Modélisation de connaissances et fouille d'informations par la cartographie dynamique : applications de veille technologique avec Matheo Analyzer™

Mylène Leitzelman, Henri Dou, Jacky Kister

Modélisation de connaissances et fouille d'informations par la cartographie dynamique : applications de veille technologique avec Matheo Analyzer™

M. Leitzelman, H. Dou et J Kister

Adresse :
UMR CNRS 6171
Faculté des Sciences et Techniques de St Jérôme
13397 Marseille cedex 20
mleitzelman@hotmail.com, dou@crrm.u-3mrs.fr, jacky.kister@u-3mrs.fr

Mots-clés : cartographie dynamique, représentation des connaissances, analyse réseau, bibliométrie, infologiciel, exemple de veille technologique.

Résumé :
Le foisonnement des informations aujourd'hui accessibles en ligne est tel qu'il ne peut être abordé que sous l'angle d'un traitement automatisé et avancé afin d'extraire du sens de données brutes et de créer de la connaissance. Les infologiciels sont donc plus que jamais nécessaires à l'analyse poussée de grands corpus informationnels, dont les outils de représentation des connaissances notamment cartographiques forment une sous-partie importante. M Analyzer™, outil complet de bibliométrie, issu des recherches du CRRM, propose un module de cartographie dynamique des données bibliographiques dont l'utilisation dans des études concrètes de veille a permis de déceler des signaux faibles et des tendances importantes. Un panel d'exemples présente les diverses conclusions qui ont pu être tirées de l'analyse de ces cartographies dynamiques.

Entrée en matière : l'évolution des besoins de veille

Il y a quelques années encore, nous pouvions tabler sur une définition stricte de la veille comme l'observation et l'analyse de l'environnement scientifique, technique, technologique et économique de l'entreprise pour contrer les menaces et saisir les opportunités de développement [Jakobiak 92].

Depuis, avec le phénomène Internet, la montée en puissance des Technologies de l'Information et de la Communication et la globalisation du commerce mondial, plusieurs concepts, comme l'intelligence économique ou l'intelligence compétitive, se sont superposés au concept de Veille Technologique, qui nécessitent un certain éclairage.

Internet poursuit une croissance exponentielle d'offre de contenus, organisés en sources formelles (bases de données techniques & scientifiques privées ou publiques) et en sources d'informations informelles, non structurées et diffuses (chat, forum,infomercials).
Tous les agents économiques (partenaires, concurrents, clients, fournisseurs, acteurs de domaines connexes proches ou éloignés) y ont recours à tout instant, pour rechercher des solutions de problèmes complexes, pour vendre ou simplement se faire connaître.
Plongée dans un environnement mondialisé, chacun doit décider plus vite et au moins à partir de toute l'information susceptible d'être considérée par ses concurrents.

Désormais, grâce à l'accessibilité individuelle à la connaissance via le réseau mondial, c'est presque chaque humain qui peut « se payer » l'accès au savoir.

Selon nous, la masse grandissante d'informations déformera la problématique ainsi décrite. Les axes actuels (« qualité / quantité des résultats », « obtention immédiate ou différée des résultats ») se déplaceront et les caractéristiques significatives (et donc à privilégier) deviendront :

- la pertinence des réponses par rapport au questionnement (non ambiguïté de la requête, validité des sources,…),

et

- l'élaboration plus ou moins grande des traitements appliqués aux résultats

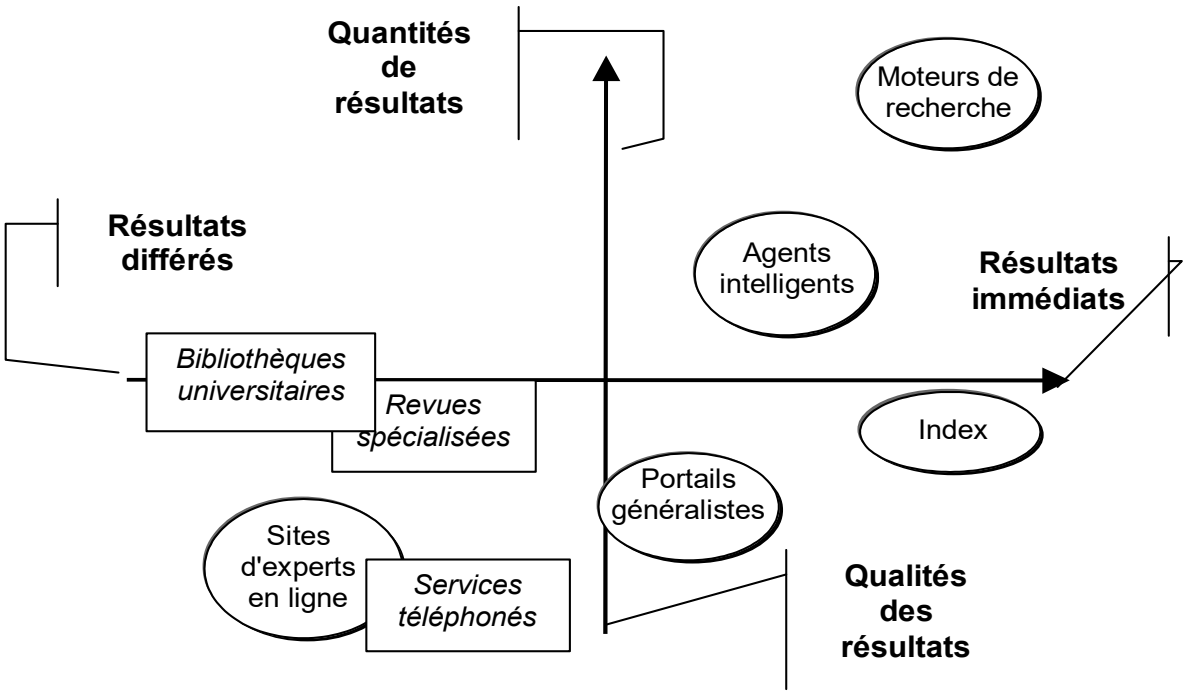


Figure 1 Positionnement des solutions de recherche d'informations

Du fait de la masse croissante du réseau Internet, les moteurs et index ne répondent que partiellement aux besoins des utilisateurs, en ne permettant l'accès qu'à moins de 30% du Web. Les informations les plus pertinentes sont diluées dans un bruit documentaire foisonnant d'où le veilleur ne retrouve plus que très rarement les signaux originaux facteurs d'avantages concurrentiels. Pour trouver des solutions capables de prendre en compte toutes les phases du cycle informationnel, c'est plutôt vers des solutions logicielles haut de gamme qu'il faut aller chercher. Selon une étude IDC, le marché des « infologiciels » comme les nomment [J. Chaumier, 2003] pourrait atteindre plus de 2,5 millions d'euros en 2005.

M Analyzer™ : un solution de gestion bibliométrique des informations

a) Rappel des fondements de la bibliométrie

Avant toute chose, nous nous basons sur la définition suivante du terme « bibliométrie », issue du rapport OCDE/GD(97)41 sur les méthodologies d'évaluation de la recherche européenne:

« La mesure bibliométrique est un outil qui permet d'observer l'état de la science et de la technologie à travers la masse des publications scientifiques, dans un contexte plus ou moins large. C'est un moyen de situer un pays dans le monde, une institution dans un pays, et même la place d'un scientifique dans sa communauté. Ces indicateurs scientifiques se prêtent – avec les précautions qui s'imposent – aussi bien à des analyses "macro" (par exemple, la part d'un

pays donné dans la production mondiale des publications scientifiques pendant une période donnée) qu'à des études "micro" (par exemple, le rôle d'un institut dans la production de textes dans un domaine scientifique très précis) ».

La bibliométrie est donc une discipline qui fait appel à la mesure, notamment à la mesure de données numériques, et qui est fondée sur trois principales lois [Rostaing, 1996], dont s'inspire l'infologiciel M Analyzer™ :

- La loi de Lotka [Lotka, 1926] qui constate la régularité d'une relation inverse entre le nombre de publications dans un domaine scientifique et le nombre des membres de la communauté en question
- La loi de Bradford [Bradford, 1934] : Loi publiée en 1934 dans le but d'aider les bibliothécaires à optimiser leur gestion de fonds documentaires, et qui s'est intéressé à la répartition des articles scientifiques, pour un domaine précis, dans les périodiques.
- La loi de Zipf [Zipf, 1949] qui analyse les régularités sur la fréquence d'apparition de mots dans l'étude de corpus de données textuelles.

Les évaluations bibliométriques de M Analyzer™ sont calculées à partir de l'état condensé des occurrences des données étudiées, c'est-à-dire, la fréquence. Il utilise également la fréquence des cooccurrences, de ce fait nous pouvons exploiter avec cet outil :

- des listes, soit de répartition des fréquences d'occurrences ou de cooccurrences,
- des tableaux, soit des matrices présences-absences d'occurrences, matrices de fréquences de cooccurrences,
- des réseaux, soit la visualisation de fréquences d'occurrences, fréquences de cooccurrences et de connectivités.

b) Fonctionnement technique de l'infologiciel M Analyzer™

L'infologiciel M Analyzer™, exploité par la société MatheO Software, est issu de la recherche du CRRM et notamment propose une version optimisée du logiciel Dataview [H. Rostaing 96], premier logiciel de bibliométrie existant au début des années 90. Dataview, comme M Analyzer™, est un logiciel bibliométrique multifonction [H. Rostaing 93] : il accepte plusieurs types de formats provenant de sources d'informations les plus variées, il effectue lui-même la phase d'importation et de transformation des données bibliographiques, il est capable de créer plusieurs types d'indicateurs bibliométriques et techniques statistiques.

M Analyzer™ est un logiciel de veille axé sur le traitement statistique de données bibliographiques, réalisant des analyses bibliométriques et du mapping d'informations à partir d'importations effectuées sur des bases de données diverses. M Analyzer™ est un outil logiciel capable de fabriquer des informations à haute valeur ajoutée extraites de volumes complexes et importants d'information, permettant ainsi d'en tirer des synthèses, de créer des indicateurs ou de cartographier des ensembles.

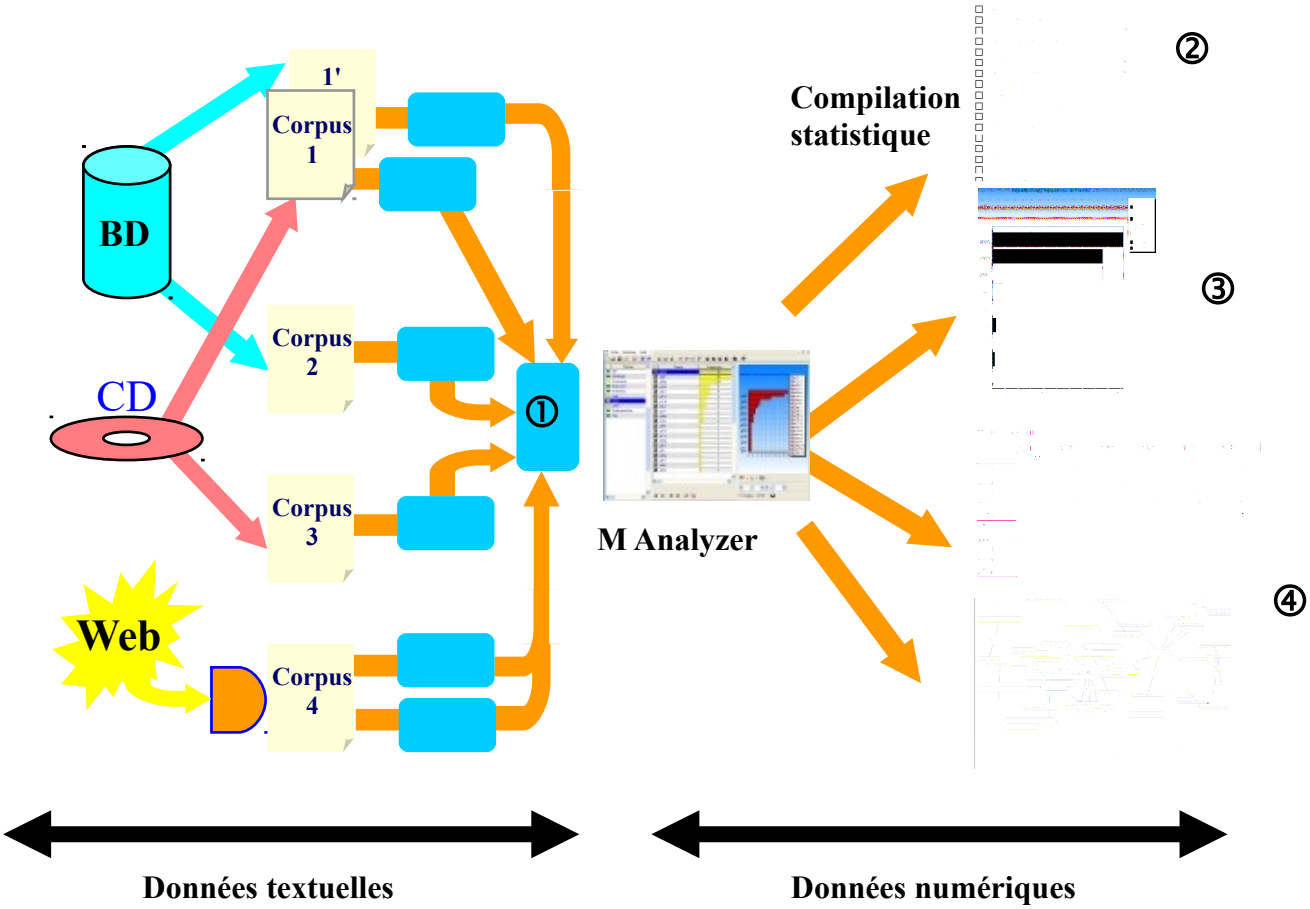


Figure 2 Schéma de fonctionnement de M Analyzer

① **Importation intuitive d'un fichier source**
> Utilisation d'une IHM (Interface Homme Machine) pour la phase d'importation > Possibilité d'utiliser des règles pré-établies d'importation > Sélection d'une information particulière dans un champ
② **Navigation intégrale dans les notices**
> Présentations des champs sélectionnés > Présentation des formes extraites > Outils de reformatage rapide (manuel et automatique) > Accès aux notices intégrales
③ **Présentation graphique des informations présentes dans les notices**

> Génération automatique de cartographies de réseaux de formes > Création automatique d'histogrammes > Propositions nombreuses de matrices d'informations à doubles entrées > Possibilités d'insérer des formes bloquantes
④ **Synthèses statistiques et bibliométriques**
Accès aux informations sur: > L'ensemble des formes sélectionnées > Le croisement des différents champs > Les notices associées à une forme

Figure 3 Principales fonctionnalités de M Analyzer™

Le tableau ci-dessus résume les principales fonctionnalités de ce logiciel (cf. le site http://www.matheo-software.com), mais dans cet article nous focalisons notre attention sur le module de cartographie dynamique de représentation des informations qu'offre M Analyzer™, inspiré par le travail de recherche de M Eric Boutin [E. Boutin 96] et la réalisation du logiciel d'infographie Matrisme.

Intérêts d'une représentation cartographique de l'information

Les outils de représentation graphique des connaissances constituent une sous-partie importante des infologiciels. Si la majorité de ces logiciels repose sur des traitements linguistiques de données textuelles (traitement sémantique, morpho-syntaxiques, ...), comme Sampler, Semiomap ou encore TermWatch qui proposent des méthodologies de clustering de termes sous forme de cartes sémantiques, M Analyzer™ se situe dans la famille des outils de traitement statistique, dans la lignée de Coevision ou Dynatools (pour une analyse comparée de ces outils, voir le rapport de Maîtrise NTIDE de S Goarin [S. Goarin, 2003]).

Si l'on se réfère à la définition suivante de ce qu'est la veille comme « …l'art de repérer, collecter, traiter, stocker des informations et des signaux pertinents (faibles, forts) qui vont irriguer l'entreprise à tous les niveaux de rentabilité, permettre d'orienter le futur (technologique, commercial, …) et également de protéger le présent et l'avenir face aux attaques de la concurrence » [D. Rouach 96], il est clair que ce type d'outils offre un moyen sûr de détecter les signaux faibles, ou précurseurs de tendances, et seule l'utilisation d'outils de traitements bibliométriques de l'information permet de repérer ces signaux dans un flux croissant d'informations.

Nous présentons dans cet article comment une traitement cartographique de l'information permet de détecter ces fameux signaux faibles, ou du moins comment on arrive à créer de la connaissance à partir de données connectées entre elles au moyens de graphes de réseaux, comme l'explique le schéma suivant.

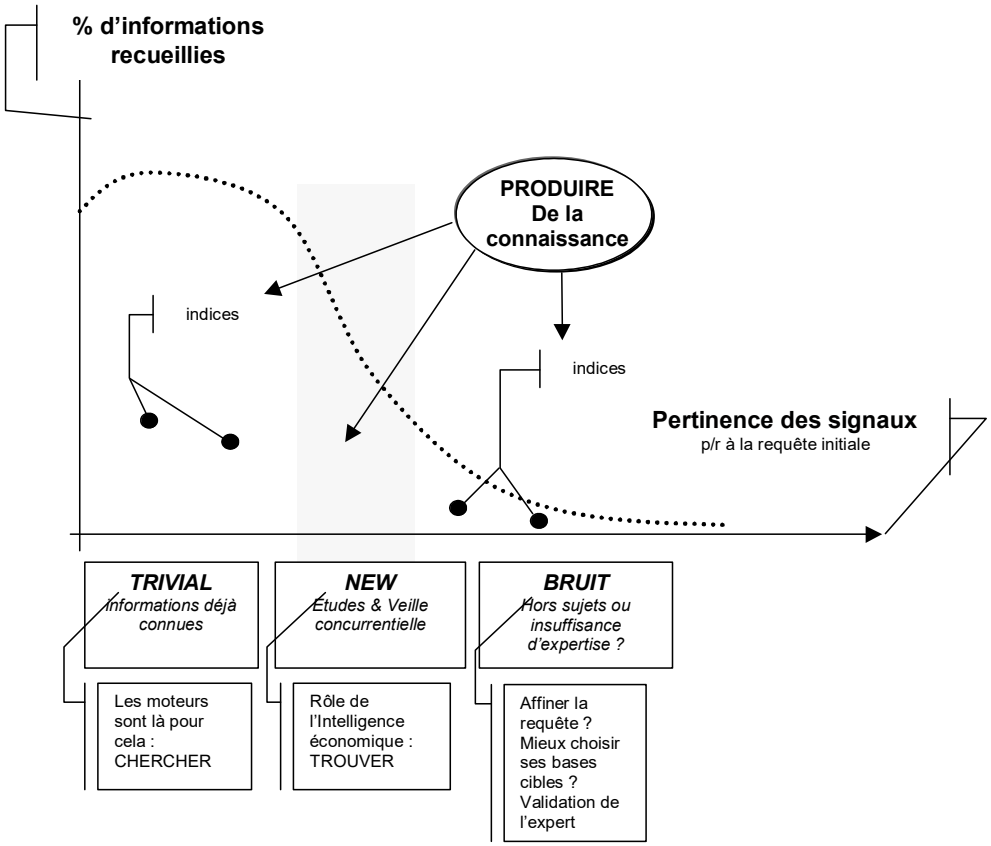


Figure 4 La détection des signaux faibles et la création de connaissances

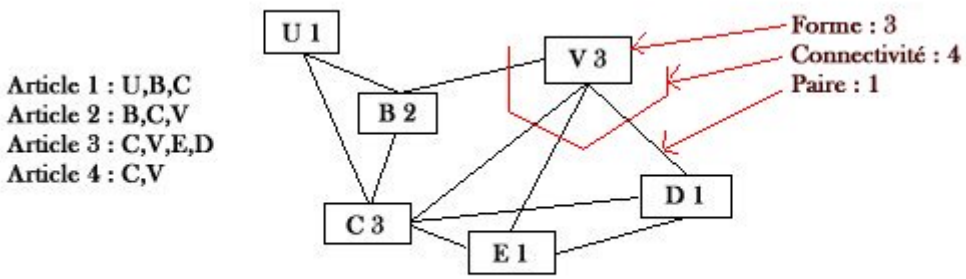
Avant tout, une explication de la fabrication des réseaux sous M Analyzer™ s'impose. Un réseau est une représentation cartographique des relations qui existent entre les éléments d'une population. Un réseau est composé de sommets et de liens, les sommets sont représentés par des boites qui sont les éléments analysés (auteurs, code CIB, mots-clés, …) et les liens entre deux sommets signifient qu'il existe au moins un lien en commun entre eux (soit présence dans la même référence des termes étudiés).

Plusieurs réseaux sont aussi disponibles à partir du traitement matriciel réalisé en sous-tâche par M Analyzer™ :

- les réseaux dit de Condorcet : représentation cartographique des relations qui existent entre les références elles-mêmes par rapport à un champ ou un ensemble de formes. Les sommets représentent uniquement des numéros de références,
- les réseaux dit symétriques : ils présentent les relations existantes entre les formes d'un même champ ou d'un même ensemble de formes. Les sommets représentent uniquement les formes du champ ou de l'ensemble de formes sélectionné,
- les réseaux dit asymétriques : ils présentent les relations existantes entre deux champs ou ensembles de formes distincts. Les sommets représentent des formes

M Analyzer™ propose trois types de filtres pour représenter les données, on peut faire varier la fréquence des formes individuelles, celle des paires et enfin la connectivité existante entre les sommets connectés.

L'exemple suivant est issu du manuel d'utilisation même du logiciel, soit le cas de l'analyse d'articles scientifiques et plus précisément les liens existant entre les différents auteurs {B,C,D,E,U,V}. Nous aurons donc la définition suivante pour les 3 filtres :



Réseau de collaboration entre auteurs d'articles

où les filtres suivants sont :

- **Formes** : Nombre d'articles qu'à publié un auteur.
- **Connectivités** : Nombre d'autres auteurs avec lesquels un auteur a collaboré.
- **Paires** : Nombre de fois où une collaboration est effective.

Nous illustrons dans la paragraphe suivant nos propos à partir d'exemples de veille concrets, qui nous ont permis de détecter des informations intéressantes à partir de l'analyse des graphes obtenus.

Panel d'informations révélées par l'analyse des graphes de réseaux

La première suite d'exemples est issue d'une étude sur la technologie de l'Airbag, à partir de la base de données gratuite Esp@cenet (http://www.espacenet.com) de l'Office Européen des Brevets (OEB).). Cette base propose un contenu relativement complet (textes des brevets au format pdf, dessins, revendications, rapport de recherche, etc…) bien que non exhaustif, cette source d'information est cependant suffisante pour effectuer des analyses sur l'innovation technologique mondiale, le positionnement d'une entreprise, les principaux concurrents, l'environnement d'une technologie ou d'une application via la Classification internationale, etc…

L'analyse des réseaux entre inventeurs et sociétés déposantes d'une part et entre l'ensemble des déposants entre eux d'autre part, nous a permis de mettre en évidence des stratégies d'acteurs économiques leaders.

➲ **Flux de compétences entre entreprises mis en valeur :**

Sur les principaux inventeurs allemands en période 2001/02, le graphe suivant nous fait remarquer que Olaf Mueller a travaillé avec les sociétés Volkswagen et Inova. Une analyse minutieuse des brevets déposés par Olaf Mueller montre qu'il a travaillé de 1999 à 2000 chez Inova et qu'il est passé ensuite chez Volkswagen en 2001.

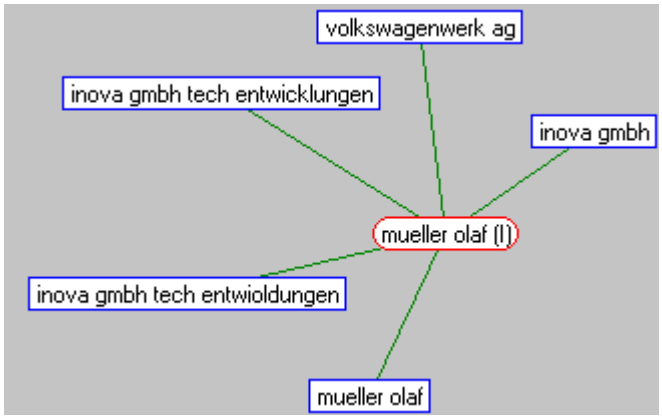
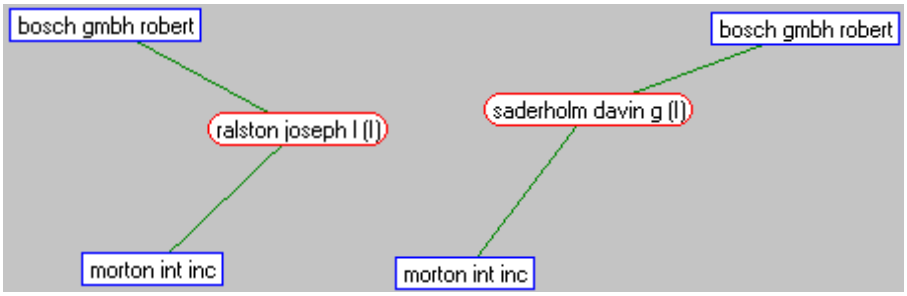


Figure 5 Le réseau entre l'inventeur et les déposants

➲ **Relations de R&D commune mis en lumière :**

Sur les principaux inventeurs américains en période 1996/97, le graphe suivant met en lumière que les inventeurs Joseph Ralstom et Davin Saderholm travaillent avec les sociétés

Morton Inc et Bosch Gmbh. Aurait-il été débauché par le concurrent ou existe-t-il une relation d'affaire entre ces deux sociétés ? Une recherche sur Internet ne nous a pas permis de trouver une relation économique entre ces deux sociétés, il semblerait donc que la première interprétation soit la bonne.



En regardant les brevets déposés de plus près, nous trouvons bien un brevet co-déposé entre ces deux sociétés, le brevet EP0800967 "Variable mass flow airbag module". On peut donc en déduire une relation étroite existante entre ces deux sociétés. Nous voyons ici clairement que l'analyse cartographique peut amener à trouver des informations stratégiques et économiques dépassant le cadre de l'analyse technologique.

Le réseau des déposants extrait de la période 1996/97 permet aussi de mettre en lumière aussi les relations entre déposants :

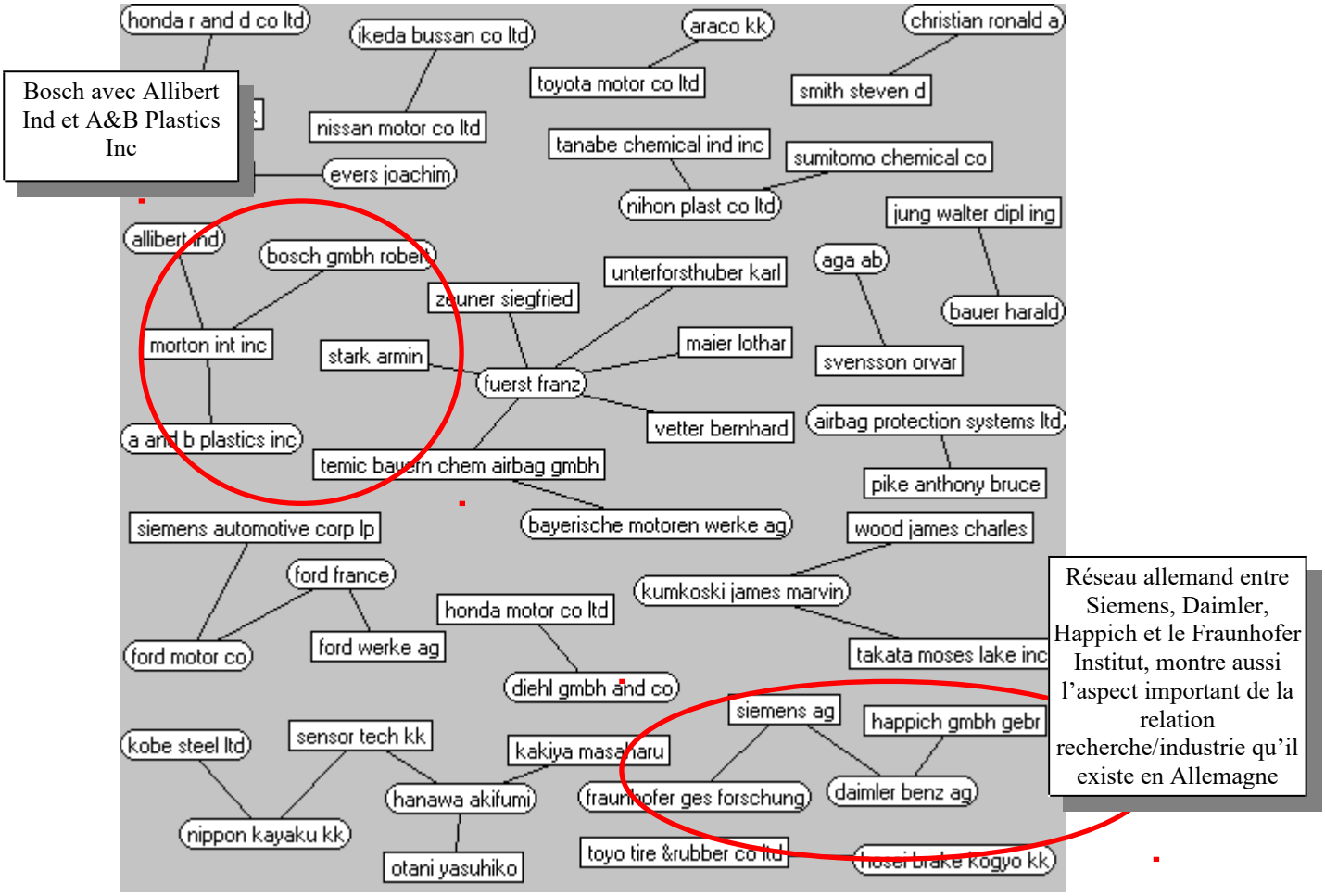
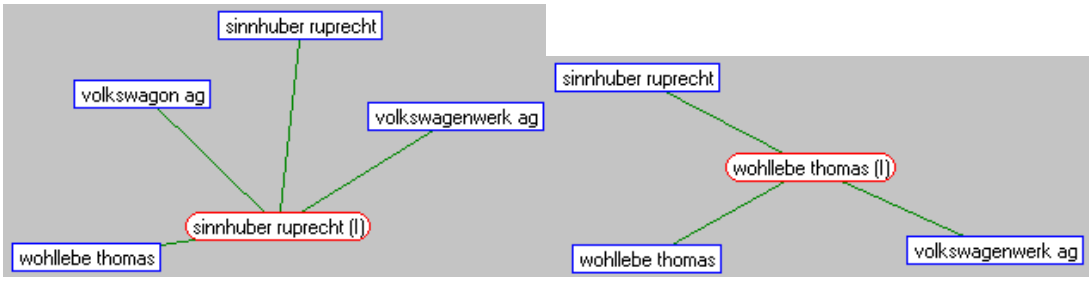


Figure 6 Graphe du réseau des déposants pour 1996/97

Le réseau des déposants s'interprète de la façon suivante : les déposants dans les boites ovales sont les 1er de la liste dans le brevet, c'est à dire le déposant principal, ceux qui sont dans les boites carrées sont les suivants. Donc un ovale avec en étoile des boites carrées représente **Un** brevet, deux boites carrées reliées signifie qu'il y a **Deux** brevets déposés en commun entre ces déposants.

➲ **Débaucher des employés au cœur des technologies phares de l'entreprise :**

Concernant les deux inventeurs allemands Ruprecht Sinnhuber et Thomas Wohllebe, respectivement 14 et 11 brevets déposés à leur actif en 2001/2002, ils travaillent tous deux pour la société allemande Volkswagen :



Ces informations peuvent être capitales, au cas où la stabilité de leur emploi dans la société peut être menacée par un concurrent venant débaucher ces employés qui sont au cœur de la technologie de l'Airbag dans la société Volkswagen.

La seconde suite d'exemples provient d'une étude sur l'utilisation de bio-carburants dans le cas de la technologie des piles à combustible et de la production d'hydrogène. Deux interrogations de bases de données scientifiques ont été menées en parallèle pour obtenir les articles scientifiques traitant du sujet, d'une part sur la base Pascal et d'autre part sur la base ISI (respectivement 54 et 57 références obtenues pour l'analyse bibliométrique).

Lorsqu'on regarde de plus près les réseaux liés à cette liste d'auteurs les plus significatifs dans le domaine, nous obtenons les graphes suivants en ne mettant en avant que les paires d'auteurs de plus de 4 en fréquences, ce qui permet de justifier d'une réelle collaboration durable. De plus, la comparaison des réseaux d'auteurs obtenus avec les données de la base Pascal et celles de la base ISI permet d'éclairer la vision plutôt européenne ou plutôt américaine des groupes d'expertises scientifiques dans le domaine de recherche traité.

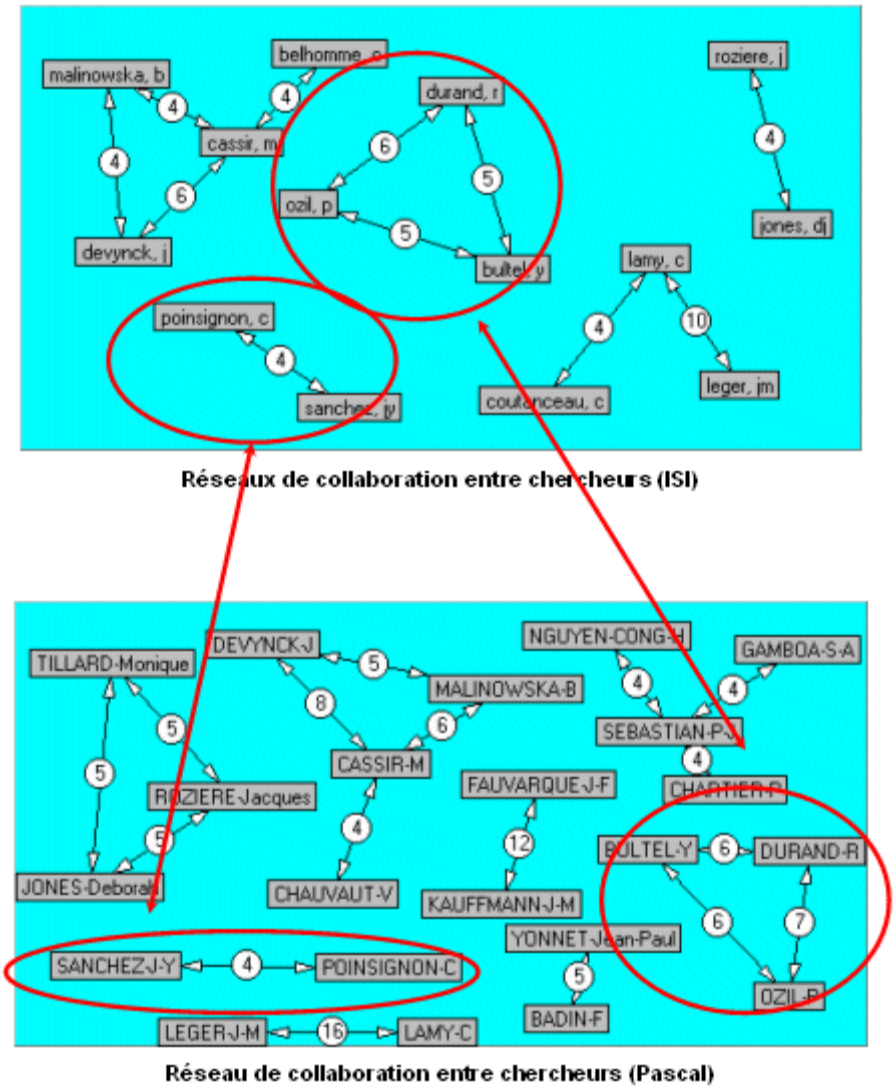
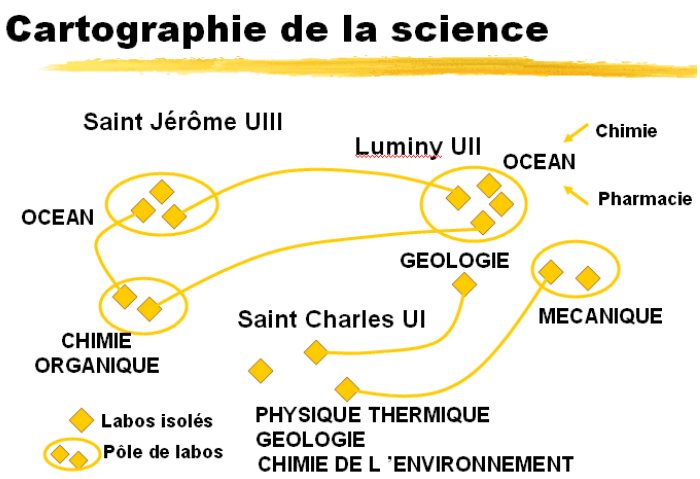


Figure 7 Comparaison de réseaux d'auteurs par bases de données interrogées

Nous retrouvons des réseaux similaires, la base Pascal nous donnant plus de détails sur d'autres réseaux que l'on ne retrouve pas dans celle de l'ISI.

Une démarche identique avait présidé une réflexion similaire sur la détection des pôles d'expertise dans les sciences de l'Environnement sur les universités d'Aix-Marseille, réalisée en mémoire de DEA et réactualisée en thèse par [M. Leitzelman, 98], évidement avec l'aide des premiers outils de traitement bibliométrique du CRRM soit Dataview et Matrisme. L'analyse des réseaux de co-auteurs avait permis de mettre en lumière les relations entre universités, comme le montre le schéma ci-dessous.



Notamment, lors de la réactualisation de cette étude avec l'ajout des bases de données Environline de Questel-Orbit et Pascal à celle de l'ISI utilisée initialement, on se rend compte que c'est avec la comparaison des trois analyses des réseaux que l'on a pu se faire une idée relativement complète du visage des sciences de l'Environnement à Marseille. Les pôles Océan avec le sous-domaine chimie marine, le pôle chimie/environnement et celui de géologie/géoscience/écologie sont ressortis dans les trois analyses, ils témoignent d'une relative constance quelle que soit la spécificité des bases choisies. Cette étude nous a aussi permis de mettre en évidence toutes les spécificités en terme de choix d'indexation de revues que font les bases de données scientifiques. C'est avec les analyses des réseaux obtenus que l'on s'est rendu compte que pour connaître de façon globale un sujet, il faut interroger plusieurs sources de données pour fabriquer une image fidèle de la réalité.

Conclusion

Nous avons la volonté de démontrer dans cet article l'importance d'une vue graphique et dynamique de l'information dans le cadre d'étude de veille technologique, voire au delà. Sans la mise à plat des connexions de données étudiées sous forme de graphes de réseaux, nous n'aurions pas pu mettre en lumière certaines des conclusions que nous avons pu avancer dans nos diverses études de veille réalisées par l'intermédiaire de l'outil bibliométrique M Analyzer™.

Le champ de recherche pour optimiser cette approche est encore ouvert et offre des perspectives larges. Notamment plusieurs réflexions portent aujourd'hui sur la transposition de cartographies de données structurées issues de références bibliographiques provenant de

bases de données formelles à l'information non structurée provenant du Web, dans la mesure où ce média est devenu incontournable pour étudier les tendances et la concurrence mondiale. Plusieurs logiciels proposent notamment une cartographie dynamique des données web, on pense à WebRain de la société The Brain ou encore Internet Cartographer d'Inventix. Mais nous pensons que l'avenir dans ce domaine se situe plutôt sur des outils tels que Human-Links (http://www.human-links.com), logiciel de cartographie de recherche Web et documents divers basé sur la technologie Peer-To-Peer et une analyse cognitive des centres d'intérêts des utilisateurs. D'ailleurs, plusieurs recherches impliquent notre laboratoire dans l'exploration de pistes exploitant les connexions possibles qu'il peut y avoir entre la scientométrie et la bibliométrie, les réseaux neuronaux et cognitifs, le peer-to-peer et la gestion des connaissances.

bibliographie :

F. Jakobiak
Exemples commentés de veille technologique
Les éditions d'organisation, 1992

A. J. Lotka
The frequency distribution of scientific productivity.
Journal of the Washington Academy of Sciences, 16, 1926, p. 317-323

S. C. Bradford
Sources of information on specific subject
26 janvier 1934, Engineering, p. 85-86.

G.K Zipf
Human Behavior and the Principle of least effort: An introduction to Human Ecology
Reading, Mass: Addison-Wesley, 1949.

J. Chaumier, M. Dejean
Recherche et analyse de l'information textuelle
Documentaliste, Sciences de l'information 2003, vol 40 n°1

T. Powell

The Information Metabolism
Competitive Intelligence Review, pp. 41-45 1995

S. Quazzotti, C. Dubois, H. Dou
Veille technologique - Guide des bonnes pratiques en PME PMI
ISBN 2-9599776-0-2, CRPHT, Elsch sur Alzette, Luxembourg, 1999.

H. Rostaing, H. Dou, P. Hassanaly, C. Paoli
"Dataview : bibliometric software for analysis of downloaded data"
Actes du colloque : Fourth congress on bibliometrics, scientometrics, informetrics , organisé à Berlin par l'ISSI, 11-15 Septembre , (1993)

H. Rostaing
« La bibliométrie et ses techniques »
Sciences de la Société / CRRM, ISSN 1168-1446, 1996

E. Boutin
Matrism : logiciel de traitement infographique
Thèse passée au CRRM, 1996.

S. Goarin
Etat du marché des logiciels de recherche et de traitemebt de l'information : Positionnement de la suite logicielle MathéO
Rapport de maîtrise NTIDE, 2003.

D. Rouach
La veille technologique et l'intelligence économique
Que sais-je ? 3086, PUF, 1996