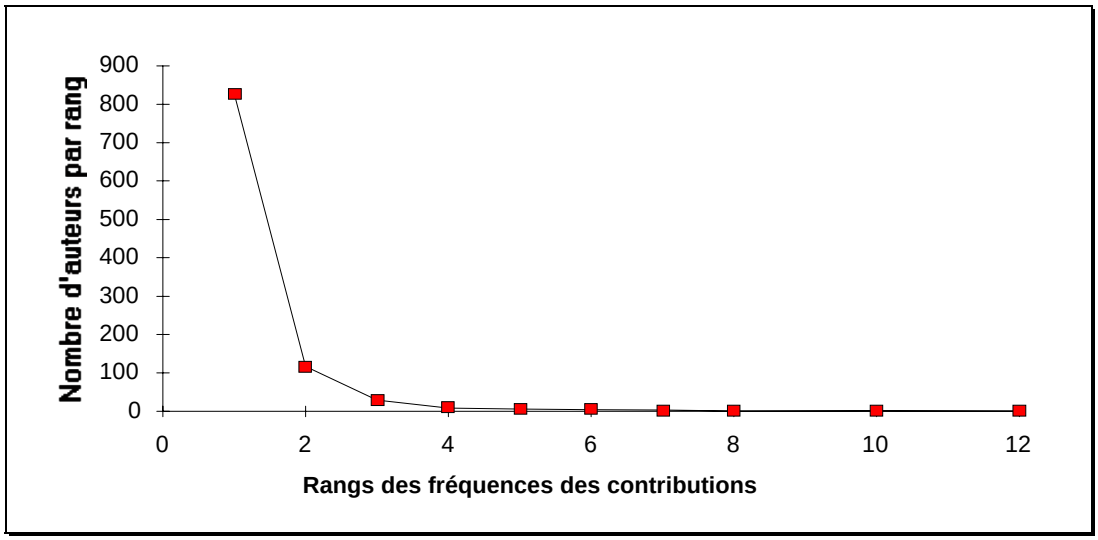


Figure 56: Répartition à partir des données brutes

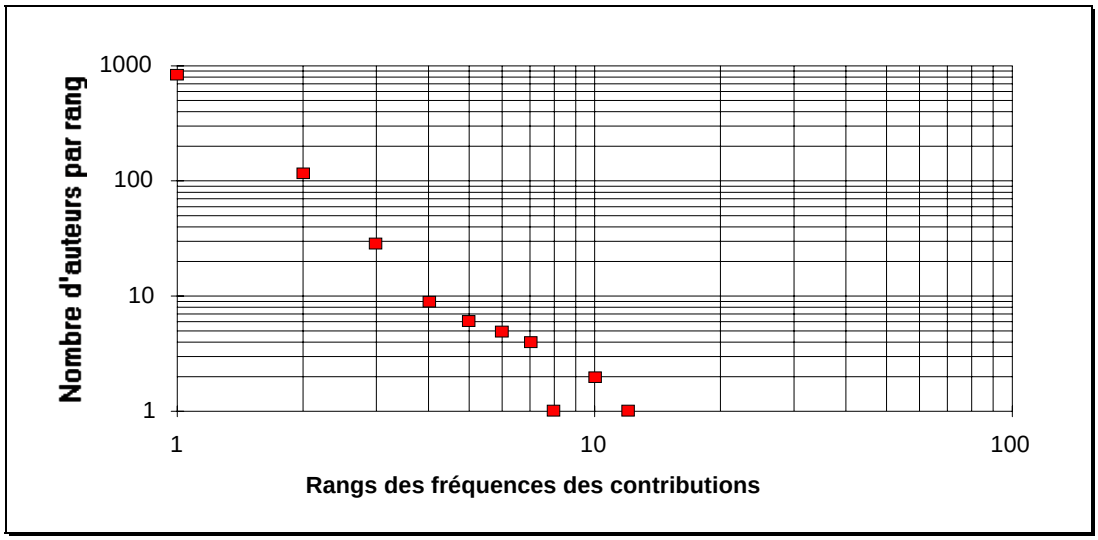


Figure 57: Présentation logarithmique de la répartition

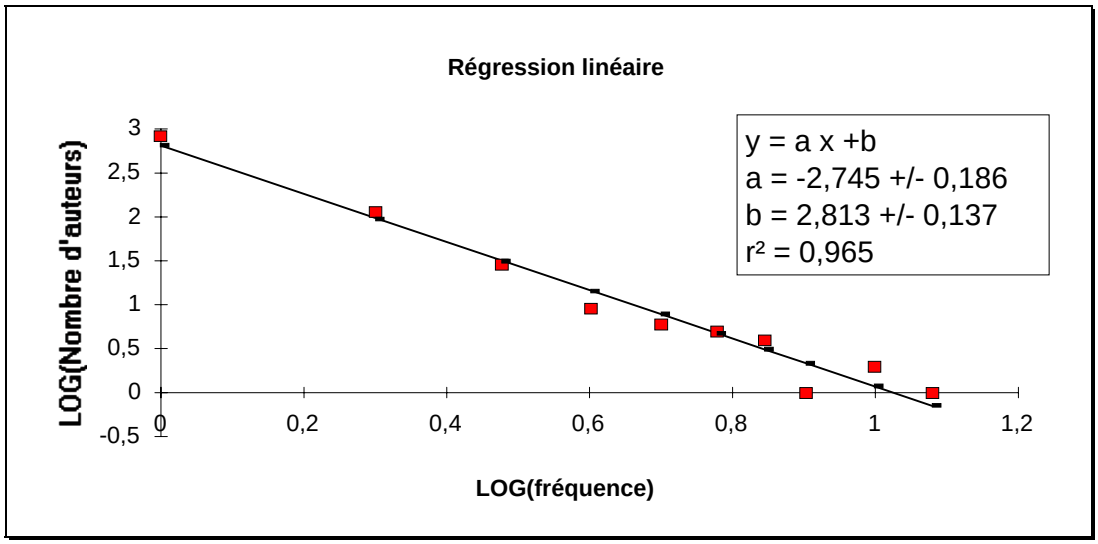


Figure 58: Régression linéaire de la représentation logarithmique

Cette loi initialement pensée pour les auteurs de recherche fondamentale s'ajuste parfaitement aux distributions d'inventeurs de brevets. Les deux exemples suivants démontrent bien cette affirmation:

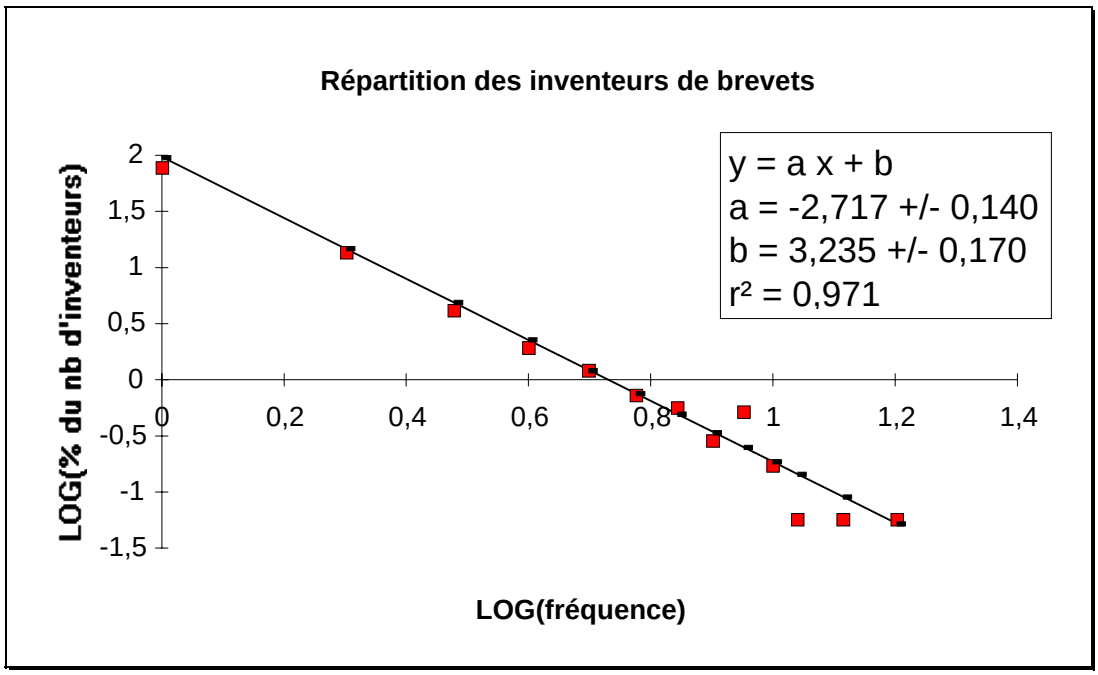


Figure 59: Inventeur de brevets dans un domaine de pharmacologie

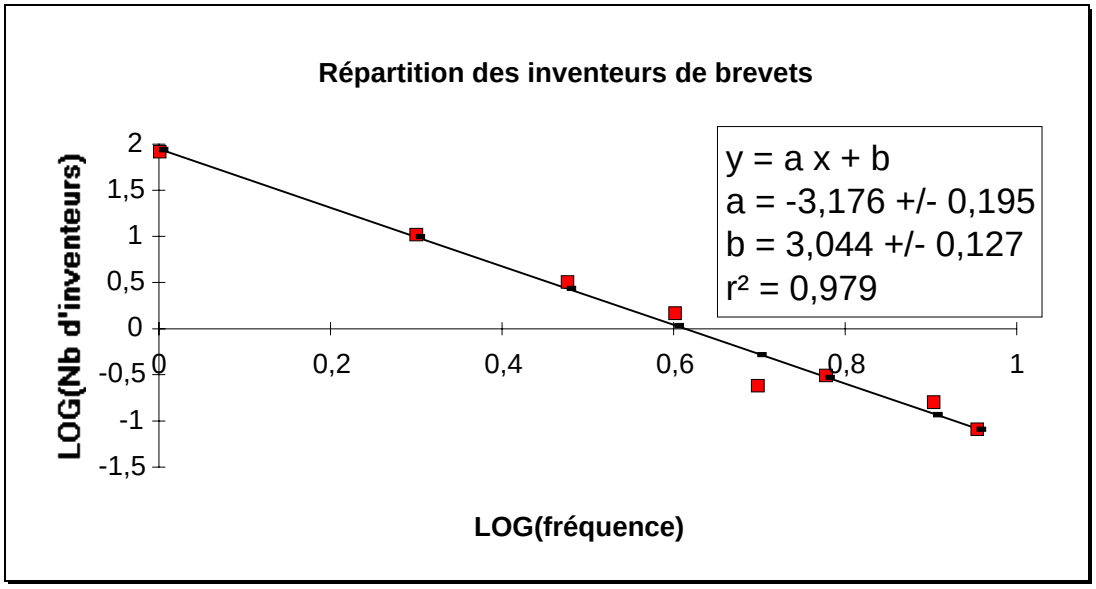


Figure 60: Inventeurs de brevets dans un domaine de l'électroménager

Ces deux corpus de références dans des domaines aussi différents ont des données qui semblent suivre la loi de Lotka en ce qui concerne leurs aspects. Par contre, la pente de ces droites varie.

On peut aussi remarquer que la forme exponentielle s'approcherait plus d'une valeur cubique que d'une valeur carrée. En cela, ces exemples confirment la remarque de Price [PRIC63].

⇨ Loi de Bradford:

La représentation des distributions selon la formulation de Bradford n'est pas plus complexe à mettre en place que celle de Lotka. La création de 3 colonnes supplémentaires dans la feuille de calcul du tableur (colonnes III, IV et V précisées sur le tableau 9) suffiront à construire la courbe.

--- Répartitions des Formes ---			Colonnes rajoutées		
I	II		III	IV	V
Valeur du X	Nb Formes à Freq X	Nb Formes à Occ X	Somme cumulée de II	LOG(III)	Somme cumulée de I
57	1	1	1	0,000	57
16	1	1	2	0,301	73
15	2	2	4	0,602	103
10	2	2	6	0,778	123
9	4	4	10	1,000	159
8	1	1	11	1,041	167
7	4	4	15	1,176	195
6	3	3	18	1,255	213
5	4	4	22	1,342	233
4	3	3	25	1,398	245
3	17	17	42	1,623	296
2	39	39	81	1,908	374
1	133	133	214	2,330	507

Tableau 9: Image des données dans la feuille de calcul d'Excel

Les représentations graphiques peuvent être ensuite générées (figure 61 et 62).

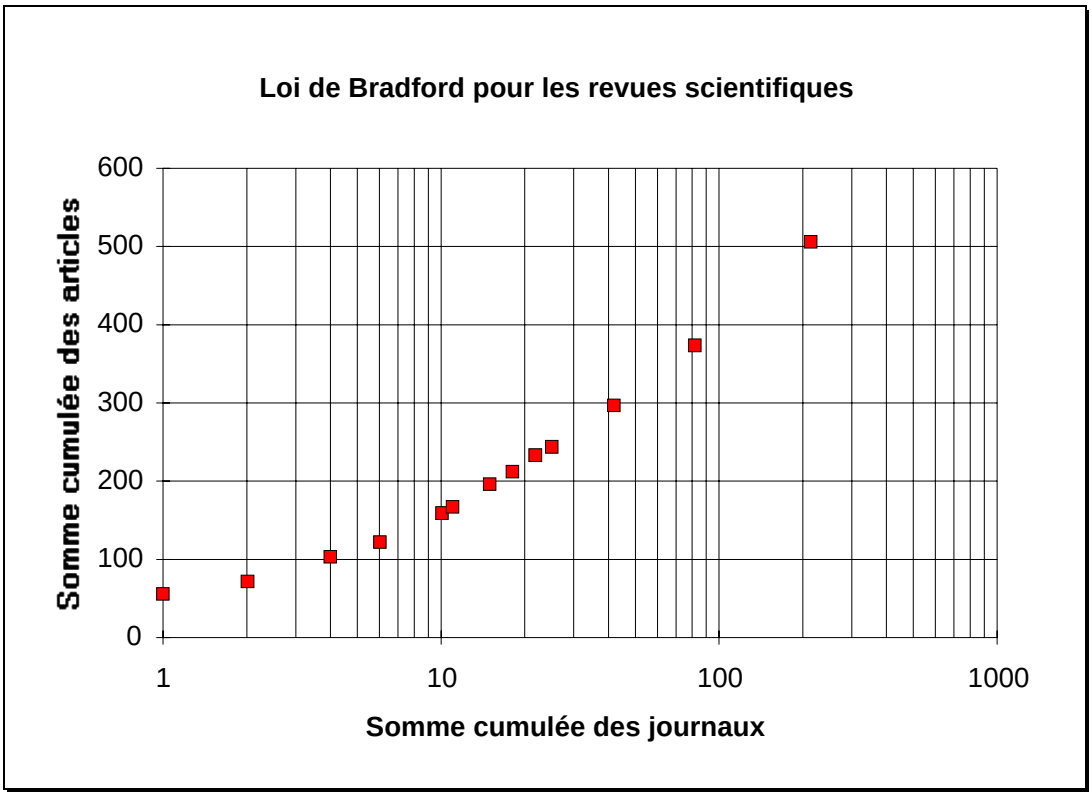


Figure 61: Distribution des articles des périodiques du corpus "Marqueurs Génétiques"

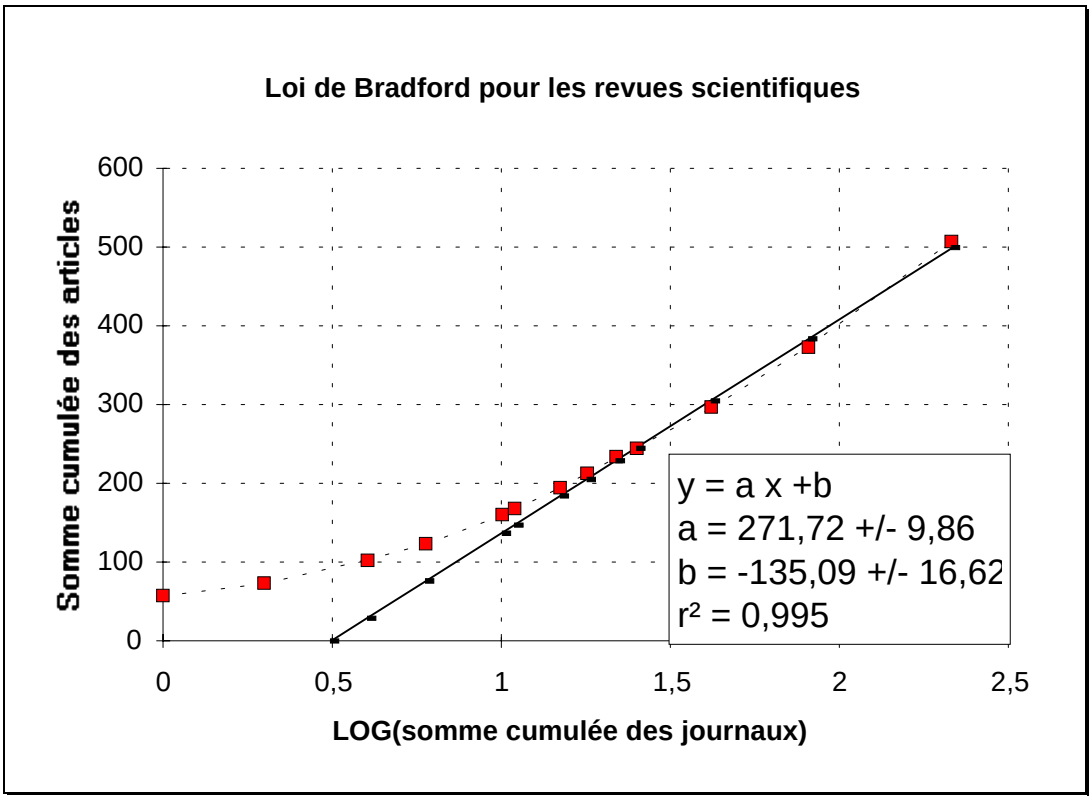


Figure 62: Recherche de la partie linéaire de la courbe

Cette loi parait aussi s'appliquer aux données des bibliographies brevets. Bradford, lors de l'invention de cette loi, cherchait à connaître les principales revues scientifiques qui couvraient un domaine. Selon la même idée, on peut vouloir connaître les principaux déposants de brevets sur le sujet étudié. Pour dégager ce "noyau" de la dispersion des autres déposants, pourquoi ne pas présenter les donner comme l'a fait Bradford en son temps.

Les courbes (figure 63 et 64) qui suivent sont établies à partir du corpus des brevets d'un thème pharmacologique. La distribution des fréquences de dépôts de brevets pour les sociétés déposantes donne l'impression de suivre parfaitement la loi de Bradford.

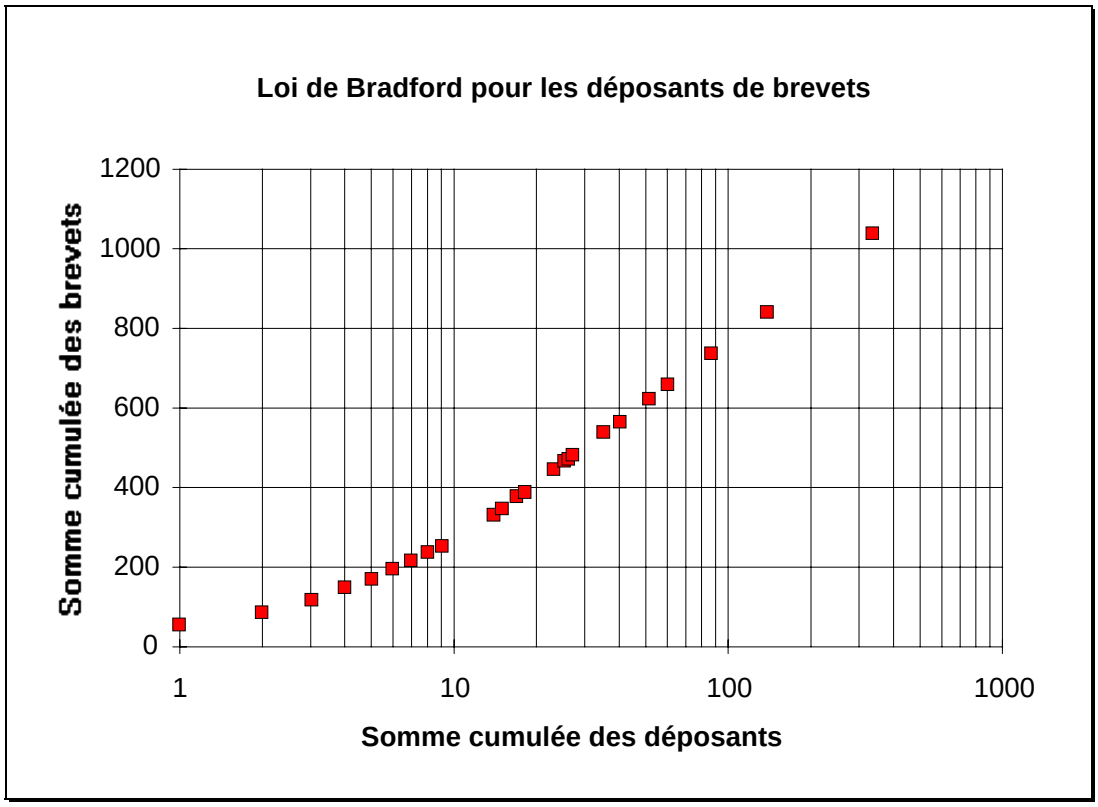


Figure 63: Distribution des déposants dans un domaine pharmacologique

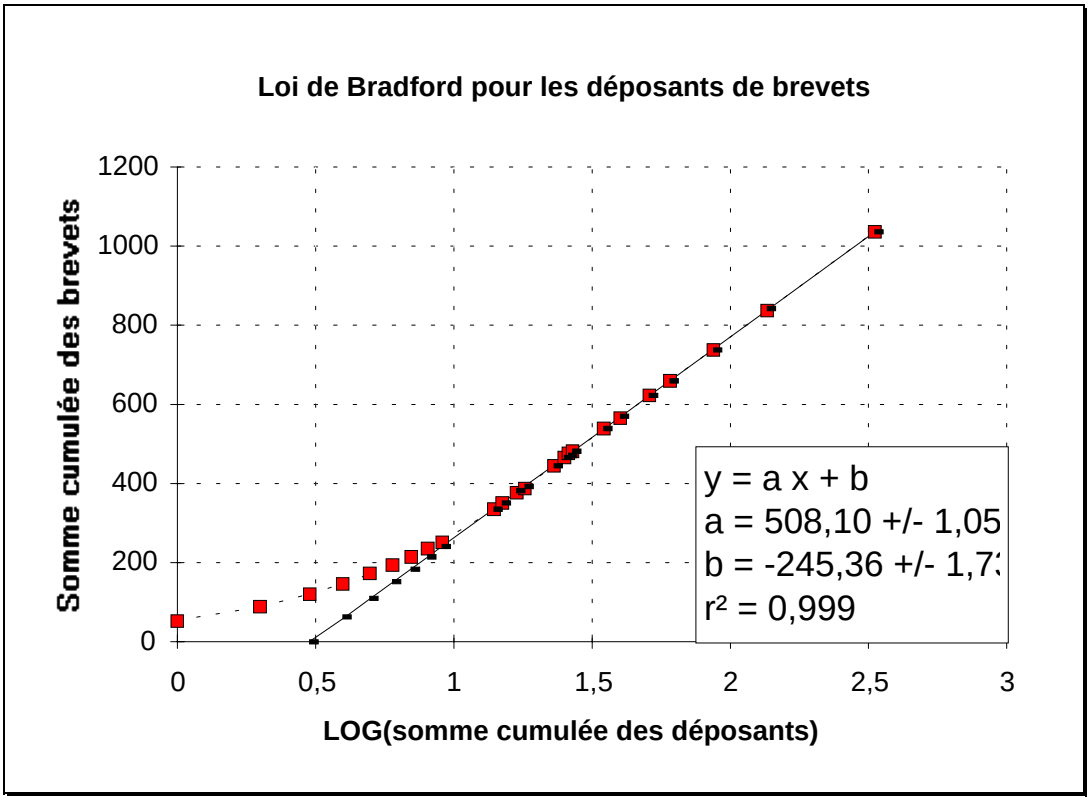


Figure 64: Recherche de la partie linéaire de la courbe

Comme Bradford, on peut découper cette courbe en zones comportant un nombre de brevets égaux (ou à peu près):

	Cumul des déposants	Cumul des brevets	
9 sociétés	1	55	254 brevets
	2	89	
	3	121	
	4	148	
	5	172	
	6	195	
	7	216	
	8	236	
	9	254	
26 sociétés	14	334	284 brevets
	15	349	
	17	377	
	18	390	
	23	445	
	25	465	
	26	474	
	27	482	
	35	538	
61 sociétés	40	568	204 brevets
	51	623	
	60	659	
	87	740	
274 sociétés	137	840	298 brevets
	335	1038	

Tableau 10: Découpage du corpus en zones égales en nombre

On trouve bien un facteur multiplieur entre les cardinaux des regroupements de déposants correspondant. Dans ce cas ce facteur serait proche de 3 ($26 \approx 9x3$, $61 \approx 9x3^2$, $274 \approx 9x3^3$).

⇨ Loi de Zipf:

Les études bibliométriques, qui s'attachent à retranscrire le sens du contenu scientifique ou technique d'un ensemble de références, traitent les champs qui contiennent des descripteurs. **Ces champs descripteurs sont de deux types: les champs en langage libre et les champs en langage contrôlé**. On peut qualifier de champs libres, ceux qui contiennent le titre et le résumé de la publication. Tandis que les champs en langage contrôlé sont ceux créés par les producteurs de bases de données lors de la phase d'indexation: champs de mots-clés et champs de codes documentaires.

Les distributions des formes graphiques pour ces deux types de champs descripteurs sont semblables dans l'apparence de leurs courbes. Par contre le sens qu'il faut donner aux différentes zones de fréquences divergent.

❑ Pour le **vocabulaire contrôlé** la répartition des formes peut se découper en **trois zones**:

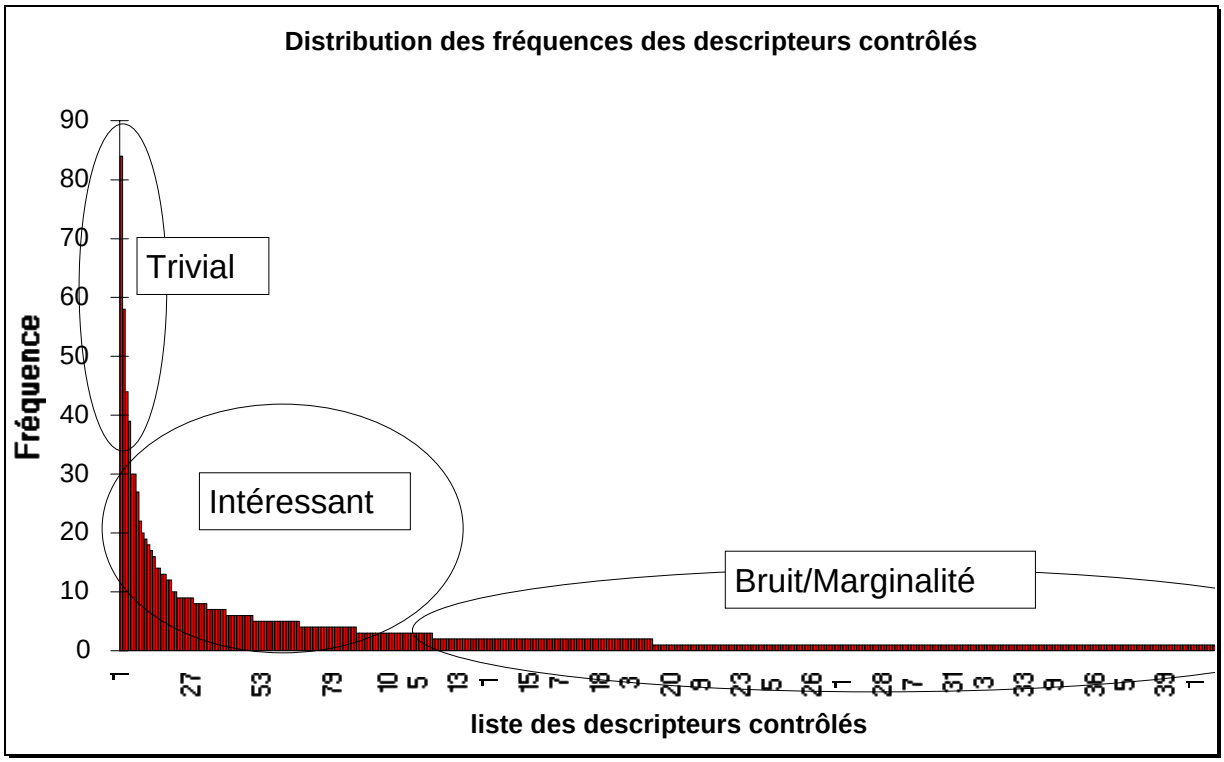


Figure 65: Distribution des fréquences des descripteurs de champs contrôlés

L'information triviale comporte les descripteurs qui définissent les thèmes principaux du sujet examiné. C'est donc une information déjà connue. On doit y retrouver les descripteurs synonymes aux concepts utilisés lors de la recherche des références.

Le bruit ou la marginalité rassemble les descripteurs rares soit parce qu'ils sont non pertinents soit parce qu'ils représentent une approche marginale ou émergente. Il est très difficile d'en discerner les descripteurs appartenant plus à l'aspect originalité qu'à l'aspect bruit.

L'information intéressante est celle qui permet de construire les structures des relations entre les différentes approches et les différentes applications faites dans le domaine étudié. Les disciplines et les techniques connexes y figurent.

❑ Dans le cas, d'une analyse basée sur des **descripteurs en langage libre** les termes se répartissent en **4 zones**. Le nouvel ensemble de formes apparaît en tête de la courbe aux plus fortes fréquences. Il contient des **descripteurs vides de sens**. Ce sont les mots outils employés dans les langages naturels pour construire les phrases: pronoms, conjonctions, adjectifs...

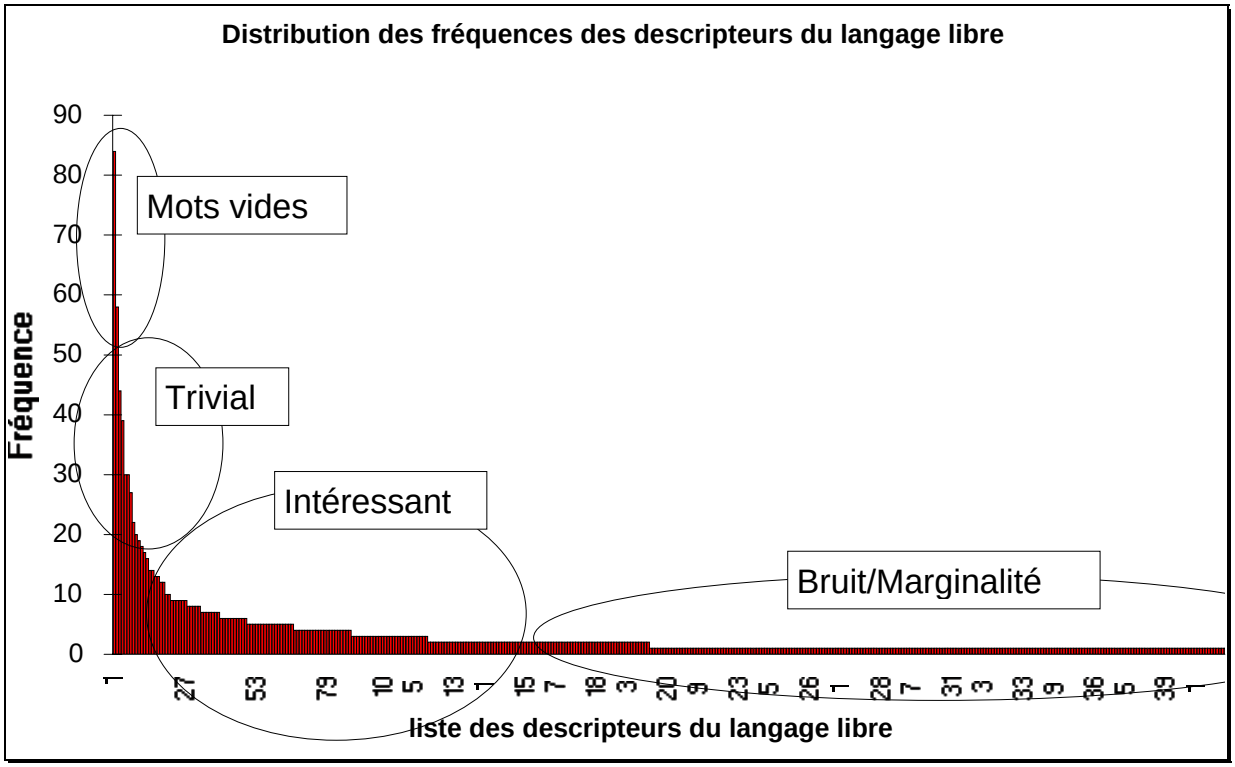


Figure 66: Distribution des fréquences des descripteurs de champs en langage libre

L'existence de ces zones pour ces deux cas de vocabulaire de descripteurs est bien connue, mais il n'existe pas de techniques pour déterminer automatiquement leurs frontières. Malgré cela, la simple prise de conscience de ces découpages de répartitions associée à la consultation des listes des fréquences de formes va permettre de choisir habilement les intervalles de fréquences à considérer pour les analyses statistiques ultérieures.

b) Listes des fréquences de formes

Pour les autres champs, la représentation graphique des éléments des listes des fréquences de formes est utile pour mieux représenter la position de certains éléments clés, vis à vis de l'ensemble des données. Tous les types de graphiques traditionnels sont envisageables:

☞ histogrammes:

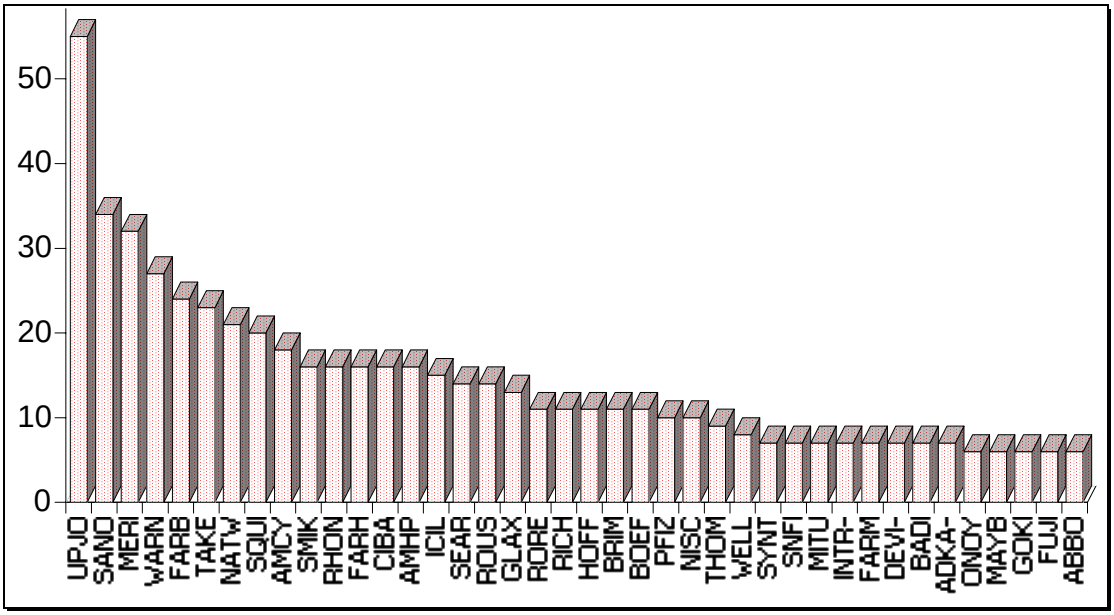


Figure 67: Principales sociétés déposantes de brevets

☞ secteurs:

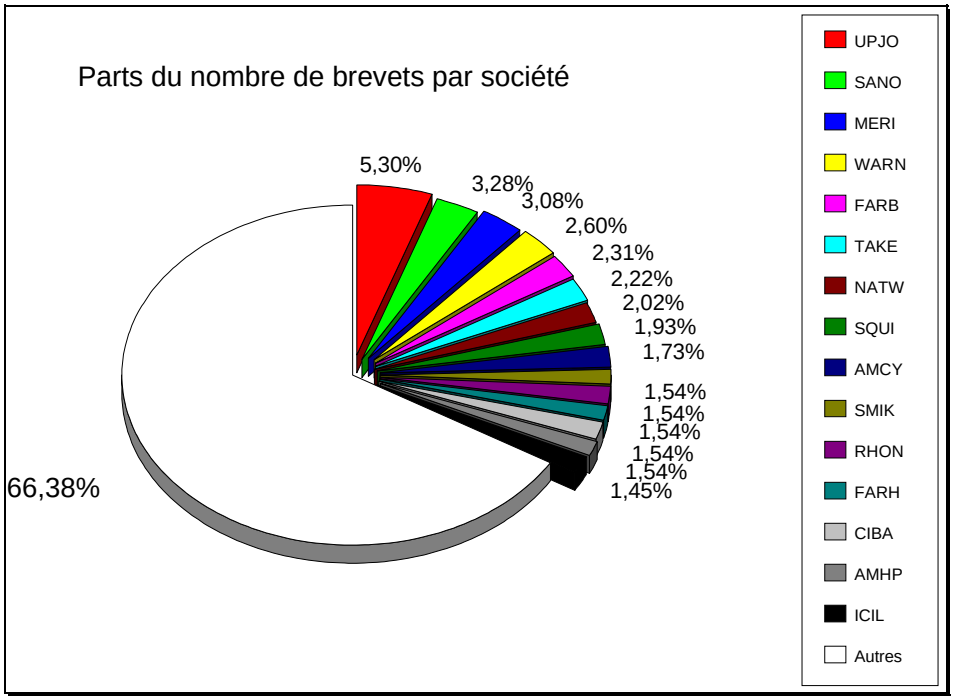


Figure 68: Part du nombre de brevets par société

☞ courbes:

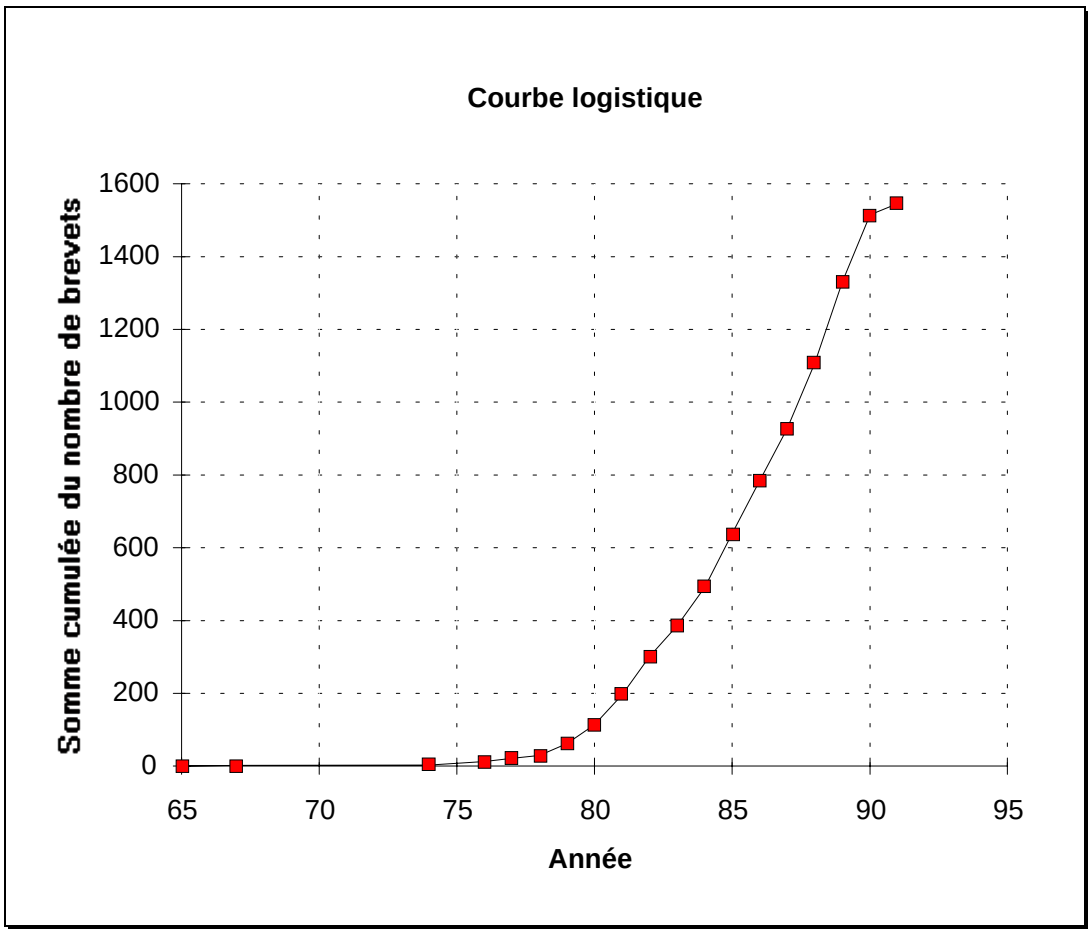


Figure 69: Courbe logistique de l'évolution d'une technologie dans le temps

❑ Remarque: Il faut faire attention de ne pas interpréter trop hâtivement l'affaissement de la fin des courbes logistiques construites sur des données collectées à partir de serveurs. **La baisse du nombre de dépôts s'explique simplement par le laps de temps qui s'écoule entre la date du dépôt d'un brevet et son introduction dans la base de données** (durée de la procédure juridique + durée de collecte et d'indexation du producteur). Donc, les dernières années de cette courbe ne sont certainement pas représentatives de la réalité.

☞ typologie par damiers

Dans le cas où la diversité des formes est fixée dans l'absolu, il est préférable d'établir un **panorama de cet ensemble de formes selon une représentation fixe** (la position de chaque élément est établie définitivement). **La comparaison des résultats en dynamique** (dans le temps, par thèmes, par groupes de travail, par revues...) **en est grandement facilitée**. Les graphiques ci-dessous représentent les pôles de recherche privilégiés restitués dans le panorama des codes de la Classification Derwent. L'ensemble de la classification ne tenant pas sur un seul graphique, les codes sont disposés sur plusieurs graphes selon les grands domaines

de la Classification Derwent (cas de la base de travail DC.JOB, Cf *Extraction et homogénéisation des champs étudiés*).

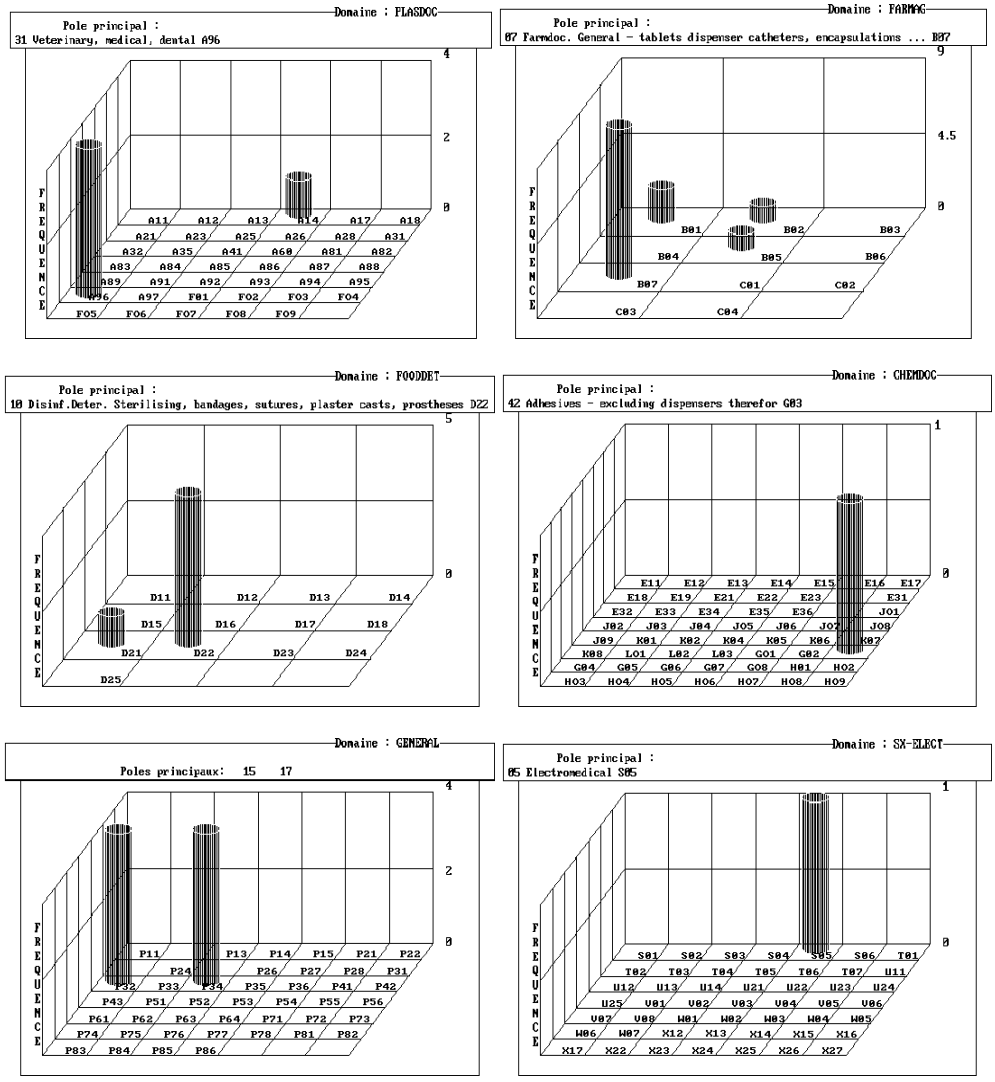


Figure 70: Typologie des Derwent Codes sous forme de damier

Ces graphes ont été créés par le logiciel DATACODE *(CRRM)* développé spécialement pour le traitement en automatique des codes documentaires (CAS, INSPEC, Derwent). On peut tout aussi bien envisager de les construire sous Excel par manipulation du fichier d'édition des fréquences de formes soit manuellement soit par macro-programmation.

c) Listes des fréquences de paires

Les fréquences de paires servent essentiellement à la construction de réseaux de paires.

Un réseau se représente sous forme d'un graphe plan qui relie par des traits toutes les formes qui forment conjointement des paires:

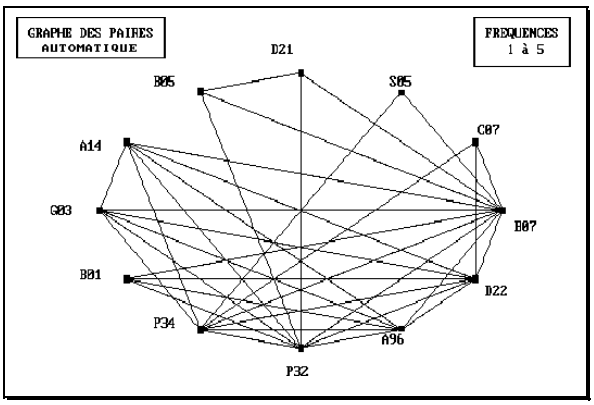


Figure 71: Réseau de paires construit en automatique

Cette première figure 71 a été construite automatiquement (Logiciel *DATACODE du CRRM*) et dispose les formes comprises dans un intervalle de fréquence sur un cercle. En l'occurrence, ce réseau est celui des paires de codes Derwent de la base de travail "DC.JOB".

Pour tendre vers une plus grande réalité des liens, la représentation doit disposer les formes non seulement en fonction de leur centralité dans le réseau (nombre de paires constituées avec une forme) mais aussi en fonction de leur intensité de liens (fréquences des différentes paires constituées avec la forme). A l'heure actuelle, de tels réseaux ne sont construits que manuellement (figure72).

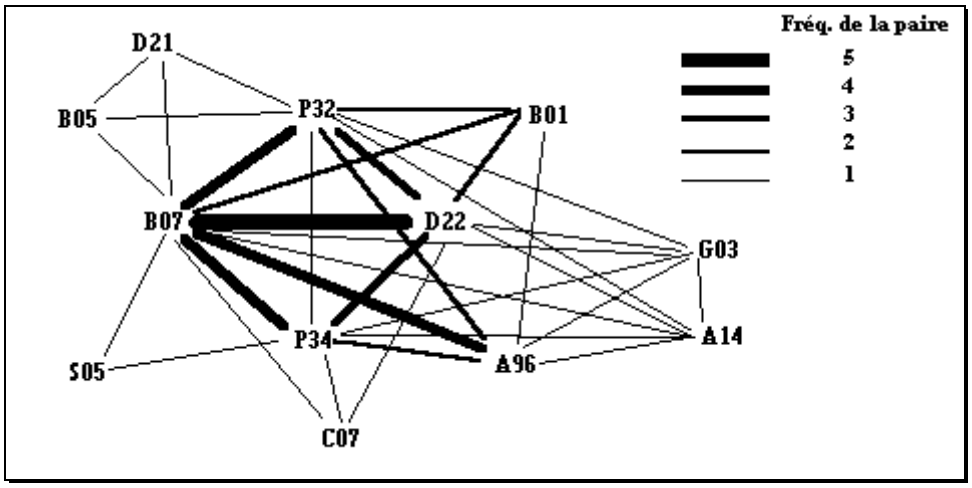


Figure 72: Réseau de paires construit manuellement

2. Traitements des tableaux:

C'est la partie des post-traitements la plus riche que ce soit en nombre de techniques existant ou que ce soit pour l'intérêt des résultats fournis.

a) Représentation graphique du tableau en lui-même

Lorsque le tableau n'est pas trop important, la représentation de l'ensemble des valeurs qu'il contient sous forme graphique peut grandement aider sa lecture.

❑ Le premier exemple est la présentation en **courbe 3D d'une matrice de fréquences** de paires qui croise les années des priorités des brevets avec les principales sociétés du domaine étudié.

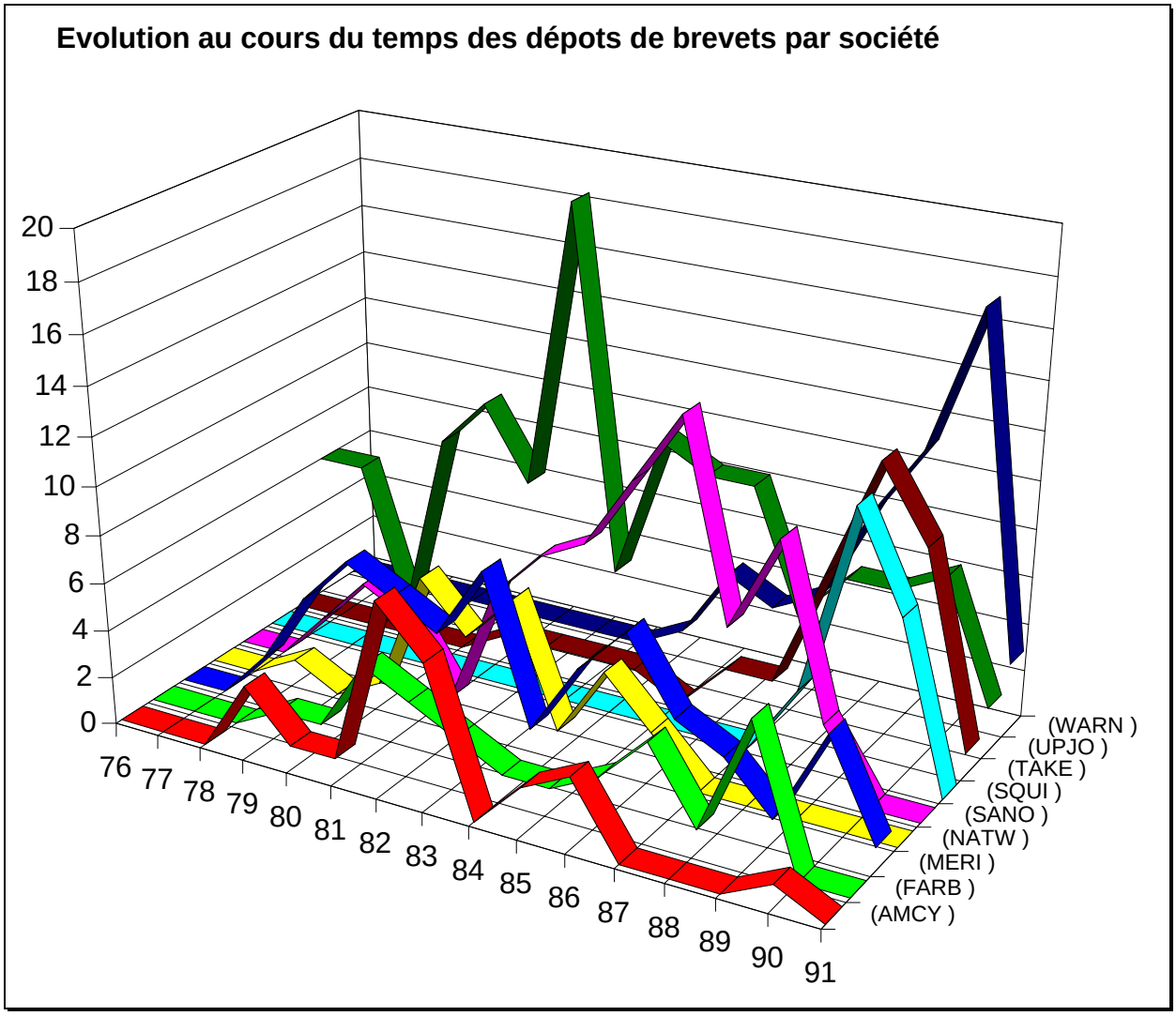


Figure 73: tableau de fréquence sous la forme de courbes 3D

❑ De la même façon, les **matrices d'indice d'association** peuvent subir les mêmes traitements graphiques. Dans ce cas, une telle représentation est très vite appréciée; les valeurs numériques des mesures d'association sont comprises dans un intervalle très petit (généralement entre -1 et 1), donc elles se différencient très peu les unes des autres. La lecture de ces valeurs dans un tableau ne permet pas de cerner les données d'un seul coup d'oeil. La représentation graphique d'un tel tableau est appréciable pour l'interprétation:

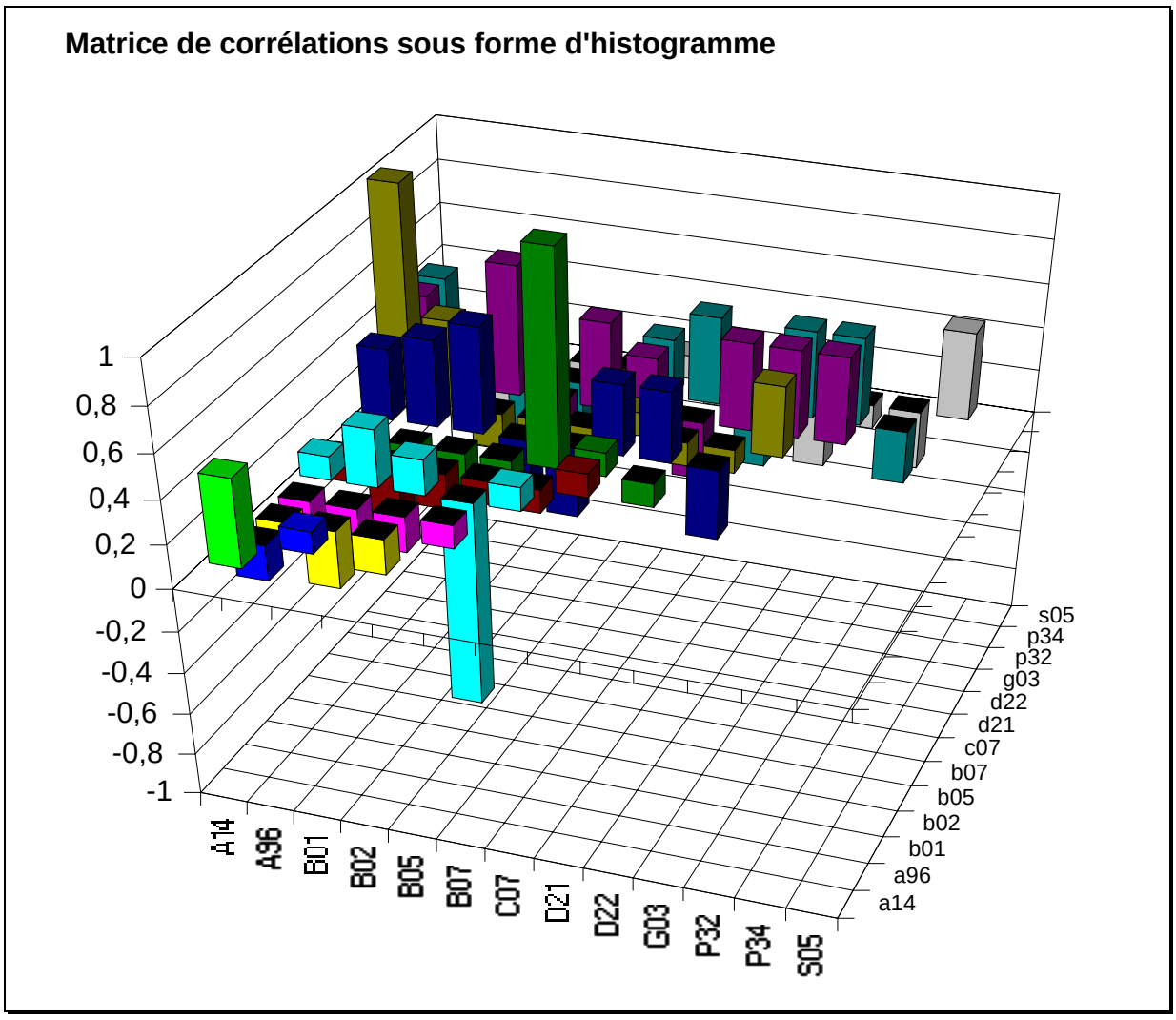


Figure 74: matrice de corrélations sous forme d'histogrammes 3D

Lorsque la taille du tableau le permet, ce genre de représentations est parfait car il a l'avantage de ne rien perdre des données. Les valeurs sont traitées graphiquement in extenso. Ces graphes conservent l'intégrité des mesures bibliométriques réalisées.

Bertin a beaucoup travaillé sur ces aspects de représentations graphiques de tableaux. Son ouvrage [BERT77] est une vrai mine d'informations à ce sujet. La plupart des techniques qu'il présente ne font pas appel à des connaissances statistiques. **Il cherche uniquement à**

transformer la présentation numérique d'un tableau en une image symbolique. La représentation graphique d'un tableau se construit en symbolisant les valeurs numériques de ces croisements par une figure nuancée (histogramme, nuances de grisé, rond de plus ou moins grands rayons...). Cette façon de représenter un tableau lui permet de **réorganiser visuellement son contenu** afin de clarifier sa lecture. Il réalise l'organisation du tableau par **permutation des lignes et des colonnes** de manières à les regrouper par similarité de profils.

Cet ordonnancement étant fastidieux et très rapidement trop complexe pour des grands tableaux, nous estimons que cette étape doit être automatisée. C'est à ce stade qu'il faut faire appel à des techniques statistiques.

❑ La figure ci-dessous est un exemple de **complémentarité entre une méthode statistique et une méthode de représentation graphique**. La méthode statistique a permis de proposer la meilleure organisation des lignes et des colonnes pour regrouper le long de la diagonale les caractères spécifiques aux lignes et aux colonnes. Cette technique statistique est nommée ***sériation***.

Puis, la méthode de représentation symbolique de Bertin est employée pour présenter le résultat. Ce n'est finalement qu'une représentation graphique du tableau sérié (le tableau traité est le tableau de présence/absence des codes DC dans *Tableaux binaires*)

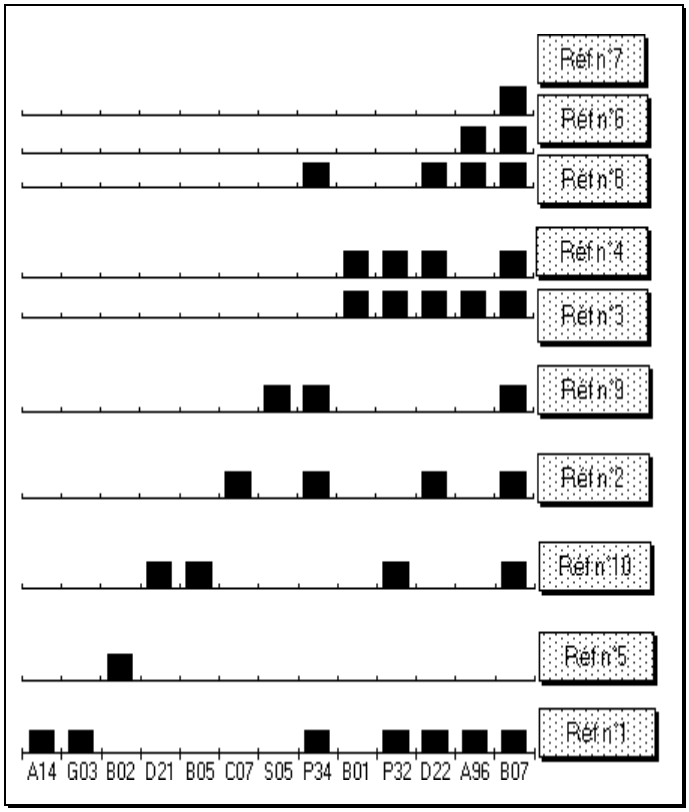


Figure 75: Représentation graphique d'une matrice binaire sériée

❑ Bien avant Bertin, Czékanowsky avait déjà travaillé dans le même ordre d'idée. Sa méthode s'applique cette fois-ci aux **matrices d'association** [CZEK09]. **Le réarrangement a pour objectif de concentrer les similarités les plus fortes au voisinage de la diagonale principale**. Ensuite pour une plus grande qualité visuelle, les valeurs des indices sont remplacées par des nuances de gris (ou de couleurs). A la fin de l'opération, les similarités élevées (grisées sombres), près de la diagonale, indiquent les groupes d'objets similaires.

Pour illustrer cette méthode nous avons représenté la matrice des corrélations entre les codes DC, exposée plus haut (Cf *Tableaux d'indice d'association*):

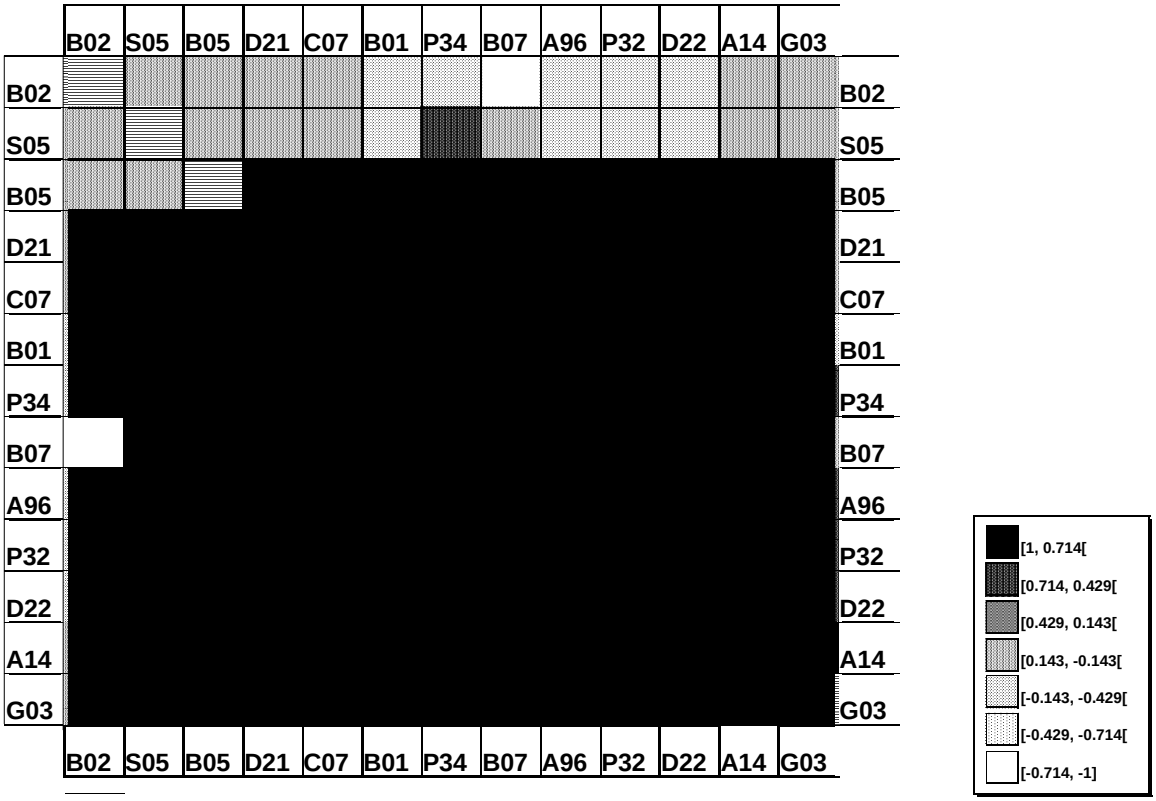


Figure 76: Matrice de corrélation traitée par la méthode de Czékanowsky Intervalle de corrélations

Ce résultat est obtenu automatiquement, à partir de la matrice de DATAVIEW au format Excel, par l'exécution d'une Macro-commande Excel développée pendant cette thèse. L'algorithme de réarrangement utilisé dans cette Macro est une solution analytique développée par Beum et Brundage en 1950.

Cette méthode, relativement simple, donne des résultats faciles à interpréter. Mais cette interprétation devient ardue dès que l'on dépasse une dizaine d'éléments.

b) Les méthodes d'analyse des données

Comme nous venons de le voir, les simples techniques de représentations graphiques des données matricielles ne sont pas suffisantes. **La bibliométrie, étant consommatrice d'une grande quantité de données, il nous faut des méthodes qui puissent réduire cette masse d'informations à l'essentiel**. **Les méthodes d'analyses de données ont spécialement été mises au point dans cette optique**. Elles ne se contentent pas de présenter les données de façon ce que ce soit agréable à l'oeil. Elles vont traiter les données initiales pour en dégager les structures sous-jacentes: trouver les points communs à la grande majorité des éléments, regrouper les éléments par ressemblance de caractères, isoler les éléments marginaux...

Nous allons décrire très rapidement les principales méthodes statistiques appliquées aux études bibliométriques:

⇨ Les techniques de groupement:

La finalité de ces méthodes est de regrouper des objets sur le principe d'évaluation d'un critère d'association entre objets (notion de similarité des caractères des objets).

❍ classifications hiérarchiques [ROUX85], [LERM]

La classification des objets qui résulte de groupements par méthodes hiérarchiques, contrairement aux autres méthodes, **produit plusieurs partitions hiérarchisées**. Cette hiérarchie des partitions est **représentée sous la forme d'une arborescence**. Chaque branche de l'arbre symbolise l'agrégation binaire de deux objets (ou de deux groupes d'objets). Ces agrégations sont distribuées selon une échelle de dissemblance. Les objets les plus semblables sont regroupés aux plus fortes valeurs de l'échelle (en bas de l'échelle). Plus les regroupements s'effectuent à des valeurs d'échelles faibles (en haut de l'échelle) et plus les objets (ou groupes d'objets) réunis ont des caractéristiques divergentes.

De telles représentations graphiques sont nommées ***Dendrogrammes***.

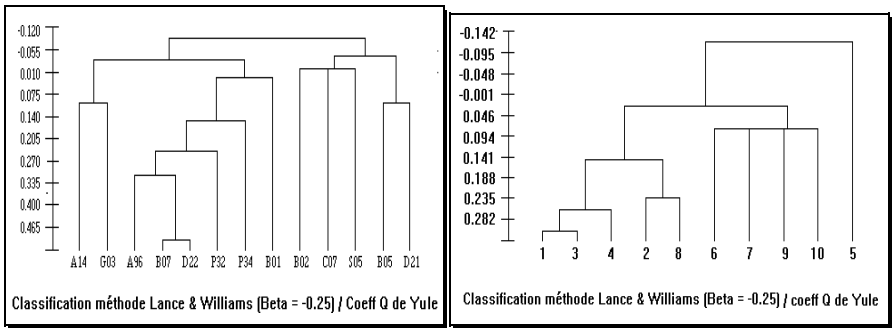


Figure 77: Représentation par dendrogrammes de classifications hiérarchiques

Ces deux graphes sont les résultats de deux classifications hiérarchiques d'une même matrice (matrice binaire des Derwent Codes déjà présentée.). Pour le premier graphe, on a considéré les éléments distribués dans les colonnes de la matrice comme étant les objets à classer (on regroupe les codes par similarité de profils de répartitions dans les références). Tandis que pour le second, ce sont les lignes de la matrice qui sont les objets à classer (on regroupe les références par similarité de profils de présences des codes dans les références). Dans les deux cas, le critère de similarité employé est le coefficient de Yule et la stratégie d'agrégation et celle de Lance et Williams. Les classifications et les graphes ont été établis par le logiciel statistique CLUSTAN [CLUSTA].

❍ sériations [MARC]

Les méthodes de sériations **classifient simultanément les lignes et les colonnes d'une matrice**. Elles aboutissent donc à une partition pour les éléments des colonnes et une partition pour les éléments des lignes. Ces partitions se définissent automatiquement par un algorithme itératif de permutation des lignes et des colonnes de façon à **regrouper les croisements caractéristiques des deux ensembles d'éléments le long de la diagonale principale**. Ces méthodes suivent un principe similaire à celui de la méthode de Czékanowski avec comme avantage de travailler non plus sur des matrices symétriques d'indice d'association mais sur des matrices de données "brutes" rectangulaires (matrices de type présence/absence ou de type fréquence). Les représentations graphiques sont là encore semblable à celle de Czékanoswki. Elles offrent une interprétation visuelle très simple puisqu'elles présentent la matrice originelle des données après réorganisation.

Comme exemple, nous avons sérié la matrice de présence/absence de Codes Derwent (paragraphe *Tableaux binaires*). Cette sériation est obtenue par un algorithme développé au CRRM et inspiré des travaux de Marcotorchino et Michaux [MARC87].

	A14	G03	B02	D21	B05	C07	P34	S05	B01	P32	D22	A96	B07	
Réf. n° 1	1	1					1			1	1	1	1	Réf. n° 1
Réf. n° 5			1											Réf. n° 5
Réf. n°10				1	1					1			1	Réf. n°10
Réf. n° 2						1	1				1		1	Réf. n° 2
Réf. n° 9							1	1					1	Réf. n° 9
Réf. n° 3									1	1	1	1	1	Réf. n° 3
Réf. n° 4									1	1	1		1	Réf. n° 4
Réf. n° 8							1				1	1	1	Réf. n° 8
Réf. n° 6												1	1	Réf. n° 6
Réf. n° 7													1	Réf. n° 7
	A14	G03	B02	D21	B05	C07	P34	S05	B01	P32	D22	A96	B07	

Tableau 11: Tableau binaire après sériation

⇨ Les analyses d'inerties:

En analyse factorielle, **les objets à grouper par similarité de caractères sont plongés dans des espaces multidimensionnels**. Le principe de construction de ces espaces géométriques consiste à trouver le système d'axes qui déforme le moins les nuages de points lors de leurs projections sur les premiers plans du système (tout en conservant les réelles distances entre les points). La constitution des groupes d'objets s'effectue alors visuellement en fonction de la proximité des points dans cet espace.

❍ Analyse Factorielle des Correspondances simples et multiples [BENZ]

L'analyse factorielle des correspondances, développée en France par Benzecri, traite des **tableaux de données de type qualitatif**. L'analyse des correspondances "simple" travaille sur les **tableaux de contingence** tandis que l'analyse des correspondances multiples exige un **tableau d'entrée disjonctif complet**. Elles permettent de faire l'**évaluation des corrélations entre les différentes modalités de caractères qui servent à décrire une population** (l'AFC simple ne peut étudier simultanément que deux catégories de caractères alors que l'AFCM en étudie autant que l'on souhaite). Ces méthodes d'analyse ne sont donc pas censées étudier le regroupement de la population.

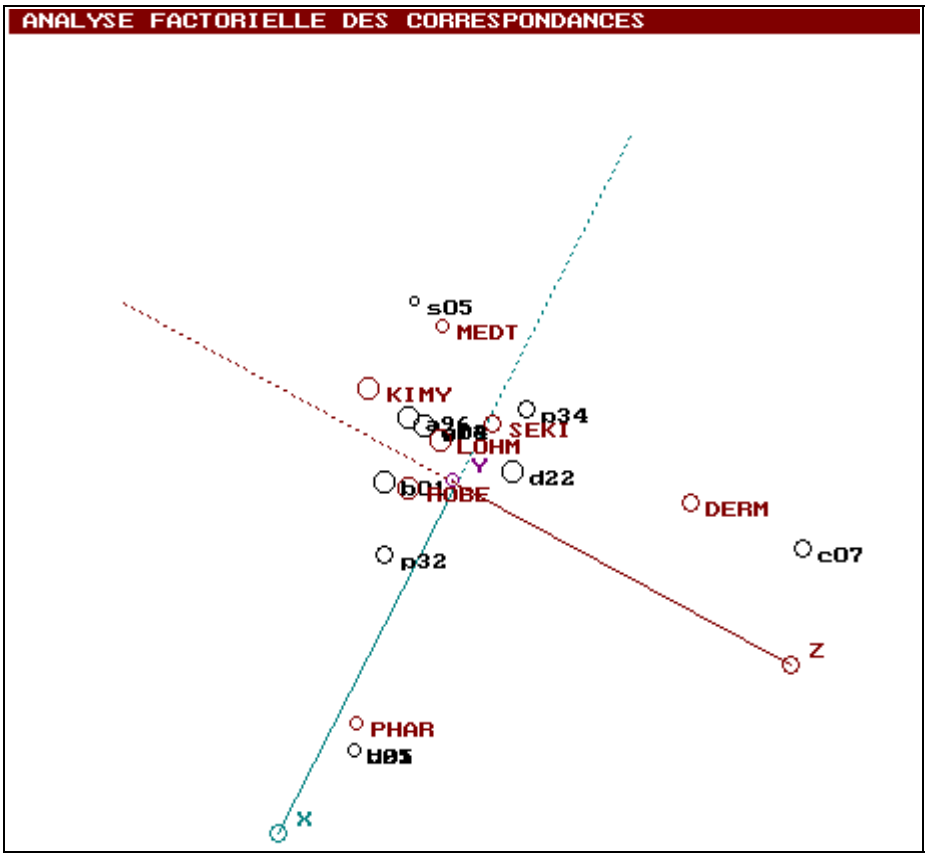


Figure 78: Présentation du nuage de points après AFC sur un plan

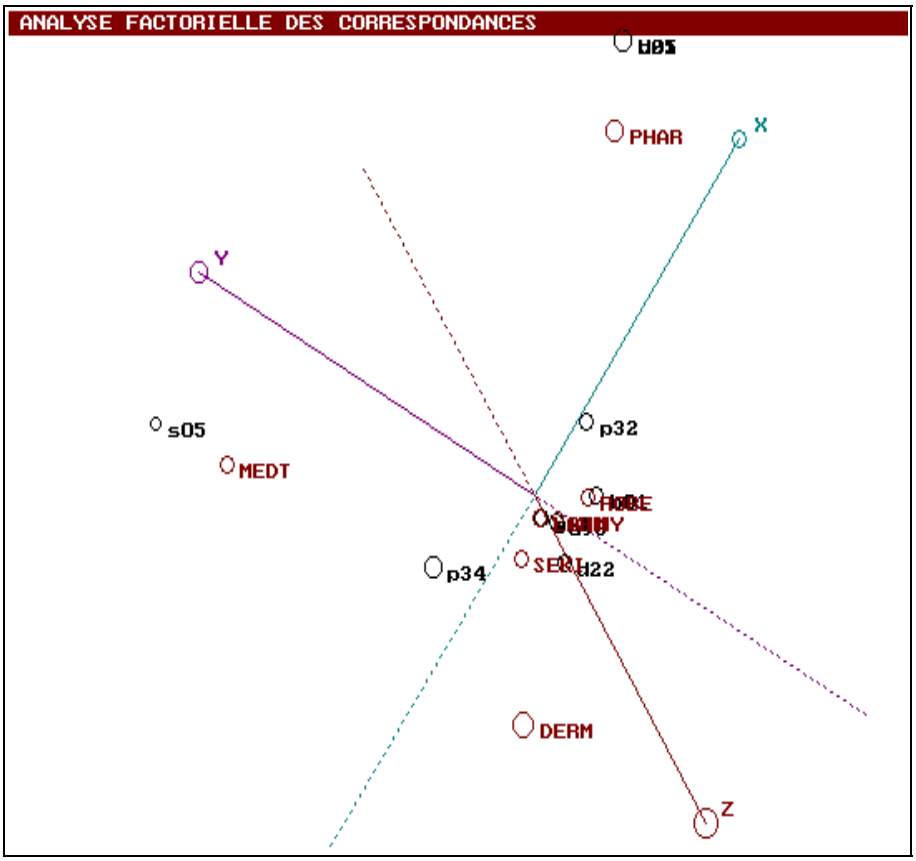
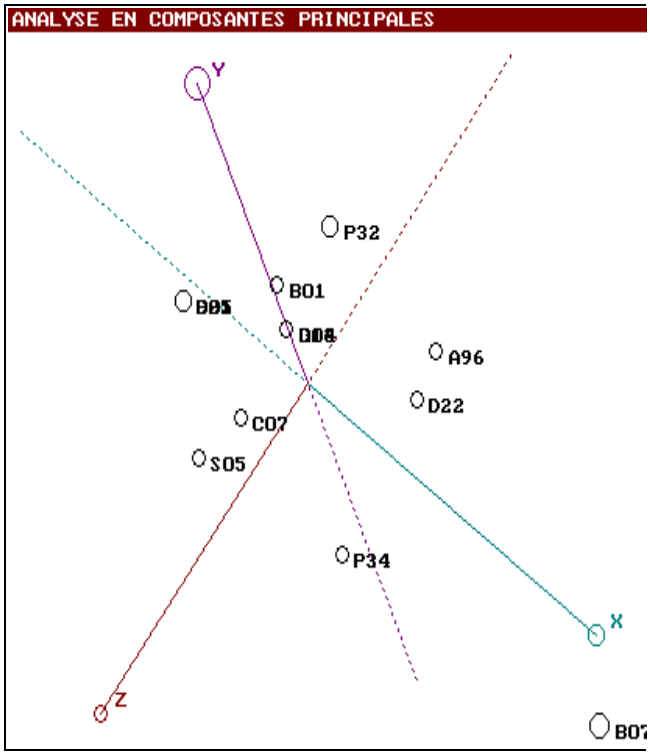


Figure 79: Présentation du nuage de points après AFC en 3D

Les deux graphes (figure 78 et 79) représentent, sous deux angles différents, le nuage de points dans l'espace des trois premiers axes factoriels calculés par AFC. Ces points sont les modalités des deux variables croisées dans la matrice de contingence (Sociétés *x* Codes DC) construite dans le paragraphe *Tableaux des fréquences de paires*. Cette espace en trois dimensions présente donc les proximités des modalités selon leurs répartitions dans la population (dans ce cas la population est l'ensemble des 10 références). On a donc le rapprochement des sociétés déposantes et des thèmes abordés dans les brevets (codes Derwent).

- Analyse en Composantes Principales [SAPO90]

  L'ACP a été mise au point, par H. Hotteling en 1933, **pour cerner les caractéristiques d'une population (les lignes) décrite par un ensemble de facteurs (les colonnes) mesurables par des données quantitatives**. Cette méthode cherche donc à représenter les individus de la population sous la forme d'un nuage de points dans un espace qui les positionne en fonction des caractéristiques mesurées. En sortie d'analyse, les graphes ne présentent que la population sous forme de points. L'information concernant les corrélations entre les facteurs n'est pas offerte dans ce même espace mais sous forme de cercles de corrélations et/ou sous forme de données chiffrées.



ANALYSE EN COMPOSANTES PRINCIPALES

VALEURS PROPRES	% D'INERTIE	% CUMULE
3.31	41.37	41.37
1.45	18.14	59.51
1.21	15.13	74.64
0.74	9.28	83.91
0.53	6.60	90.51
0.44	5.47	95.99
0.28	3.53	99.51
0.04	0.49	100.00

VECTEURS PROPRES

AXE 1	AXE 2	AXE 3
0.34	-0.33	0.27
0.42	0.19	-0.14
0.33	-0.24	-0.45
0.30	-0.11	0.63
0.07	0.73	0.11
0.30	0.24	-0.47
0.47	-0.28	-0.11
0.44	0.35	0.23

Figure 80: Représentation graphique du nuage de points après une ACP

En supposant que la population à étudier n'est plus l'ensemble des références mais l'ensemble des thèmes abordés par les brevets. Et en supposant, qu'il serait intéressant de les caractériser par l'emploi qu'en font les sociétés déposantes. Alors on va vouloir analyser par ACP la même matrice que celle précédemment traitée par AFC. On obtient alors le graphe ci-dessus où les codes Derwent sont disposés dans un espace d'axes caractéristiques des sociétés déposantes.

Remarque: Ces deux analyses factorielles (AFC et ACP) ont été réalisées par le logiciel Tétralogie développé au laboratoire I.R.I.T.(1) sous la direction de Mr Doucet. Les représentations graphiques de Tétralogie possèdent de nombreuses qualités: représentation dynamique (rotation des axes pour déterminer le meilleur angle de vision), aspect de perspective (taille des points selon l'angle de fuite) et donne des informations de la 4ème dimension (nuances de couleurs des points disposés dans l'espace à 3 dimensions) [DOUC91].

(1) I.R.I.T., Université Paul Sabatier 118, route de Narbonne, 31062 Toulouse

Exemples d'études bibliométriques pouvant entrer dans un processus de veille technologique

VI. Exemples d'études bibliométriques pouvant entrer dans un processus de Veille Technologique

Cette partie pratique du mémoire expose deux études mettant en oeuvre des techniques bibliométriques ainsi que l'exploitation de l'outil informatique. Cet outil informatique, par l'automatisation de traitements textuels, a permis l'élaboration de résultats statistiques ne pouvant être obtenue autrement que par cette intervention informatique (le premier exemple présente l'exploitation d'un ensemble de près de 6500 références bibliographiques).

Cet exposé pratique se restreint à des études bibliométriques menées pour des communications scientifiques. Elles ont donc un caractère principalement méthodologique et "académique".

D'autres études, dans le cadre de contrats industriels, ont été réalisées en appliquant cet outil bibliométrique sur des données et selon des objectifs correspondant à des applications en veille technologique. Le caractère confidentiel des données manipulées ainsi que la nature stratégique des résultats ne nous permettent pas de les exposer dans ce mémoire. Leur absence est regrettable car elles auraient probablement été bien plus convaincantes que celles présentées ici. Nous avons affaire au paradoxe auquel les chercheurs de notre discipline sont contraints pour garder toute la confiance des partenaires industriels: plus les résultats d'une étude bibliométrique en veille technologique sont pertinents et plus l'interdiction de les divulguer est implicite.

Cependant, en ce qui concerne les traitements mis en oeuvre, les méthodologies sont identiques que les études soient purement "académiques" ou qu'elles répondent à des objectifs industriels stratégiques.

A. Evaluation d'un secteur scientifique: étude de la production Scientifique en Chimie en France

Cette étude a été publiée sous forme d'un article:

L'analyse des données au service de la bibliométrie
Outils de Veille Technologique à la dimension des moyennes entreprises

Dou H., Quoniam L., Rostaing H., Nivol W.

Revue Française de Bibliométrie Appliquée V. 8, 12/1990

L'ANALYSE DES DONNEES AU SERVICE DE LA BIBLIOMETRIE

OUTILS DE VEILLE TECHNOLOGIQUE A LA DIMENSION DES MOYENNES ENTREPRISES

Dou Henri, Quoniam Luc, Rostaing Hervé, Nivol William
Centre de Recherches Rétrospectives de Marseille,
Université Aix-Marseille III,
13397 Marseille CEDEX 13

La Veille Technologique est au goût du jour. Et pour cause, le besoin en France se fait sentir. Les entreprises prennent conscience de l'indispensable nécessité de gérer l'information qui les environne. C'est à ce titre qu'un Système d'Information Scientifique et Technique, développé par une entreprise, prend toute sa fonction [1, 2].

Certains ont déjà compris que pour survivre aux "attaques" de ses concurrents il faut intégrer cette activité à celles déjà existantes. Seule de grandes entreprises ont pu se le permettre jusqu'à présent et souvent elles ont orienté ce service vers une unique gestion des portefeuilles de brevets. Actuellement, les entreprises d'importance moyenne veulent aussi accéder à ces modes de surveillance et de prévision. La bibliométrie(1) est devenu un outil accessible à tous grâce à l'évolution des nouvelles technologies de l'ordinateur [3, 4, 5]. Le développement de traitements automatiques établis sur des méthodes d'analyse des données (ACP, AFC, Classification Automatique, AFD...) offre une nouvelle dimension au traitement de l'information.

Le travail exposé est l'illustration même du panel de traitements d'analyse des données que l'on peut mettre en oeuvre pour traduire les principales tendances d'un ensemble d'information. Ces méthodes ont permis de mener à bien une étude sur la recherche en Chimie dans les universités françaises: grouper les villes universitaires suivant leur activité scientifique. Cette étude s'inscrit fortement dans une optique Scientométrique(2) car elle a été traitée dans un cadre Universitaire. Mais elle pourrait totalement s'appliquer à un thème ou un domaine sensible à la stratégie d'une entreprise.

(1) Système d'analyses statistiques et d'évaluation à partir de bibliographie de publications scientifiques ou de brevets.
(2) La scientométrie permet d'évaluer les divers domaines de la science et d'estimer leurs évolutions.

I. LES SOURCES: LES BASES DE DONNEES

L'exploitation des bases des données est la ressource première lors de ce type d'analyse. Le gisement qu'elles constituent donne accès (par leurs exhaustivités géographiques, temporelles et thématiques) à l'information recherchée. On peut être certain, tout au moins dans le monde universitaire, que l'étude des publications que ce soit d'un pays, d'une ville ou d'un laboratoire reproduit parfaitement leurs activités scientifiques.

On peut donc bâtir notre travail à partir de cette source en toute quiétude à condition que le choix de la base ait été bien réfléchi. Elle doit avoir une bonne couverture du thème et ne doit favoriser aucun des acteurs ou des critères du phénomène à l'étude. Ce choix s'est porté sur *Chemical Abstract*(3) qui reste incontestablement la base la plus complète et la plus exhaustive concernant l'information scientifique et technique en Chimie [6].

Pour des raisons bien compréhensibles nous nous sommes restreint aux publications, pour l'année 1985, des 17 principales villes universitaires (Liste 1 p 3). Le nombre de références télédéchargées sur micro-ordinateur s'est élevé à environ 6500. Le traitement d'une telle masse de données ne pourrait s'envisager sans des procédures automatiques d'analyse statistique.

Une référence est constituée de champs. Un champ représente une part de l'information globale de la référence mais est aussi, en lui même, une information. Le champ qui reproduit le mieux l'information "thème de l'article" est le champ section code de *Chemical Abstract*(4) pour les raisons suivantes:

- La condensation de l'information en un code
- La non-existence de termes synonymes comme dans le langage
- La pérennité dans l'espace et le temps
- Le traitement informatique rapide: gain de temps

Un tel champ est constitué d'un code qui indique le thème principal de l'article et d'une succession de sections secondaires retraçant les thèmes connexes abordés dans l'article.

A priori la simple étude des sections codes principales pourrait induire une perte d'information par rapport à une analyse plus systématique de l'ensemble des sections codes. Nous estimerons la part d'information apportée lors de l'introduction des sections secondaires pour notre exemple.

---

(3) Base de données interrogée sur le serveur *Orbit Search Service.* Nous remercions l' organisme *ORBIT* pour son aide lors de la réalisation de ce travail.

(4) Les thèmes de l'article sont classés parmi 80 domaines scientifiques (Liste 2 p 4). Ainsi le champ sections codes d'une référence contient autant de codes qu'il y a de thèmes connexes abordés dans l'article (thèmes principaux et secondaires).

Nom de la ville	Abréviation utilisée
Bordeaux/Talence	BOR
Clermont-Ferrand	CLE
Dijon	DIJ
Grasse	GRA
Grenoble/St Martin d'Hères	GRE
Lille/Villeuneuve d'Ascq	LIL
Lyon/Villeurbanne	VIL
Marseille	MAR
Montpellier	MON
Nice	NIC
Orsay	ORS
Paris	PAR
Poitiers	POI
Rennes	REN
Strasbourg	STR
Toulon	TON
Toulouse	TOU

Liste 1: Liste des 17 villes universitaires étudiées

1 pharmacology
2 mammalian hormone
3 biochemical genetics
4 toxycolagy
5 agrochemical bioregulators
6 general biochemistry
7 enzymes
8 radiation biochemistry
9 biochemical methods
10 microbial biochemistry
11 plant biochemistry
12 nonmammalian biochemistry
13 mammalian biochemistr
14 mammalian pathological biochemistry
15 immunochemistry
16 fermentation and bioindustrial chemistry
food and feed chemistry
18 animl nutrition
19 fertilizers; solids; and plant nutrition
20 history; education; and documentation
21 general organic chemistry
22 physical organic chemistry
23 aliphatrc compounds
24 alicyclic compounds
25 benzene; its derivate: and condensed benzoid compounds
26 biomolecules and theif synthetic analogs
27 heteracyclic compounds(one hetero atom)
28 heterocyclic compounds(more than one hetero atom)
29 organometallic and or gaometalloidal compounds
30 terpenes and terpeno ids
31 alkaloids
32 stcroids
33 carbohydtatcs
34 amino acids: peptides; and proteins
35 chemistry of synthetic high polymers
36 physical properties of synthetic high polymers
37 plastics manufacture and processing
38 plastics fabrication ans ures
39 synthetic elastomers and natural rubber
40 textiles
41 dyes; organic pigments; fluorescent brighteners; and photographic sensitirers
42 coating; inks and related products
43 *cellulose; lignin; paper; and other wood products
44 industrial carbohydrates
45 industrials organrc chemicals; leathers; fats; and waxes
46 surface-active agents and detergents
47 apparatus and plant equipment
48 unrt operations and proceses
49 industrial inorganic chemicals
50 propellants and explosfves
51 fossil fuels; derlvatives; and telated products
52 electrochemical; radiational; and terminal energy technology
53 mineralogical and geological chemis try
54 extractive aetallurg
55 ferrous metals and aYlays
56 nonferrous metals and alloys
57 ceramics
58 cement; concrate; and related building materials
59 air pollution and industrial hyglene
60 waste treatment and disposal
61 water
62 essentfal oils and cosmetics
63 pharmaceuticals
64 pharmaceutical analysis
65 general physical chemistry
66 surface chrmistry and colloids
67 catalysis; reaction kinetics; and inorganfc reaction mechanisms
68 phase equilibriums; chemical equilibriums; and solutions
69 thermodynamics; thermochemistry; and terminal properties
70 nuclear phenomena
71 nuclear technology
electrochemistry
33 optical; electron; and mass spectroscopy and other related properties
74 radiation chemistry; photochemistry; and photographics and other reprographic prmesse
75 crystallography and liquid crystals
76 electric phenomena
77 magnetic phenomena
78 inorganic chemicals and reactions
79 inorganic analytical chemistry
80 organic analytical chemistry

Liste 2: **Liste des sections codes du Chemical Abstract et de leurs significations**

II. LES METHODES MISES EN OEUVRE: L'ANALYSE DES DONNEES

Les méthodes d'analyse des données permettent de dégager, derrière une grande masse d'informations, des structures d'organisation entre ces dernières. Celles qu'on a mises en application engendrent des modèles à caractères plus descriptifs que prédictifs (méthodes situationnistes). Mais Ces évaluations recoupées en dynamique (état des lieux à des périodes différentes) permet d'estimer les tendances majeures de l'évolution du thème à l'étude.

Les méthodes que nous avons utilisées ont toutes pour principes de replacer les objets étudiés les uns par rapport aux autres en fonction de leurs ressemblances ou leurs dissemblances, ceci pour pouvoir les regrouper suivant leur similarité de caractères. Deux approches mathématiques différencient ces méthodes :

- méthodes d'analyse factorielle [7, 8]: ACP, AFC, AFD:
  l'objectif est de représenter dans un espace réduit l'information contenue dans un tableau de données. Elles renseignent, par conséquent, sur les deux entrées du tableau. Donc le but est de condenser l'information dans un espace non fini. Le critère d'optimisation de la méthode est de construire l'espace conservant la plus grande part de la diversité de l'information (variance).

- méthodes de classification automatique [9, 10, 11]:
  elles ont pour objet d'établir une partition parmi les individus étudiés suivant des critères de proximité et d'agrégation choisis. L'espace obtenu est un espace fini mais il ne représente qu'une des deux entrées du tableau de données (une entrée du tableau se classe en fonction de la répartition de l'autre entrée).

III. RESULTATS

Tous les résultats présentés ont été obtenus grâce à une version améliorée par le CRRM du logiciel STAT-ITCF(5).

Le tableau initial de données (Tableau 1 p 7), obtenu à partir du fichier télédéchargé par traitement automatique de chaîne de caractères (logiciel DATRANS *@CRRM*), se constitue pour les colonnes des 17 villes Universitaires et pour les lignes des sections codes principales (suivies des secondaires pour le traitement global). Nous avons donc à traiter un tableau (matrice) à 17 colonnes et 78 (ou 158) lignes puisque 2 sections principales étaient présentes dans aucune des villes(6).

La répartition des villes en colonnes et des sections en lignes nous est imposée par les analyses d'inerties qui, par leurs principes mathématiques, exigent une matrice comportant plus de colonnes que de lignes. Alors qu'un comportement normal consiste dans ce genre d'analyse à placer les individus étudiés sur les lignes et à distribuer les variables qui permettent de les caractériser sur les colonnes.

III.1. La classification automatique (Cf ANNEXE 1 p 18):

Connaissant le sujet de l'étude il n'est pas étonnant que la première méthode utilisée soit la classification automatique. De nombreux paramètres entrent en jeu lors de l'élaboration de la stratégie classificatoire. Celle qui a été choisie est la suivante:
- classification ascendante hiérarchique
- avec calcul des distances euclidiennes
- le critère d'agrégation étant la moyenne des distances pondérées

Le critère d'agrégation offre un bon compromis entre une bonne qualité d'inertie intergroupe et intragroupe [8], ceci pour que la marginalité entre les villes ne soit pas trop influente. La méthode ascendante hiérarchique donne des résultats satisfaisants pour une matrice de taille acceptable [11]. La raison du choix de la distance euclidienne s'explique plus loin pour des raisons de compatibilité entre méthodes d'analyse(7).

---

(5) Logiciel statistique développé par l'*Institut Technique des Céréales et des Fourrages*. Nous les remercions pour leur collaboration.

(6) Ces deux sections ont pour codes :

41, dyes; organic pigments; fluorescent brightners; and photographic sensitizers

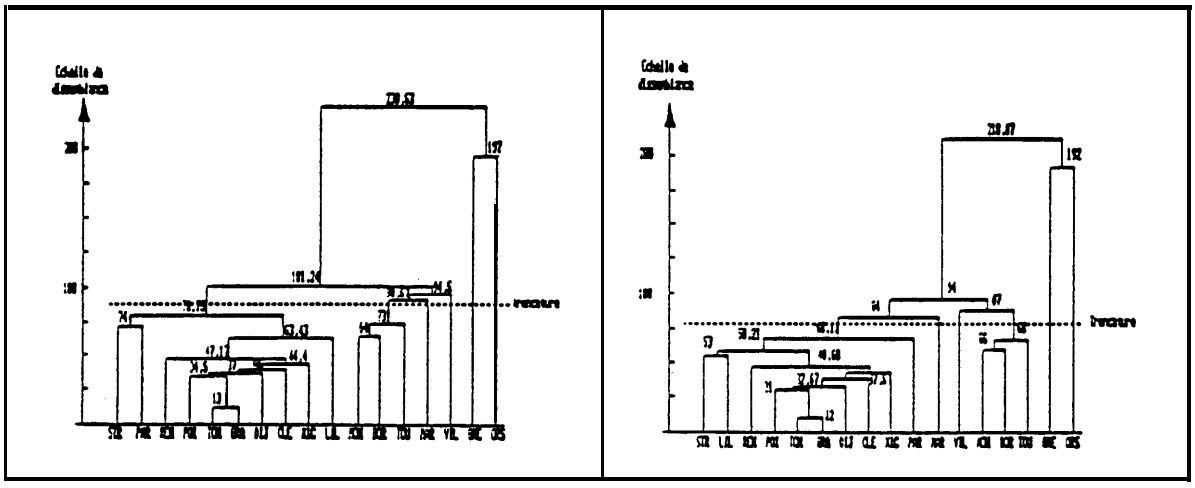
44, industrial carbohydrates

(7) La métrique euclidienne est la seul distance de proximité acceptée pour des coordonnées calculées par AFC et donc permettant une classification automatique sur les résultats obtenus par AFC

	1 STR	2 REX	3 POI	4 PAR	5 ORS	6 NIC	7 NON	8 LIL	9 GRE	10 CLE	11 TOU	12 TON	13 VIL	14 DOR	15 NAR	16 DIJ	17 LAA
1	17.00	9.00	0.00	15.00	3.00	10.00	47.00	10.00	2.00	15.00	53.00	4.00	36.00	35.00	31.00	9.00	0.00
2	10.00	9.00	3.00	53.00	16.00	23.00	31.00	9.00	2.00	11.00	53.00	0.00	36.00	41.00	3.00	4.00	0.00
3	35.00	2.00	0.00	26.00	21.00	0.00	4.00	0.00	4.00	1.00	13.00	0.00	14.00	2.00	13.00	0.00	0.00
4	1.00	4.00	1.00	4.00	12.00	6.00	5.00	14.00	1.00	2.00	14.00	2.00	36.00	12.00	9.00	4.00	0.00
5	1.00	0.00	0.00	0.00	1.00	0.00	5.00	1.00	0.00	0.00	1.00	0.00	5.00	1.00	3.00	0.00	0.00
6	22.00	3.00	0.00	9.00	12.00	6.00	20.00	17.00	24.00	1.00	0.00	0.00	16.00	9.00	17.00	1.00	0.00
7	11.00	0.00	0.00	5.00	7.00	3.00	15.00	3.00	14.00	0.00	15.00	1.00	10.00	11.00	10.00	0.00	0.00
8	1.00	0.00	0.00	1.00	6.00	3.00	3.00	3.00	5.00	0.00	2.00	0.00	3.00	1.00	0.00	0.00	0.00
9	10.00	7.00	2.00	6.00	0.00	0.00	12.00	13.00	0.00	3.00	10.00	0.00	24.00	0.00	16.00	3.00	0.00
10	5.00	2.00	2.00	7.00	10.00	2.00	5.00	12.00	0.00	9.00	25.00	0.00	21.00	10.00	20.00	4.00	0.00
11	2.00	1.00	9.00	4.00	12.00	6.00	16.00	5.00	6.00	6.00	21.00	0.00	13.00	5.00	5.00	9.00	1.00
12	3.00	14.00	3.00	2.00	4.00	3.00	10.00	7.00	1.00	2.00	7.00	0.00	10.00	10.00	6.00	4.00	0.00
13	6.00	3.00	3.00	16.00	11.00	21.00	14.00	0.00	0.00	3.00	16.00	1.00	32.00	5.00	17.00	3.00	0.00
14	3.00	1.00	0.00	14.00	14.00	6.00	7.00	0.00	2.00	1.00	12.00	0.00	10.00	10.00	11.00	3.00	0.00
15	12.00	2.00	1.00	22.00	15.00	2.00	23.00	22.00	12.00	1.00	2.00	0.00	21.00	7.00	23.00	4.00	0.00
16	0.00	0.00	0.00	6.00	0.00	0.00	12.00	0.00	0.00	5.00	3.00	0.00	2.00	9.00	11.00	5.00	0.00
17	0.00	5.00	0.00	4.00	0.00	1.00	17.00	3.00	2.00	2.00	0.00	0.00	2.00	2.00	10.00	15.00	5.00
18	3.00	2.00	2.00	5.00	6.00	0.00	0.00	2.00	0.00	0.00	3.00	0.00	3.00	7.00	5.00	0.00	0.00
19	0.00	0.00	1.00	0.00	3.00	0.00	7.00	1.00	1.00	0.00	4.00	0.00	4.00	0.00	0.00	0.00	0.00
20	0.00	1.00	0.00	0.00	1.00	0.00	2.00	0.00	1.00	0.00	1.00	2.00	1.00	2.00	1.00	1.00	0.00
21	0.00	3.00	1.00	5.00	10.00	1.00	2.00	0.00	3.00	0.00	3.00	0.00	1.00	1.00	1.00	0.00	0.00
22	10.00	9.00	0.00	2.00	20.00	4.00	14.00	10.00	7.00	3.00	11.00	0.00	20.00	17.00	10.00	1.00	0.00
23	5.00	4.00	1.00	1.00	4.00	3.00	4.00	1.00	0.00	2.00	3.00	0.00	7.00	3.00	2.00	0.00	0.00
24	3.00	3.00	0.00	0.00	9.00	0.00	1.00	0.00	1.00	0.00	2.00	0.00	3.00	3.00	7.00	0.00	0.00
25	5.00	4.00	0.00	1.00	6.00	0.00	2.00	0.00	1.00	0.00	1.00	0.00	7.00	3.00	4.00	1.00	0.00
26	2.00	1.00	0.00	4.00	4.00	0.00	2.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00
27	1.00	0.00	4.00	0.00	4.00	2.00	6.00	2.00	1.00	2.00	5.00	0.00	7.00	5.00	1.00	2.00	0.00
28	2.00	6.00	1.00	1.00	4.00	0.00	2.00	5.00	0.00	3.00	2.00	0.00	6.00	2.00	7.00	0.00	0.00
29	7.00	13.00	2.00	3.00	10.00	5.00	7.00	0.00	3.00	0.00	47.00	0.00	7.00	10.00	0.00	13.00	0.00
30	0.00	0.00	0.00	0.00	3.00	0.00	1.00	0.00	2.00	0.00	0.00	0.00	2.00	4.00	2.00	0.00	0.00
31	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.00	1.00	0.00	2.00	0.00	0.00	0.00	0.00
32	2.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	4.00	5.00	1.00	0.00	0.00
33	0.00	0.00	0.00	1.00	0.00	0.00	1.00	3.00	16.00	[illegible]	7.00	0.00	7.00	0.00	1.00	0.00	0.00
34	0.00	0.00	0.00	3.00	1.00	0.00	15.00	4.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	2.00	0.00
35	3.00	0.00	0.00	3.00	1.00	1.00	7.00	2.00	7.00	5.00	1.00	0.00	3.00	0.00	1.00	0.00	0.00
36	9.00	0.00	0.00	0.00	14.00	4.00	7.00	5.00	0.00	1.00	2.00	0.00	6.00	2.00	1.00	0.00	0.00
37	1.00	0.00	0.00	1.00	0.00	0.00	2.00	1.00	1.00	1.00	12.00	0.00	10.00	1.00	0.00	0.00	0.00
38	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
39	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
40	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00	3.00	0.00	1.00	0.00	0.00
41	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	[illegible]	0.00	0.00
42	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	11.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
43	0.00	0.00	2.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	4.00	0.00	9.00	0.00	2.00	0.00	0.00
44	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	2.00	0.00	0.00	0.00	1.00	0.00	0.00
45	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
46	1.00	0.00	1.00	3.00	4.00	0.00	2.00	1.00	4.00	0.00	24.00	0.00	2.00	0.00	2.00	2.00	0.00
47	0.00	0.00	0.00	0.00	0.00	9.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00
48	0.00	0.00	3.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	3.00	1.00	0.00
49	0.00	0.00	6.00	3.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00	3.00	4.00	7.00	0.00	0.00
50	1.00	3.00	2.00	0.00	5.00	0.00	14.00	3.00	15.00	0.00	0.00	0.00	6.00	4.00	7.00	2.00	0.00
51	2.00	13.00	12.00	6.00	16.00	6.00	14.00	7.00	7.00	0.00	13.00	0.00	0.00	7.00	5.00	0.00	1.00
52	0.00	0.00	0.00	1.00	0.00	1.00	1.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53	0.00	0.00	3.00	1.00	0.00	0.00	0.00	2.00	6.00	0.00	4.00	0.00	0.00	2.00	1.00	4.00	0.00
54	4.00	2.00	12.00	0.00	14.00	0.00	2.00	3.00	38.00	1.00	7.00	0.00	10.00	3.00	10.00	4.00	0.00
55	1.00	11.00	2.00	0.00	6.00	0.00	4.00	0.00	3.00	0.00	4.00	0.00	5.00	7.00	0.00	0.00	0.00
56	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	4.00	0.00	5.00	0.00	0.00	2.00	0.00
57	0.00	1.00	0.00	3.00	1.00	0.00	1.00	2.00	5.00	0.00	3.00	0.00	4.00	1.00	1.00	0.00	0.00
58	0.00	2.00	1.00	1.00	0.00	0.00	1.00	3.00	0.00	0.00	4.00	0.00	2.00	0.00	1.00	0.00	0.00
59	0.00	10.00	5.00	5.00	0.00	0.00	4.00	0.00	3.00	1.00	7.00	0.00	0.00	11.00	10.00	0.00	0.00
60	0.00	1.00	0.00	0.00	0.00	0.00	2.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	2.00	0.00	10.00
61	4.00	2.00	0.00	1.00	0.00	1.00	16.00	0.00	0.00	5.00	0.00	0.00	10.00	3.00	6.00	1.00	0.00
62	0.00	0.00	0.00	1.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	3.00	0.00	0.00
63	7.00	3.00	3.00	6.00	66.00	4.00	4.00	5.00	43.00	1.00	11.00	0.00	20.00	13.00	10.00	2.00	0.00
64	6.00	1.00	1.00	12.00	19.00	2.00	20.00	3.00	21.00	2.00	4.00	0.00	17.00	7.00	41.00	3.00	0.00
65	0.00	0.00	11.00	7.00	5.00	0.00	5.00	5.00	1.00	0.00	0.00	0.00	46.00	10.00	7.00	0.00	0.00
66	0.00	2.00	1.00	15.00	4.00	2.00	6.00	3.00	3.00	13.00	2.00	0.00	10.00	2.00	7.00	0.00	0.00
67	2.00	0.00	0.00	1.00	3.00	0.00	1.00	1.00	11.00	20.00	1.00	0.00	3.00	0.00	9.00	1.00	0.00
68	5.00	0.00	0.00	14.00	115.00	7.00	16.00	0.00	46.00	4.00	1.00	0.00	19.00	2.00	36.00	0.00	0.00
69	1.00	2.00	5.00	5.00	26.00	0.00	1.00	0.00	32.00	1.00	2.00	0.00	10.00	0.00	0.00	0.00	4.00
70	5.00	16.00	0.00	4.00	2.00	0.00	3.00	3.00	19.00	2.00	15.00	1.00	7.00	3.00	5.00	6.00	0.00
71	10.00	14.00	3.00	13.00	190.00	15.00	22.00	37.00	45.00	3.00	20.00	1.00	37.00	34.00	37.00	7.00	1.00
72	4.00	0.00	0.00	0.00	26.00	1.00	3.00	0.00	2.00	3.00	0.00	0.00	5.00	2.00	0.00	1.00	0.00
73	5.00	16.00	0.00	7.00	65.00	3.00	15.00	12.00	65.00	4.00	19.00	0.00	16.00	32.00	12.00	9.00	0.00
74	11.00	6.00	2.00	5.00	80.00	1.00	37.00	25.00	126.00	0.00	42.00	0.00	16.00	34.00	14.00	2.00	0.00
75	14.00	1.00	0.00	1.00	42.00	2.00	9.00	0.00	111.00	1.00	3.00	0.00	2.00	14.00	4.00	0.00	0.00
76	12.00	11.00	1.00	2.00	0.00	3.00	3.00	4.00	0.00	2.00	21.00	0.00	10.00	11.00	2.00	0.00	0.00
77	2.00	0.00	0.00	0.00	10.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	5.00	1.00	1.00	0.00	0.00
78	0.00	0.00	1.00	0.00	1.00	1.00	2.00	2.00	1.00	0.00	2.00	0.00	2.00	4.00	1.00	0.00	0.00

Tableau 1: Tableau des sections principales

La comparaison de l'arbre hiérarchique de la matrice globale avec celui de la matrice des sections codes principales nous porte à croire que l'introduction des sections secondaires dans la matrice des données n'apporte que peu d'information supplémentaire (cet état de fait a été vérifié par d'autres méthodes d'analyse). Les liens entre villes ne varient pratiquement pas et la troncature effectuée (6 classes) fournit dans les deux cas la même partition (Figures ci-dessous).



Arbre hiérarchique de la matrice brute sections principales et secondaires

Arbre hiérarchique de la matrice brute sections principales

Classe	Villes appartenant à la classe
1	Strasbourg, Rennes, Poitiers, Paris Nice, Lille, Clermont-Ferrand Toulon, Dijon, Grasse.
2	Montpellier, Toulouse, Bordeaux
3	Marseille
4	Lyon/Villeneuve d'ascq
5	Grenoble/St Martin d ' Hères
6	Orsav

Tableau 2: Classement commun aux deux classifications (suivant les troncatures indiquées sur les figures précédentes)

Comme lors des analyses la présence d'une section secondaire dans un champ n'est pas affectée d'un poids relatif à son importance dans le champ (importance inférieure à la section principale et partagée avec les autres sections secondaires) il est probable que l'information globale en soit faussée. Il est inutile de tenter un tel risque. Donc l'analyse se poursuivra en ne considérant que les sections principales.

La méthode de classification automatique regroupe effectivement les villes suivant leurs similarités mais il est difficile d'estimer à partir de la stratégie de classification quels sont les critères de ressemblance ou de dissemblance qui sont reproduit. Il paraîtrait intéressant de savoir quel type d'activité scientifique se cache derrière une classe ?

III.2. Analyse Factorielle Discriminante (Cf ANNEXE 2 p 25):

Pour se faire on a essayé de savoir si une Analyse Factorielle Discriminante nous renseignerait sur cette notion. Notre espoir de réussite était faible. L'AFD permet de grouper les individus (lignes) d'une matrice entre eux ainsi que d'affecter les caractères (colonnes) aux groupes selon leur importance d'influence sur ces groupes. Mais auparavant, les individus doivent avoir subi un pré-ordre.

Dans notre cas il faut donc, cette fois-ci, tout d'abord classer la matrice sur les sections codes, et non pas sur les villes, avec les mêmes principes classificatoires et la même troncature pour réaliser le pré-ordre indispensable à l'AFD. Ainsi la matrice affectée de sa nouvelle colonne peut être analysée(8). Si le dépouillement des résultats livre la même partition des villes que pour la classification automatique alors les deux partitions se recouvreraient et on pourrait associer un groupe de villes à un groupe de sections codes.

Les résultats sont probants, notre exemple ne vérifie pas le cas précédemment exposé: Poitiers, Lyon/Villeneuve d'Ascq et Toulouse se retrouvent à proximité; Marseille s'est lié à Bordeaux; et Clermont est totalement isolé.

III.3. Analyse Factorielle (Cf ANNEXE 3 p 34):

La solution à tous ces désagréments est l'analyse d'inertie. Le principe mathématique est d'analyser les corrélations pour les espaces duaux(9). Ces méthodes représentent les deux espaces duaux de telle façon qu'on puisse les mettre graphiquement en relation. On peut donc avoir à la fois l'information sur les individus et sur les variables. Ceci résout notre petit problème.

---

(8) Cette colonne établit à chaque ligne le groupe auquel appartient l'objet que représente la ligne.

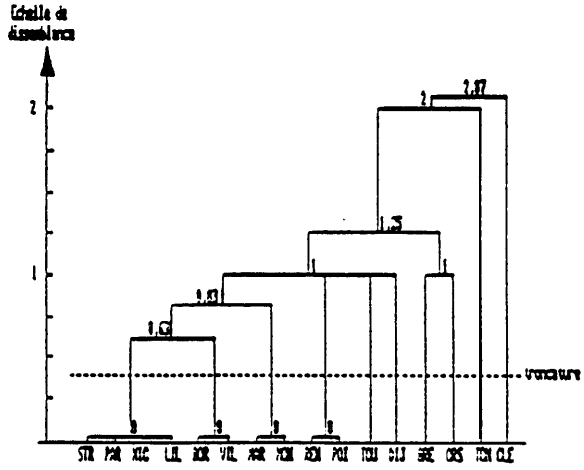
(9) La représentation vectorielle d'un tel tableau est fait soit dans l'espace des lignes soit dans l'espace des colonnes. On est donc dans l'obligation de représenter ces deux espaces (indissociables) pour connaître la totale information du tableau.

Les deux principales méthodes d'analyse factorielle utilisées en France sont l'analyse des composantes principales (ACP) et l'analyse factorielle des correspondances (AFC).

Seule la dernière méthode est applicable à notre problème. La matrice constituée a un caractère plus. proche d'un tableau de contingence que d ' un tableau variables-individus. Elle représente bien le croisement de deux variables définissant chacune une partition sur la population étudiée: chaque case du tableau équivaut aux références dénombrées ayant le caractère "ville" et le caractère "sections codes" commun. Ainsi dans une AFC les lignes et les colonnes jouent le même rôle. Elles sont successivement considérées comme les individus puis les variables d'une ACP, ce qui inhibe la transposition obligatoire effectuée entre les lignes et les colonnes qui nous ait imposé par ces méthodes (C f p 6).

Cette méthode mesure la ressemblance (ou dissemblance) en utilisant la distance du Khi-2, qui a l'avantage par rapport à la distance Euclidienne (de 1' ACP) , d'accéder à la représentation des individus des colonnes et des lignes dans un même espace. Ce qui offre l'opportunité de conserver les coordonnées des villes, en fin d'analyse d'AFC, pour enchaîner sur leur classification automatique(10). La répartition en groupe des villes est faite automatiquement (Figure ci-dessous), ce qui n'est pas pour nous déplaire car le classement d'objets à partir des seuls résultats d'une AFC n'est pas chose aisée.

La succession des deux méthodes permet d'une part d'obtenir l'arbre hiérarchique de la classification des villes et d'autre part de retrouver les sections codes les plus proches des classes d'une troncature. La classification automatique fournit le centre de gravité de chaque classe dans l'espace de 1'AFC. Il suffit de reporter ces points sur les graphes de 1'AFC pour estimer les sections les plus influentes pour chacune des classes.



Arbre hiéarchique sur les coordonnées de I'AFC

Il faut remarquer qu'une précaution a été prise dans 1'AFC pour un meilleur résultat. La ville de Grasse a été introduite en variable supplémentaire car elle était trop marginale auprès des autres villes. Le point de cette ville placé dans l'espace réduit calculé, déformait

(10) Pour I'ACP les variables ne sont pas représentables dans l'espace où sont positionnés les objets. Seules leurs directions dans cette espace sont connues.

le nuage de points et par conséquent réduisait et distordait le champ d'information (Cf p 39). Le principal centre d'intérêt des recherches en chimie de Grasse est facilement déduit de la première AFC. On peut considérer que Grasse constitue, à elle seule, une classe qui est à ajouter à la liste de celles déterminées par la classification.

La qualité de la classification automatique après l'AFC est très bonne. A la seule vue de la hiérarchie le classement des villes est évident: la troncature sur le dendrogramme ou le regroupement sur le graphe à plat s'effectuent sans aucune hésitation (Cf p. 41 et 42). Ce qui facilite l'interprétation.

Classe	Villes	Sections qui caractérisent la classe
1	Strasbourg, Nice Paris, Lille	2,3,6,9,13,14,15 23,26,32,40,66
2	Bordeaux, Lyon	19,22,23,25,40,65,80
3	Marseille, Montpellier	6,7,10,15,18,28 52,62,66
4	Rennes, Poitiers	20,27,29,48,57,61
5	Toulouse	12,27,48,51,62,78
6	Dijon	16,17,18,20,50 51,61,62
7	Grenoble	33,56,76,77
8	Orsay	65,70,73,74
9	Clermont	39,68,69
10	Toulon	5,20,50
10	Grasse	62

Tableau 3: Récapitulatif des résultats de l'AFC suivie de la Classification Automatique

IV. COMPARAISON DES RESULTATS

Si on se penche de plus près sur les résultats des différentes analyses certains paraissent contradictoires. En fait, l'enchaînement des méthodes, comme nous venons de le faire, n'a rien de recommandable. Il a été réalisé uniquement pour démontrer que toutes ces méthodes sont applicables à des traitements bibliométriques. L'intérêt de l'analyse des données est de fournir un condensé d'information interprétable rapidement. Il ne faut donc pas retomber dans l'excès inverse en cumulant les résultats de toutes les méthodes mises à sa portée. Car même si elles sont utilisées à bon escient, elles ne feront qu'apporter des vues différentes de l'information mais pas nécessairement complémentaires. On ne peut pas recréer une information "complète" en synthétisant des "résumés" d'information établis suivant des "rhétoriques" différentes.

Pour mieux comprendre ces affirmations, nous choisissons d'exposer un exemple assez flagrant. Etudions les résultats concernant la ville de Clermont-Ferrand.

- ❑ Pour l'analyse par classification automatique (ANNEXE 1 p 18), Clermont-Ferrand est positionné au sein d'un agrégat de sept villes fortement liées entre elles (Cf p 23 et 24): Rennes, Poitiers, Toulon, Grasse, Dijon, Nice et Clermont-Ferrand.

- ❑ Par contre la méthode de l'AFD (ANNEXE 2 p 25) marginalise la ville de Clermont-Ferrand des autres villes. Les graphes l'isolent nettement et la rattachent fortement aux deux sections principales 68 et 69 (Cf p 31 à 33).

- ❑ La méthode de l'AFC (ANNEXE 3 p 34) livre le même résultat que l'AFD pour cette ville (p 38 et 39). La classification automatique sur les coordonnées des villes dans l'espace de l'AFC affirme encore mieux cet isolement (p 41 et 42)

Que dire de ces deux tendances apparemment contradictoires ?

C'est à la confrontation de tels résultats qu'on ressent qu'une analyse trop rapide peut déboucher sur des conclusions totalement faussées.

En fait les résultats n'ont rien de contradictoires. Il faut toujours garder en tête que chaque méthode mathématique va favoriser l'émergence de certains caractères plus que d'autres.

Ainsi la classification automatique sur les données brutes, pour le critère et la stratégie d'agrégation employés, a principalement classé les villes par l'importance de leur taux de publications. Clermont-Ferrand, étant une ville qui publie assez peu (170) , est positionné dans le groupe des villes à faible taux de publications : Rennes (235), Poitiers (145), Toulon (12), Grasse(19), Dijon (228), Nice (183). Mais plus important, il est à remarquer que toutes ces villes décrivent le même phénomène: elles ont peu de publications mais elles ont toujours un domaine de la chimie concentrant la plus grande part des publications. Ce domaine prend donc une importance considérable, sauf s'il correspond à un domaine à fort taux de publications nationales (dans ce cas banalisation de son émergence dans l'ensemble). Donc, bien que chacune de ces villes ait des pôles d'activités différents, la classification automatique les a regroupées pour leurs phénomènes communs de marginalité.

Inversement les deux analyses factorielles ont moins tendance à représenter les marginalités, mais plutôt de concentrer l'information pour présenter le maximum de liens entre les éléments. C'est pour cette raison que, les éléments qui ont des caractères communs se regroupent, et que plus l'élément présente l'ensemble des caractères communs à tous plus il se rapproche de l'origine du repère de l'espace. Hors Clermont-Ferrand se détache de tout nuage de points formé, ce qui indique que son activité est foncièrement différente de celle de l'ensemble des villes. Ceci s'explique par le fait que Clermont-Ferrand ne publie pratiquement pas dans les domaines de la chimie qui sont sources de grandes quantités de publications, tandis que vingt pour-cent de ses publications concernent les sections principales 68 et 69 où il représente vingt-six pour-cent des publications nationales.

On peut donc imaginer qu'une analyse trop rapide des résultats d'une classification automatique aurait abouti à la conclusion erronée qui énoncerait Clermont-Ferrand comme faisant partie d'un groupe de villes ayant le même type d'activité scientifique.

Cet exemple permet de comprendre qu'il faille avant toute chose rechercher la méthode qui maximise les critères qu'on cherche à faire émerger de l'ensemble des données.

Notre étude n'a pas pour finalité de dégager l'innovation ou la marginalité de certains éléments mais on cherche plutôt une information organisée et faisant état de chacune des observations. Donc pour cette étude l'analyse factorielle des correspondances est celle qui répond le mieux à ces critères. La classification automatique sur les résultats de l'AFC doit alors être considérée comme une aide à l'interprétation des résultats.

L'UTILISATION DE CES METHODES DANS UN CENTRE D'I.S.T.

Chacune de ces méthodes est à prendre comme un outil à fonctionnalité pointue. Elles sont d'un emploi facile, lorsqu'on a pris connaissance de leurs limites et de leurs domaines d'application, car toutes sont automatisées par des systèmes informatiques. Leurs analyses demandent de grandes précautions et de solides connaissances. Mais sous le flot grandissant d'information, toute personne chargée de traiter ces flux, pour en dégager les renseignements vitaux à son entreprise, est dans l'obligation de manipuler ce genre de traitements d'information au quotidien. La profession de Veille Technologique fait appel à de nombreuses compétences. La Veille Technologique devient un métier à part entière et comme toute profession elle ne peut être mise en place et exécuté efficacement que par des spécialistes. La formation de spécialiste et le développement de systèmes informatiques de traitement de l'information donneront accès très prochainement aux moyennes entreprises à cette activité d'éveil envers son environnement.

[1] Maîtriser l'information critique
F. Jakobiak
Editions d'organisation, 1988

[2] La veille technologique, concurrentielle et commerciale
B. Martinet, J.M. Ribault
Editions d'organisation

[3] Veille technologique et information documentaire
De l'usage de la bibliométrie dans les services de documentation.
H. Dou, Parina Hassanaly, L. Quoniam, Albert La Tela
Documentaliste, vol.27, n°3, mai-juin 1990

[4] Infographic Analytical tools for decision makers
Analysis of the research production in sciences. Application to Chemistry.
Comparison between Marseille and Montpellier.
H. Dou, P. Hassanaly, L. Quoniam
Scientometrics 17 (1989)

[5] Teaching bibliométric Analysis and MS/DOS commands
H. Dou, Luc Quoniam and P. Hassanaly
Education for information 6 (1988) 411-423
North-Holland

[6] Les banques de données en Chimie
L'évolution au cours des dix dernières années
A. Deroulede, C. Dutheuil
Informations Chimie n°315, mars 1990

[7] Analyse des données (3ème édition)
M. Volle
Economia

[8] L'analyse des données
J.M. Bouroche et G. Saporta
Que sais-je, Presse Universitaires de France

[9] Classification et analyse des données
I.C. Lerman
Dunod

[10] Classification automatique des données
G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, H. Ralamboudrainy
Dunod informatique

[11] Algorithme de classification
M. Roux
Masson

[12] Comment interpréter les résultats d'une analyse factorielle Discriminante
STAT-ITCF
R. Tomassone
Institut Technique des Céréales et des Fourrages

[13] Comment interpréter les résultats d'une analyse en composantes principales
STAT-ITCF
G. Philipeau
Institut Technique des Céréales et des Fourrages

[14] Comment interpréter les résultats d'une analyse factorielle des correspondances
STAT-ITCF
R. Dervin
Institut Technique des Céréales et des Fourrages

ANNEXES 1

CLASSIFICATION AUTOMATIQUE SUR LES VILLES DE LA MATRICE GLOBALE

OPTIONS DEMANDEES

Classification Sur les colonnes

Classification Ascendante Hiérarchique

Distance Utilisée: Distance EUCLIDIENNE

Critère d'Agrégation: Moyenne des Distances Pondérées

********** CLASSIFICATION AUTOMATIQUE **********

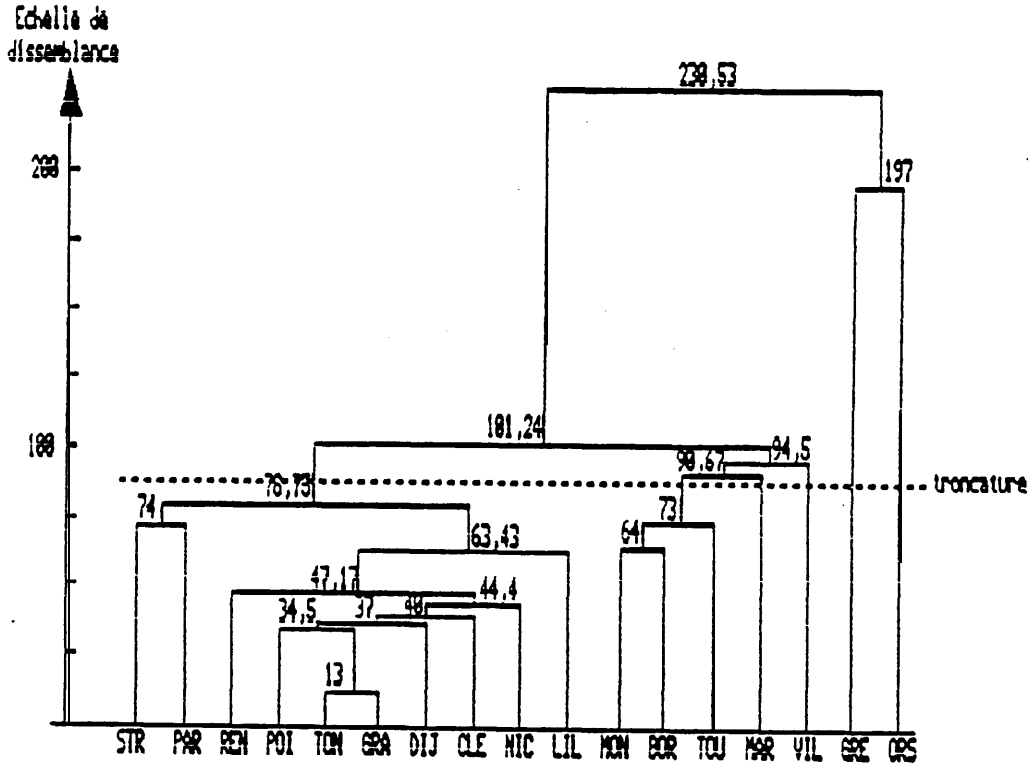
MATRICE DES DISTANCES

	STR	REN	POI	PAR	ORS	HIC	MON	LIL	GRE	CLE	TOU	TON	VIL	BOR	MAR	DIJ
REN	71															
POI	36	43														
PAR	74	77	88													
ORS	252	237	270	255												
HIC	77	49	46	56	259											
MON	90	96	108	78	240	98										
LIL	74	58	65	71	23s	56	81									
GRE	201	218	219	218	197	228	195	198								
CLE	84	58	48	71	268	42	99	61	218							
TOU	107	102	119	100	242	107	78	97	202	114						
TON	91	53	33	33	279	44	117	71	228	38	131					
VIL	113	121	129	99	217	114	93	98	211	128	101	144				
BOR	86	71	88	78	227	77	64	64	185	85	68	99	89			
MAR	83	91	97	86	222	88	02	73	197	93	108	106	95	82		
DIJ	83	48	35	77	266	44	101	61	218	42	111	33	126	81	9i	
GRA	92	54	36	84	279	46	110	72	228	48	133	13	146	101	198	38

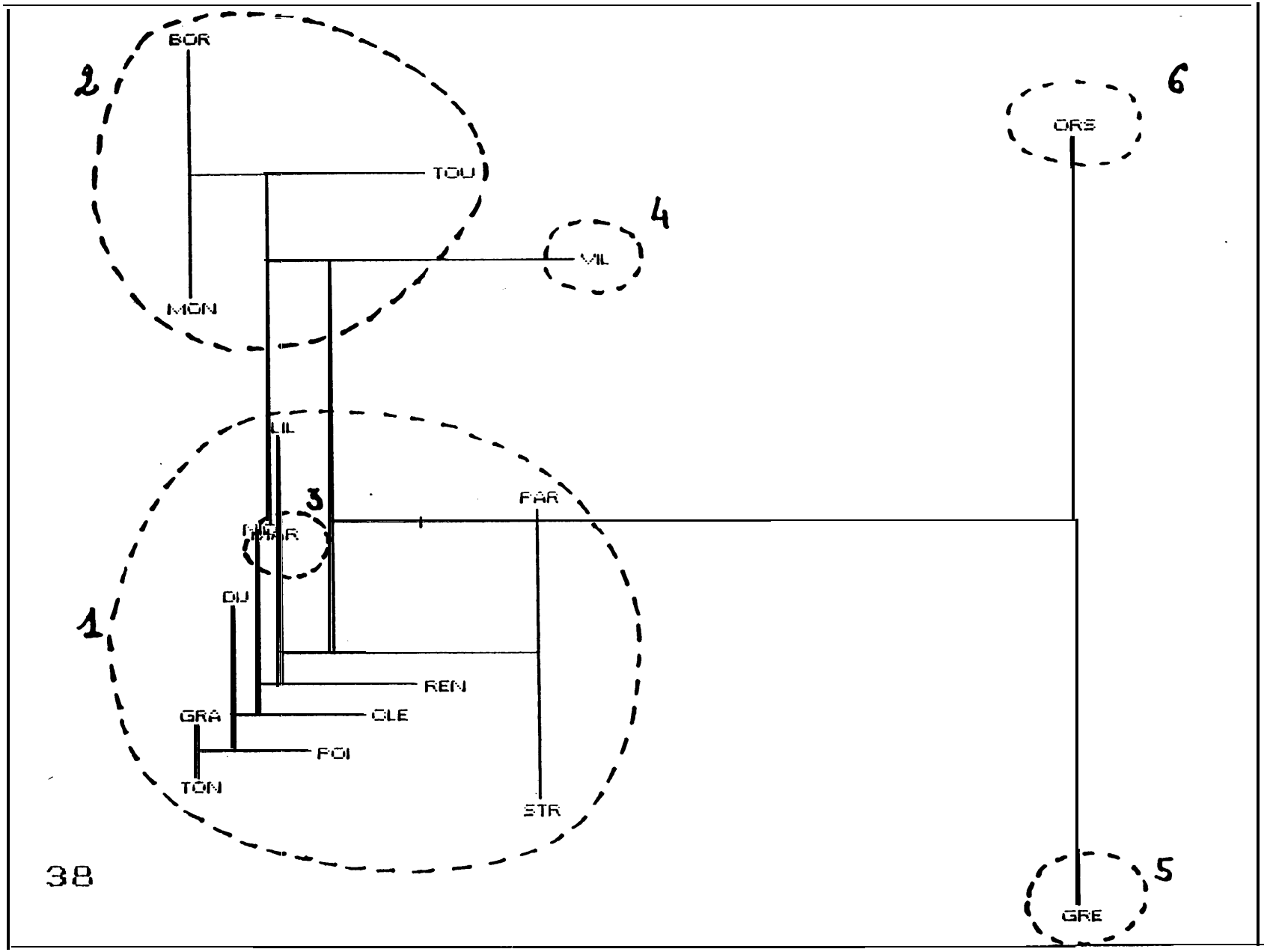
DESCRIPTION DE LA HIERARCHIE

NOEUDS	A INES	BENJAM.	POIDS	NIVEAUX
N°18	TON	GRA	2	13.00
N°19	POI	N° 18	3	34.50
N°20	N° 19	OIJ	4	37.00
N°21	N° 20	CLE	5	40.00
N°22	N° 21	HIC	6	44.40
N°23	REH	N° 22	7	48.17
N°24	N° 23	LIL	8	63.43
N°25	MON	BOR	2	64.00
N°26	N° 25	TOU	3	73.00
N°27	STR	PAR	2	74.0?
N°28	N° 27	N° 24	10	78.75
N°29	N° 26	MAR	4	98.57
N°30	N° 29	VIL	5	94.50
N°31	N° 20	HI 30	13	101.24
N°32	GRE	ORS	2	197.0?
N°33	N° 31	HI 32	17	230.53

ARBRE HIERARCHIQUE DES VILLES A PARTIR
DE LA MATRICE BRUTE GLOBALE



Arbre hiérarchique de la matrice brut
sertions principales et secondaires



REPRESENTATION A PLAT DE LA HIERARCHIE PRECEDENTE

CLASSIFICATION AUTOMATIQUE SUR LES VILLES DE LA MATRICE DES SECTIONS PRINCIPALES

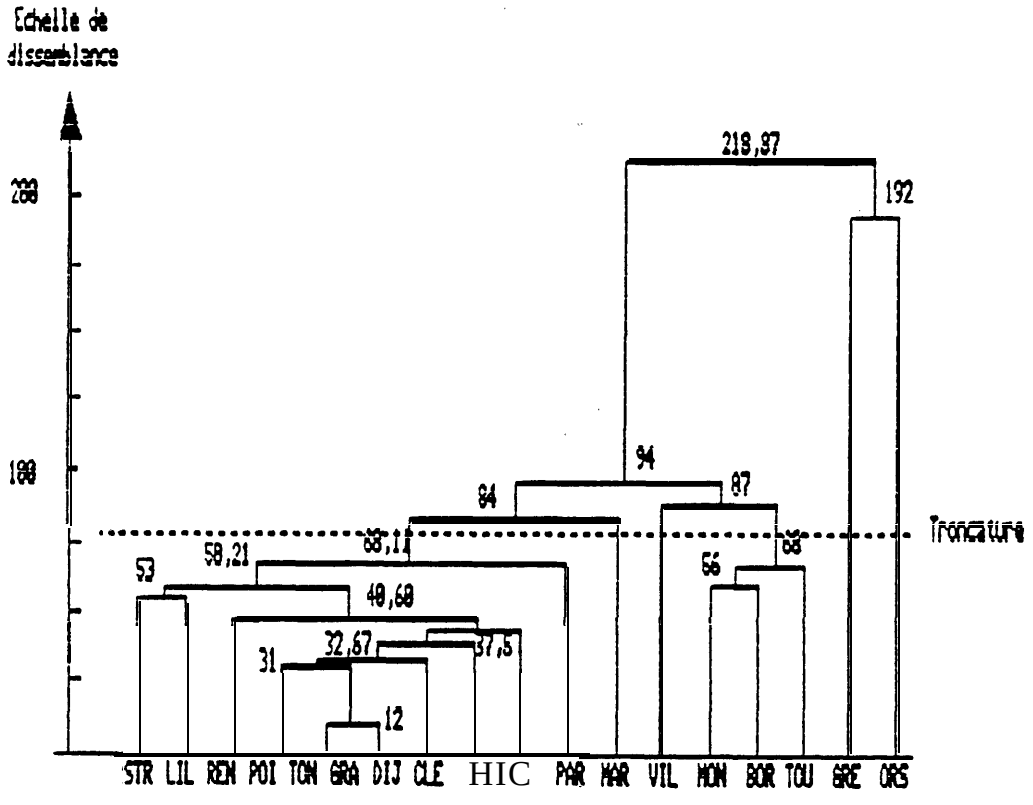
********** CLASSIFICATION AUTOMATIQUE **********

MATRICE DES DISTANCES

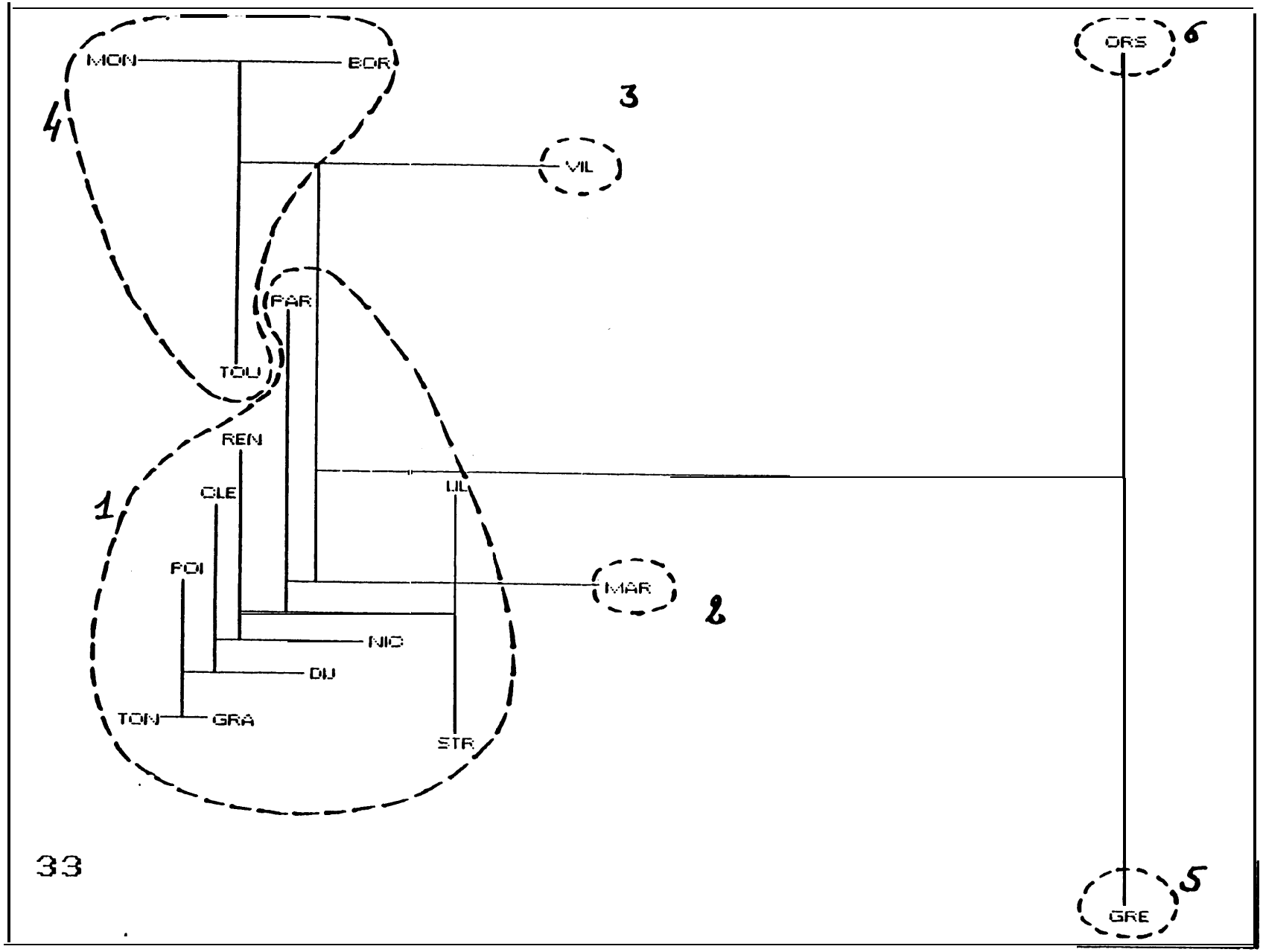
	STR	REN	POI	PAR	ORS	HIC	MON	LIL	GRE	CLE	TOU	TON	VIL	BOR	MAR	DIJ
REN	56															
POI	57	36														
PAR	58	71	73													
ORS	246	249	260	246												
HIC	53	43	41	53	249											
MON	82	88	96	69	234	82	.									
LIL	53	54	68	66	227	52	71									
GRE	188	196	291	282	192	292	181	183								
CLE	56	45	37	66	239	39	87	58	281							
TOU	99	98	111	93	237	100	73	92	198	106						
TON	60	46	38	77	267	40	105	66	219	36	122					
VIL	106	114	128	90	211	105	85	98	199	119	95	134				
BOR	73	67	a1	71	221	71	56	50	171	71	63	92	81			
MAR	74	a4	a9	ai	214	ai	77	66	183	a5	193	97	98	79		
DIJ	56	37	31	71	250	40	92	58	283	39	104	33	119	77	as	
GRA	62	47	32	70	268	43	106	67	210	30	124	12	136	94	90	34

NOEUDS	AINES	BENJAM.	POIDS	NIVEAUX
N018	TON	GRA	2	12.00
N019	WI	N0 18	3	31.00
N020	N0 19	DIJ	4	32.67
N021	N0 29	CLE	5	37.59
N022	NI 21	HIC	6	40.60
N023	REN	N0 22	7	42.33
N024	STR	LIL	2	53.00
N025	MON	BOR	2	56.00
N026	N0 24	N0 23	9	58.21
N027	N0 25	TOU	3	60.00
N028	N0 26	PAR	10	60.11
N029	N0 28	MAR	11	64.00
N030	VIL	N0 27	4	87.00
N031	N0 29	N# 38	13	94.00
N032	GRE	ORS	2	192.00
N033	N0 31	NI 32	17	210.07

ARBRE **HIERARCHIQUE DES VILLES A PARTIR DE LA MATRICE DES SECTIONS PRINCIPALES**



Arbre hiérarchique de la matrice brute des sections principales



REPRESENTATION A PLAT DE LA HIERARCHIE PRECEDENTE

ANNEXES 2

CLASSIFICATION AUTOMATIQUE SUR LES SECTIONS DE LA MATRICE DES SECTIONS PRINCIPALES

HIERARCHIE DECOUPEE EN 15 CLASSES

!N°CLAS!	EFFECTIF	! DESCRIPTION DES CLASSES
! 1 !	43	! P1 P2 P4 P6 P7 P9 P10 P11 P12 P13 P14 P15 P18 P19.
!		! P21 P22 P23 P25 P27 P28 P29 P32 P37 P38 P40 P42 P45 P48
!		! P51 PS2 PS3 PS7 PS0 PS9 P60 P61 P62 P63 P66 P67 P72 P78
!		! P80
! 2 !	18	! PS P8 P24 P26 P30 P33 P36 P47 PS5 P56 P65 P70 P71 P73
!		! P74 P75 P76 P79
! 3 !	3	! PS P17 P20
! 4 !	1	! P16
! 5 !	1	! P31
! 6 !	1	! P34
! 7 !	1	! P35
! 8 !	2	! P39 P69
! 9 !	1	! P43
! 10 !	2	! P46 P49
! 11 !	1	! P58
! 12 !	1	! PS4
! 13 !	1	! P64
! 14 !	1	! P68
! 15 !	1	! P77

AIDE A L'INTERPRETATION DE LA PARTITION

CONTRIBUTIONS DE VARIABLES QUANTITATIVES

VARIAHCE TOTALE = 73

VARIANCE INTERCLASSE = 52.27
INTER/TOTAL = 8.71

ANALYSE FACTORIELLE DISCRIMINANTE DR LA MATRICE DES SECTIONS PRINCIPALES

NOMBRE D'OBSERVATIONS : 78 NOMBRE DE-VARIABLES : 18

NOMBRE DE GROUPES : 18
NOMBRE DE VARIABLES QUANTITATIVES : 17

NOMBRE D ' AXES DEMANDES : 5

ETUDE PAR GROUPE

GROUPE	EFFECTIF	VARIABLES	MOYENNES	ECARTS-TYPES DES SERIES
1 (11	43	STR	4,116	5.863
		REN	4,870	4,627
		POI	2,442	3.142
		PAR	5,878	8.940
		ORS	6.209	6.810
		NIC	2,884	4,947
		MON	8.558	10.906
		LIL	4.698	5,471
		GRE	4.558	5.986
		CLE	1,977	3.246
		TOU	18.721	12.829
		TON	8.186	9,691
		VIL	11,868	12,891
		BOR	6,791	8.351
		MAR	7,558	8.994
		DIJ	2,442	3.194
		GRA	8,279	1.513
2 (21	18	STR	5,444	7,967
		REN	2,722	4,628
		POI	2.000	3.266
		PAR	5.111	6.624
		ORS	37.000	48,431
		HIC	2,111	3,693
		MON	6.500	9,500
		LIL	3,722	9.700
		GRE	24,444	31.897
		CLE	1.611	2,831
		TOU	7,889	11.005
		TON	0.056	0.229
		VIL	11.222	12.634
		BOR	7,444	L1.931

		MAR	8.111	11.175
		DIJ	1.667	2.683
		GRA	8,111	0.316
(3)	5	STR	8.290	0.400
		REN	1.280	1,939
		POI	0.000	0.000
		PAR	1.400	1.744
		ORS	0.400	8,498
		NIC	8.280	0.400
		MON	10.280	5.776
		LIL	1.400	1.744
		GRE	0.000	0.748
		CLE	1.480	1.969
		TOU	2.000	2.713
		TON	0.400	0.000
		VIL	1.000	1.833
		BOR	3.000	3.833
		MAR	5.000	4.684
		DIJ	6.200	3.036
		GRA	1.000	2.000
4 (4)	1	STR	0.000	0.000
		REN	1.000	0.000
		POI	0.000	0.000
		PAR	0.000	0.000
		ORS	0.000	0.000
		NIC	0.000	0.000
		MON	0.000	0.000
		LIL	0.000	0.000
		GRE	0.000	0.000
		CLE	4.000	0.000
		TOU	1.000	0.000
		TON	0.000	0.000
		VIL	2.000	0.000
		BOR	0.000	0.000
		MAR	0.000	0.000
		DIJ	0.000	0.000
		GRA	0.000	0.000
5 (5)	1	STR	3.000	0.000
		REN	0.000	0.000
		POI	0.000	0.000
		PRR	3.000	0.000
		ORS	1.000	0.000
		HIC	1.000	0.000
		MON	7.000	0.000
		LIL	2.000	0.000
		GRE	7.000	0.000
		CLE	5.000	0.000
		TOU	1.000	0.000
		TON	0.000	0.000
		VIL	3.000	0.000
		BOR	0.000	0.000
		MAR	1.000	0.000
		DIJ	0.000	0.000
		GRA	0.000	0.000
6 (6)	3	STR	2.000	1.633
		REN	8.667	8,943
		POI	8,333	0.471
		PAR	5.333	6.848

		ORS	3.333	1.700
		NIC	0.667	8,943
		MON	2.667	2.357
		LIL	1.333	1.247
		GRE	4.667	4.643
		CLE	11.333	7.846
		TOU	1.000	*0.816*
		TON	0.000	*8,888*
		VIL	4.667	3.859
		BOR	*0.667*	*0.943*
		MAR	5.333	3.859
		DIJ	0.333	0.471
		GRA	0.000	0.000
7 (7)	2	STR	7.000	7.000
		REN	0.500	0.500
		POI	0.000	0.000
		PAR	1.000	0.000
		ORS	21.000	21.000
		HIC	1.000	1.000
		MON	4.500	4.500
		LIL	0.000	0.000
		6RE	61.000	50.000
		CLE	a.508	0,598
		TOU	2.000	1.000
		TON	0.000	0.000
		VIL	1.000	1.000
		EOR	7.500	6.500
		MAR	2.598	1.598
		DIJ	0.000	0.000
		GRA	0.000	0.000
8 (8)	2	STR	0.000	0.000
		REN	0.000	0.000
		POI	0.000	0.000
		PAR	0.000	0.000
		ORS	0.000	0.000
		HIC	0.500	0.500
		MON	0.000	0.000
		LIL	0.000	0.000
		GRE	0.500	a.500
		CLE	0.000	0.000
		TOU	2.000	0.000
		TON	0.000	0. 000
		VIL	0.000	0.000
		BOR	0.000	0.000
		MAR	0.000	0.000
		DIJ	0.000	0.000
		GRA	0.000	0.000
9 (9)	1	STR	0.000	0.000
		REN	0.000	0.000
		POI	3.000	0.000
		PAR	0.000	0.000
		ORS	0.000	0.000
		HIC	0.000	0.000
		MON	0.000	0.000
		LIL	1.000	0.000
		GRE	0.000	0.000
		CLE	0.000	0.000
		TOU	0.000	0.000
		TON	1.998	0.000

		VIL	0.000	0.000
		BOR	0.000	0.000
		MAR	3.000	0.889
		013	1.000	0.008
		GRA	0.000	0.000
10 (10)	2	STR	0.000	0.000
		REN	0.000	0.000
		POI	0.000	0.000
		PAR	1.000	0.000
		ORS	0.000	0.000
		NIC	0.500	0.500
		non	1.500	0.500
		LIL	1.500	0.500
		GRE	0.000	8.888
		CLE	0.000	0.000
		TOU	0.000	0.000
		TON	0.000	0.000
		VIL	9,580	0.500
		BOR	0.000	0.000
		MAR	1,580	1.500
		DIJ	8.808	0.000
		GRA	0.000	8.888

Axe	Valeur propre	Inntir	Pseudo F	WILKS	ddl	Proba Z	Corrél
1	1.4067	37.7%	11.23	174.40	153	11.35	0.5979
2	0.9943	23.2%	7.51	116.53	128	75.69	0.4986
3	0.9114	23.1%	6.89	72.72	105	99.30	0.4768
4	0.3159	8.0%	2.39	31.58	84	100.00	0.2401
5	0.1427	3.6%	1.08	14.15	65	0.00	a.1249

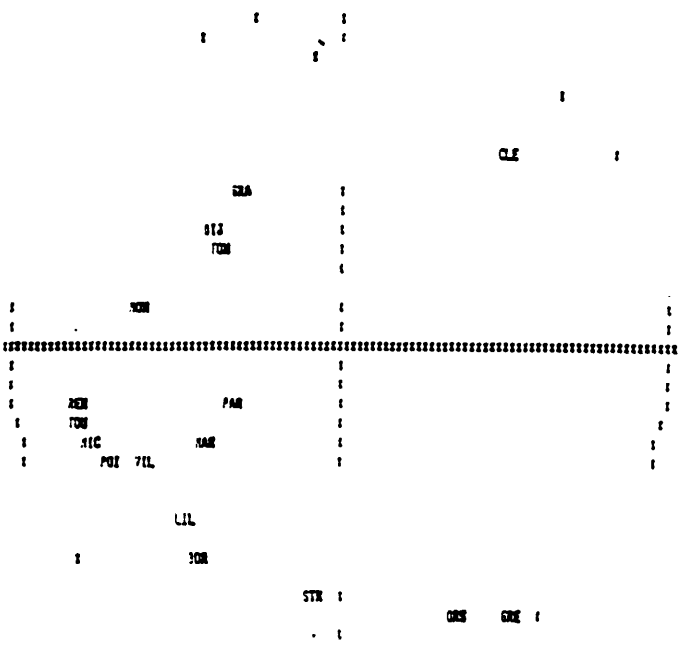
ETUDE DES CENTRES DE GRAVITE DES GROUPES

Pour chaque AXE :
IRE COLONNE : COORDONNEES DES INDIVIDUS SUR LES AXES DISCRIMINANTS
2E COLONNE : COSINUS CARRES (QUALITE DE LA REPRESENTATION)

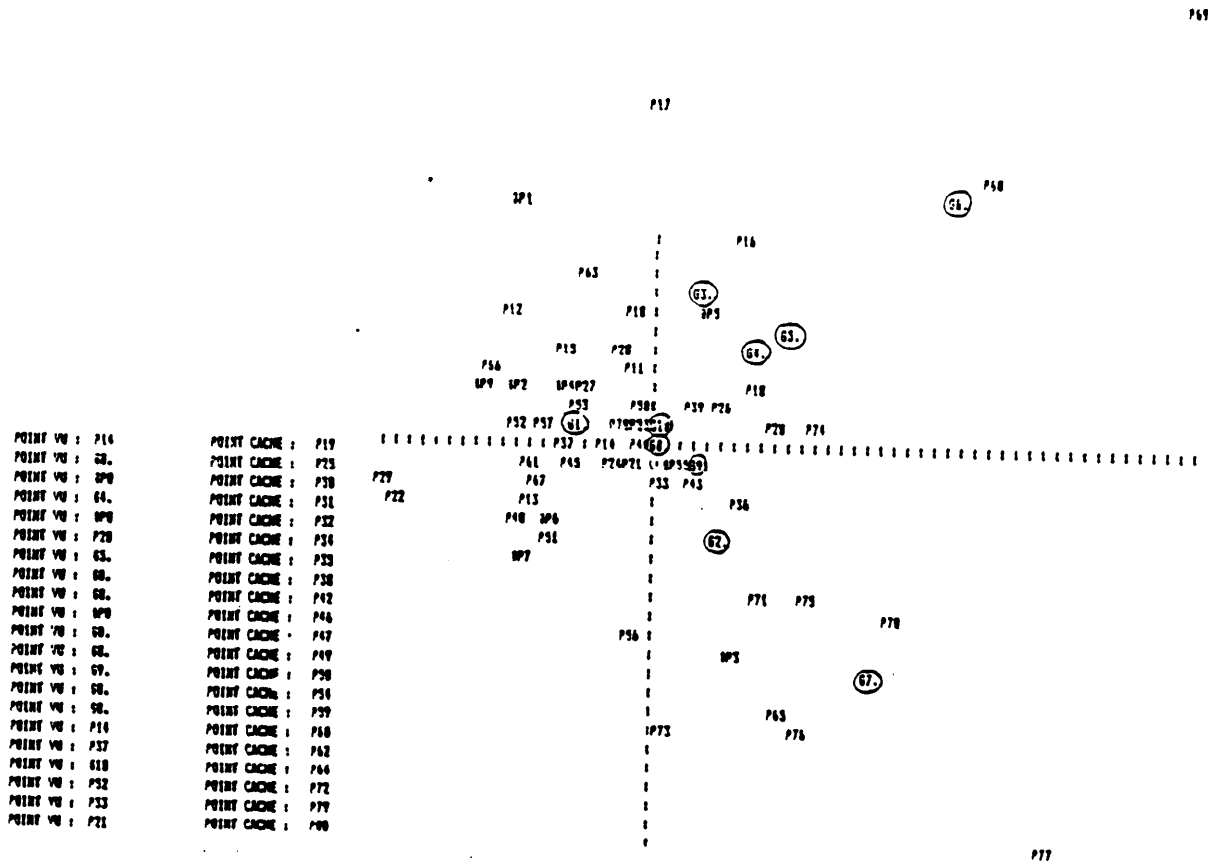
GROUPE	AXE 1		AXE 2		AXE 3		AXE 4		AXE 5	
1(1)	-0.5884	0.8163	0.1217	0.6349	0.2072	0.1012	-0.1255	0.0371	-0.0519	0.0063
2(2)	0.5554	0.2327	-0.8328	a.5232	-0.8433	0.0014	0.5477	fi.2263	-0.1448	0.0158
3(3)	0.4873	0.0245	1.2392	8.2268	-2.2323	0.7261	0.0688	0.0007	-0.2743	0.0111
4(4)	fi.8283	0.2867	8.7382	1.2277	0.6187	0.1558	-0.0594	0.0015	0.3790	0.0608
5(5)	1.0692	8.3277	0.0281	8.1920	a.1472	0.0062	-0.1058	0.0032	-0.0002	0.0000
6(6)	2.4013	0.4748	1.0899	8.2936	1.6635	0.2275	-0.0145	0.0000	-0.0716	0.9884
7(7)	1.7387	0.2392	-1.7468	0.2414	-0.5666	8.8234	-2.4946	0.4924	-0.0003	0.0005
8(8)	0.0778	0.0039	-9.8797	0.0041	-a.8794	8.0841	-0.0356	0.0008	0.7841	0.6269
9(9)	0.4185	0.0213	-0.1688	0.0036	-0.9698	9.1192	8.4917	0.0284	2.1318	0.5759
10(10)	0.0625	0.0023	0.0771	0.0035	-0.2282	0.3385	0.8324	0.0006	1.8334	8.6584

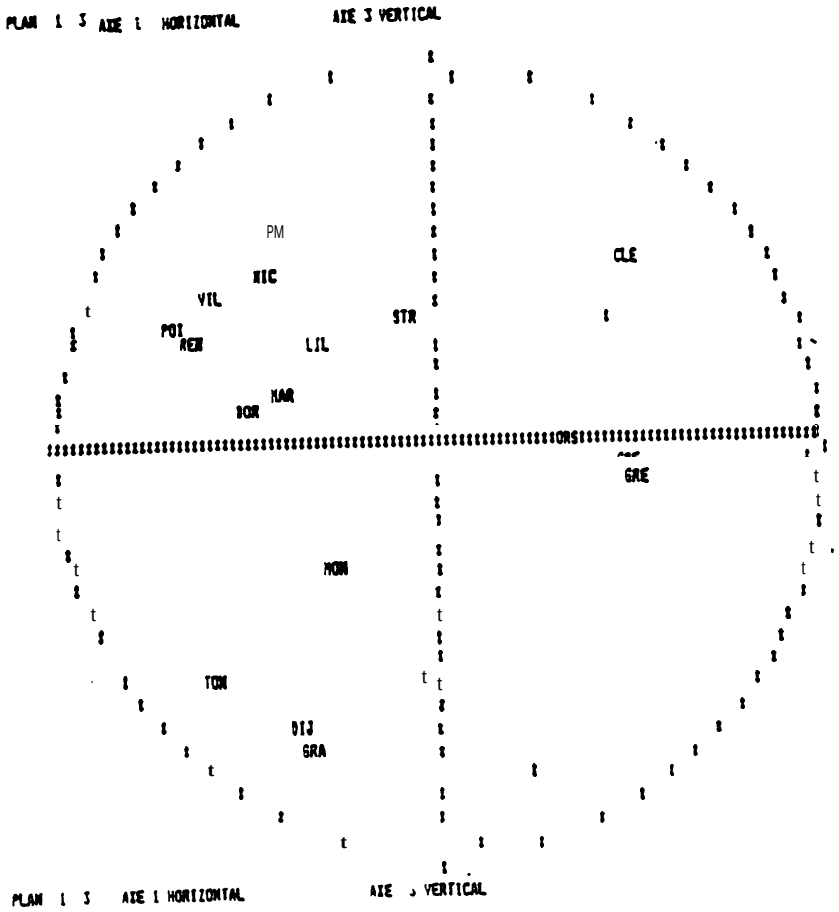
CERCLE DE CORRELATION

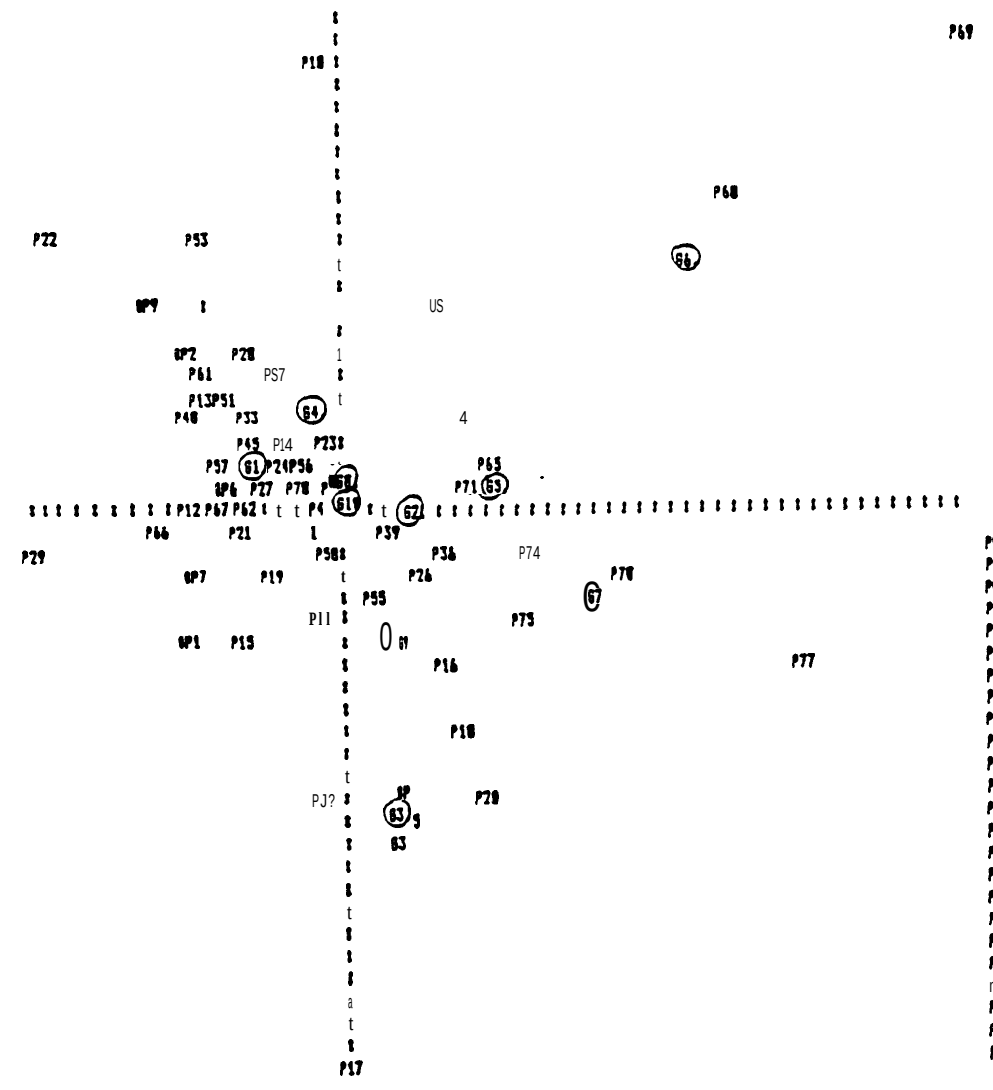
PLAN : 2 AXE 1 HORIZONTAL AXE 2 VERTICAL



PLAN : 1 2 AXE 1 HORIZONTAL AXE 2 VERTICAL



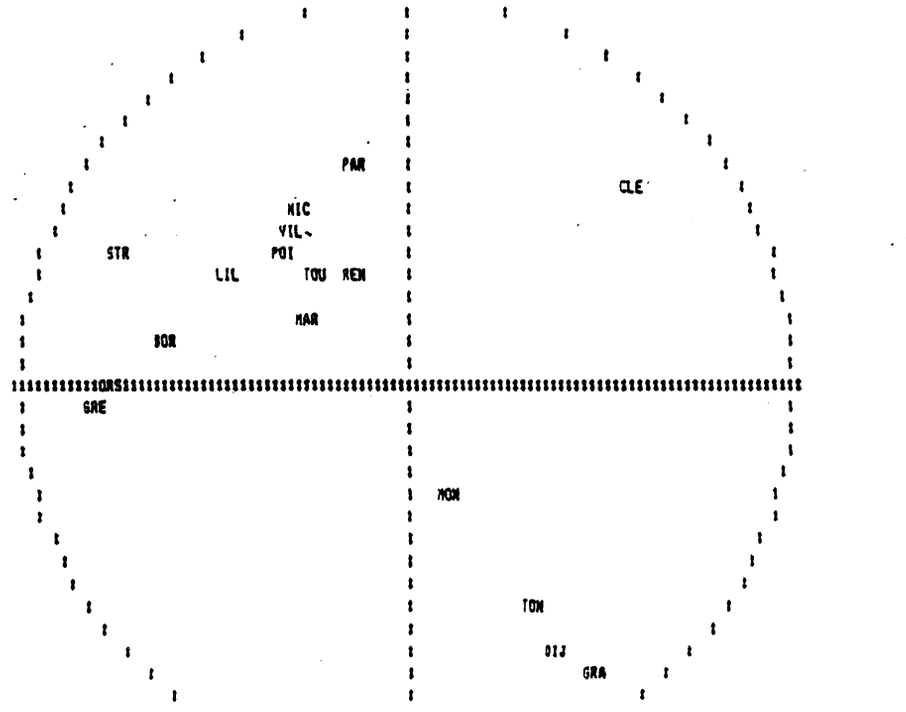




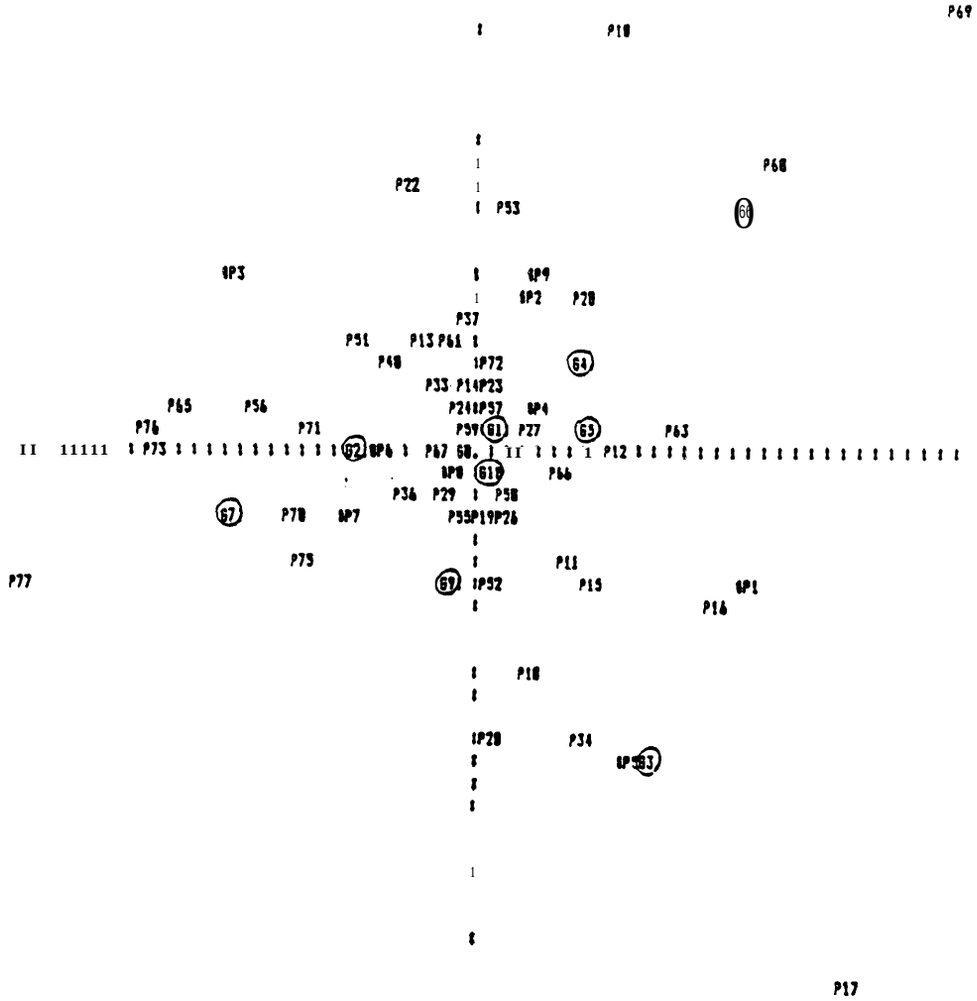
POINT VU : 61.
POINT VU : 618
POINT VU : 618
POINT VU : 618
POINT VU : 64.
POINT VU : 68.
POINT VU : 65.
POINT VU : 618
POINT Vu : 618
POINT VU : P39
POINT Vu : 68.
POINT VU : 618
POINT VU : a .
POINT VU : 69.
POINT Vu : 0P1
POINT VU : 618
POINT VU : P U
POINT Vu : 61.
POINT VU : 618
POINT VU : P48
m a t VU : 68.
POINT VU : 65.
POINT VU : 68.
POINT VU : P48

POINT CACHE : 0P4
POINT CACHE : 0P8
POINT CACHE : P25
POINT CACHE : P30
POINT CACHE : P31
POINT CACHE : P S 2
POINT CACHE : P35
POINT CACHE : P38
POINT CACHE : P42
P O I N T CACHE : P43
POINT CACHE : P46
POINT CACHE : P47
POINT CACHE : P49
POINT CACHE : P50
POINT CACHE : P52
POINT CACHE : P54
POINT CACHE : P60
POINT CACHE : P63
POINT CACHE : P64
POINT CACHE : P72
POINT CACHE : P73
POINT CACHE : P76
POINT CACHE : P79
POINT CACHE : P80

PLAN 2 3 AIE 2 HORIZONTAL AIE 3 VERTICAL



PLAN 2 3 AXE 2 HORIZONTAL AXE 3 VERTICAL



POINT VU : 6P6	POINT CACHE : P21
POINT VU : 610	POINT CACHE : P25
POINT VU : 6P8	POINT CACHE : P38
POINT VU : 64.	POINT CACHE : P31
POINT VU : 68.	POINT CACHE : P32
POINT VU : 65.	POINT CACHE : P35
POINT VU : 610	POINT CACHE : P38
POINT VU : 610	POINT CACHE : P39
POINT VU : 60.	POINT CACHE : P40
POINT VU : 610	POINT CACHE : P42
POINT VU : 6P0	POINT CACHE : P43
POINT VU : P23	m n CACHE : P45
POINT vu : 68.	POINT CACHE : P46
POINT vu : 610	POINT CACHE : P47
POINT VU : 60.	POINT CACHE : P49
POINT vu : 69.	POINT CACHE : P50
POINT VU : 610	POINT CACHE : P54
POINT VU : P57	ta111 CACHE: P60
POINT vu : 60.	POINT CACHE : P62
POINT VU : 610	POINT CACHE : P64
POINT vu : P58	POINT CACHE : P74
POINT vu : 61.	POINT CACHE : P78
POINT Y : P67	POINT CACHE : P79
POINT VU : 60.	POINT CACHE : P80

ANNEXE 3

. ANALYSE FACTORIELLE DES CORRESPONDANCES
DE LA MATRICE DES SECTIONS PRINCIPALES

NOMBRE DE VARIABLES (Colonnes) ACTIVES OU TABLEAU : 17
NOMBRE DE VARIABLES (Colonnes) SUPPLEMENTAIRES : 0

NOMBRE D' AXES DEMANDES : 5

VALEURS PROPRES ET VECTEURS PROPRES

1ère LIGNE : VALEURS PROPRES (VARIANCES SUR LES AXES PRINCIPAUX)
2ème LIGNE : CONTRIBUTION A L'INERTIE TOTALE (POURCENTAGES EXPLIQUES PAR LES AXES PRINCIPAUX)

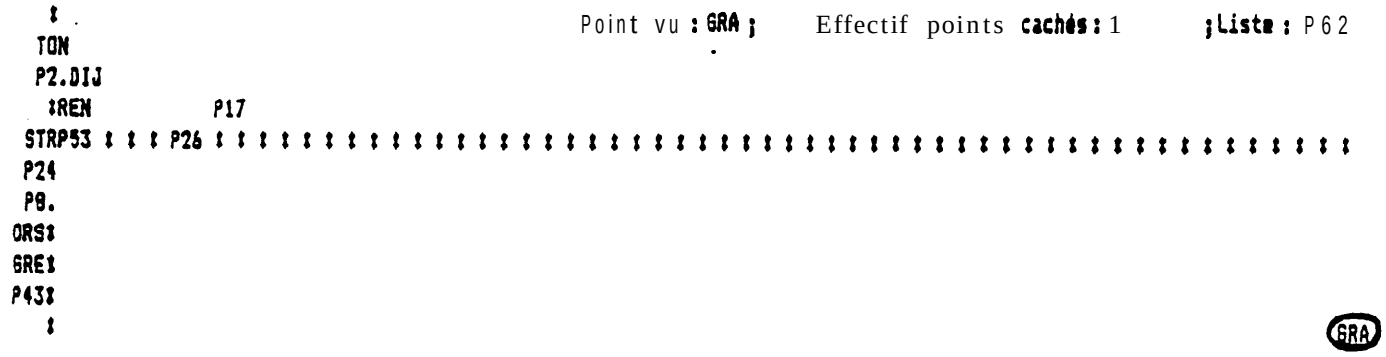
0.3163	0.2326	0.1073	8.3896	0.0730
25.9 %	19.8 %	8.6 %	7.3 %	6.0 %

VECTEURS PROPRES (COEFFICIENTS DES VARIABLES DANS L'EQUATION LINEAIRE DES AXES PRINCIPAUX)

STR	-0.1331	8.2279	8.9936	0.3015	-1.3670
REN	0.2114	0.6719	-1.8644	-0.8125	1.7160
POI	-0.0296	0.5913	-2.855%	-0.7362	1.6481
PAR	-0.9921	0.7431	1.4676	0.4677	-1.2237
ORS	-0.2888	-1.2869	0.8640	-1.1131	0.9314
HIC	-0.0441	0.7658	0.9751	-0.4588	-0.4604
MON	0.2116	0.6485	0.0665	0.7374	-0.7311
LIL	0.0428	0.4085	8.6091	-0.0384	-0.4738
GRE	-0.3148	-1.9083	-1.0507	1.0414	-0.8346
CLE	0.0271	8.7251	8.7726	4.5068	3.6744
TOU	0.0651	8.0206	-1.2416	-0.2787	-0.6240
TON	0.0356	1.7068	-1.2883	-1.0000	0.9866
VIL	-0.0926	0.6317	0.5672	-0.5365	8.0973
BOR	-0.0005	0.3766	-0.6683	-0.3917	0.0507
MAR	0.0902	0.2739	4.6356	0.2757	0.3512
DIJ	0.4278	1.1311	-1.6525	-0.6649	0.4037
GRA	18.2297	-2.9148	0.7166	-0.3709	-0.0353

REPRESENTATION SIMULTANEE DES LIGNES (Observations) ET COLONNES (Variables)

PLAN 1 2 AXE 1 HORIZONTAL AXE 2 VERTICAL



ANALYSE FACTORIELLE DES CORRBSPONDANCRS DE LA MATRICE DES *SECTIONS* PRINCIPALES AVEC GRASSE EN VARIABLE SUPLEMENTAIRE

NOMBRE DE VARIABLES (Colonnes) ACTIVES DU TABLEAU : 16
NOMBRE DE VARIABLES (Colonnes) SUPPLEMENTAIRES: 1

NOMBRE D' AXES DEMANDES: 5

VALEURS PROPRES ET VECTEURS PROPRES
****

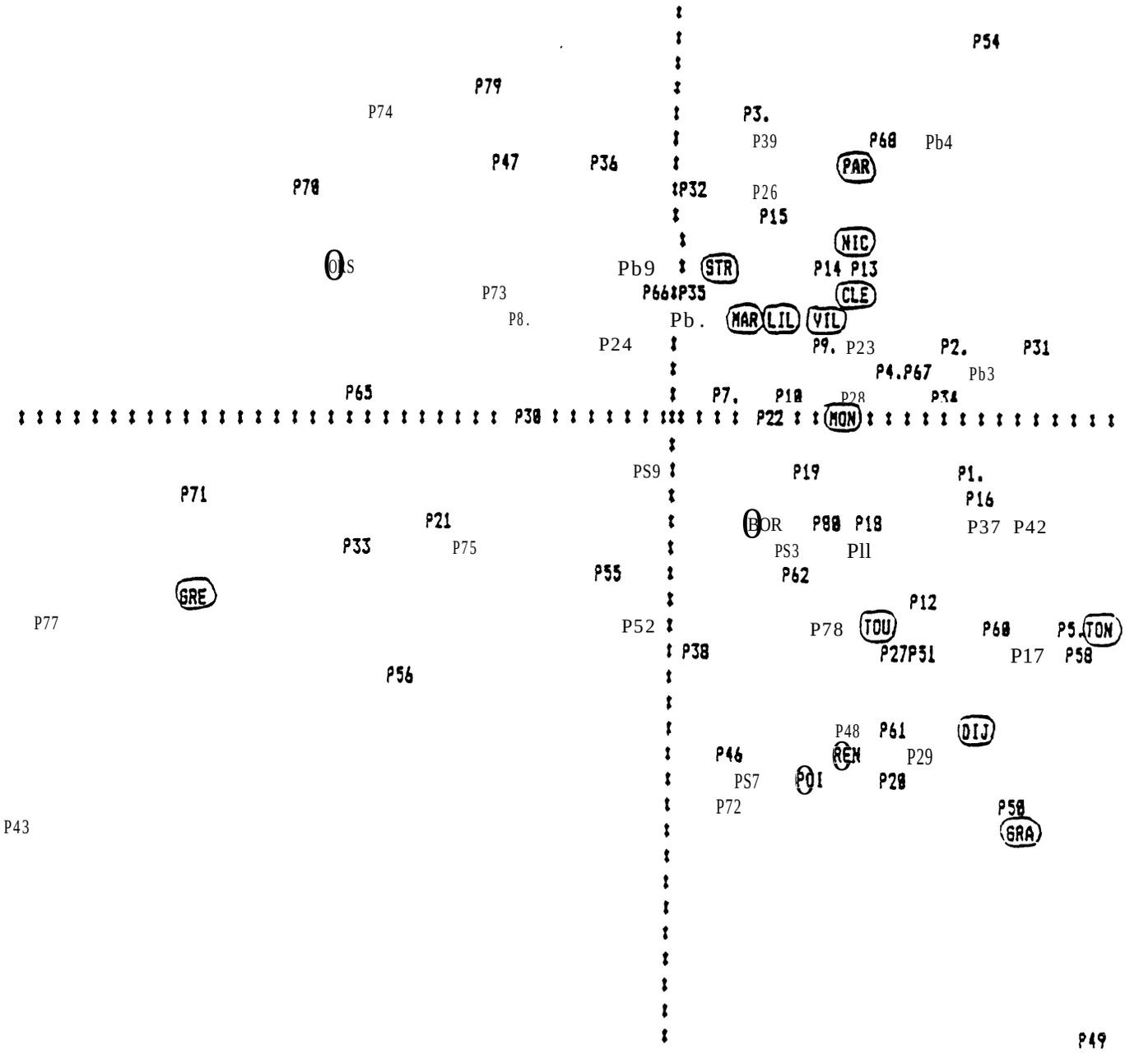
1ère LIGNE: VALEURS PROPRES (VARIANCES SUR LES AXES PRINCIPAUX)
2ème LIGNE: CONTRIBUTION A L'INERTIE TOTALE (POURCENTAGES EXPLIQUES PAR LES AXES PRINCIPAUX)

0.2358	0.1878	0.0980	0.0738	0.0638
25.7 Z	11.8 z	9.9 z	8.0 Z	6.9 Z

VECTEURS PROPRES (COEFFICIENTS DES VARIABLES DANS L'EQUATION LINEAIRE DES AXES PRINCIPAUX)

STR	8.1099	0.9960	0.2922	1.3772	-0.3857
REN	0.7065	-1.0714	-0.0344	-1.7027	-0.8120
POI	0.5729	-2.0041	-0.7780	-1.6744	0.5375
PAR	0.7251	1.4736	0.4034	1.2129	-0.6214
ORS	-1.3189	0.8666	-1.0927	-0.9259	-0.1783
NIC	0.7265	1.0213	-0.4643	0.5020	-1.1964
MON	0.6920	0.0243	0.7675	0.7208	0.9697
LIL	0.4110	0.6031	-0.0120	0.4073	0.3445
GRE	-1.9240	-1.0753	1.0202	0.0323	0.0074
CLE	0.7137	0.7613	4.4811	-3.6565	-1.4355
TOU	8.8209	-1.1977	-4.3095	0.6439	-1.5704
TON	1.7000	-1.2963	-1.0300	-1.0661	6.6034
VIL	8.5934	0.6001	-0.5609	-0.0036	-0.2323
BOR	0.3495	-0.6422	-4.4162	-0.0529	0.0106
MAR	0.2915	0.5920	8.3832	-0.3491	1.7633
DIJ	1.2234	-1.7256	-8.6178	-4.4276	207323

REPRESENTATION SIMULTANEE DES LIGNES (Observations) ET COLONNES (Variables) ###
PLAN 1 2 AXE 1 HORIZONTAL AXE 2 VERTICAL



Point vu :P7.; Effectif points cachés : 1 ; Liste :P25

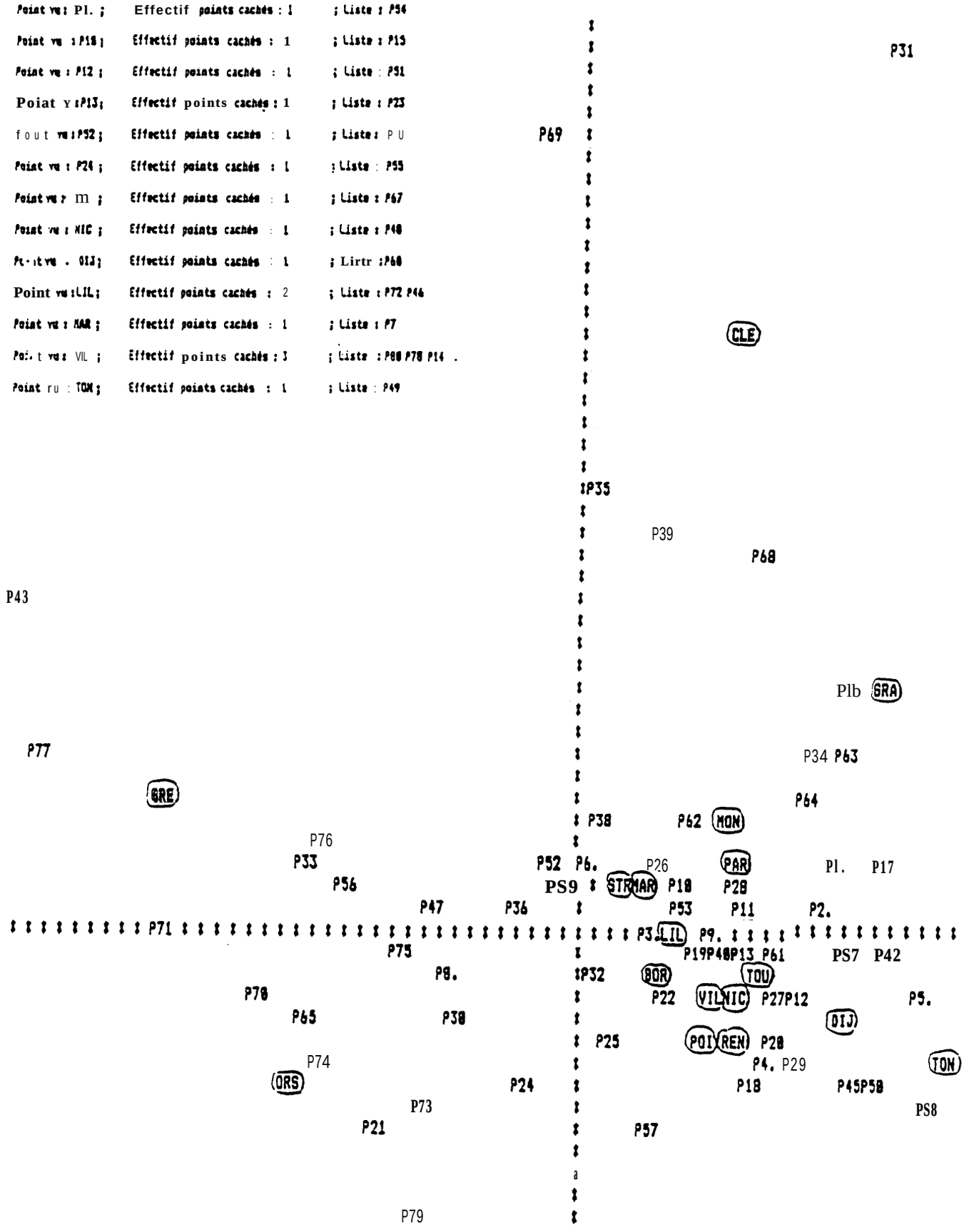
Point vu :P33; Effectif points cachés : 1 ; Liste : P76

Point vu :P16; Effectif points cachés : 1 ; Liste :P45

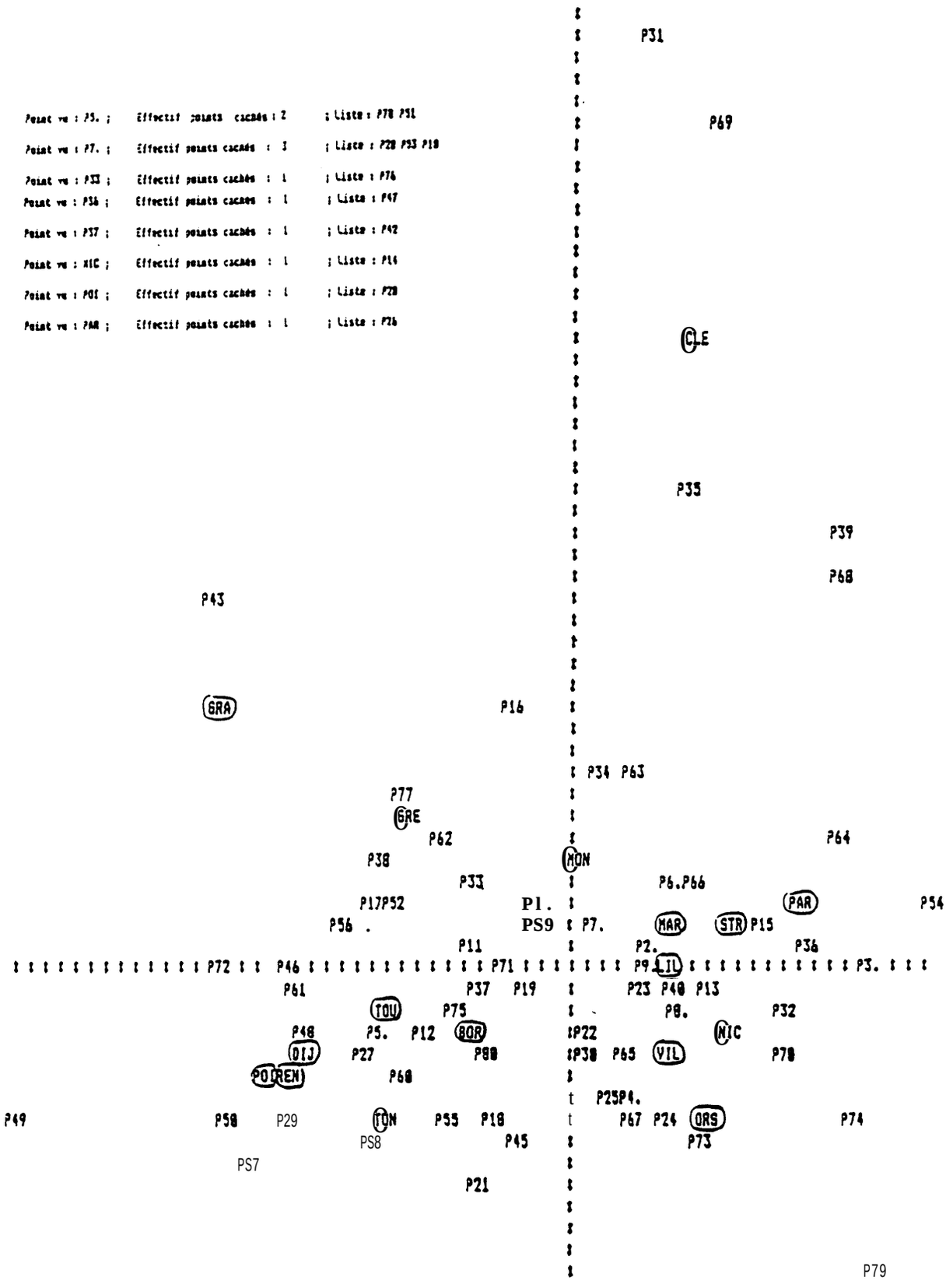
Point vu :VIL; Effectif points cachés : 1 ; Liste :P48

REPRESENTATION SIMULTANEE DES LIGNES (Observations) ET COLONNES (Variables) ***
PLAN 1 3 AXE 1 HORIZONTAL AXE 3 VERTICAL

Poiat vu : Pl. ; Effectif points cachés : 1 ; Liste : P56
Poiat vu : P18 ; Effectif points cachés : 1 ; Liste : P15
Poiat vu : P12 ; Effectif points cachés : 1 ; Liste : P51
Poiat vu : P13 ; Effectif points cachés : 1 ; Liste : P23
fout vu : P32 ; Effectif points cachés : 1 ; Liste : P U
Poiat vu : P24 ; Effectif points cachés : 1 ; Liste : P55
Poiat vu : m ; Effectif points cachés : 1 ; Liste : P67
Poiat vu : NIC ; Effectif points cachés : 1 ; Liste : P40
Point vu : DIJ ; Effectif points cachés : 1 ; Liste : P60
Point vu : LIL ; Effectif points cachés : 2 ; Liste : P72 P46
Poiat vu : NAR ; Effectif points cachés : 1 ; Liste : P7
Point vu : VIL ; Effectif points cachés : 3 ; Liste : P60 P70 P14 .
Poiat vu : TON ; Effectif points cachés : 1 ; Liste : P49



REPRESENTATION SIMULTANEE DES LIGNES (Observations) ET COLONNES (Variables)!!!
PLAN 2 3 AIE 2 HORIZONTAL AIE 3 VERTICAL



CLASSIFICATION AUTOMATIQUE SUR LES COORDONNEES DES VILLES CALCULES PAR L'AFC PRECEDANTE

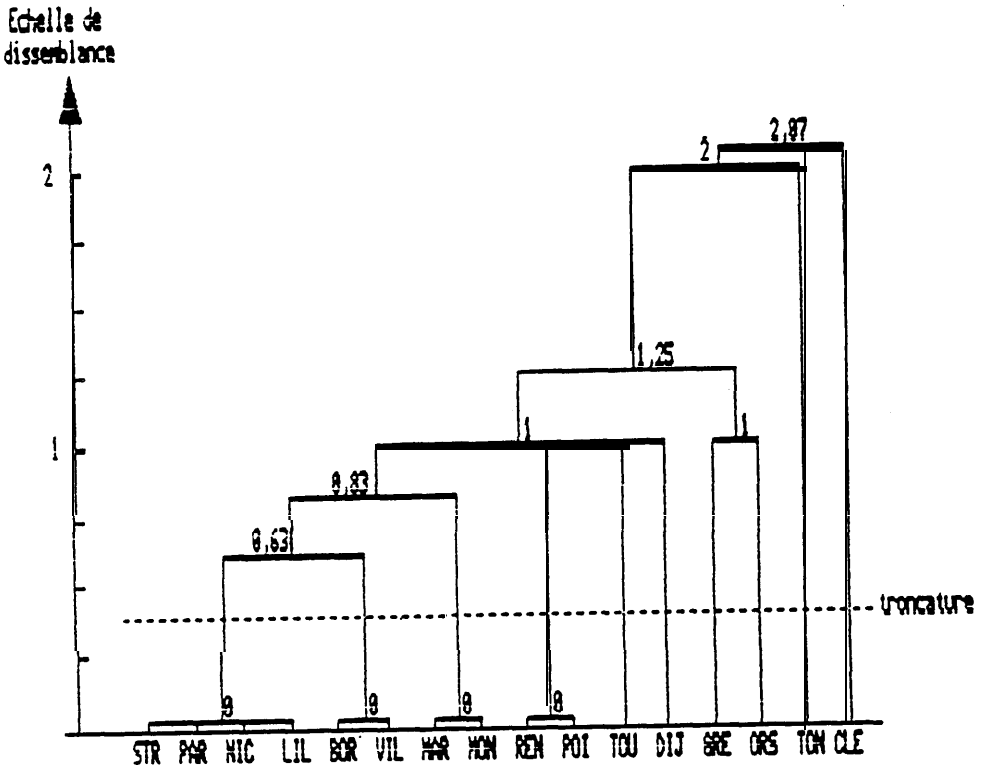
********** C L A S S I F I C A T I O N A U T O M A T I Q U E **********

MATRICE DES DISTANCES

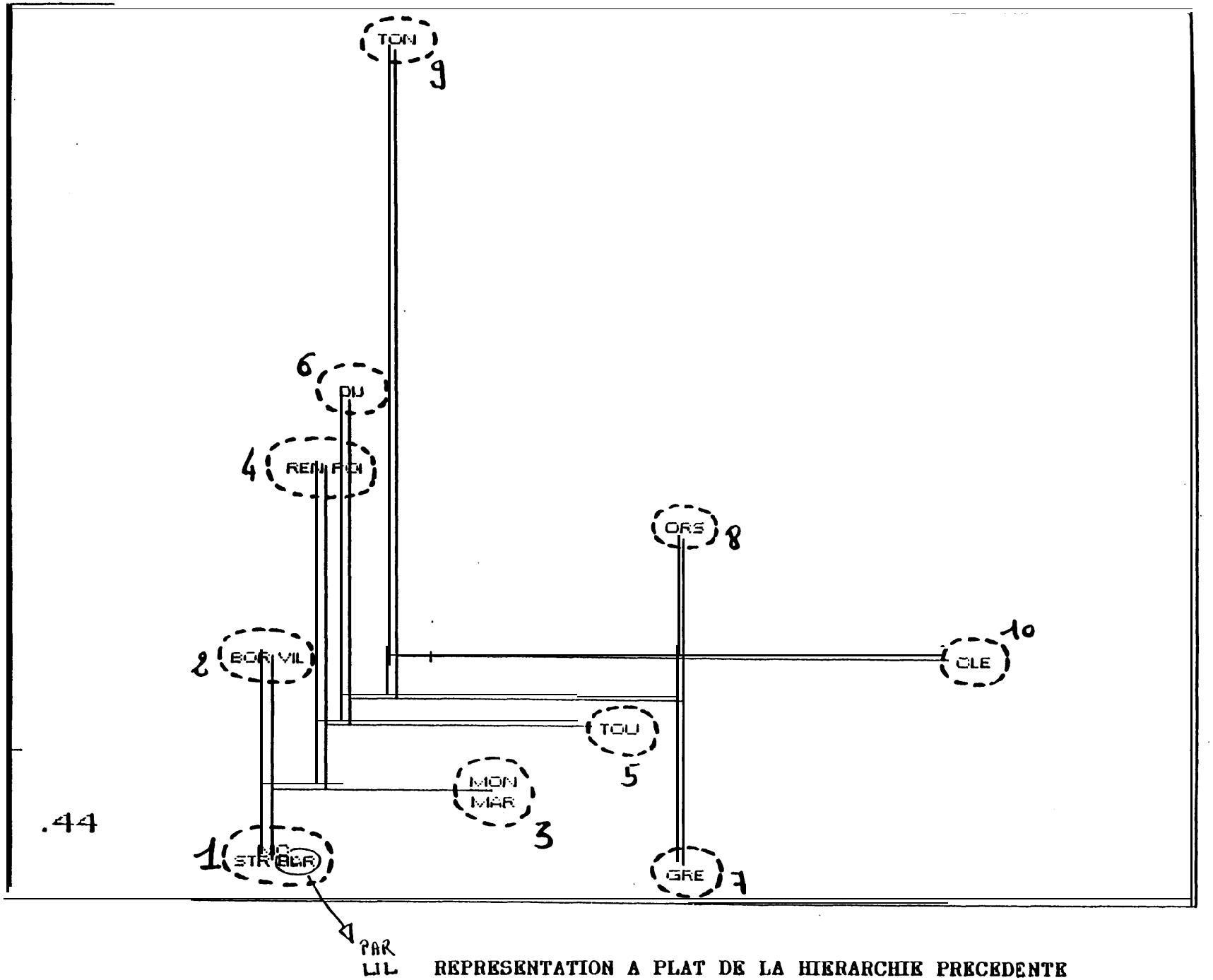
	STR	REN	POI	PAR	ORS	HIC	MON	LIL	GRE	CLE	TOU	TON	VIL	BOR	MAR
REN	1														
POI	1	0													
PAR	0	1	1												
ORS	1	1	1	1											
HIC	0	1	1	0	1										
MON	1	1	1	1	1	1									
LIL	0	1	1	0	1	0	0								
GRE	12	2	2	12	1	1									
CLE	2	2	2	2	2	2	2	2	2						
TOU	1	1	1	1	1	1	1	1	1	2					
TON	2	2	2	2	2	2	2	2	3	3	2				
VIL	1	1	1	1	1	a	1	0	1	2	1	2			
BOR	1	1	1	1	1	l-l	e	1	2	1	2	0			
MAR	1	1	1	1	1	1	8	0	1	2	1	2	1	1	
DIJ	1	1	1	1	2	1	1	1	2	2	1	1	1	1	1

NOEUDS	A	INES	BENJAM.	POIDS	NIVEAUX
N017	REN		POI	2	0.00
N018	STR		PAR	2	0.00
N019	NI	18	NIC	3	0.00
N020	MAR		MON	2	0.00
N021	BOR		VIL	2	0.00
N022	N0	19	LIL	4	0.00
N023	N0	22	N0	21	6	0.63
N024	N0	23	N0	20	8	0.83
N025	N0	24	N0	17	10	1.00
N026	N0	25	TDU	11	1.00
N027	N0	26	DIJ	12	1.00
N028	GRE		OR9	2	1.00
N029	N0	27	N0	28	14	1.25
N030	N0	29	TON	15	2.00
N031	N0	30	CLE	16	2.07

ARBRE HIERARCHIQUE DES VILLES A PARTIR
DES COORDONNBES'DE L'AFC PRECEDENTE



Arbre hiérarchique sur les coordonnées de l'A.F.C.



.44

REPRESENTATION A PLAT DE LA HIERARCHIE PRECEDENTE

B. Evaluation d'un secteur technique: Etude simultanée de trois codifications documentaires brevets

Cette étude a été publiée sous forme de communication:

L'exploitation systématique des bases de données:
des analyses stratégiques pour l'entreprise

Rostaing H., Nivol W., Quoniam L., BEDECARRAX C., HUOT C.

Journée d'étude ADEST 1-2/06/1992

Le texte de cette communication sera publié dans un numéro spécial des cahiers de l'ADEST

L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise

ROSTAING Hervé, NIVOL William, QUONIAM Luc [1]
BEDECARRAX Chantal, HUOT Charles [2]

[1] Centre de Recherche Rétrospective de Marseille
Aix-Marseille III, Faculté St Jérome, 13397 Marseille CEDEX 13

[2] Centre Européen Scientifique en Mathématiques Appliquées
IBM, 68/76 quai de la Rapée, 75592 Paris CEDEX 12

INTRODUCTION

Aujourd'hui tout industriel se pose des questions sur l'environnement scientifique, technologique et concurrentiel de son entreprise. Pour répondre à ces questions il met en place une structure dite de veille technologique ou de veille stratégique. Cette structure doit faire preuve d'une vigilance de tous les instants et être à même de répondre à des questions assez générales sur l'état de la concurrence dans un domaine, d'établir des cartographies d'une technologie ou encore d'analyser des évolutions à travers le temps. Le système de surveillance à instaurer dans une entreprise doit donc assurer deux fonctions (*illustration 1*):

- contrôle en continu de l'environnement pour alerter en cas de menaces émergentes
- monter des dossiers thématiques pour connaître les tendances d'évolution et se positionner par rapport à ses concurrents

La première fonction demande essentiellement un contrôle de l'information informelle et floue tandis que la seconde impose une maîtrise de l'information formalisée et confirmée. La bibliométrie est un outil qui peut s'avérer très utile dans cette seconde tâche [1]. Son application va permettre l'élaboration d'indicateurs de tendances à partir du maximum de connaissances scientifiques ou technologiques que l'on puisse avoir sur un thème à un moment précis.

La bibliométrie est de plus en plus connue comme une méthode d'analyse des bases de données accessibles en ligne. La raison de ce recours aux bases de données est simple; elles représentent la première et la plus importante source d'information scientifique, technique et technologique [2]; elles offrent de plus l'avantage d'être analysables par des systèmes informatiques automatisés [3].

Illustration 1

VEILLE TECHNOLOGIQUE - BIBLIOMETRIE

LE BESOIN INDUSTRIEL

CONNAISSANCE DE SON ENVIRONNEMENT
(technologique, scientifique, concurrentiel,...)

SYSTEMES DE SURVEILLANCE

- EN CONTINU : **ALERTER**
- THEMATIQUES : **POSITIONNEMENT / CONCURRENCE**

OUTIL : LA BIBLIOMETRIE

- ANALYSE DE REFERENCES :
  - * INDICATEURS SYNTHETIQUES
- MASSE ETCOMPLEXITE DES DONNEES :
  - * OUTILS RAPIDES, SYSTEMATIQUES, APPROPRIES :
    - . SPECIFIQUES (Préparation des données)
    - . STATISTIQUES (Détermination d'une grille de lecture)
- INFORMATION CIBLEE = INFORMATION BREVETS
  - * NECESSAIRE AUX ENTREPRISES
  - * FIABILITE

Nous allons présenter dans cet article une étude mettre en oeuvre deux outils adaptés au traitement automatique des données bibliographiques (*illustration 2*):

- Le premier automatise la manipulation des références pour restituer les caractéristiques bibliométriques du corpus étudié (distributions bibliométriques, listes de fréquences de termes, listes de fréquences de cooccurrences de termes, matrices d'occurrences, matrices de cooccurrences, matrices d'associations):

  DATAVIEW (développé au CRRM).

- Le second est l'outil statistique que l'on appliquera à ces caractéristiques bibliométriques pour discriminer et structurer le contenu informationnel du corpus:

  L'ANALYSE RELATIONNELLE (développée au CESMAP)

L'information concernant l'innovation industrielle est une information stratégique, aux yeux des entreprises. Il est crucial de pouvoir la maîtriser pour conserver sa performance dans l'art d'innover. L'information brevet en fait partie. Elle est la source de multiples renseignements sur l'état des innovations. Nous avons donc choisi de cibler cette étude sur l'information brevet à travers la base Derwent (WPIL).

Le thème choisi pour ce travail est une technologie charnière entre la médecine et la pharmacie: *Les systèmes transdermiques thérapeutiques sous forme de patchs (T.T.S).* On peut facilement envisager qu'un tel sujet puisse être sensible pour une entreprise pharmaceutique. Dans un système de veille, on pourrait imaginer que ce thème soit sélectionné comme étant l'un des facteurs critiques, c'est-à-dire un sujet stratégique pour garantir la pérennité de l'entreprise.

● OBJECTIF DE L'ETUDE:

Une évaluation méthodologique de nouveaux traitements bibliométriques

Une référence signalétique dans une base de données contient diverses informations réparties en plusieurs rubriques. Ces rubriques, nommées champs, ont des portées significatives différentes. Il est rare que la richesse de cette diversité d'information soit pleinement exploitée dans les études bibliométriques de corpus de références.

Traditionnellement, en bibliométrie, les analyses statistiques ne traitent que le contenu d'un champ à la fois. Les méthodes d'analyse de co-citations [4] établissent leurs cartes sur la rubrique concernant les citations faites par les auteurs scientifiques. La méthode de mots-associés [5] est développée pour exploiter un champ indexé. La méthode des citations-

LE CAS ETUDIE

LE CORPUS

- LE THEME : **SYSTEMES TRANSDERMIQUES THERAPEUTIQUES SOUS FORME DE PATCH (T.T.S.)**

  SYSTEMES ADHESIFS QUI ASSURENT UNE DIFFUSION CONTROLEE D'UN PRINCIPE ACTIF PAR VOIE TRANSDERMIQUE

- LA SOURCE : LA BASE BREVETS DERWENT WPIL

LES TRAITEMENTS AUTOMATIQUES DES DONNEES TEXTUELLES

- DATAVIEW (C.R.R.M.)

LES TRAITEMENTS STATISTIQUES

- L'ANALYSE RELATIONNELLE (C.E.M.A.P. - F.MARCOTORCHINO, P.MICHAUD)

croisées de journaux [6] est une approche vers cette combinaison d'informations de champs différents. Cette dernière se base sur une matrice croisant des journaux "citants" (champ source) avec des journaux cités (champ citation).
Cette déficience dans les traitements bibliométriques est due à la complexité des relations engendrées par toutes les combinaisons de ces informations.

Cet article tente d'apporter une réponse à cette carence. Il traite de la complémentarité d'informations apportées par la prise en compte simultanée de trois champs de la base brevets de Derwent.
Tout brevet référencé dans la base est qualifié, au sens informationnel, par les codes de trois classifications différentes. Une classification documentaire découpe les domaines scientifiques en sections. Si la classification est construite à partir d'un principe de hiérarchie, ces sections sont alors elles-mêmes découpées en sous-sections, classes, sous-classes, groupes, sous-groupes... Chaque niveau dans cette hiérarchie de découpage est représenté par une codification.

Une référence de brevet dans la base Derwent reçoit donc l'affectation, dans plusieurs champs, des différents codes qui illustrent les thèmes abordés par l'invention.
Les trois "champs codes" que nous allons exploiter sont les champs:

- DC (Derwent Codes) : Classification documentaire établie par Derwent
- MC (Manuel Codes) : Autre classification documentaire établie par Derwent
- IC (International Patent Classification) : Classification établie par les instituts officiels de dépôts de brevets.

Pour estimer l'apport spécifique de chacune de ces classifications, nous allons les confronter pour le corpus de références brevets établi sur le thème des patchs transdermiques thérapeutiques. Les méthodes d'Analyse de Données Relationnelles font partie du panel des analyses statistiques élaborées pour mieux cerner les phénomènes complexes. Leur exploitation va nous permettre d'évaluer deux caractéristiques des relations que ces champs peuvent entretenir (*illustration 3*):

- les **complémentarités** de ces classifications documentaires en qualité de descripteurs de brevets: nous allons estimer si le fait d'utiliser ces trois classifications simultanément nous permet de mieux décrire les réels liens entre les brevets.

- les **correspondances** ou les **similarités** entre les codes de ces classifications au niveau de leur sens: on pourrait ainsi connaître les recouvrements de signification, les codes synonymes et les complémentarités entre les classifications.

OBJECTIFS

LES BASES DE DONNEES BREVETS ACCESSIBLES EN LIGNE OFFRENT UNE DIVERSITE D'INFORMATION

→ INFORMATION

* STRUCTUREE
* RICHE
* DIFFICILE A CORRELER

A PARTIR D'UN OU PLUSIEURS CHAMPS DE DESCRIPTION

- ETUDE DE LA COMPLEMENTARITE DES DESCRIPTEURS

- ETUDE DE LA CORRESPONDANCE OU DE LA SIMILARITE DES DESCRIPTEURS

CONSTITUTION DU CORPUS DES REFERENCES

Cette première étape, dans une analyse bibliométrique, est certainement celle qui influence le plus la validité des résultats. Cette validité est capitale lorsque les résultats rentrent dans un processus de décision.

Pour notre étude, l'objectif principal est de présenter une méthode permettant de prendre en compte la diversité du sens apporté par trois classifications utilisées sur les brevets. Néanmoins, il est indispensable, pour conclure sur la cohérence des résultats, de constituer un ensemble homogène de références tout en assurant une couverture acceptable du sujet.

Le corpus dégagé doit, autant que possible, être suffisamment large pour couvrir le thème étudié et suffisamment étroit pour présenter un "bruit" aussi faible que possible. Nous avons volontairement réduit les risques de bruits pour qu'ils ne viennent pas perturber l'interprétation des résultats statistiques. L'ensemble des références, qui a été collecté pour cette étude, n'est donc probablement pas exhaustif mais devrait posséder une propriété d'homogénéité.

Dans le cadre d'une étude de veille, cette étape serait obligatoirement conduite en présence d'experts du thème pour garantir la validité de la base des connaissances construite.

La stratégie d'interrogation a été affinée selon une méthode itérative de type "coups de sonde". Chaque itération permet, après lecture d'échantillons de références, de dégager de nouvelles pistes pour enrichir la stratégie. Cette itération est répétée tant que les échantillons, obtenus par les croisements des nouvelles pistes, laissent apparaître des références non pertinentes. Les échantillons ont été estimés pertinents pour la stratégie d'interrogation suivante:

QUESTION 1 :	PATCH ou PATCHES ou PATCHS	(2106)
QUESTION 2 :	1 et TRANSDERM:	(103)
QUESTION 3 :	1 et THERAPEUTIC:	(36)
QUESTION 4 :	1 et PERCUTANE:	(15)
QUESTION 5 :	1 et (DRUG ou DRUGS)	(102)
QUESTION 6 :	1 et MEDICIN:	(14)
QUESTION 7 :	2 ou 3 ou 4 ou 5 ou 6	(160)
QUESTION 8 :	7 sans (CARDIAC# à coté de PATCH##)	(159)
QUESTION 9 :	8 sans (VASCULAR# à coté de GRAFT#)	(158)
QUESTION 10:	9 sans (PATCH## à coté de GRAFT#)	(156)
QUESTION 11:	10 sans (FASTENING# à coté de PATCH##)	(155)
QUESTION 12:	11 sans (CARRY à coté de PATCH##)	(155)
QUESTION 13:	12 sans (CARRIES à coté de PATCH##)	(154)
QUESTION 14:	13 sans PROSTHES:	(148)
QUESTION 15:	14 sans CAMERA	(147)
QUESTION 16:	15 sans (TEST# à coté de PATCH##)	(146)

Remarque: les questions sont posées sur l'Index de Base de la base Derwent, c'est à dire sur les champs: Résumé, Résumé équivalent, Titre et Titre normalisé de Derwent. ({#} = Troncature courte. {:} = Troncature large)

L'examen de cette stratégie finale montre que la simple utilisation des termes "PATCH" et "THERAPEUTIC" ne nous permettait par de couvrir la totalité des brevets du domaine. Par contre l'emploi d'autres termes, pour élargir la recherche, laissait apparaître des références hors-sujet car le terme PATCH a des significations multiples et variées (rustine, greffon, pièce de prothèse, chute de film, test d'allergie...). Par conséquent, les brevets, qui détiennent pour ce terme un sens autre que celui recherché, sont désélectionnés.
Cette stratégie d'interrogation permet de dégager un corpus de 146 brevets. Le télédéchargement de ces références est réalisé non seulement pour les trois champs des classifications mais aussi pour tous les champs qui fournissent des renseignements pouvant aider à la compréhension des résultats des analyses statistiques.

● TRAITEMENT DES CARACTERISTIQUES BIBLIOMETRIQUES DU CORPUS

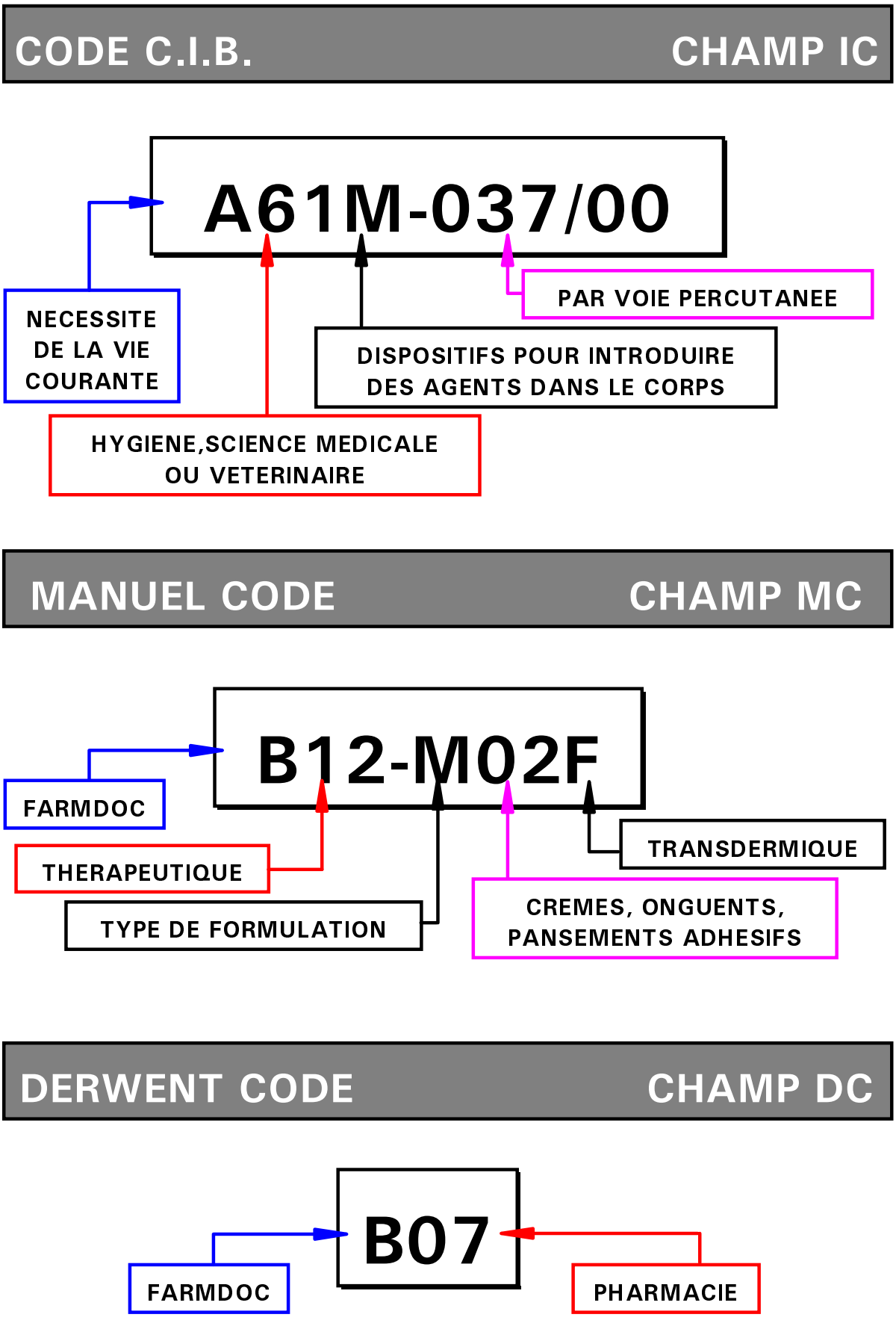
Pour confronter les différentes informations apportées par chacune des classifications nous allons reproduire leurs interactions par la construction de tableaux. Ces tableaux également appelés "matrices" sont le point d'entrée des analyses statistiques.

Choix des niveaux hiérarchiques:

Les trois classifications brevets possèdent leur propre hiérarchie de codes. Un code est assimilable à un chemin pris parmi les branches de la hiérarchie. Dans cette hiérarchie, les branches sont réparties à partir d'un niveau de signification très large vers des niveaux de signification de plus en plus fins. Plus on descend dans les branches de la hiérarchie, plus le code a une représentation complexe et plus son sens est précis. Cette notion est représentée par l'*illustration 4* où l'on a expliqué la signification de chaque niveau de hiérarchie.

Quel niveau hiérarchique considérer pour cette étude?
Décider de prendre les niveaux hiérarchiques les plus précis pour avoir les descriptions les plus fines est contestable. En effet, tous les codes affectés aux références ne sont pas forcément renseignés jusqu'au dernier niveau de la hiérarchie. Par exemple sur notre corpus de références, 66 % des codes IPC sont renseignés jusqu'au dernier niveau contre 12 % pour les Manuels Codes.
Parallèlement, plus le niveau hiérarchique est fin, plus la diversité des codes augmente. Ceci est un argument en faveur du choix d'un niveau assez fin.

Illustration 4



Ce choix de niveaux de hiérarchie, impose donc un compromis entre la perte d'information et le gain de signification (pour un choix très fin, perte de certains codes mais codes restants plus précis).

Une pré-étude a été menée pour déterminer quels niveaux hiérarchiques satisfaisaient la meilleure solution statistique. Les critères étudiés pour chaque niveau hiérarchique étaient:

- le nombre de codes restants
- le nombre de brevets encore renseignés
- la qualité d'agrégation des brevets par l'Analyse Relationnelle pou ce niveau de la hiérarchie (nombre d'agrégats et nombre de codes non agrégés)

Remarque: Les codes à fréquence 1 ne sont pas considérés puisqu'ils n'établissent aucun lien entre les brevets.

Le choix s'est porté sur les niveaux hiérarchiques dont les nombres de codes et les nombres d'agrégats pour les différentes classifications documentaires sont proches (*illustration 5*). Les niveaux hiérarchiques choisis pour l'étude ont donc été:

- Les Derwent Codes à 3 caractères
- Les Manuel Codes à 3 caractères
- Les codes IPC à 7 caractères

L'absence, parmi ces critères purement statistiques, de critères qualitatifs pour comparer les degrés de sens à chaque niveau des hiérarchies est critiquable. En l'absence d'experts du domaine étudié, il était difficile de s'investir dans de telles considérations. Nous sommes conscients que ce choix demande à être confirmé par des critères plus qualitatifs mais nous voulions en premier lieu évaluer si cette utilisation "novatrice" de plusieurs codifications était bénéfique.

CHOIX DES NIVEAUX HIERARCHIQUES DES CODES

HIERARCHIES DES CODES (Nb de digits)	NB DE CODES	NB DE CODES APRES ELIMINATION DES CODES A FREQUENCE 1	NB DE BREVETS	NB DE BREVETS APRES ELIMINATION DES CODES A FREQUENCE 1	NB MOYEN DE CODES PAR BREVET	NB DE CLASSES	% DE CLASSES A UN ELEMENT
DC 3	52	32	146	146	3,6	27	30
IC 4	40	21	146	143	2,2	23	22
IC 7	94	35	146	140	2,9	42	36
IC 11	133	36	146	113	2	67	52
MC 3	51	42	143	143	4,6	41	34
MC 5	139	90	143	143	6,9	80	50
MC 7	315	158	142	141	8,7	106	74
MC8	144	69	135	134	4,3	80	55
MC 9	43	12	46	34	1,7	124	90

Construction des tableaux de l'étude

Le choix des niveaux étant fait pour chacune des hiérarchies, vient alors la phase de construction des tableaux pour l'analyse.

Plusieurs types de tableaux sont exploitables par les méthodes d'analyse statistique. En ce qui concerne l'Analyse Relationnelle des Données, les tableaux d'entrée sont du type matrice de présence-absence. Ces matrices croisent un ensemble *d'individus*, noté *I*, et un ensemble de *variables* descriptives, noté *J*. Les individus sont naturellement ici les 146 brevets et l'ensemble J des variables est constitué des codes. Le croisement de *I x J* fait donc figurer, à l'intersection de la ligne *i* et de la colonne *j*, la valeur 1 si le code *j* est présent dans la référence du brevet *i* et la valeur 0 dans le cas contraire. Ce tableau est la simple restitution des données élémentaires contenues dans les références, rien n'est omis, ni ajouté, ni transformé.

Pour l'étude des complémentarités des codes, nous avons mené parallèlement l'analyse des regroupements des références:

- lorsqu'elles sont décrites par un des trois ensembles de codes
- lorsqu'elles sont décrites par la réunion des trois ensembles.

Nous avons donc construit quatre matrices; trois pour les ensembles isolés de codes et une par l'union de ces codes (*illustration 6*).

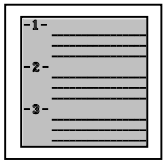
Ce passage des données textuelles (références) aux données tabulées (matrices) est réalisé par le logiciel DATAVIEW du CRRM.
Les matrices sont construites, en automatique, avec DATAVIEW selon la logique suivante:

- extraire tous les codes
- éliminer les codes qui ne sont pas renseignés jusqu'aux niveaux des hiérarchies choisies
- tronquer les codes restants à ces niveaux de hiérarchie
- construire les matrices de présence-absence

D'autres caractéristiques bibliométriques

L'*illustration 6 (à l'extrême droite)* présente aussi sous forme symbolique d'autres traitements bibliométriques que permet DATAVIEW. Ces résultats sous forme de distributions de fréquences ou de rangs de fréquences peuvent déjà apporter de nombreux renseignements. Il est possible d'examiner l'évolution des brevets en fonction du temps, ce qui permet de situer le niveau de maturité du sujet étudié. De même, nous pouvons établir la répartition par pays de dépôts, ce qui peut permettre de dégager les pays dont les marchés sont convoités. On peut aussi connaître les sociétés leaders en nombre de dépôts de brevets. Ces résultats demandent souvent des prétraitements de reformatage, d'homogénéisation, d'élimination d'ambiguïté

CONSTRUCTION ET ANALYSE DES MATRICES



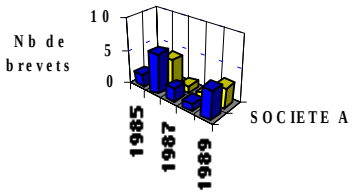
DATAVIEW

* DISTRIBUTION
* LISTES DE FREQUENCES

DC | MC 3 | IC 7 | DC +MC 3 +IC 7

Ref

REPARTITION PAR SOCIETE



REPARTITION PAR PAYS

ANALYSE RELATIONNELLE

MATRICES PERMUTEES

COMPLEMENTARITE DES CODES

DC +MC 3 +IC 7

Ref

DC | MC 3 | IC 7

Ref

CORRESPONDANCE DES CODES

DC +MC 3 +IC 7

Ref

difficilement envisageables sans outil informatique. DATAVIEW facilite ce travail préliminaire, souvent fastidieux, mais néanmoins fondamental pour garantir la qualité des données et donc la validité des résultats.

● L'ANALYSE STATISTIQUE

Pour évaluer la complémentarité des codes en qualité de description des brevets, l'Analyse Relationnelle est appliquée sur l'espace des brevets pour les regrouper par similarité de description de codes.
Pour évaluer les correspondances entre les codes, l'Analyse Relationnelle servira à regrouper les codes par leurs ressemblances de répartition dans l'ensemble des brevets.

La technique que nous allons mettre en oeuvre pour construire la partition des brevets en classes disjointes s'inscrit dans le cadre méthodologique général de l'Analyse Relationnelle. Nous avons pris l'option, dans cet article, de bannir toute formule mathématique et de réduire au minimum les explications et les justifications méthodologiques. On trouvera dans [1] et [7] l'essentiel de ces informations.
Nous nous contenterons de parler, ici, du point fondamental qui préside au déroulement du traitement, à savoir: le critère de classification qui mesure les similarités entre les objets que l'on compare. C'est incontestablement la question qu'il faut à tout prix se poser pour appliquer la procédure de classification dans un cadre clairement défini. Faute de quoi, on ne sait pas, a posteriori, expliquer la structure que l'on a mise en évidence.

L'agrégation des références brevets:

On va chercher ici à dégager des classes de brevets qui s'apparentent par les codes descripteurs qu'ils ont en commun, autrement dit des familles de brevets caractérisées par leurs similitudes en termes de technologies partagées.
Les spécialistes de la bibliométrie, de la scientométrie, de l'infométrie ou encore de la lexicographie mathématique, ont depuis fort longtemps mis en évidence et étudié les caractéristiques des distributions que l'on rencontre dans ces domaines (lois de Zipf, Bradford et Lotka).
Les données extraites des bases de données telles que celles dont nous disposons sont de cette nature. Ainsi, un certain nombre de codes se retrouvent dans la grande majorité des références (ce sont les thèmes qui ont présidé à la construction du corpus), d'autres apparaissent de façon moins systématique et d'autres enfin ne figurent que dans un faible, voire très faible, nombre de références. Il importe que le critère de classification qui permette de mesurer les ressemblances entre les brevets tienne compte de ces phénomènes.

Notre choix s'est porté sur le critère de Burt pondéré, qui outre ses bonnes propriétés axiomatiques, se base sur un indice de présence-rareté répondant parfaitement à notre problème. En effet, ce critère a pour effet d'attribuer aux codes un poids inversement proportionnel à leur présence dans le corpus. Ainsi, deux brevets qui partagent un code rare seront "plus similaire" que deux brevets qui auraient en commun un code présent dans de nombreuses références.

Les classes de brevets issues de la classification pilotée par le critère de Burt pondéré sont donc formées de brevets qui se ressemblent non seulement parce qu'ils partagent les mêmes codes mais encore parce que ces codes sont absents (ou peu "typiques") dans les autres brevets. On garantit le découpage du corpus en familles de brevets homogènes et discriminantes. L'analyse du résultat permet en outre d'expliquer chacune de ces classes en fonction des codes ou groupes de codes qui ont présidé à leur création, autrement dit, d'attacher à chacune d'elle une *étiquette* synthétique résumant les thèmes technologiques qu'elle recouvre.

La classification des codes

L'objectif de ce traitement est de découper l'ensemble des codes descripteurs en classes homogènes sur la base de leurs cooccurrences dans les références. Autrement dit, on cherche ici à dégager les pôles technologiques autour desquels s'articule le corpus.

Cette fois ce sont les colonnes du tableau de départ, c'est-à-dire l'ensemble des codes descripteurs, qui sont soumis à la procédure de classification.
Les considérations (lois Zipf-Bradford-Lotka) qui avaient induit le choix du critère de classification sur les brevets sont encore valides dans ce contexte. Il n'est toutefois plus possible d'utiliser le critère de Burt pondéré. Celui-ci aurait en effet pour conséquence de faire jouer un rôle à la richesse de description des brevets. Or le fait qu'un brevet possède un plus ou moins grand nombre de codes ne doit pas être pénalisant. En revanche, il convient toujours de prendre en compte la fréquence d'apparition des codes qui, elle, reste tout à fait pertinente.
Notre choix s'est finalement porté sur le critère de Burt, très couramment utilisé en Analyse Relationnelle pour ses bonnes propriétés de règle d'agrégation.
A l'issue de ce traitement, nous obtenons une partition des codes descripteurs. Chacune des classes de cette partition regroupe un certain nombre de codes qui ont pour caractéristique d'apparaître conjointement dans les références de brevets.
On a donc mis en évidence des combinaisons de technologies qui présentent un fort taux de corrélation à l'échelle du corpus étudié.

● DISCUSSION DES RESULTATS

Etude de la complémentarité des codes pour décrire les brevets

Pour montrer l'utilité de la combinaison des trois ensembles de codes, nous décrirons les résultats pour un exemple de sous-thème paraissant convaincant. Ce sous-thème correspond aux brevets revendiquant l'application de patchs transdermiques thérapeutiques pour la diffusion de composés actifs de la famille des stéroïdes, soit 15 brevets.

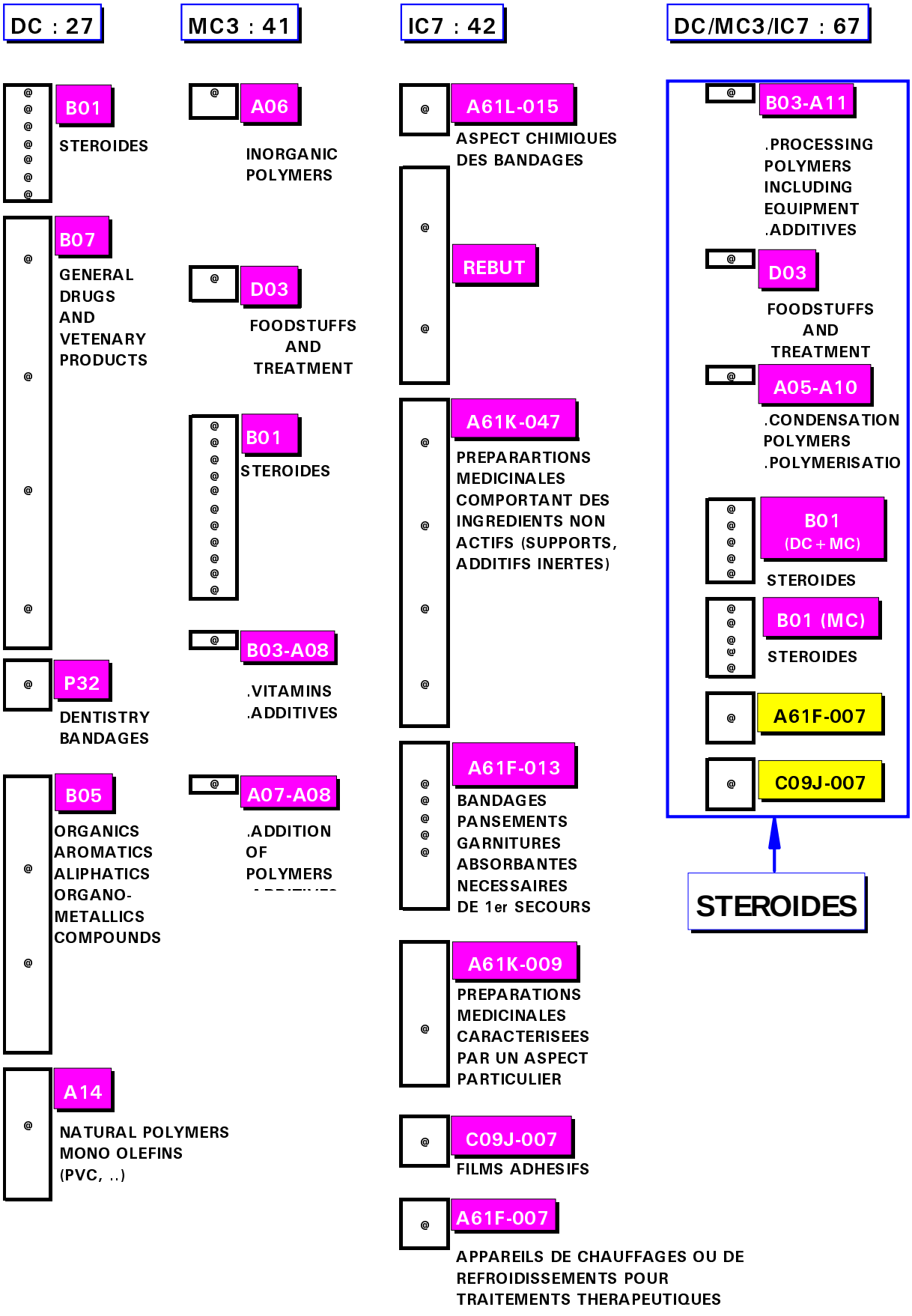
Comparons la répartition des brevets de ce sous-thème parmi les classes obtenues, pour les quatre matrices construites. La disposition des 15 brevets dans leurs classes d'appartenance est présentée symboliquement sur l'*illustration 7*.

Comment lire cette illustration?
Pour la matrice des DC à 3 caractères, les 15 brevets se distribuent dans 5 des 27 classes créées. Pour la matrice des MC à 3 caractères, ils se distribuent dans 5 classes parmi les 41... etc...
Sur cette illustration est indiqué, à côté de chaque classe, le code qui a le plus contribué à la construction de la classe. Ce sont les fameuses *étiquettes* décrites précédemment. Elles indiquent que le thème est spécifique aux brevets de cette classe et qu'il est pratiquement absent dans les brevets des autres classes.

Que nous dit cette illustration?
- Pour l'analyse faite à partir de la matrice des DC, une classe libellée *stéroïde* rassemble la moitié des brevets, les autres étant répartis dans d'autres spécialités. Donc, les codes DC ne décrivent correctement que la moitié des brevets *stéroïdes*.
- Pour l'analyse MC3, une classe *stéroïde* est constituée de 11 brevets; 2 autres brevets sont dans des classes singletons; et les 2 derniers appartiennent à de toutes petites classes. Ici, les MC3 ont très bien regroupé ces brevets dans des classes très homogènes.
- Pour l'analyse IC7, il n'existe aucune classe spécifique aux brevets *stéroïdes* et les brevets sont disséminés dans 7 classes de grandes tailles. Les codes IC7 ne permettent pas de regrouper les brevets concernés par les composés stéroïdes.

- Par contre, l'analyse réunissant les trois ensembles a parfaitement bien caractérisé ces brevets (une partie de cette matrice résultante est présentée par l'*illustration 8*). Ils sont tous dans des classes très homogènes. Les deux plus importantes de ces classes sont spécifiques aux codes descripteurs *stéroïde.* L'une est caractérisée par la co-présence des deux codes B01 qui représente les composés stéroïdes pour les deux codifications MC3 et DC, l'autre par la simple présence du B01 de la codification MC3. Cinq brevets *stéroïdes* se retrouvent isolés dans

Illustration 7



CLASSES	REFERENCES	B	d c _ B	B	d c _ A	A	d c _ P	A K -	B	A L -	A K -	D	d c _ D	d c _ P	A K -	A F -	B	B	d c _ B	B	A M -	d c _ B	B	d c _ B	C	A	d c _ S	d c _ D	d c _ C	S	A N -	A M -	C	d c _ B	d c _ A
4	17	X	.	X	.	.	.	.	X	.	X	.	.	.	.	.	X	.	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	25	X	X	.	.	.	.	.	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	42	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	44	X	.	X	.	.	.	.	X	.	X	.	.	.	.	.	X	.	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	75	X	.	.	.	.	X	.	X	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	##	X	.	.	X	X	.	X	X	.	X	.	.	.	.	.	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	##	X	.	X	.	.	.	.	X	.	X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	6	X	X	X	X	X	.	X	X	.	.	.	.	.	.	.	X	X	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
6	33	X	X	X	X	X	X	X	.	X	.	X	X	.	X	.	.	.	.	X	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
6	57	X	X	X	X	X	.	.	X	.	X	X	X	X	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
6	76	X	X	X	.	.	.	.	X	.	.	X	X	X	.	X	X	.	.	X	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
6	##	X	.	X	.	.	X	X	X	X	X	.	.	X	X	X	.	X	X	X	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
7	3	X	X	X	X	X	.	.	.	.	.	X	X	X	.	X	.	X	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
7	4	X	X	.	.	.	.	.	.	.	.	X	X	X	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
7	43	X	X	X	X	X	.	.	X	.	.	.	.	X	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
7	67	X	X	X	.	.	.	.	X	.	X	.	.	.	X	.	.	.	X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
7	##	X	.	X	.	.	X	.	X	X	X	.	.	.	.	.	.	X	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
#	38	X	X	X	X	X	X	X	.	X	X	.	.	X	.	X	.	.	.	.	X	.	X	.	.	.	.	.	.	.	.	.	.	.	X
#	45	X	X	X	X	X	X	.	.	X	.	X	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.
#	58	X	.	X	X	X	X	X	X	X	X	.	.	.	X	.	.	.	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.
#	##	X	X	.	.	.	X	X	.	X	.	.	.	X	.	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.
#	##	X	X	X	.	.	X	X	X	X	.	X	X	X	X	.	X	.	X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
#	26	X	X	X	X	X	X	X	.	X	.	.	.	.	X	.	.	.	.	.	.	.	X	.	.	X	.	.	.	.	.	.	.	.	.
#	13	X	X	X	X	X	X	.	.	X	.	X	X	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.
#	##	X	.	X	X	X	X	.	X	X	X	.	.	.	.	.	.	.	.	.	.	.	X	.	.	X	.	.	.	.	.	.	.	.	.

CLASSES	REFERENCES	d c _ P	d c _ P	B	d c _ B	D	A	A	d c _ A	d c _ P	A K -	B	A B -	d c _ G	G	B	A	A	A	A F -	d c _ P	C J -	B B -	C	C	A	A	A J -	d c _ S	d c _ A	C K -	A F -	d c _ P	C G -	D		
4	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	C A M P
4	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	N I H A
4	42	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	N O V A
4	44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	B E R C
4	75	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	E L B A
4	##	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	S C H E
4	##	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	M O M E
6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	K I M Y
6	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	R I K L
6	57	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	R U T F
6	76	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	Y A M A
6	##	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	A M H P
7	3	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	R O B E
7	4	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	R O B E
7	43	.	.	.	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	P H A R
7	67	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	E A S T
7	##	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	#	K E Y P
#	38	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	#	L O H M
#	45	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	#	B A D I
#	58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	#	L O H M
#	##	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	#	C I B A
#	##	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	#	Y A M A
#	26	.	.	X	.	.	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	#	N I T L
#	13	.	.	.	X	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	#	N O V E
#	##	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	#	K E Y P

des classes ou presque seuls. Cet isolement est dû au fait qu'ils contenaient d'autres codes très rares. Ces codes rares les ont marginalisés du reste du sous-thème. Deux de ces brevets ont été placés dans deux classes dont la spécificité est créée par un code IC7. Ces deux codes sont rares dans les brevets car ils ont des revendications très accentuées sur un caractère atypique: le code C09J-007 revendique une grande qualité d'adhérence, le code A61-007 concerne des patchs dont le composé actif se libère de la matrice de diffusion par différence de température avec la peau.

Donc grâce à l'utilisation des trois ensembles de codes réunis, l'agrégation établie fait ressortir simultanément les caractéristiques décrites par les différents ensembles de codes:

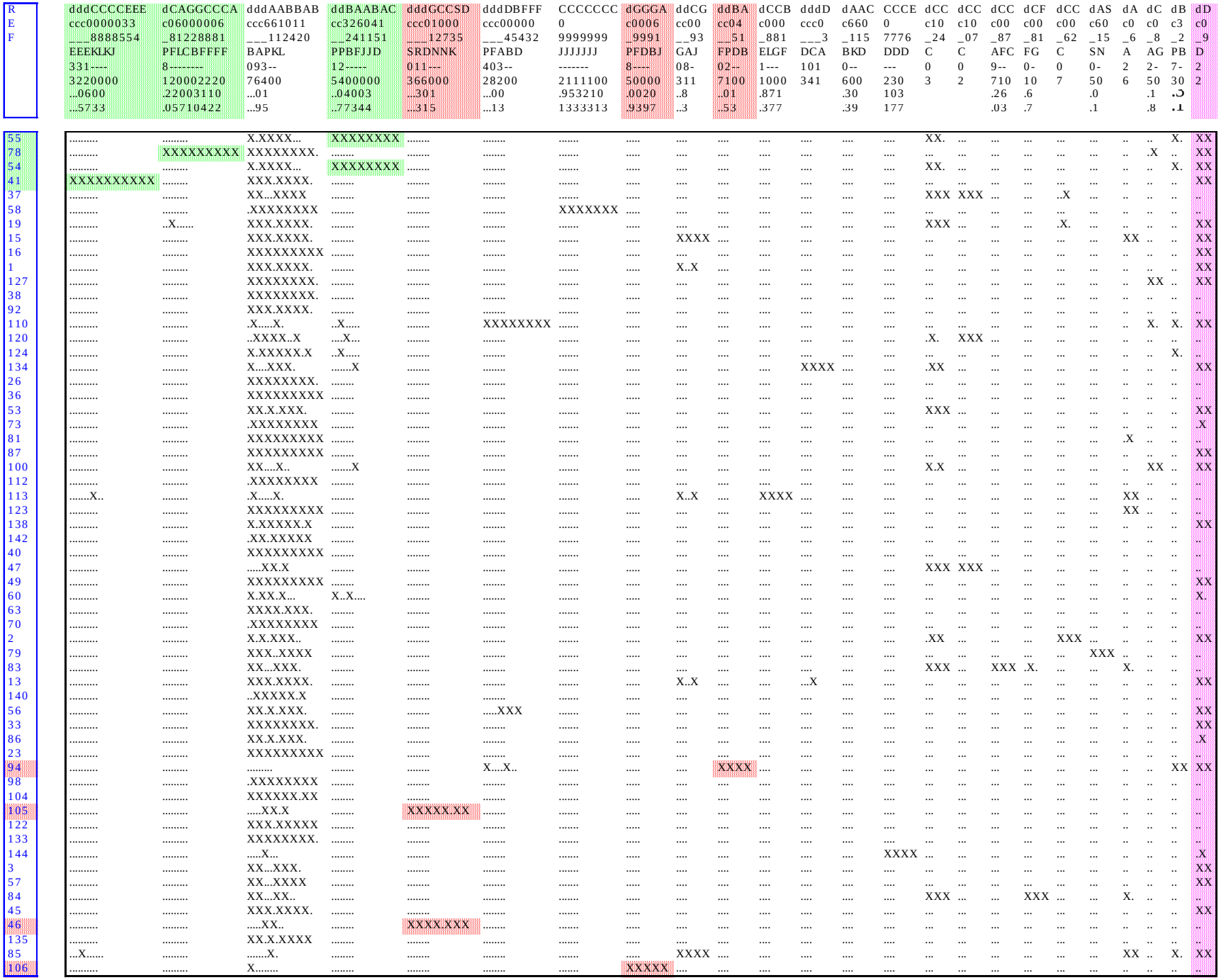
- le caractère *composés stéroïdes* décrit par les codes DC et MC3 et qui disparaît dans l'analyse de la matrice des codes IC7
- les caractères spécifiques plus rares uniquement décrits par la codification plus fine IC7 et qui sont noyés dans d'autres thèmes dans les analyses des matrices des codes DC et MC3

Etude des correspondances entre codes

Les résultats obtenus par cette analyse s'ils n'ont pas répondu à toutes nos espérances, ont toutefois révélé des informations inattendues et très prometteuses.

Nous espérions pouvoir faire apparaître les synonymies et les déficiences entre les trois ensembles de codes. L'analyse n'a pas livré ces correspondances entre les ensembles de codes, probablement parce que le nombre de brevets de notre échantillon était trop restreint. Au final, seuls les codes, soit très rares, soit présents presque partout, ont été regroupés dans des classes de tailles raisonnables. Les codes entre ces deux extrêmes, c'est-à-dire les codes dont les fréquences ne sont ni fortes ni faibles, ont créé une multitude de petites classes sans cohésion entre elles. Or, ce sont précisément ces codes qui sont porteurs de l'information la plus intéressante puisque les deux autres corps de codes se rapportent pour le premier au bruit de fond et pour le second à l'information triviale. Pour une taille de corpus bien plus élevée, la plage intermédiaire de codes porteurs d'information augmenterait vers des valeurs de fréquences plus élevées et, les poids des codes grandissant, des agrégats pourraient se former.

Cette analyse nous a tout de même apporté des satisfactions. Le fait que les codes très rares et toujours présents ensemble soient très bien discriminés n'est pas uniquement un inconvénient. Comme on le voit sur l'*illustration 9,* il est très facile de détecter les brevets


R E F
dddCCCCEEE ccc0000033 ___8888554 EEEKLKJ 331---- 3220000 ...0600 ...5733
dCAGGCCCA c06000006 _81228881 PFLCBFFFF 8-------- 120002220 .22003110 .05710422
dddAABBAB ccc661011 ___112420 BAPKL 093-- 76400 ...01 ...95
ddBAABAC cc326041 __241151 PPBFJJD 12----- 5400000 ..04003 ..77344
dddGCCSD ccc01000 ___12735 SRDNNK 011--- 366000 ...301 ...315
dddDBFFF ccc00000 ___45432 PFABD 403-- 28200 ...00 ...13
CCCCCCC 0 9999999 JJJJJJJ ------- 2111100 .953210 1333313
dGGGA c0006 _9991 PFDBJ 8----- 50000 .0020 .9397
ddCG cc00 __93 GAJ 08- 311 ..8 ..3
ddBA cc04 __51 FPDB 02-- 7100 ..01 .53
dCCB c000 _881 ELGF 1--- 1000 .871 .377
dddD ccc0 ___3 DCA 101 341
dAAC c660 _115 BKD 0-- 600 .30 .39
CCCE 0 7776 DDD --- 230 103 177
dCC c10 _24 C 0 3
dCC c10 _07 C 0 2
dCC c00 _87 AFC 9-- 710 .26 .03
dCF c00 _81 FG 0- 10 .6 .7
dCC c00 _62 C 0 7
dAS c60 _15 SN A 0- 50 .0 .1
dA c0 _6 A 2 6
dC c0 _8 AG 2- 50 .1 .8
dB c3 _2 PB 7- 30 .3 .1
dD c0 _9 D 2 2
55 .... X.XXXX... XXXXXXXX .... XX. .. .. X. XX
78 ..... XXXXXXXXX XXXXXXXX. .... .X .. XX
54 .... X.XXXX... XXXXXXXX .... XX. .. .. X. XX
41 XXXXXXXXXX ..... XXX.XXXX. XX
37 .... XX...XXXX .... XXX XXX ... .X ... .
58 .... .XXXXXXXX .... XXXXXXX ....
19 ....X..... XXX.XXXX. .... XXX ... ... .X. XX
15 .... XXX.XXXX. .... XXXX .... XX XX
16 .... XXXXXXXXX XX
1 .... XXX.XXXX. .... X..X XX
127 .... XXXXXXXX. .... XX .. XX
38 .... XXXXXXXX. ....
92 .... XXX.XXXX. ....
110 .... .X.....X. ..X..... ..... XXXXXXXX .... X. X. XX
120 ....XXXX..X ....X... .... .X. XXX ... ...
124 .... X.XXXXX.X ..X..... .... X. ..
134 .... X....XXX. .....X .... XXXX .... .XX XX
26 .... XXXXXXXX. ....
36 .... XXXXXXXXX ....
53 .... XX.X.XXX. .... XXX XX
73 .... .XXXXXXXX X
81 .... XXXXXXXXX .... .X .. .. ..
87 .... XXXXXXXXX XX
100 .... XX....X.. .....X .... X.X ... ... XX .. XX
112 .... .XXXXXXXX ....
113 .....X.. ..... .X.....X. .... X..X .... XXXX .... XX .. .. ..
123 .... XXXXXXXXX .... XX .. .. ..
138 .... X.XXXXX.X XX
142 .... .XX.XXXXX ....
40 .... XXXXXXXXX ....
47 ....XX.X .... XXX XXX ... ...
49 .... XXXXXXXXX XX
60 .... X.XX.X... X..X.... .... X.
63 .... XXXX.XXX. ....
70 .... .XXXXXXXX ....
2 .... X.X.XXX.. .... .XX ... ... XXX XX
79 .... XXX..XXXX .... XXX .. ..
83 .... XX...XXX. .... XXX ... XXX .X. ... ... X. .. .. ..
13 .... XXX.XXXX. .... X..X ....X XX
140 ....XXXXX.X ....
56 .... XX.X.XXX. ....XXX XX
33 .... XXXXXXXX. XX
86 .... XX.X.XXX. X
23 .... XXXXXXXXX ....
94 .... X....X.. .... XXXX .... XX XX
98 .... .XXXXXXXX ....
104 .... XXXXXX.XX ....
105 ....XX.X ..... XXXXX.XX ....
122 .... XXX.XXXXX ....
133 .... XXXXXXXX. ....
144 ....X... .... XXXX X
3 .... XX...XXX. XX
57 .... XX...XXXX XX
84 .... XX...XX.. .... XXX ... ... XXX ... ... X. .. .. ..
45 .... XXX.XXXX. XX
46 ....XX.. ..... XXXX.XXX ....
135 .... XX.X.XXXX ....
85 ...X..... ....X. .... XXXX .... XX .. X. XX
106 .... X..... .... XXXXX ....
DOCUMENTS ORIGINAUX
DOCUMENTS NON PERTINENTS
CORRESPONDANCE DE CODES

décrits par une famille de codes très rares. Ces brevets représentent deux catégories:

- des brevets non pertinents:
  ils ont très peu de codes dans la classe des codes "communs" à tous les brevets (la 3ème classe)

- des brevets pertinents dont les revendications sont très originales:
  ils possèdent pratiquement tous les codes de la classe de codes "communs".
  Quelques exemples de ces brevets originaux:
  - La référence 41: utilisation d'oxydes métalliques dans un précipité aqueux qui peut être moulé et mis en forme pour rentrer dans la constitution de patchs.
  - La référence 78: polymère ayant des caractéristiques de très grande transparence, haute perméabilité, flexibilité et doux en état hydraté. Outre les applications en comme membrane de séparation de gaz, il peut servir de garniture de blessure perméable à l'oxygène, de lentille de contact, de patch buccal ou d'implant dans le corps.
  - La référence 55: patch qui délivre une haute dose initiale de médicament suivi d'une dose basse et régulière (système à plusieurs réservoirs d'agent actif).

● CONCLUSION

Cette méthodologie nous a permis de mettre en évidence plusieurs caractéristiques:

Grâce à la complémentarité des codes :

On relève des détails techniques très précis qui sont, en général, noyés au travers d'aspects très généraux et par conséquent très difficiles à détecter; mais ceci sans perdre de vue les aspects généraux auxquels ils se rapportent.

Pour l'exemple du sous-thème *principes actifs de type stéroïde*

* l'utilisation d'un seul type de codification permet:
  - soit de connaître plus ou moins rapidement quels sont les documents qui traitent des stéroïdes de façon générale, sans aller beaucoup plus loin dans le détail.

  - soit de déterminer plus ou moins rapidement leurs aspects spécifiques, sans savoir qu'ils parlent de stéroïdes.

* L'utilisation simultanée des différents types de codifications permet non seulement de déterminer très facilement les aspects généraux traités dans les documents (ici l'aspect stéroïde), mais aussi de révéler les aspects spécifiques développés dans ces mêmes documents (par exemple des problèmes de température ou d'adhérence).

Donc, choisir d'exploiter plusieurs champs de descriptions pour un traitement bibliométrique permet d'aboutir à une **MEILLEURE CARACTERISATION** du corpus

Grâce à la correspondance des codes :

On met en regard des ensembles de codes qui sont fortement dépendants (ou proches) les uns des autres parce que très souvent employés conjointement dans les références. Ceci nous a permis de:

- déceler très aisément les documents non pertinents de notre corpus c'est à dire ceux qui représentent le **BRUIT**
- reconnaître très facilement les **DOCUMENTS ORIGINAUX** (au sens innovateur du terme) qui sont basés sur l'emploi ou la description de techniques, de méthodes, de procédés qui se démarquent des autres documents.

La représentation de l'information obtenue par les outils d'analyses

L'expertise et l'interprétation des résultats ne sont fonction que de l'information qui résulte des traitements analytiques auxquels on a recours.

De ce fait, afin que l'ensemble des experts puisse interpréter de façon objective les résultats, il est nécessaire d'utiliser des méthodes de traitements qui permettent non pas de résumer l'information de départ, mais de la restructurer afin d'en dégager les faits marquants sans perdre le reste. Trop souvent négligé par les méthodes d'analyses traditionnelles, cet ensemble rebut constitue une part considérable de l'information initiale (lois de Zipf- Bradford). Bien que ces éléments soient présents à des fréquences très faibles, ils n'en contiennent pas moins une information intéressante. En effet, on y relève **L'INFORMATION MARGINALE** qui selon le cas peut se traduire en termes d'information **INNOVANTE** ou **DISCORDANTE** (le bruit).

L'ANALYSE RELATIONNELLE parait, de ce fait, parfaitement appropriée à l'analyse bibliométrique. Elle **NE NEGLIGE AUCUNE INFORMATION** même si sa faible présence lui donne a priori un caractère mineur.

Le rapport temps d'analyse - temps d'expertise

Afin que les experts du domaine puissent se consacrer pleinement à l'expertise du sujet au travers des résultats livrés par le traitement, celui-ci doit prendre le moins de temps possible.
Pour minimiser ce rapport temps d'analyse - temps d'expertise, on se doit d'utiliser des méthodes de traitement, de calcul, d'analyse, qui permettent d'obtenir des résultats dans des délais extrêmement brefs et qui permettent de s'utiliser de façon systématique afin de réorienter les analyses selon les interprétations que l'on obtient.
C'est la première caractéristique qu'un système de surveillance doit vérifier pour assurer un fonctionnement performant. L'accélération de l'obsolescence des technologies impose à la veille technologique d'être le pourvoyeur d'une **INFORMATION ELABOREE DANS DES TEMPS RESTREINTS**. Pourquoi dans ces conditions ne pas profiter des avantages que nous offre l'informatique pour le traitement des données. C'est dans cet état d'esprit que le logiciel **DATAVIEW** a été conçu.

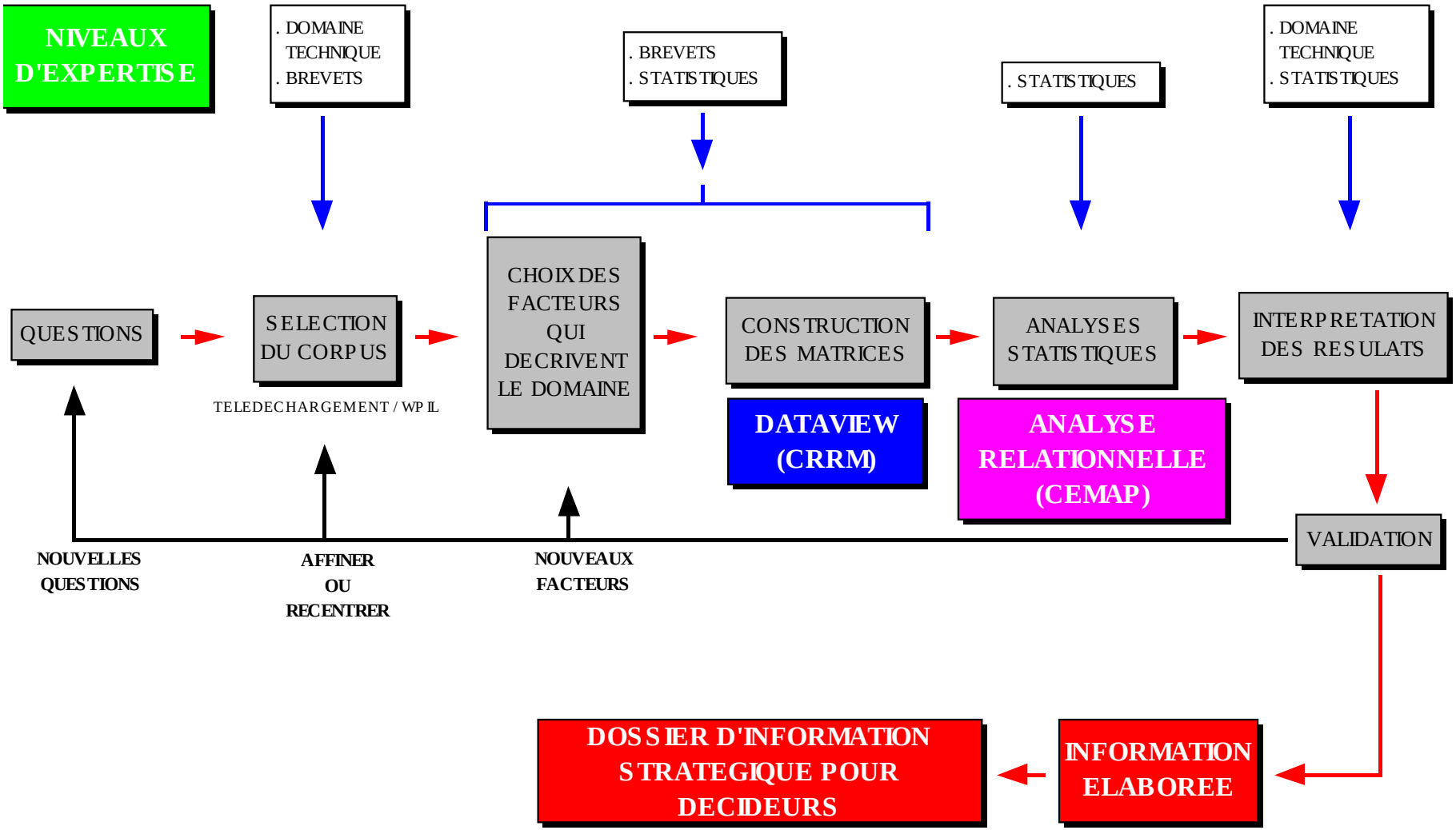
Le recours permanent à des experts différents

Pour que l'analyse délivre une **INFORMATION FIABLE** à but stratégique, il est indispensable de valider chaque étape dans l'élaboration du dossier, depuis la sélection du corpus jusqu'à l'interprétation des résultats, par des **NIVEAUX D'EXPERTISES** différents et adaptés.

- Expert du domaine technique étudié
- Expert brevet
- Expert de l'information
- Expert en statistiques

Ce dernier point est illustré sur le synoptique des traitements effectués pour cette étude (*illustration 10).*
Ce synoptique est volontairement replacé dans le contexte d'une veille technologique industrielle. Le processus démarre sous la manifestation de questions sur des sujets sensibles à l'entreprise (facteurs critiques). L'aboutissement de la chaîne du traitement est l'élaboration de dossiers stratégiques pour informer les décideurs de la situation présente et des évolutions de tendances.
Ce synoptique montre les exigences et les compétences qu'impose l'insertion de l'outil bibliométrique dans un système de surveillance industriel. Ce sont les contraintes que doit respecter un système de veille pour permettre de traiter des sujets dont la masse et la complexité des connaissances ne peuvent être appréhendés par de simples traitements manuels, dans des temps acceptables.

SYNOPTIQUE


NIVEAUX D'EXPERTISE
. DOMAINE TECHNIQUE
. BREVETS
. BREVETS
. STATISTIQUES
. STATISTIQUES
. DOMAINE TECHNIQUE
. STATISTIQUES
QUESTIONS
SELECTION DU CORPUS
TELEDECHARGEMENT / WPIL
CHOIX DES FACTEURS QUI DECRIVENT LE DOMAINE
CONSTRUCTION DES MATRICES
DATAVIEW (CRRM)
ANALYSES STATISTIQUES
ANALYSE RELATIONNELLE (CEMAP)
INTERPRETATION DES RESULATS
VALIDATION
NOUVELLES QUESTIONS
AFFINER OU RECENTRER
NOUVEAUX FACTEURS
DOSSIER D'INFORMATION STRATEGIQUE POUR DECIDEURS
INFORMATION ELABOREE

BIBLIOGRAPHIE

[1] La veille technologique
sous la direction de H Dou, H Desval
Dunod, 1992

[2] Pratique de la veille technologique
F Jakobiak
Les éditions d'organisation, 1991

[3] L'analyse des données au service de la bibliométrie. Outils de veille technologique à la dimension des moyennes entreprises.
H Dou, L Quoniam, H Rostaing, W Nivol
Revue Française de biliométrie, Vol 8, p 27-67, déc. 1990

[4] The relationship of information science et the social sciences- a cocitation analysis
H Small
Information porcessing & management, Vol 17, N°1, p39-50, 1981

[5] L'analyse des associations
B Michelet
Thèse de doctorat, Université de Paris VII, 26 oct 1988

[6] An analysis of citations in statistical journals
K Fz Agirre, J M Piris, F Tusell
Proceedings of the first international symposium on applied stochastic models and data analysis, p 15-24, 23-26 april 1991, Granada, Spain

[7] Application de l'analyse relationnelle à la veille technologique: des outils d'analyse de l'information documentaire
C Bédécarrax, C Huot
Revue française de bibliométrie, Vol 9, p 66-80, sept 91

Conclusion

VII. Conclusion

La gestion de l'information, lorsqu'elle entre dans un processus de décision, ne peut se concevoir sans une maîtrise complète de trois facteurs:

- la manipulation d'une grande masse de données
- la fourniture d'informations de qualité comme aide à la décision
- le traitement des deux premiers facteurs dans un laps de temps réduit

La gestion de l'information scientifique et technique, que ce soit dans un processus de veille technologique industrielle ou dans un tout autre contexte, ne manque pas à cette règle. Au terme de ce mémoire, nous estimons que l'informatique est un outil indispensable à cette tâche. Il répond parfaitement à ces trois conditions:

- c'est un moyen d'accès privilégié à de colossales masses d'informations
- les applications statistiques et graphiques sont des procédés adaptés à l'esprit de synthèse auquel l'individu a recours lors du processus prise de décision
- l'informatique multiplie considérablement la rapidité de toutes opérations de traitement d'informations

La conception et le développement de l'application informatique présentée dans ce mémoire s'inscrivent totalement dans ce contexte d'aide à la maîtrise de l'information scientifique et technique.

L'analyse de la chaîne des traitements, que requiert cette gestion de flux d'information, laisse apparaître une absence d'automatisation informatique pour l'une des étapes: le passage des données textuelles aux données numériques.
La collecte des données s'effectue dans des temps records par l'intermédiaire des télécommunications. Par ailleurs, les logiciels de traitement des données numériques sont devenus d'un emploi courant. **L'outil informatique présenté dans ce mémoire est donc d'une utilité considérable pour l'automatisation et la systématisation du traitement de l'information textuelle**.

Une analyse consciencieuse du besoin en veille technologique (second chapitre) et des méthodes à mettre en oeuvre pour répondre à ce besoin (quatrième chapitre) a permis **la conception d'un outil informatique très comple**t. Pour répondre au mieux à sa fonction d'aide dans la manipulation des données textuelles, il est conçu de façon à:

- **accepter tout type de données textuelles en entrée:**
  tout fichier texte au format ASCII peut être étudié dès lors qu'il contient des repères délimitant les unités statistiques à considérer pendant les calculs. Ainsi tout ensemble de références bibliographiques peut être assimilé à la population statistique. Il en est de même pour des fichiers rapportant, soit des résultats d'enquêtes, soit des renseignements sur des clients ou des fournisseurs, soit des textes... Cet outil peut être appliqué à d'autres domaines qu'à la bibliométrie.

- permettre une diversité de croisement d'information:
  comme le type de renseignements le plus utile est la mise en évidence de relations entre deux catégories d'éléments, le logiciel est volontairement conçu pour offrir toutes les possibilités de recoupement entre les différentes données d'une même population statistique.

- **fournir une grande diversité de résultats numériques:**
  les données peuvent être présentée selon divers critères d'arrangement. Chacun d'eux donne un éclairage différent des données. Nous avons volontairement focaliser notre attention sur le fait que ces données puissent ensuite faire l'objet d'analyses statistiques plus complexes ou de traitements infographiques.

- **être agréable à l'emploi:**
  la maniabilité, la souplesse et l'ergonomie d'un logiciel sont devenus des atouts indispensables pour remporter l'approbation des utilisateurs. Cet aspect informatique n'a pas été négligé.

Un outil informatique comme celui-ci peut répondre à d'un double emploi:

- s'**insérer dans un processus de veille technologique**, sa mise en oeuvre peut permettre l'élaboration d'indicateurs stratégiques essentiels à la préparation et au suivi des programmes de recherches et de développements. L'implantation de cet outil au sein de plusieurs sociétés (CEDOCAR, CETIM, L'OREAL, MOULINEX et en cours de négociation pour GAZ DE FRANCE et THOMSON) nous a prouvé sa parfaite

adaptation aux réalités industrielles. Son utilisation dans des études stratégiques industrielles a déjà livré des résultats plus qu'encourageants. Le fait de savoir que des programmes de recherche aient été révisés, à la suite d'études bibliométriques le mettant en oeuvre, ne peut que nous conforter sur le bien-fondé de cet outil.

- le caractère "ouvert" aux divers traitements bibliométriques et textuels fait de cet outil une **plate-forme de recherche idéale**.

  Le secteur de recherche, spécifique à la gestion automatisée de l'information scientifique et technique, est récent. De nombreuses améliorations sont encore à apporter aux méthodes existantes. La recherche de nouvelles méthodes d'analyses statistiques ainsi que le développement de représentations graphiques plus adaptées aux données bibliographiques parait maintenant être un axe prioritaire de la recherche en bibliométrie. Le Centre Recherche Rétrospective de Marseille s'engage fermement dans cette direction en collaborant avec le CESMAP d'IBM et Clustan pour l'aspect statistique, avec l'IRIT pour l'aspect infographique et avec le CETIM pour l'aspect applicabilité.

  L'application de DATAVIEW à d'autres traitements que ceux des textes bibliographiques laisse un grand champ de recherche et d'application à explorer.

En France, outre le développement d'outils informatiques, l'activité de veille technologique nécessite encore de nombreuses améliorations pour rivaliser avec des pays comme le Japon ou les Etats Unis. Sur ce point, une volonté doit s'imposer chez les industriels, et auprès des dirigeants nationaux ou régionaux. De nombreux aspects restent à développer tel que:

- la mise en place de **systèmes informatiques opérationnels** pour automatiser toute la chaîne des traitements de données scientifiques et techniques. Dans notre domaine, la notion de station de travail commence seulement à prendre forme.
- le développement **d'outils informatiques adaptés à l'aide à la gestion des informations informelles**. L'aspect multimédia prenant une place de plus en plus importante dans le monde informatique il serait dommage pour ne pas en trouver une application dans l'activité de veille industrielle.
- **l'ouverture aux réseaux informatiques d'information de la recherche**. Là encore, la France à pris un retard considérable par rapport aux Etats Unis ou à l'Angleterre.
- à l'instar des japonais il serait bien venu d'**établir des réseaux de collecte d'informations à une échelle nationale** ou même régionale par catégorie d'activité industrielle.

- **insuffler des aides aux petites et moyennes entreprise**s pour leur permettre de profiter elles aussi d'une activité de surveillance de leurs technologies.
- et finalement le plus important, **parfaire la formation de spécialistes** dans le domaine du traitement de l'information scientifique et technique. Il faut aussi qu'une sensibilisation générale au problème dans tous les enseignements scientifiques se mette en place

Nous conclurons par une citation, tirée de l'ouvrage de Bertin [BERT77], qui résume très bien l'idée générale de ce mémoire:

"Décider c'est choisir et choisir c'est d'abord s'informer"

Bibliographie

VIII. Bibliographie

1. ADIS86 Evaluation of the national performance in basic research
   Adisory board for the research councils
   The royale society, economic and social research council, Londres, 1986

2. AGIR91 An analysis of citations in statistical journals
   Agirre K Fz, Piris J M, Tusell F
   Proceedings Ist int. symp. applied stochastic models & data analysis, 23-26/04/1991, Granada, Sp.

3. AIYE77 Bradford distribution theory - compounding of Bradford periodical literatures in geography
   Aiyepeku W O
   Journal of Documentation, Vol 33, N°3, p 210-219, 1977

4. ALAB79 Bradford's law and its application
   Alabi G
   International Library Revue, Vol 11, p 151-158, 1979

5. ASHO92 Laser research in india: scientometric study end model projections
   Ashok J, Garg K C
   Scientometrics, Vol 23, N°3, 1992, p 395-415

6. ASHT83 Patent analysis as a technology forcasting tool
   Ashton W B, Campbell R S, Levine L O
   Fall conference, Atlantic city, NJ, sept 18-21, 1983

7. BALM91 Who's doing what in humane genome research?
   Balmer B, Martin B R
   Scientometrics, Vol 22, N°3, 1991, p 369-377

8. BARE91 Clustering research fields for macro-strategic analysis: a comparative specialisation approch
   R. Barré
   Scientometrics, Vol 22, N°1, 1991, p 95-112

9. BARR92 Analyse macro-bibliométrique des relations science-technologie: les "balances scientifiques" comparées des pays
   Barré R, Zitt M
   Actes du colloque Journées d'études ADEST, 1-2 juin 1992, p 171-774

10. BASS69 A new product growth model for consumer durables
Bass F M
Management science, Vol 15, 1969, p 215

11. BATELL Battelle Europe
7 route de Drize
CH-1227 Carrouge-Genève, Suisse

12. BAUI92 Les français dans la base de publications scientifiques SCI de l'ISI
Bauin S, Crance M, Sigogneau M, Quinault L
Actes du colloque Journées d'étude ADEST, 1-2 juin 1992, p 41-44

13. BEDE91 L'application de l'analyse relationnelle à la veille technologique
Bédécarrax, Huot; CEMAP
Revue française de bibliométrie, Vol 9, 1992, p 64-80

14. BEDE92 Analyse relationnelle: des outils pour la documentation automatique
Bédécarrax C, Huot C
Dans: La veille technologique, sous la direction de H Desval et H Dou, Dunod, 1992, 436 p.

15. BELO Etude bibliométrique sur la productique
Bélot J M
Technologie mécanique, N°11, p 2-9

16. BELO92 Analyse bibliométrique et avis d'experts: le cas des revêtements mécaniques
Bélot J-M, Saint-Etienne A, Lieurade H-P
Actes du colloque Journées d'études ADEST, 1-2 juin 1992, p 175-178

17. BENN84 Multivariate regression models for estimating journal usefulness in physics
Bennion B, Karschamroon S
Journal of documention, Vol 40, 1984, p 217-227

18. BENZ L'analyse des données
J.P. Benzécri
Tome 1: La taxinomie
Tome 2: L'analyse des correspondances
Dunod, Paris

19. BERT77 La graphique et le traitement graphique de l'information
Bertin J

Flammarion, 1977, 273 p.

20. BILL89 Medline vue par l'analyse factorielle et la classification automatique
Billard P , Dousset B, Hilaire A, Laurent D, Paoli C, Longevialle C
Revue française de bibliométrie, Vol 7, 1990, p 61-74

21. BRAD34 Sources of information on specific subjects
Bradford S C
Engineering, January 1934, Vol 137, p 85-86

22. BRAD48 Documentation
Bradford S C
Crosby Lockwood & Son LTD., London, 1948, pp.156

23. BROA87 Early approches to bibliometrics
Broadus R N
Journal of the american society for information science, march 1987, 38 (2)

24. BROO68 The derivation and the application of the Bradford-Zipf distribution
Brookes B C
Journal of documentation., Vol 24, N°4, p 247-267, 1968

25. BROO88 Biblio-, sciento-, info-métrics ??? What are we talking about?
Brooks B C
Informetrics 87/88, Proceedings of the diepenbeek conference, Amsterdam: Elsevier, 1988

26. BUDD91 Superstring theory: information transfert in an emerging field
Budd J, Hurt C D
Scientometrics, Vol 21, N°1, 1991, p 87-98

27. BUFF91 Etude de l'impact de la revue de l'Institut Français du Pétrole par des méthodes bilbiométriques
Buffeteau A
Revue française de biliométrie, Vol 9, p 293-323, 1991

28. BURR91 The bradford distribution and the gini index
Burrel Q L
Scientometrics, Vol 21, N°2,1991, p 181-194

29. CALL87 La recherche française est-elle en bonne santé?
Callon M, Leydesdorff L
La recherche, N°186, Mars 1987, p. 412-419

30. CALL91 Co-word analysis as a tool for describing teh network of interactions between basic and technological research: the case of polymer chemistry
Callon M, Courtial J-P, Laville F
Scientometrics, Vol 22, N°1, 1991, p 155-205

31. CARD92 Une station de travail de lecture des contenus documentaires pour la veille scientifique et technique
Cardine P, Muller E, Turner W A
Actes du colloque Journée d'étude ADEST, 1-2 juin 1992, p 29-38

32. CARP73 Clustering of scientific journals
Carpenter M P, Narin F
Journal of teh amercain society for information science, Vol 24, 1973, p 425-436

33. CARP79 Similarity of Pratt's measure of class concentration to the Gini indice
Carpenter M
Journal of the american society for information science, Vol 30, p108-110, 1979

34. CARP81 The adequacy of the science citation index (SCI) as an indicator of international scientific activity
Carpenter M P, Narin F
Journal of american society for information science, Vol 32, N°6, 1981, p 430-439

35. CHAU88 Le traitement linguistique de l'information
Chaumier J
Entreprise moderne d'édition, 1988, 186 p.

36. CHEN85 Statistical models of text: a system theory approach
Chen Y S
Ph.D. dissertation, Purdue University, 1985

37. CLUSTA Cluster analysis software
Développé par D. Whishart
Clustan Limited, 16 Kingsburgh Road, Edingburg EH12 6DZ, Scotland

38. COIL77 Lotka's freequency distribution of scientific productivity
Coile R C

Journal of the american society for the information science, Novembre 1977, Vol 28, N°6, p. 366-370

39. COLE17 The history of comparative anatomy. Part I: astatistical analysis of the literature
Cole F J., Eales N B
Science progress, London, April 1917, Vol 11, p.578-596

40. COMM89 Rapport de la commision Europe technologique, Industrielle et commerciale
Commissariat général du Xème plan
Juillet 1989

41. COUR76 Analyse et langage documentaire
Courrier Y
Documentaliste, Vol 13, N°5-6, 1976, p 178-189

42. COUR90 Introduction à la scientométrie. De la bibliométrie à la veille technologique.
Courtial J-P
Anthropos - Economica, 1990, 137 p.

43. CREE84 Bibliometric analysis of ethnomusicology
McCreery L S, Miranda L
Proceedings of the american society for information science (ASIS) 47th annual meeting, , Philadelphia, PA White plains, NY: Knolewdge Industrie Publications, 21-25 oct 1984, p 212-216

44. CZER90 Scientometric indicator for a speciality in theoretical high-enrgy physics
Czerwon H-J
scientometrics, Vol 18, N°1-2, 1990, p 5-20

45. CZEK09 Zur Differentialdiagnose der Neandertalgruppe
J. Czékanowski
Korrespondenz-Blatt deutsch. ges. Anthropol. Ethnol. Urgessch. 40, p.44-47, 1909
Allemagne

46. DEMA92 De la pratiqueet du bon usage des processus d'évaluation des chercheurs
Demazure M
Pour la science, N°117, juil 1992, p 7

47. DERWEN Derwent Publications LTD
Rochdale House, Theodbalds Road
London WC1 X3RP, GB

48. DEVA89 Veille technologique par la bibliométrie: une image statistique de la banque de données bibliographique du CETIM
Devalan P, Belle F
Technologie mécanique, N°9, 1989, p I-XI

49. DEVA90 La bibliométrie. un outil de veille technologique pour l'entreprise
Devalan P, Candoret J P, Bouvet C,, Lion J C
CETIM-informations, N°116, juin 1990, p 89-95

50. DEVA91 Les marchés de la productique
Devalan P, Belot J-M, Frémaux P
CETIM-Informations, N°124, oct 1991, p 35-41

51. DIMO90 Méthodologie pour l'étude de l'évolution scientifique et technologique
Ioana Dimo
Revue française de bibliométrie, Vol 6, 1990, p 302-330

52. DOBR69 The information basis of scientometrics
Dobrov G M, Korennoi A A
A I Michailov et al. (eds), On theoretical problems of informatics, Moscow VINITI for FID, 1969, p 165-191

53. DORE85 Structure equivalence in psychology journal network
Journal of american society for information science, Vol 36, N°6, nov 1985, p 411-417

54. DORE87 Banques de données et analyses multivariables
Dore J C, Gilbert J, Miquel J-F, Dutheuil C
Revue française de bibliométrie, Vol 1, mai 1987, p14-25

55. DOU87 New aspect of online retrieval, get more from your dowloaded data. Noise can be useful!
Dou H, Hassanaly P, La Tela A
Colloque Online meeting, déc 1987, Londre

56. DOU88 Infographics analytical toold for décision makers.
Dou H, Hassanaly P, Quoniam L
Scientometrics, Vol 17, N°1-2, p 133-149

57. DOU90 Informations stratégiques en chimie. Analyse topologique automatique de la base Chemical Abstract
Dou H, Hassanaly P, Quoniam L
Revue francaise de bilbiométrie, Vol 7, 1990, p 14-

58. DOU91 The scientific dynamic of a city: a study of chemistry in Marseille from 1981 to the present
Dou H, Quoniam L, Hassanaly P
Scientometrics, Vol 22, N°1-2, 1991, p 83-93

59. DOU92 La veille technologique
Sous la direction de Dou H, Desvals H
Dunod, Paris, 1992, 436 p.

60. DOUC91 Qu'apportent la représentation de la 4ème dimension en analyse de données multidimensionnelles
B. Doucet, T. Dkaki, S. Koussoubé, C. Longevialle, A. Hilaire
Congrés de la S.F.B.A., Les systèmes d'informations élaborées, Ile Rousse, Juin 1991

61. DROT80 Telephone communication
Drott M C
1 april 1980

62. DUCL92 D'une boîte à outils à la description du domaine des cognisciences
Ducloy J, Polanco X
Actes du colloque Journées d'études ADEST, 1-2 juin 1992, p 65-74

63. EGGH86 A characterization of distribution which satisfy Price's Law and consequences for the laws of Zipf and Mandelbrot
Egghe L, Rousseau R
Journal of information science, North-Holland, 1986, Vol 12, p. 193-197

64. EGGH87 Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule
Egghe L
Journal of the american society for information science, Vol 38, N°4, p 288-297, july 1987

65. EGGH88 The relative concentration of a journal with the respect to a subject and the use of online services in calculating it
Egghe L
Journal of the american society for information science, Vol 39, N°4, p 281-284, july 1988

66. EGGH90 Elements of concentrations theory
Egghe L, Rousseau R

L. Egghe, R. Rousseau (Eds), Informetrics 89/90, Elsevier, Amsterdam, 1990

67. EGGH91 The exact place of zipf's and pareto's law amongst the classical informatrics laws
Egghe L
Scientometrics, Vol 20, N°1, 1991, p 93-106

68. ESTI69 La statistique bibliographique
Estivals R
Bulletins des bibliothèques de France, N°12, déc 1969, p 481-502

69. ESTO16 Gammes stenographiques
Estoup J B
4th edition, 1916

70. FAIR69 Progress in documentation
Fairthorne R A
Journal of documentation, Vol 25, N°4, dec 1969, p 319-343

71. FANO56 Documentation in action
Fano R N
New York, Reinhold, 1956, pp 238-244

72. FEDO82 A Zipfian model of an automatic bibliographic system: an application to MEDLINE
Fedorowicz J
Journal of the american society for information science, july 1982, p 223-232

73. FEDO82 The theoretical foundation of Zipf's law and its application to the bibliographic database environment
Fedorowicz J
Journal of the american society for information science, September 1982, p. 285-293

74. GARF79 Citation indexing - its theory and application in science, technology, and humanities
Garfield E
John Willey & sons, New York, 1979, 274 p.

75. GARF81 The 1,000 contemporary scientists most-cited 1957-1978. Part I. The basic list and introductiion
Current contents, N°41, oct 1981, p 5-14

76. GARF82 Journal citation studies: 36 pure and applied mathematics journals: what theys cite and vice-versa

Garfield E
Currents Contents, Vol 15, 1982, p 5-13

77. GOFF64 Generalisation of epidemic theory; an application to the transmission of ideas
Goffman W, Newill V. A.
Nature, 1964, Vol 204, p225-228

78. GRIF79 The aging of scientific literature: a citation analysis
Griffith B C, Anker A, Servi P, Drott M C
Journal of documentation, Vol 25, N°3, sept 1979, p 179-196

79. GRIF83 The structure of the social and behavioral sciences literature
Griffith B C, Small H G
Stockholm, Sweden: Royal institute of technology library, 1953, 53 p.

80. GROO67 Bradford's law and the Keenan-Atherton data
Groos O V
American Documentation, Vol 18, p 46, 1967

81. GROS27 College libraries and chemical education
Gross P L K, Gross E M
Science, October 1927, V 66, p 1229-1234

82. HARG80 Research areas and stratification process in science
Hargens L L, Mullins N C, Hecht P K
Social studies of science, Vol 10, N° 1, feb 1980, p 55-74

83. HAWK77 Unconventional use of one-line information retrieval systems: one-line bibliometrics studies
Hawkins D T
Journal of American society for information science, 1977, Vol 28, N°1, p.13-18

84. HE86 A discipline-specific journal selection algorithme
He C, Pao M
Informationn processing and management, Vol 22, 1986, p 405-416

85. HEAL86 An experiment in science mapping for research planning
Healey P, Rothman H, Hoch P K
Research policy, North-Holland, Vol 15, 1986, p 233-251

86. HOPK84 New causal theory and ethnomethodology: cocitation patterns across decade
Hopkins F
Scientometrics, Vol 6, N°1, jan 1984, p 33-53

87. HUBE78 A relationship between two forms of Bradfro's law
Hubert J J
Journal of the american society for information science, 1978, Vol 29, N°2, p. 159-161

88. HULM23 Statistical bibliographie in relation to the growth of modern civivlization.
Hulme E W
London: Grafton; 1923, 44 p.

89. HUOT92 New method concerning analysis of downloaded data for strategic decison
Huot C, Quoniam L, Dou H
Scientometrics, à paraitre

90. HUST78 Identifying a set of inequality measures for science studies
Hustopecky J, Vlachy J
Scientometrics, Vol 1, p 85-98, 1978

91. INFO Information und kommunication
I+K France
9 avenue Ville Preux, 78340 Clayes-sous-bois

92. JAGO90 Etude d'un front de recherche identifié par les co-citations à partir de la base PASCAL
Jagodzingski-Sigoneau M, Bauin S,Courtial J-P, Turner W A,
Cahiers de l'ADEST,1990

93. JAKO91 Pratique de la veille technologique
Jakobiak F
Les éditions d'organisation, Paris, 1991, 232 p.

94. JAKO92 Exemples commentés de veille technologique
Jakobiak F
Les éditions d'organisation, 1992, 199 p.

95. KAMI89 Aide à la décison. Les neufs commandements
Kami M
McGraw-Hill, Paris, 1989

96. KEND60 The bibliography of operational research
Kendall M. G.
Operational research quartrely, 1960, Vol 2, p 31-

97. KESS63 Bibliographic coupling between scientific papers
Kessler M M
American documentation, 1963, Vol 14, p 10-15

98. KESS65 Comparison of the results of bibliographic coupling and analytic subject indexing
Kessler M M
American documention, 1965, Vol 16, p223-233

99. LAFO91 Les distributions bibliométriques
Lafouge T, Quoniam L
Revue française de bibliométrie, Vol 9, p 128-138

100. LAFO91 Problématique de la circulation de l'information
Lafouge T
Documentaliste, Vol 28, N°3, 1991, p 132-134

101. LAIN91 La veille technologique - De l'amateurisme au professionnalisme
Lainé F
Eyrolles, 1991, 138 p.

102. LAW88 Policy and the mapping of scientific change: a co-word analysis of research into environmental acidification
Law J, Bauin S, Courtial J P, Whittaker J
Scientometrics, Vol 14, 1988, p 251-264

103. LAW92 Mapping acidification research: a test of the co-word method
Law J, Whittaker J
Scientometrics, Vol 23, N°3, 1992, p 417-461

104. LAWA73 Bradford's law and the literature of agricluture
Lawani S M
International Library Revue, Vol 5, p 341-350, 1973

105. LAWA82 On the heterogeneity and classification of author self-citations
Lawani S M
Journal of the american society for information science, Vol 33, N°5, sept 82, p 281-284

106. LAWS80 A bibliometric study of the nex subject field: energy analysis
Lawson J, Korstrewski B, Oppenheim C
Scientometrics, Vol 2 , N°3, may 1980, p 227-237

107. LEBA88 Analyse statistique des données textuelles
Lebart L, Salem A
Dunod, Paris, 1988, 209 p.

108. LECR90 Système d'accès à des ressources documentaires - vers des antéserveurs intelligents -
Le Crosnier H
Thèse de doctorat, Université d'Aix-Marseille III, Faculté St Jérome, 355 p., 21 déc 90

109. LEGE84 Ecologie numérique. I - Le traitement multiple des données écologiques. II - La structure des données écologiques
Legendre L, Legendre P
Masson, Presses de l'université du Québec, 1984

110. LEIM67 The Bradford distribution
Leimkuhler
Journal of Documentation, Vol 23, N°3, p197-207, 1967

111. LERM Classification et analyse ordinale des données
I. C. Lerman
Dunod, Paris

112. LESC86 Système d'information pour le management stratégique de l'entreprise
Lescat H
McGraw-Hill, 1986

113. LEYD86 The development of frames of reference
Leydesdorff L
Scientometrics, Vol 9, N° 3-4, 1986, p 103-125

114. LEYD87 Co-words and citations relations between documments sets and environments
Leydersdorff L
First international conference on bibliometrics an theoetical aspects of information retrieval, August 24-28, 1987, Diepenbeek, Belgium

115. LEYD87 Various methods for mapping of science
Leydesdorff L
Scientometrics, Vol 11, N°5-6, may 1987, p 295-324

116. LOTK26 The frequency distribution of scientific productivity
Lotka A J
Journal of the Washington academiy of sciences, 1926 June, Vol 16, N° 12, p. 317-323

117. MAIA84 On the unity of Bradford's law
Maia M F, Maia M D
Journal of the documentation, Vol 40, N°3, p 206-216

118. MARC Optimisation en analyse ordinale des données
Marcotorchino F, Michaux P
Masson, Paris

119. MARC87 Block seriation problems: a unified approach
F. Marcotorchino
Applied stochastic models ans data analysis, Vol. 3, p 73-91, 1987

120. MARS81 Citation networks in information science
Marshakova I V
Scientometrics, Vol 3, N°1, jan 1981, p 13-26

121. MART83 Methods for evaluating the number of relevant documents in a collection.
Martin W A
Journal of information science, Vol 6, N°5, feb 1983, p 173-177

122. MART89 La veille tecnologique, concurrentielle et commerciale
Martinet B, Ribault J-M
Les éditions d'organisation, Paris, 1989, 304 p.

123. MCCA83 The author cocitation structure of macroeconomics
Mac Cain K
Scientometrics, Vol 5, N°5, sept 1983, p 227-289

124. MCCA86 The paper trails of scholarship: mapping the literature of genetics
Mac Cain K
Library quatrely, Vol 56, N°3, july 1986, p 258-271

125. MCCA89 Citation context analysis in genetics
Mac Cain K W, Turner W A
Scientometrics, Vol 17, 1989

126. MCRO89 Problems of citation analysis: a critical review
Mac Roberts M, Mac Roberts B
Journal of the american society for information science, Vol 40, N°5, p 342-349, 1989

127. MEND53 An information theory of the statistical structure of language
Mendelbrot B
Proceedings of the symposium on applications of

communication theory, Butterworth, 1953, London, p 486-500

128. MENS88 La technique en crise
Mensch G
Recherche innovation industrie, Mars 1988

129. MICH88 L'analyse des associations
Michelet B
Thèse de doctorat, Université de Paris VII, 26 oct 1988

130. MOED85 The application of bibliometrics indicators: important field and time dependant factors to be considered
Moed H, Burger W J M, Frankfort J G, Van Raan A F J
Scientometrics, Vol 8, N°3-4, 1985, p 177-204

131. MOED86 Observations and hypotheses on the phenomenon of multiple citation to a research group's oeuvre
Moed H F, Van Raan A F J
Scientometrics, Vol 10, N°1-2, july 1986, p 17-34

132. MOED91 International scientific co-operation and awareness within the european community: problems and perspectives
Moed H F, De Bruin R E, Nederhof A J, Tijssen R J W
Scientometrics, Vol 21, N°3, 1991, p 291-311

133. MORI85 L'excellence technologique
Morin J
Publi Union, Editions Jean Picollec, Paris, 1985

134. MOU87a Utilisation des banques de données sur les brevets
M Moureau, A Girard
Revue française de bibliométrie, Vol 1, N°2, 1987, p 9-20

135. MOU87b Patents and statistical analysis; a user view
Moureau M, Girard A
Derwnt online news, N°4, sept 87, p 5-8

136. MULL84 Group structure of co-citation clusters - comparative study
Mullins N C, Hargens L L, Hetch P K, Kick E L
American sociology review, Vol 42, N°4, aug 1977, p 552-562

137. MURP73 Lotka's law in the humanities?
Murphy L J
Journal of the american society for information

science, Vol 24, p 461-462, 1973

138. NADE83 Commitment and co-citation - an indicator of incommensurabiblity in patterns of formal communication
Nadel E
Social studies of science, Vol 13, N°2, may 1983, p 255-283

139. NIYA83 A technique of two-stage clustering applied to environmental and civil engineering and related methods of citation analysis
Miyamoto S, Nakayama K
Journal of the society for information science, Vol 34, 1983, p 192-201

140. OBER88 Some statistical aspects of co-citation cluster analysis and a judgement by physicists
Oberski J E L
Handbook of quantitative studies of science and technology, A F J Van Raan (ed), Elsevier science publishers B.V., North-Holland, 1988, p 431-462

141. PAIS86 The convergence of cpmmunication and information science
Paisley W
Eldeman, Hendrik, ed. Librairies and information science in the electronic age, Philadephia, PA: ISI Press, 1986, p 122-153

142. PAOL92 La station d'analyse bibliométrique "ATLAS"
Paoli C, Laville F, Longevialle C
Actes du colloque Journées d'études ADEST, 1-2 juin 1992, p 129-138

143. PARK56 A theory of word-frequency distribution
Parker-Rhodes A F, Joyce T
Nature, 1956, Vol 178, p 1308

144. PENA92 Analyse des citations: application à la théorie microéconomique
Penan H
Dans: La veille technologique, sous la direction de Héléne Desvals et Henri Dou, Dunod, 1992, p 313-330

145. PERI83 Are methodological papers more cited than theoretical or empirical ones? The ces of sociologie
Peritz B C
Scientometrics, Vol 5, N°4, 1983, p 211-218

146. PETE91 Structuring scientific activities by co-author analysis
An exercice on a university faculty level
Peters H P J, Van Raan A F J
Scientometrics, Vol 20, N°1, 1991, p 235-255

147. POLA91 A la recherche de la diversité perdue
Xavier Polanco, Laurent Schmitt, Dominique Besagni, Luc Grival
Revue française de bibliométrie, Vol 9, p 273-292, 1992

148. PONT86 Qualitative aspects of the Bradford distribution
Pontigo J, Lancaster F W
Scientometrics, Vol 9, N°1-2, 1986, p 59-70

149. PRAT77 A measure of class concentration on bibliometrics
Pratt A D
Journal of the american society for information science, Vol 28, p285-292, Sept 1977

150. PRAV91 Distribution of scientific productivity: ambiguities in the assignement of author rank
Pravdic N, Oluic-Vukovic V
Scientometrics, Vol 20, N°1, 1991, p 131-144

151. PRIC63 Little science, big science
De Solla Price D
Columbia, New York, 1963, 118 p

152. PRIC66 Collaboration in an invisible college
De Solla Price D J, Beaver D
American psychologist, Vol 21, 1966, p 1011-1018

153. PRI81a The analysis of scientometric matrices for policy implications
De Solla Price D
Scientometrics, Vol 3, N°1, 1981, p 47-54

154. PRI81b The analysis of square matrices of scientometric transactions
De Solla Price D
Scientometrics,Vol 3, N°1, 1981, p 55-63

155. PRIT69 Statistical bibliogrphy or bibliometrics?
Pritchard A
Journal of publication, Vol 25, 1969, p 348-349

156. QUON88 Bibliométrie informatisée et information stratégique
Quoniam Luc

Thèse de doctorat, Université Aix-Marseille III, 1988

157. RADH79 Lotka's law and computer science literature.
Radhakrishnan T, Kernizan R
Journal of the american society for information science, Vol 30, p 51-54, 1979

158. RAIS62 Statistical bibliography in the health science
Raisig L M
Bulletin of medical library association, July 1962, Vol 50, N°3, p. 450-461

159. REMY91 Analyse bibliométrique appliquée à la recherche fondamentale: une expérience sur 10 ans
Remy D, Vergnes G,Mossetti M
Revue française de bibliométrique, Vol 9, p 227-237

160. RICE89 Journal-to-journal citation data: issues of validity and reliability
Rice R, Borgman C L, Bednarski D, Hart P J
Scientometrics, Vol 15, N°3-4, 1989, p 257-282

161. ROCK79 Chief executive define their own data needs
Rockart J F
Hervards bussiness review, March-April 1979

162. ROUX85 Algorithmes de classification
M. Roux
Masson, 1985, Paris

163. SAIT84 Indentification of the specialities in library and information science using co-citation analysis
Saito Y
Library and information science, Vol 22, 1984, p 61-74

164. SALT79 A citation study of computer science literature
Salton G, Bergmark D
IEEE transactions on professional communicatoin, Vol 22, N°3, sept 1979, p 146-158

165. SAPO90 Probabilités, analyse des données et statistique
G. Saporta
Editions technip, 1990, Paris

166. SCHU86 Relative indicators and relational charts for comparative assessement of publication output an citation impact
Schubert A, Braun T

Scientometrics, Vol 9, N°5-6, 1986, p 281-291

167. SMAL73 Co-citation in the scientific literature: a mew measure of the relationship between two documents
Small H G
Journal of american society for information science, Vol 24, N°4, jul-aug 1973, p 265-269

168. SMAL74 The structure of scientific literature I: indentifying a grahing specialities
Small H G, Griffith B C
Science studies, Vol 4, N°1, janv 74, p 17-40

169. SMAL81 The relationship of information science to the social sciences - a cocitation analysis
Small H G
Information processing & management, Vol 17, N°1, p 39-45

170. SMAL85 Clustering the science citation index using co-citations. II. Mapping science
Small H G, Sweeney E, Greenlee E
Scientometrics, Vol 8, N°5-6, 1985, p 321-340

171. SOMM92 La propriété industrielle, outil de management pour la stratégie de l'entreprise
Sommier J-L
Dans: La veille technologie, Sous la direction de H Desvals et H Dou, Dunod, 1992, 436 p

172. STAC92 Méthodes pratiques de mesure des performances dans la gestion de la technologie
Stacey G
Colloque Journée d'études ADEST, 1-2 juin 1992, p 99-110

173. STER85 The growth of knowledge: testing a theory of scientific revolutions with a formal model
Sternam J D
Technological forecasting and social change, Vol 28, 1985, p 93

174. STEV89 National Citation indicators based on citing year: the citation time anomaly
Stevens K, Narin F
CHI, Haddon Heights, New Jersey, 1989

175. SUBR79 Lotka's law and the literature of computer science
Subramanyam K
IEEE transactions of professional communications,

Vol 22, p 187-189, 1979

176. TAGU81 The law of exponentiel growth: evidence, implications and forecasts
Tague J
Dans: Library trands, Biliometrics, Vol 30, 1981

177. TODO87 Journal citation measures: a concise review
Todorov R, Glänzel W
North-Holland Information & Business, 1987

178. TODO90 Mapping australian geographics: a co-heading analysis
Todorov R, Winterhager M
Scientometrics, Vol 19, N°1-2, 1990, p 35-56

179. TODO91 An overview of mike Moravcsik's publication activity in physics
Todorov R, Winterhager M
Scientometrics, Vol 20, N°1, 1991, p 163-172

180. TODO92 Displaying content of scientific journals: a co-heading analysis
Todorov R
Scientometrics, Vol 23, N°2, Fev 1992, p 317-334

181. TURN90 De la bibliométrie à l'infométrie: des axes de recherche nouveaux pour la veille scientifique et technologique
Turner W A
Revue française de bibliométrie, Vol 6, 1990, p 161-179

182. VANR88 Handbook of quantitative studies of science and technology
Van Raan A F J
Van Raan (ed), Elsevier, 1988

183. VANR89 Dynamic of a scientific field analysed by co-subfield structures
Van Raan A F J, Peters H P F
Scientometrics, Vol 15, N°5-6, 1989, p 607-620

184. VICK48 Bradford's law of scattering
Vickery B C
Journal of documentation, Vol 4, N°3, p 198-203, 1948

185. VILA89 L'entreprise au aguets
Villain J

Masson, Paris, 1989, 192 p.

186. VINK88 An attempt of surveying and classifying bibliometric indicators for scientometric purposes
Vinler P
Scientometrics, Vol 13, N°5-6, 1988, p 239-259

187. VLAC85 Citation histories of scientific publications. The data sources.
Vlachy J
Scientometrics, Vol 7, N°3-6, 1985

188. VOOH74 Lotka and information science
Voohs H
Journal of the american society for information science, Vol 25, p 270-272, 1974

189. WALL86 The relationship between journal productivity and obsolescence.
Wallace D P
Journal of the american society for information science, Vol 37, N°3, may 1986, p 136-145

190. WHIT81 Author cocitation: a literature measure of intellectual structure
White H D, Griffith B C
Journal of american society for information science, Vol 32, N°3, may 1981, p 163-172

191. WHIT81 Bradfordizing search output - How it would help online users
White H D
Online review, Vol 5, N°1, feb 1981, p 47-54

192. WHIT87 Quality of indexing in online data bases
White H D, Griffith BC
Information processing & management, Vol 23, N°3, 1987, p 211-224

193. WHIT89 Bibliometrics
White H D, Mc Cain K W
Annual review of information sience and technology (ARIST), Vol. 24, 1989

194. WILK72 The ambiguity of bradford's law
Wilkinson E A
Journal of documentation, Vol 28, p122-130, 1972

195. ZIPF49 Human behaviour and the principale of least effort
Zipf G K
Addison Wesley, 1949

Annexes

IX. Annexes

A. Annexe 1: Liste des indices d'association statistique calculés dans DATAVIEW

Listes des indices d'association disponible dans DATAVIEW

		Forme X	
		Présence	Absence
Forme Y	Présence	N_A	N_B
	Absence	N_C	N_D

et $N_A + N_B + N_C + N_D = M$

Nom	Formule	Indications
Bray & Curtis	$(N_B + N_C) / (2 * N_A + (N_B + N_C))$	dissimilitude [0, 1]
Concordance	$(N_A + N_D) / M$	similitude [0, 1] (Sokal & Mich. 85)
Corrélation de Pearson	$((N_A * N_D)-(N_B * N_C))/\sqrt{((N_A+N_B)*(N_A+N_C)*(N_B+N_D)*(N_C+N_D))}$	similitude [-1, 1]
Czekanowski	$(2 * N_A) / (2 * N_A + N_B + N_C)$	similitude [0, 1]
Différence de Forme	$(M * (N_B + N_C) - (N_B - N_C)^2) / M^2$	dissimilitude [0, 1]
Différence de Modèle	$N_B * N_C / M^2$	dissimilitude [0, 1]
Différence de Taille	$(N_B + N_C)^2 / M^2$	dissimilitude [0, 1]
Dispersion	$((N_A * N_D) - (N_B * N_C)) / M^2$	similitude [-1, 1]
Euclidienne	$(N_B + N_C) / M$	dissimilitude [0, 1]
Equivalence	$N_A^2 / (N_A + N_B) * (N_A + N_C)$	similitude [0, 1]
Faith	$(N_A + N_D / 2) / M$	(Faith 83)
Hamman	$((N_A + N_D) - (N_B + N_C)) / M$	similitude [-1, 1]
Inclusion	$N_A / \min \{(N_A + N_B), (N_A + N_C)\}$	
Jaccard	$N_A / (N_A + N_B + N_C)$	similitude [0, 1] (Jaccard 1900)
Kulczynski 1	$N_A / (N_B + N_C)$	similitude [0, ∞] (Kulczynski 28)
Kulczynski 2	$(N_A / (N_A + N_B) + N_A / (N_A + N_C)) / 2$	similitude [0, 1] (Sokal Sneath 63)
Moyenne²	$(N_B + N_C) / M$	dissimilitude [0, 1]
Ochiai 1	$N_A / \sqrt{((N_A + N_B) * (N_A + N_C))}$	similitude [0, 1] (Ochiai 1957)
Ochiai 2	$N_A / \sqrt{((N_A + N_B)*(N_C + N_D)*(N_A + N_C)*(N_B + N_D))}$	
Q de Yule	$((N_A * N_D)-(N_B * N_C)) / ((N_A * N_D)+(N_B * N_C))$	similitude [-1, 1]
Rogers & Tanimoto	$(N_A + N_D) / ((N_A + N_D) + 2*(N_B + N_C))$	similitude [0, 1] (Rogers Tanim. 60)
Russel & Rao	N_A / M	similitude [0, 1] (Russel Rao 40)
Shannon	$2 * (N_B + N_C) * \text{Log}(2)$	dissimilitude [0, ∞]
Sokal & Sneath 1	$2 * (N_A + N_D) / (2 * (N_A + N_D) + (N_B + N_C))$	similitude [0, 1] (Sokal Sneath 63)
Sokal et Sneath 2	$N_A / (N_A + 2 * (N_B + N_C))$	similitude [0, 1] (Sokal Sneath 63)
Sokal et Sneath 3	$(N_A + N_D) / (N_B + N_C)$	similitude [0, ∞]
Sokal et Sneath 4	$(N_A/(N_A+N_B) + N_A/(N_A+N_C) + N_D/(N_B+N_D) + N_D/(N_C+N_D))/4$	similitude [0, 1]
Sokal et Sneath 5	$N_A * N_D / \sqrt{((N_A+N_B) * (N_A+N_C) * (N_B+N_D) * (N_C+N_D))}$	similitude [0, 1]
Variance	$(N_B + N_C) / (4 * M)$	similitude [0, 1]

B. Annexe 2: Exemple de références bibliographiques de documents brevets

Echantillon de 10 références de la base Derwent utilisé pour les exemples de manipulations bibliométriques de DATAVIEW (DATAVIEW: logiciel bibliométrique):

```
-1-
AN  - 92-009829/02
TI  - Patches for topical or transdermal drug delivery - with adhesive layer
      contg. polyacrylate adhesive and film former
TT  - PATCH TOPICAL TRANSDERMAL DRUG DELIVER ADHESIVE LAYER CONTAIN
      POLYACRYLATE ADHESIVE FILM FORMER
PR  - 90.06.25 90DE-020144
PN  - EP-464573-A  92.01.08 (9202)
      DE4020144-A  92.01.09 (9203)
AP  - 91.06.24 91EP-110409  90.06.25 90DE-020144
DS  - AT BE CH DE DK ES FR GB GR IT LI LU NL SE
PA  - (LOHM ) LTS LOHMANN THERAPI
IN  - MULLER W,MINDEROP H,TEUBNER A
LA  - G
CT  - (G)DE3843238 DE3843239 EP-305758 EP-379933
IC  - A61L-015/16 A61F-013/02 A61M-037/00
DC  - A96 B07 D22 G03 A14 P34 P32
MC  - A04-F06E5 A08-P01 A12-V03A B04-C03B B12-M02F D09-C04B G03-B02D1 G03-B04
AB  - (EP-464573)
      Topical or transdermal patches comprise a backing layer, an adhesive
      layer and a release liner. The adhesive layer comprises 100 pts.wt. of a
      polyacrylate adhesive (I), 5-150 pts.wt. of a polyacrylate-compatible
      film former (II), 0.250 pts.wt. of non-plasticising active agents and/or
      additives, and 10-250 pts.wt. of plasticising active agents and/or
      additives.
             ADVANTAGE - Inclusion of (II) overcomes consistency problems
      associated with high levels of plasticising components.   (10pp
      Dwg.No.0/0)

-2-
AN  - 92-009106/02
TI  - Transdermal admin. of drugs to humans and animals - uses laminar pref.
      multilayer matrix for direct application to ear
TT  - TRANSDERMAL DRUG HUMAN ANIMAL ADMINISTER LAMINA PREFER MULTILAYER MATRIX
      DIRECT APPLY EAR
PR  - 91.02.11 91US-653393  90.06.14 90IL-094737  90.08.28 90IL-095508
      91.02.11 91IL-097215
PN  - EP-463454-A  92.01.02 (9202)
AP  - 91.06.11 91EP-109534
DS  - AT BE CH DE DK ES FR GB IT LI LU NL SE
PA  - (DERM-) DERMAMED
IN  - GERTNER A,RUBINSTEIN Y
LA  - E
CT  - (E)No-SR.Pub
IC  - A61L-015/16
DC  - B07 C07 D22 P34
MC  - B02-Z B04-B01C1 B04-B02D B06-F03 B12-B02 B12-B04 B12-D02 B12-D06 B12-D07
      B12-F01 B12-J01 B12-K02 B12-L09 B12-M02F C02-Z C04-B01C1 C04-B02D C06-F03
      C12-B02 C12-B04 C12-D02 C12-D06 C12-D07 C12-F01 C12-J01 C12-K02 C12-L09
      C12-M02F D09-C04
AB  - (EP-463454)
      A pharmaceutical compsn. for use in the transdermal admin. of a
      medicament, comprises: (i) a medicament adapted for transdermal admin.;
      (ii) a carrier, which is semi-solid or liq.  at ambient temps.,
      comprising at least one of esters of 8-24C fatty acids, aliphatic
      polyhydroxy cpds. or non-volatile paraffins; and (iii) opt., an
      antiinflammatory agent and/or an antihistamine, to mitigate any skin
      uncompatibility.
             Also claimed is a matrix for transdermal admin. of the compsn.,
      comprising a porous, absorbent, perforate and flexible laminar solid
      support, with the compsn. absorbed on it.
             USE/ADVANTAGE - The device provides admin. of a wide variety of
      drugs, either singly or in combination, transdermally to humans or
```

animals. Skin irritation and sensitisation is avoided, and shaving the hair, or selecting a non-hair skin surface, is unnecessary. Unlike adhesive patches, the device cannot be rubbed off, and is therefore suitable for animals (partic. hairy animals), although the matrix can also be adapted for human use by means of an impervious bandage. (50pp Dwg.No.1A/23)

-3-
AN - 92-007165/01
TI - Device for transdermal administration of drug - avoids problems associated with devices which adhere to skin, by attachment to carrier e.g. watch
TT - DEVICE TRANSDERMAL ADMINISTER DRUG AVOID PROBLEM ASSOCIATE DEVICE ADHERE SKIN ATTACH CARRY WATCH
PR - 90.06.01 90US-532216
PN - WO9118572-A 91.12.12 (9201)
AP - 91.05.24 91WO-U03698
DS - *CA *JP AT BE CH DE DK ES FR GB GR IT
PA - (ROBE/) ROBERTSON D N
IN - MOOYOUNG A,ZEPEDAORTE A,CROXATTO HB
LA - E
CT - (E)US4592753 US4883669
IC - A61F-013/02
DC - A96 B07 D22 B01 P32
MC - A12-V01 A12-V03A B01-C06 B04-C03 B11-C04 B12-K03 B12-M02F D09-C04B
AB - (WO9118572)
A device for transdermally administering a drug has a drug-releasing surface and an attachment surface and comprises in sequence a drug-polymer matrix, adhered to this an impermeable layer substantially impervious to the drug, and attachment means adhered to the impermeable layer and located on the attachment surface, the attachment means being capable of attachment to a carrier for holding the device.

USE/ADVANTAGE - The drug-polymer matrix releases the drug through the drug releasing surface to the skin. The device avoids the problems associated with devices such as a patch which adhere to the skin, such as dermatitis. The device is esp. useful for the administration of the contraceptive agent 16-methylene-17alpha acetoxy-19 -nor-4-pregnene-3,20-dione (ST1435). (23pp Dwg.No.1/9)

-4-
AN - 92-007163/01
TI - Transdermal admin of progesterone analogue ST1435 - 16-methylene-17-acetoxy-19-nor-4-pregnene-3,20-di:one in vehicle, to provide effective contraceptive levels etc.
TT - TRANSDERMAL ADMINISTER PROGESTERONE ANALOGUE METHYLENE ACETOXY NOR PREGNENE DI ONE VEHICLE EFFECT CONTRACEPTIVE LEVEL
PR - 90.06.01 90US-532215
PN - WO9118570-A 91.12.12 (9201)
AP - 91.05.24 91WO-U03697
DS - *CA *JP *US AT BE CH DE DK ES FR GB GR IT LU NL SE
PA - (ROBE/) ROBERTSON D N
IN - MOOYOUNG A,ZEPEDAORTE A,CROXATTO HB
LA - E
CT - (E)US4834978
IC - A61F-013/00
DC - B01 D22 B07 P32
MC - B01-C06 B12-E09 B12-K03 B12-M02F B12-M10A D09-C04
AB - (WO9118570)
A method of administering 16-methylene-17alpha-acetoxy-19-nor-4-pregnene-3,20-dione (ST1435) comprises topically applying the cpd. to the skin in a suitable vehicle. A compsn. for topical application comprises ST1435 in a pharmaceutically acceptable vehicle.

Pref. the vehicle does not occlude the skin. The vehicle may be liquid, semisolid or solid and suitable vehicles include ointments, creams, rinses and gels. Alternatively the vehicle may be a device such as a band, disc, patch or bracelet. The ST1435 is suitably present in the vehicle in an amount sufficient to attain serum levels of at least 50 pmol/l. Other active ingredients may be also be present, esp. hormones and their anologues and derivatives, particularly oestrogen.

USE/ADVANTAGE - ST1435 is used as a contraceptive and in the treatment of various gynaecological problems and is usually administered orally. However, it has now been found that, unlike progesterone, ST1435

diffuses through the skin to achieve pharmaceutically effective serum levels. Topical delivery has advantages over other dosage forms, e.g. convenience of application and removal, avoidance of hepatic first-pass metabolism and gastrointestinal incompatibility, controlled substained release, maintenance of a steady state plasma level, and reduced frequency of dosing. (40pp Dwg.No.0/9)

-5-
AN - 91-376550/51
TI - Calcium channel blocking 5H-pyrano-(3,2-C) quinoline-5-one derivs. - used for treating ischaemia, congestive heart failure, peripheral vascular disease, hypertension, CNS disorders and cancer
TT - CALCIUM CHANNEL BLOCK PYRANO QUINOLINE ONE DERIVATIVE TREAT ISCHAEMIC CONGESTED HEART FAIL PERIPHERAL VASCULAR DISEASE HYPERTENSIVE CNS DISORDER CANCER
PR - 89.12.19 89US-452999
PN - US5070088-A 91.12.03 (9151)
AP - 89.12.19 89US-452999
PA - (SQUI) SQUIBB E R & SONS INC
IN - ATWAL K
LA - E
IC - A61K-031/44 C07D-491/05
DC - B02
MC - B06-E05 B12-C06 B12-C10 B12-D01 B12-D02 B12-D04 B12-E01 B12-E02 B12-E08 B12-E09 B12-F01 B12-F02 B12-F05B B12-F07 B12-G01 B12-G02 B12-G03 B12-G07 B12-H02 B12-H03 B12-H04 B12-J04 B12-K02 B12-L04
AB - (US5070088)
Pyranyl quinoline derivs. of formula (I) and their salts are new. X = O or S. R = H, alkyl, alkenyl, alkynyl, aryl, aryl, halo, cyclo or (cycloalkyl) alkyl, -NO2, -CN, -CF3, alkoxy or halo. R1, R5 = independently H, alkyl, alkenyl, alkynyl, aryl, aryl, halo, cyclo or (cycloalkyl) alkyl. R2 = H, OH, -OCOR1 (R1 is not H). R3, R4 = independently H, alkyl or arylalkyl. R6 = R1 opt. substd. NH2 or -OR1 (R1 is not H), or R5 and R6 together form a 5-, 6- or 7-membered satd. ring. Alkyl, alkenyl, alkynyl, alkoxy = independently 1-10C. Aryl = phenyl opt. monosubstd.; cycloalkyl = 3-7C gp.; halogen = F, Cl, Br or I. substd. NH2 = -NZ1Z2. Z1 = H, alkyl or aryl-(CH2)m-; Z2 = alkyl or aryl-(CH2)m-. m = 0-2.
16 Cpds. are specifically claimed e.g. trans-4-((4-chloro-1-oxobutyl)amino)-2,3,4,6-tetra-hydro-3-hydroxy-2,2-dimethyl-6-(phenyl methyl)-5H-pyrano(3,2-c) quinolin-5-one.
USE/ADVANTAGE - Treating ischaemia (claimed). Ca entry blocking vasodilators, useful as antihypertensive-, antiarrhythmic-, anti-anginal-, anti-fibrillatory- and anti-asthmatic agents. They increase the ratio of HDL-cholesterol to total serum cholesterol and limit myocardial infarction. Also used for treating congestive heart failure, peripheral vascular disease and pulmonary hypertension, as additives to cardioplegic solns. for cardiopulmonary bypasses, as adjuncts to thrombolytic therapy, treating CNS vascular and behavioural disorders (e.g. stroke, epilepsy), diarrhoea, dysmenorrhoea, tinnitus etc., oedema, glaucoma, renal failure, hepatotoxicity, endocrine hypersecretory states, drug induced tardive dyskenesia, allergies, muscular dystrophy and cancer. Also used for reversing adriamycin resistance and regulating cell growth. A unit dose contains 10-500 mg of (I) for oral administration or administration via a transdermal patch or nasal inhalation solns. (8pp Dwg.No.0/0)

-6-
AN - 91-366716/50
TI - Transdermal patch - consists of drug reservoir layer, polymeric film, sheet coated with adhesive, protecting sheet, tip and drug-impermeable sheet
TT - TRANSDERMAL PATCH CONSIST DRUG RESERVOIR LAYER POLYMERISE FILM SHEET COATING ADHESIVE PROTECT SHEET TIP DRUG IMPERMEABLE SHEET
PR - 87.10.05 87KR-011065
PN - KR9006832-B 90.09.22 (9150)
AP - 87.10.05 87KR-011065
PA - (KIMY/) KIM Y
IN - KIM YJ
IC - A61K-009/70
DC - A96 B07
MC - A12-V01 A12-V03A B01-A02 B04-A01 B04-C02A3 B04-C03B B04-C03D B07-D09

B10-B03B B11-C04 B12-M02F
AB - (KR9006832)
A transdermal patch for transdermal admin. of a drug comprising; (a) drug reservoir layer contg. pharmaceutically active ingredients, e.g. clonidine, acopolamine, estradiol, propranolol, etc., covered with polymeric film; (b) in contact with the drug reservoir layer, adhesive-coated sheet, removable protecting sheet, drug-impermeable sheet and removable tip. The adhesive coated sheet is selected from polyethylene, polypropylene, cellulose acetate, polycarbonate, etc..

-7-
AN - 91-339533/46
TI - Sub-lingual pharmaceutical compsn. - moulds easily to contours of sub-lingual cavity giving greater efficiency of admin.
TT - SUB LINGUAL PHARMACEUTICAL COMPOSITION MOULD CONTOUR SUB LINGUAL CAVITY GREATER EFFICIENCY ADMINISTER
PR - 90.04.26 90GB-009390
PN - WO9116041-A 91.10.31 (9146)
AP - 91.04.24 91WO-GO0651
DS - *AU *CA *JP *KR *US AT BE CH DE DK ES FR GB GR IT LU NL SE
PA - (SMIK) SMITH KLINE & FRENC
IN - TOVEY GD
LA - E
CT - (E)GB2022999 GB2085299 DE-871821 FR2571253 FR2514642
IC - A61K-009/20
DC - B07
MC - B04-B01B B04-B01C B04-B02D3 B04-B02D4 B04-C02B2 B07-D04D B12-M10A
AB - (WO9116041)
Pharmaceutical compsn. for delivery of medicaments which are absorbed through the sub-lingual mucosa comprises one or more medicaments and a solid carrier which is a wafer formed from starch and is of a thickness that permits it to be moulded to the contours of the sub-lingual cavity following hydration with saliva, allowing localised delivery of the medicament.
USE/ADVANTAGE - The method overcomes problems in prior art such as inefficiency of admin. due to liq. draining away down throat, and the difficulty of retaining tablets under the tongue for a time sufficient to permit adequate absorption of the drug. Also prior art patches were not edible, i.e., not readily degraded by body so must be remvoed from the mouth for disposal. The present patch fits comfortably against the mucosal site. (16pp Dwg.No.0/0)

-8-
AN - 91-337202/46
TI - Hydrophilic transdermal patch - comprising plastic film and hydrophilic layer contg. aq. gel, adhesive and drug, for reduced skin irritation
TT - HYDROPHILIC TRANSDERMAL PATCH COMPRISE PLASTIC FILM HYDROPHILIC LAYER CONTAIN AQUEOUS GEL ADHESIVE DRUG REDUCE SKIN IRRITATE
PR - 90.01.31 90JP-023437
PN - J03227919-A 91.10.08 (9146) {JP}
AP - 90.01.31 90JP-023437
PA - (SEKI) SEKISUI CHEM IND KK
IC - A61K-009/70 A61L-015/58
DC - A96 B07 D22 P34
MC - A12-V01 A12-V03A B04-C02A2 B04-C03B B04-C03C B12-M02F B12-M03 D09-C06
AB - (J03227919)
Hydrophilic transdermal patch is composed of plastic film on which a hydrophilic layer is laminated. The hydrophilic layer consists of an aq. gel, adhesive, surface and drug.
The aq. gel is polyvinyl alcohol, polyvinyl acetone, polyethylene glycol, methylvinyl ether, carboxymethyl cellulose, etc. The plasticiser in the gel is glycerol, diglycerol, polyglycerol, sorbitol, liquid polyethylene glycol, liquid polypropylene glycol, etc.. A plastic film through which vapour passes but water and aq. soln. do not pass directly is used for the patch.
USE/ADVANTAGE - The transdermal patch provides good water absorption, elasticity and flexibility. It has high productivity and simple structure. The patch causes no skin irritation. (5pp Dwg.No.0/0)

-9-
AN - 91-324988/44
TI - Patch system for drug delivery - has sensor detecting patient activity used to control delivery
TT - PATCH SYSTEM DRUG DELIVER SENSE DETECT PATIENT ACTIVE CONTROL DELIVER
PR - 90.03.30 90US-502422
PN - WO9115261-A 91.10.17 (9144)
AP - 91.03.20 91WO-U02160
DS - *AU *CA *FI *JP *NO AT BE CH DE DK ES FR GB GR IT KR LU NL SE
PA - (MEDT) MEDTRONIC INC
IN - LATTIN GA,PADMANABHA RV,GRACE MJ,SORENSON PD,PHIPPS JB,MCNICHOLS LA
LA - E
CT - (E)WO8607269 EP-191404 FR2562800 US4146029 WO8808729 US4725263 US4406658
IC - A61N-001/30
DC - B07 S05 P34
MC - B11-C04 S05-J
AB - (WO9115261)
Appts. for the iontophoretic delivery of drugs to a patient comprises two reservoirs each contg. a charged ionic substance for delivery to the patient. A current generator is coupled to the reservoirs for supplying current in response to a control signal. An activity sensor monitors the physical activity of the patient and generating a signal indicative of this activity.
USE/ADVANTAGE - The appts. is a patch system used for the transdermal, electrotransport delivery of drugs. The drugs may be any therapeutic agent, e.g. analgesics, antidepressants, beta-blockers, vasodilators etc. It may also be used for the controlled delivery of peptides, polypeptides, proteins or other macromolecules. The drug delivery is dependent upon the physical activity of a patient. This may have benefits in certain cases where activity (or the lack thereof) is symptomatic of a specific condition, e.g. delivery of an anticonvulsant for treatment of an epiletic seizure, delivery of antiparkinson agent in response to patient shaking, delivery of an antiemetic in response to motion for the treatment of motion sickness etc. (21pp Dwg.No.2/3)

-10-
AN - 91-324917/44
TI - Transdermal admin. of (anti)cholinergic drugs e.g. physostigmine - used in memory impairment e.g. Alzheimer's disease, glaucoma, tardive dyskinesia and myasthenia gravis
TT - TRANSDERMAL ADMINISTER CHOLINERGIC DRUG PHYSOSTIGMINE MEMORY IMPAIR ALZHEIMER'S DISEASE GLAUCOMA TARDIVE DYSKINESIA MYASTHENIA GRAVIS
AW - ANTICHOLINERGIC
PR - 90.04.06 90US-506702
PN - WO9115176-A 91.10.17 (9144)
AP - 91.04.08 91WO-U02265
DS - *AT *JP AT BE CH DE DK ES FR GB GR IT LU NL SE
PA - (PHAR-) PHARMETRIX CORP
IN - KOCHINKE F,BAKER RW
LA - E
CT - (E)US4788063 US4685911 US4746515 US4965074
IC - A61F-013/00
DC - B05 D21 B07 P32
MC - B04-A01 B04-A04 B06-D04 B06-D11 B06-D16 B07-H B10-B02G B12-E02 B12-E04 B12-E05 B12-G04A B12-L04 B12-M02F B12-M10 D10-B04
AB - (WO9115176)
Compsn. for transdermal application of a cholinergic or anticholinergic drug of high intrinsic specific activity comprises the drug and the ester of a low mol.wt. alcohol and a fatty acid.
Active drug is physostigmine, naloxone, nicotine, arecoline, tetrahydro-aminoacridine, oxotremorine, pelocarpine, acceclidine, scopolamine, atropine, benztropine, aprophen, artane, trihexylphenidyl, or benactyzine.
USE/ADVANTAGE - Compsn. is used for controlled release of cpds. effective for treatment of memory impairment (e.g. in Alzheimer's disease), glaucoma, tardive dyskenesia, and myasthemia gravis. Physostigmine has a short half life in the body due to rapid metabolism, and a narrow therapeutic window of safety. Transdermal admin. avoids multiple daily dosage as only one patch is required in 24 hr. Therapy can be terminated by simple removal, and stable, controlled blood levels maintained while wearing the patch. (23pp Dwg.No.1/3)

Annexe 3: Caractéristiques informatiques de DATAVIEW

La conception et le développement du logiciel DATAVIEW ont représenté un lourd investissement en temps pendant ce doctorat. La rédaction du code informatique équivaut à un an de programmation pratiquement à temps plein, et six mois de mises au point et de tests effectués sur des études de cas réels. Pour cette seconde période, la part consacrée à la réalisation informatique occupa uniquement cinquante pour-cent du temps.

Ce développement a pu s'effectuer dans un période de temps aussi restreinte grâce à une très bonne connaissance des besoins dès le démarrage du projet. Les développements informatiques, conçus dans le passé au laboratoire du CRRM, étaient un acquis expérimental conséquent. La confrontation de ces programmes informatiques à des études concrètes d'analyses bibliométriques a permis de connaître et d'évaluer les réels besoins et les réelles contraintes qu'imposent les données bibliographiques. Le logiciel DATAVIEW est donc l'aboutissement d'une longue réflexion menée par le CRRM pendant ces dernières années. Cette réflexion s'était déjà concrétisée sous la forme de deux principaux développements informatiques: DATALINK [LATE87] et DATRANS [QUON88]. DATAVIEW est le direct descendant de cette lignée de produits informatiques. Chaque génération de cette lignée conservant l'acquis expérimental de la génération précédante, cette mutation aboutit à un outil informatique de plus en plus performant et de plus en plus adapté aux besoins. Ce dernier produit a donc été conçu pour être directement opérationnel dans un système de gestion de l'information scientifique et technique. Pour atteindre cet objectif, plusieurs exigences se sont avérées indispensables à respecter.

⌦ La configuration informatique:

Il n'est pas utile de mettre un outil bibliométrique à disposition de tous. Seuls les spécialistes de la gestion de l'information scientifique et technique ont besoin de mettre en oeuvre de tels outils informatiques. Le système informatique adapté entre donc dans une logique de type station de travail.

De plus, cette station de travail doit permettre à ces spécialistes de manipuler les données depuis l'information documentaire primaire jusqu'à la rédaction du dossier d'étude. Le poste informatique doit être considéré comme un outil d'aide et doit par conséquent pouvoir alléger la tâche à chaque étape de la constitution de ce dossier. Ce poste de travail devrait disposer d'outils pour faciliter:

- ❑ le choix des sources d'information
- ❑ l'accès à l'information
- ❑ la recherche sélective et la collecte de l'information
- ❑ le stockage de la collecte

- l'insertion d'informations informelles sous une forme informatisée
- l'analyse de l'information accumulée
- la rédaction de dossiers
- accessoirement, la communication des synthèses par un réseau informatique à des destinataires ciblés.

Pour avoir toutes ces fonctions présentes sur un même poste de travail, le système informatique que nous envisageons être le plus adapté est un environnement micro-informatique. La souplesse et la diversité des logiciels existant en micro-informatique facilitent l'installation d'une station de travail spécialisée pour la gestion de 1'IST. Il est assez facile de regrouper sur un poste de type PC 486 un ensemble de matériels informatiques (figure 81) permettant:

- une connexion aux diverses sources d'informations (réseaux informatiques internes ou externes, réseaux des télécommunications, disques optiques numériques.. .)
- **une** ouverture vers les informations textuelles non informatisées (scanner, reconnaissance de formes.. .)
- la prise en compte d'information non textuelle (son, image.. . aspect multimédia)
- une capacité de stockage importante
- une puissance de calcul modérée.

Le logiciel DATAVIEW a été développé dans cet état d'esprit.

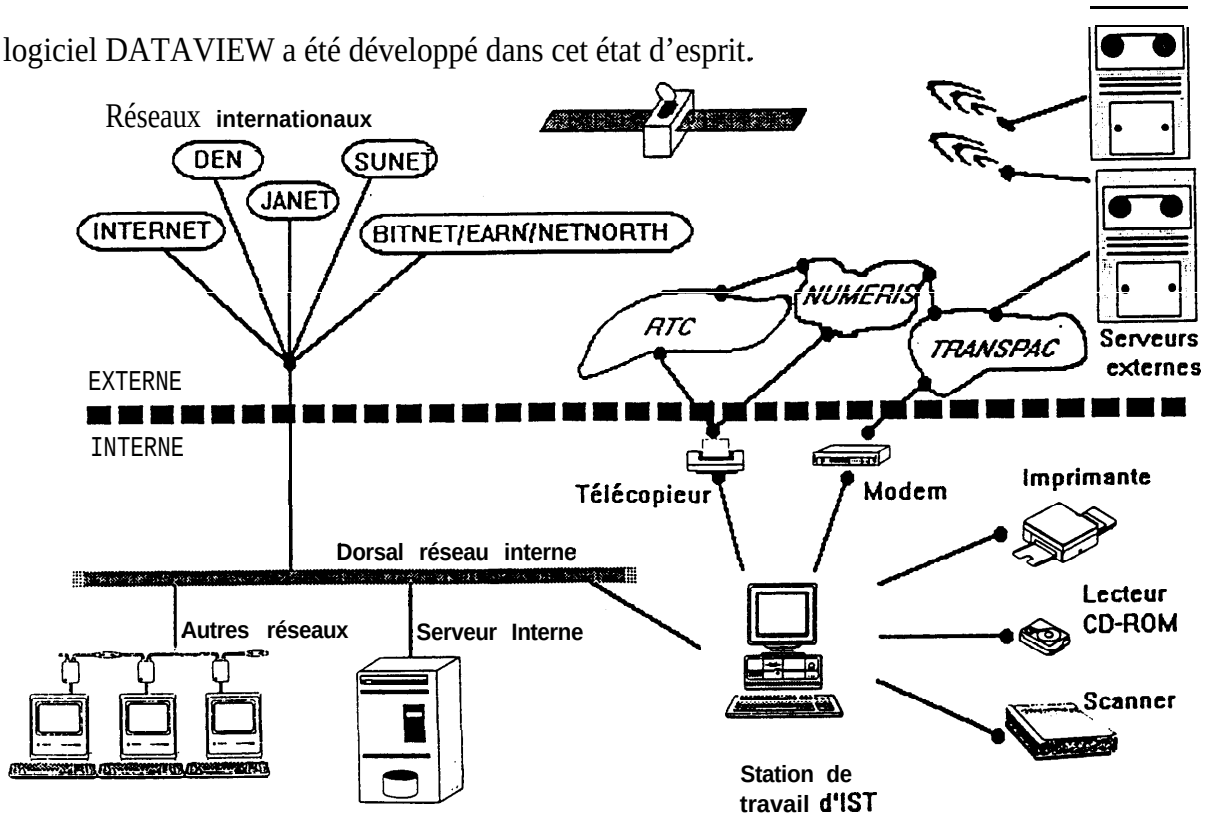


figure 81: Environnement de travail du poste d'I.S.T.

⌦ Langage de programmation

Etant convenu que l'environnement de développement est le système d'exploitation MS-DOS, nous avons opté pour un langage de développement offrant un grand confort de programmation tout en répondant parfaitement aux exigences du cahier des charges. Ce choix s'est finalement porté sur le langage Basic pour développeurs professionnels: le PDS 7.1.

Le principe de rédaction de ce langage de programmation est très évolué tout en étant d'une très grande souplesse. Sa programmation structurée et modulaire offre tous les avantages d'un langage de programmation évolué.

Les fonctions intégrées au PDS 7.1 sont parfaitement adaptées à la manipulation des chaînes de caractères, à la gestion des fichiers et à la création d'interface utilisateur conviviale. Ce langage correspond tout à fait aux besoins des traitements bibliométriques. La puissance de calcul des données numériques n'est pas primordiale (elle n'intervient qu'ultérieurement pour les analyses statistiques), par contre une parfaite gestion des périphériques de stockage et d'affichage est indispensable.

L'environnement de développement du PDS 7.1 est très agréable car il offre un grand confort de programmation et de mise au point: éditeur perfectionné (ou encore appelé éditeur "intelligent"), compilateur mémoire, débogueur intégré, gestion de la mémoire paginée pour stocker le code de programmation, gestion de la mémoire étendu pour stoker l'environnement de développement. De plus, la compilation permet la gestion de la mémoire paginée pour créer une organisation du code exécutable sous forme d'*overlay* et ainsi optimiser la place de la mémoire vive lors de l'exécution du programme.

⌦ Repousser les limites de capacité des données traitées

Selon le cas étudié, une étude bibliométrique peut nécessiter le traitement de grands ensembles de notices bibliographiques. Il est donc indispensable que le logiciel puisse accepter de traiter ces grandes quantités de données.

Pour outrepasser les limites de taille de la mémoire vive, ces données sont gérées par l'intermédiaire de création de fichiers aux diverses étapes de la procédure informatique (fichiers d'indexe, lexiques, listes, tableaux....). Les limites de DATAVIEW dépendent du langage de programmation choisi et des techniques de programmation mises en oeuvre pour

l'exploitation de ces fichiers. La valeur buttoir de ces limites est 32 000. Ainsi DATAVIEW peut traiter jusqu'à:

- ❑ 32 000 notices bibliographiques:
  Nous n'avons encore jamais dépassé plus de 10 000 références pour une étude bibliométrique

- ❑ 32 000 caractères par champs:
  Les résumés des références bibliographiques sont donc traités par DATAVIEW

- ❑ 32 000 formes différents dans l'ensemble des références:
  Le domaine étudié peut donc faire appel à un vocabulaire de 32 000 mots différents

- ❑ 2 000 000 000 de paires dans l'ensemble des références:
  Le vocabulaire peut mettre en oeuvre une combinatoire des cooccurrences s'élevant jusqu'à 2 milliards de possibilités de croisement.

- ❑ près de 30 000 paires pour une référence:
  Cette limite est la seule qui reste dépendante de la place libre en mémoire vive. Cette limite a déjà été grandement optimisée par la gestion mémoire du code exécutable par *overlay*. Nous pensons prochainement dépasser cette limite en utilisation une technique de programmation qui exploitera les informations d'une même référence sous forme de fichiers et non plus sous forme de tableaux en mémoire.

- ❑ les listes d'éditions:
  Il n'existe aucune limite propre à ce mode d'édition de résultats puisqu'une routine de tri sur fichier a été mis au point pour ne plus avoir de limites de nombre d'éléments à trier par liste.

- ❑ création de tableaux comportant 32 000 x 32 000 éléments au maximum:
  C'est dire soit un tableau de présence-absence indiquant la localisation de 32 000 éléments pour un ensemble de 32 000 références, soit un tableau de fréquence reproduisant 1 milliard de relations (32 000 x 32 000) entre deux ensembles de 32 000 éléments.

⌦ Rapidité de traitement:

La rapidité d'exécution du logiciel est un point très important pour que ce logiciel soit parfaitement opérationnel. Nous estimons que l'utilisateur ne pourra pas obtenir du premier coup le résultat qu'il escompte et qu'il sera obligé de réitérer plusieurs fois certaines étapes pour affiner ses résultats. Il est donc indispensable que le logiciel puisse répondre à l'utilisateur dans des délais raisonnables. Pour ce faire, le traitement complet d'un corpus bibliographique s'effectue en trois étapes. A chacune d'elles l'utilisateur peut donc vérifier s'il est satisfait du résultat de l'opération. Dans le cas inverse, l'utilisateur peut reprendre l'opération en l'affinant.

Les trois étapes sont:

- Extraire uniquement les informations que l'on cherche à exploiter (extraction des champs à manipuler):
  Cette étape réduit la taille du fichier à manipuler et permet de transformer les données dans un format standard.

- Construire les fichiers de codage:
  Ces fichiers sont composés de lexiques de formes et de lexiques de paires. Ces fichiers ne sont en réalité pas uniquement des lexiques ni même des fichiers inversés (fichiers d'indexe) puisqu'ils comportent, en plus de la liste des formes ou des paires, des informations sur leurs occurrences, leurs fréquences, leurs cooccurrences et leurs forces d'association. Pour accélérer le temps de création de ces fichiers un algorithme de H-Coding a été mis au point au CRRM [LATE87].

- Editions des résultats:
  L'utilisateur peut maintenant cheminer selon son bon vouloir dans les données créées par l'étape de codage. Par de multiple options d'édition (Cf la partie *Editions des Résultats*), l'utilisateur va pouvoir dégager les éléments d'information qui semblent pertinents. Les fichiers créés dans l'étape de codage sont dans un format codé, de type accès direct, qui permet des recherches d'information dans des temps très rapides. L'utilisateur aura donc la possibilité de réitérer certaines éditions, dans le seul but d'affiner ces résultats, sans que la contrainte de temps se fasse sentir.

⌦ Synergie avec les autres logiciels de la chaîne de traitement:

DATAVIEW n'est qu'un maillon dans la chaîne des traitements informatiques pour constituer le dossier stratégique. Pour que ce maillon s'adapte bien à cette chaîne, il faut qu'il puisse favoriser les liens avec les outils informatiques situés en amont ou en aval de ses propres traitements.

❑ En amont:

DATAVIEW est programmé pour appeler directement le logiciel de reformatage INFOTRANS à partir d'un menu.

DATAVIEW paramètre automatiquement le fichier d'initialisation d'INFOTRANS pour que l'utilisateur puisse retrouver immédiatement les fichiers qu'il veut traiter (positionnement automatique sur les répertoires des fichiers).

Dans un autre menu de DATAVIEW l'utilisateur peut créer automatiquement la grille principale de reformatage de INFOTRANS.
A l'origine, cette grille doit être constituée par l'utilisateur d'INFOTRANS. Il doit indiquer la structure de références bibliographiques du fichier à reformater en précisant la chaîne de caractères qui marque le début d'une référence et pour chaque champ la chaîne de caractères qui symbolise l'intitulé de ce champ. Ce travail assez fastidieux n'est pas toujours couronné de succès dès la première construction de cette grille. L'utilisateur ne peut pas se contenter de reproduire la structure de la première référence du corpus puisque les serveurs, lors du télédéchargement, ne fournissent pas les champs qui ne sont pas renseignés. Dans la première référence il se peut qu'un champ ne soit pas présent alors qu'il l'est dans d'autres références. L'utilisateur doit donc balayer toutes les références pour faire l'inventaire de l'ensemble des champs présents dans le fichier bibliographique. Cette tâche est particulièrement rebutante.
C'est pour cette raison que nous proposons de réaliser automatiquement cet inventaire des champs dans DATAVIEW avant de rentrer dans INFOTRANS. En précisant quelques paramètres de reconnaissance, DATAVIEW va permettre de créer un fichier au format INFOTRANS contenant cet inventaire des champs ainsi que le traitement des champs vides (à conserver en bibliométrie). Ce fichier peut être sélectionné comme grille de reformatage dans INFOTRANS.

❑ En aval:

DATAVIEW détermine les caractéristiques bibliométriques d'un corpus bibliograhique. Ces caractéristiques bibliométriques sont ensuite retranscrites sous forme de listes ou sous forme de tableaux par les diverses options d'éditions de DATAVIEW. Pour que ces éditions soient utilisables par des traitements infographiques ou statistiques ultérieurs, DATAVIEW propose plusieurs formats d'exportations vers des logiciels commercialisés:

☞Pour les tableurs:

- 🖫 Excel
- 🖫 Lotus 123
- 🖫 Multiplan

☞ Pour les logiciels d'analyse statistique:

- 🖫 StatItcf
- 🖫 Clustan
- 🖫 Arcade (logiciel d'Analyse Relationnelle du CESMAP)
- 🖫 Tétralogie

⌦ <u>Convivialité du logiciel:</u>

Un grand effort a été réalisé sur l'ergonomie et l'esthétique de l'interface graphique de DATAVIEW.

Au démarrage du projet, une évaluation des générateurs d'interface du commerce s'est conclue par une insatisfaction totale. Les fonctions des libraires fournies ne correspondaient pas parfaitement à nos besoins (à l'époque Visual Basic n'était pas encore sur le marché).

Une librairie regroupant un ensemble de routines d'interface a donc été développée pendant ce doctorat. Cette libraire contient à la fois des routines d'interface qui ont des fonctions très conventionnelles et d'autres routines qui répondent à des besoins particuliers à DATAVIEW:

- ☞ Générateur automatique de menus à partir d'un fichier de description
- ☞ Boîte de message d'alerte avec ou sans interventions de l'utilisateur
- ☞ Boîte de dialogue avec divers types de champs de saisies (choix logique, choix multiples, saisies numériques, saisies alphabétiques, saisies de noms de fichiers, proposition de listes d'options avec ascenseur...)

- ☞ Routine de déplacement dans les arborescences des unités pour sélectionner des fichiers (affichage de la taille, de la date, de l'heure et de l'attribut des fichiers)
- ☞ Routine de tests de fichiers (en création, en utilisation ou bien les deux)
- ☞ Boîte de sélection d'éléments dans une liste "infinie" (sélection/déselection manuelle, sélection/déselection automatique par masque de recherche, liste de paires d'éléments avec sélection que sur un des deux éléments de chaque paires)

Rq: toutes ces routines ont été développées pour pouvoir gérer la souris.