LA FOUILLE DE DONNEES POUR LA VEILLE TERRITORIALE. LE CAS DU SUD LOIRE

Sylvie CHALAYE (*,) – Christine LARGERON(*)**
sylvie.chalaye@univ-st-etienne.fr , Christine.Largeron@univ-st-etienne.fr

(*) CREUSET, Université Jean Monnet St-Etienne
6 rue basses des rives
42023 Saint-Etienne Cedex 2, France
(**) Epures, Agence d'urbanisme de la région stéphanoise
46, rue de la télématique, BP 801
42952 Saint-Etienne cedex 9, France

Mots-clés :
Fouille de données, Extraction de connaissances à partir de données, Veille scientifique territoriale
Keywords :
Data-Mining, Knowledge extraction in databases, local technology watch
Palabras clave :
Extracción de información a partir de los datos, Vigilencia scientifica y tecnologica

Résumé

Comme les entreprises, les collectivités territoriales ressentent de plus en plus le besoin d'exercer une veille économique pour mieux connaître leur environnement et pouvoir anticiper les évolutions. L'innovation est aujourd'hui au cœur de la croissance et la conduite de politiques technologiques locales est essentiellement fondée sur le rapprochement entre la sphère privée et la sphère publique. Dans ce contexte, il s'agit de pouvoir fournir aux acteurs locaux des informations sur les compétences scientifiques disponibles sur leur territoire, d'une part, et sur l'organisation de la recherche, d'autre part. La démarche de veille que nous proposons à partir de la base de données bibliographiques Pascal comporte principalement deux objectifs : la mesure de la production scientifique sur un territoire donné et l'analyse des coopérations scientifiques. Ainsi, cette démarche appliquée au Sud Loire a permis de mieux évaluer la production scientifique dans les différents pôles de compétences. Par ailleurs, l'intensité des liens entre les différents acteurs (universités, écoles d'ingénieurs, entreprises) a pu être mesurée. Enfin, les principaux territoires partenaires du Sud Loire dans les coopérations scientifiques ont été identifiés. Dans ce cadre, l'application de techniques de fouille de données sur le corpus de notices bibliographiques a permis de rendre compte de la complexité des réseaux territoriaux.

Summary

Today, economic growth depends more and more on the capacity to create new knowledge and to innovate. Moreover, local technological policies are based on the strengthening of links between public and private actors. In this context, territorial communities need information about the innovation capacity of their territory and the organisation of research activities. So, we propose, in this article, a local technology watch approach for measuring the level of knowledge creation, on the one hand, and for analysing scientific cooperations of a local area, on the other hand. Data Mining tools are also used to illustrate the complexity of territorial networks. This approach has been applied to the urban area of Saint-Etienne.

1 Introduction

L'intelligence économique a connu un développement important ces dernières années (B. Gilad, 1986 ; C. Halliman, 2001 ; L.T. Moss, 2003 ; B. Carayon, 2003). Dès 1994, elle a été définie par H. Martre comme *''l'ensemble des actions de recherche, de traitement, de distribution et de protection de l'information obtenue légalement et utile aux acteurs économiques*'' (H. Martre, 1994). Lorsque les informations à analyser sont de nature scientifique et technique telles que des brevets ou de la documentation scientifique et technique (articles, thèses,…), on parle plus spécifiquement de veille technologique (H. Desvals et H. Dou, 1992 ; F. Jakobiak, 1990 ; F. Jakobiak, 1994).

Pour répondre aux besoins des collectivités et des acteurs publics locaux, les pratiques de veille et les démarches d'intelligence économique s'exercent de plus en plus dans une logique territoriale. La veille territoriale peut être définie comme ''*le processus informationnel par lequel la collectivité se met à l'écoute anticipative des signaux de son environnement dans le but de réduire les incertitudes et de conduire des politiques locales adaptées à son contexte politique, économique et social*." (M-C. Chalus-Sauvannet, 2004). Ce processus de veille territoriale s'inscrit dans une démarche plus large d'intelligence territoriale décrite par P. Herbaux et Y. Bertacchini "*comme une culture d'organisation basée sur la mutualisation et le traitement des signaux en provenance des acteurs économiques destinés à fournir au donneur d'ordres, au moment opportun, l'information décisive.*" (P. Herbaux et Y. Bertacchini, 2003). Cette démarche répond en effet aux besoins des collectivités locales d'avoir une information en continue sur les évolutions économiques de leur territoire. Ainsi, alors que la veille économique était un thème peu familier des collectivités locales il y a encore quelques années, une enquête réalisée par l'AMF (Association des Maires de France) et ETD (Entreprises Territoire et Développement) auprès des intercommunalités montre qu'elle est largement perçue aujourd'hui comme un enjeu majeur (AMF, 2004).

Dans ce contexte, les techniques de fouille de données peuvent apporter une aide précieuse pour le management stratégique des territoires et un certain nombre d'études ont d'ores et déjà été réalisées en ce sens. Dans cet article, nous montrons comment exploiter la base de données bibliographique Pascal à l'aide de techniques de fouille de données dans une perspective de veille territoriale. Notre approche comporte trois volets. Le premier consiste à mesurer la production scientifique du territoire. Le second est consacré à l'analyse des réseaux d'acteurs présents sur le territoire et le dernier à l'identification de partenaires extérieurs au territoire. Après avoir décrit plus précisément dans une première section les données utilisées et la démarche de veille scientifique territoriale que nous proposons, nous revenons dans les sections suivantes sur chacun de ces volets, en les illustrant sur le territoire du Sud Loire[1]. Des perspectives ouvertes par ce travail sont évoquées en conclusion.

2 Une démarche de veille scientifique territoriale par la fouille des données bibliométriques de la base Pascal

La capacité à produire, diffuser et exploiter de la connaissance est aujourd'hui au cœur de la croissance (D. Foray, 2000). Dès lors, les acteurs locaux ont besoin d'information sur la production scientifique de leur territoire. Un des objectifs d'une veille scientifique exercée à un niveau local consiste donc à suivre presque en temps réel la production et les coopérations scientifiques impliquant les acteurs localisés sur un territoire, en l'occurrence dans le cadre de notre étude le Sud Loire. Face à ces besoins, il existe déjà des sources d'information comme les tableaux de bord de la Science et la Technologie[2], développés initialement au niveau des nations puis de l'OCDE et de l'Europe, dans le but de mesurer les efforts scientifiques et technologiques d'une nation, d'identifier ses forces et faiblesses et d'en

[1] Le Sud Loire est un périmètre d'observation de l'Agence d'urbanisme de la région stéphanoise. Il comprend, au 31 janvier 2005, 117 communes. Ce périmètre correspond à celui du SCOT (Schéma de Cohérence Territoriale). Le choix du périmètre repose sur un projet de territoires (qui renvoie à une logique de gouvernance) et non pas sur des limites administratives (arrondissement, département…).

[2] Cf. par exemple le Tableau de Bord STI 2003, 6ème numéro d'une série biennale lancée il y a dix ans par l'OCDE (OCDE, 2003).

suivre les évolutions[3]. Mais ces données ne sont pas suffisamment fines pour des exercices d'observation territoriale souvent confrontés à une double difficulté : le manque d'informations récentes d'une part et, le niveau d'agrégation des données, d'autre part. En effet, les données fournies par les organismes de collecte de statistiques présentent souvent un décalage de plusieurs années entre la date à laquelle la donnée a été collectée et la date à laquelle le territoire est étudié. Ce décalage s'explique par le temps nécessaire à la collecte des informations (notamment à partir d'enquêtes), à la vérification de la cohérence des données, et enfin à la mise à disposition des données collectées sous une forme exploitable statistiquement. La seconde difficulté est liée au niveau d'information diffusée. Il s'agit souvent de données agrégées à un niveau national, régional parfois départemental, mais il est rare d'avoir accès à un niveau d'observation plus fin, par exemple infra départemental. De plus, les données sont également très agrégées par discipline ; ce qui n'autorise pas les analyses par filière ou par pôle.

Le recours à des techniques de fouille de données peut permettre de palier ces difficultés grâce à l'exploitation d'autres sources d'information que celles produites par des organismes de collecte de données statistiques, comme par exemple celles disponibles dans la base Pascal gérée par l'INIST[4]. Dans cette perspective, la mise en place d'un outil de veille scientifique territorial consiste à exploiter directement les notices bibliographiques des articles publiés par les acteurs scientifiques localisés sur le territoire, à l'aide de techniques de fouille de données. L'originalité de la démarche réside dans le fait que la clé d'entrée considérée n'est plus un domaine technologique ou scientifique comme c'est le cas majoritairement dans les outils de veille internes aux structures (entreprises ou structures de recherche), ni même un niveau géographique de l'ordre d'un pays mais un niveau territorial très fin illustré ici par le Sud Loire. Cette approche permet en outre de surmonter les deux difficultés évoquées précédemment quant au niveau d'agrégation territorial et thématique. En effet, la richesse d'informations qu'offrent les notices bibliographiques de la base Pascal permet de :

- construire un corpus de références bibliographiques pour un territoire infra départemental. Le code postal et la commune des affiliations étant renseignés, il est possible, à partir de ceux-ci, de définir un périmètre géographique d'analyse très personnalisé.
- réaliser des analyses pour une discipline scientifique donnée, le code disciplinaire issu du plan de classement Pascal (champ CC) permettant d'identifier très précisément la discipline scientifique dans laquelle s'inscrit la publication. La base Pascal présente l'avantage de comporter des éléments descriptifs du contenu scientifique de la publication plus précis que ceux offerts par le Sciences Citation Index. La base Pascal dispose à la fois d'un plan de classement documentaire et de mots clés comme éléments descriptifs (DEF, IDF) contrôlés pour chaque article alors que l'Institut for Scientific Information ne dispose que d'une catégorisation attribuée selon le thème du journal où l'article a été publié[5]. Ainsi, à partir de la base de données Pascal, il est possible de repérer les publications produites pour chacun des pôles de compétences d'un territoire.

Notamment, dans le cadre de notre étude sur le Sud Loire, il a été ainsi possible d'analyser la production scientifique de trois pôles de compétences identifiés par les acteurs locaux : optique/vision, métallurgie/mécanique et technologies médicales. Plus largement, les domaines de la physique et de la santé ont été également étudiés. La science physique[6] peut être rapprochée de la métallurgie / mécanique ou de l'optique bien qu'elle relève davantage de la science fondamentale. On a élargi également les technologies médicales au secteur de la santé car l'analyse des co-publications avec les données de l'OST ont montré une spécialisation pour le département de la Loire dans ce domaine. Toutes disciplines confondues, 5421 publications allant de 1986 à 2004, ont été extraites de la base Pascal. Trois périodes

[3] C'est l'OCDE qui imagine pour la première fois l'intérêt de disposer d'indicateurs de la science et de la technologie dans les années 1960 dans le cadre du débat sur le « gap technologique » Etats-Unis/Europe. La NSF américaine dans les années 1970, Eurostat en Europe puis les différentes nations dans les années 1990 suivent ensuite l'exemple.

[4] La base Pascal fait l'objet de nombreuses analyses statistiques ou de synthèses documentaires réalisées par l'INIST sur des sujets très divers. Voir par exemple : http://www.inist.fr/DEMOS/anastat.php ou http://www.inist.fr/DEMOS/synthdoc.php

[5] De plus, la base de données bibliographiques du SCI (Sciences Citation Index) de l'Institut for Scientific Information présente également l'inconvénient d'un biais de couverture des revues en faveur de la science anglo-américaine au détriment des revues françaises (R. Barré et al., 1995).

[6] Nous avons extrait les notices qui font référence au pôle Optique/Vision.

correspondant approximativement à un même nombre de publications ont été distinguées : 1850 publications de 1986 à 1994, 1785 entre 1995 et 1999 et, 1786 entre 2000 et 2004.

A partir de ces données, l'analyse proposée dans une perspective de veille scientifique territoriale a un double objectif : premièrement, mesurer la production de la recherche scientifique et deuxièmement appréhender la manière dont cette recherche s'organise sur le territoire. Afin de répondre à ce double objectif, cette analyse comporte trois volets :
1. une étude de la production scientifique du territoire dans le but est de connaître, d'une part, le nombre de publications et son évolution et, d'autre part, les disciplines qui prédominent (Section 3).
2. une étude des réseaux d'acteurs pour mieux mesurer l'intensité des liens entre les acteurs de recherche publique et privée au niveau local et extra local (Section 4).
3. une étude territoriale destinée à identifier les principaux territoires partenaires du territoire et à appréhender notamment la dimension internationale des coopérations scientifiques. Cette analyse permet de connaître le degré d'ouverture du territoire vers l'extérieur. Grâce aux techniques de fouille de données, il s'agit de mettre en évidence les territoires les plus fortement associés (Section 5).

3 Analyse de la production scientifique d'un territoire

L'objectif premier de la mise en place d'un outil de veille scientifique territorial est de pouvoir évaluer presque en temps réel la production scientifique d'un territoire bien défini et de la caractériser. Il s'agit donc de connaître les disciplines prédominantes. Cette production scientifique est généralement mesurée à partir des publications scientifiques. En France, l'OST (Observatoire des Sciences et des Techniques) produit notamment des données de publications et les met à jour annuellement. Ces données fournies à l'échelle départementale sont agrégées en huit disciplines : recherche médicale, chimie, sciences pour l'ingénieur, physique, biologie fondamentale, sciences de l'univers, mathématiques, biologie appliquée et journaux multidisciplinaire. Or, cette nomenclature ne correspond pas au cadre de référence des acteurs locaux notamment dans le cadre du Sud Loire. Par exemple, les sciences pour l'ingénieur intègrent des publications dans le domaine aussi bien de la mécanique que de l'électricité - électronique etc.

En s'appuyant sur une étude bibliométrique pour mesurer le potentiel de recherche d'un pays, H. Rostaing et V. Leveille (2001) ont déjà souligné la difficulté de faire correspondre les nomenclatures des bases de données produites ‘‘systématiquement’’ aux cadres de référence des décideurs publics. Dans leur travail, les auteurs attachent beaucoup d'importance à l'appropriation des résultats par les commanditaires de telles études ; ce qui les conduit à effectuer une recodification des données brutes pour obtenir des résultats compréhensibles par les destinataires de ces analyses.

Dans le cadre de notre analyse du Sud Loire, un important travail de recodage des codes de classement de la base de données Pascal a aussi du être effectué pour affecter aux publications un domaine d'activité qui rejoigne la nomenclature personnalisée utilisée au niveau local. De la sorte, les publications dans chacun des pôles de compétences ont pu être comptabilisées. Ainsi, par exemple, la figure 1 montre que le poids de chacun des pôles dans le total des publications du Sud Loire a fortement évolué sur les deux dernières périodes. Alors que la part des publications dans l'optique n'était que de 2% pour la période 1986 à 1994, celle-ci atteint aujourd'hui les 11% devançant alors la métallurgie - mécanique et les technologies médicales qui enregistraient jusque là un nombre plus important de publications que l'optique. Le pôle de l'optique connaît une croissance très régulière et soutenue du nombre moyen annuel de publications (+354% entre les deux premières périodes et +64% entre les deux dernières périodes). A l'inverse, la métallurgie - mécanique a connu une baisse de son poids, conséquence directe de la baisse du nombre de publications dans ce secteur (figure 2). Le poids du secteur de la physique qui renvoie davantage aux sciences fondamentales progresse néanmoins. Par ailleurs, on observe qu'un peu plus de la moitié des publications sont produites dans la santé. Plus spécifiquement, le pôle des technologies médicales concentre 7% des publications et son poids progresse légèrement au cours des périodes.

Figure 1 : Evolution de la part des pôles de compétences ou des disciplines retenues dans le total des publications du Sud Loire

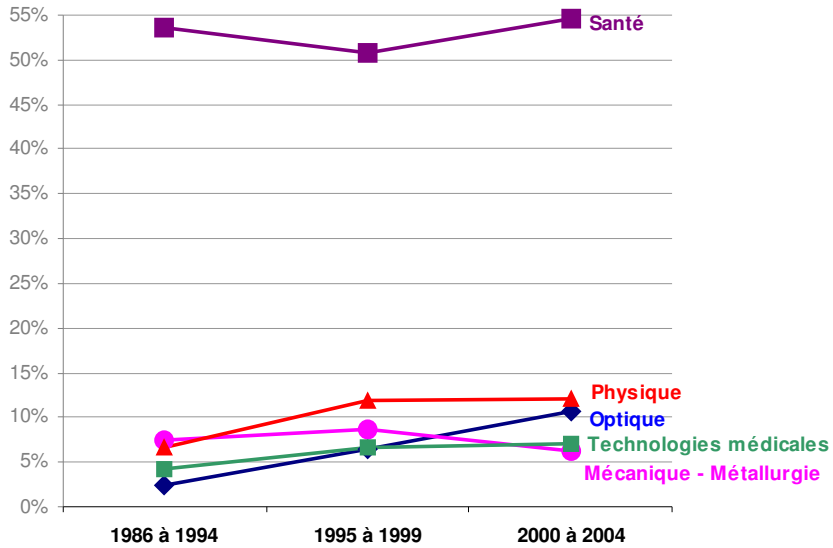
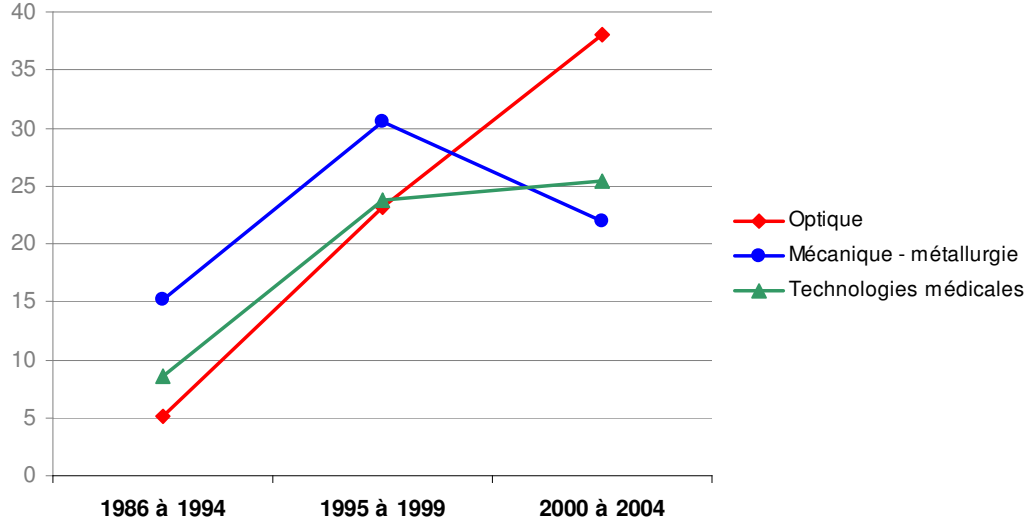


Figure 2 : Evolution du nombre moyen annuel de publications par pôles de compétences du Sud Loire



4 Identification des acteurs d'un territoire

Le second volet de l'analyse que nous proposons est consacré au repérage des réseaux d'acteurs localisés sur un territoire dans les différents domaines de collaboration. De telles études ont déjà été menées, au niveau d'un secteur d'activité, dans le but d'identifier les collaborations constituées entre les entreprises d'un même secteur. Par exemple, B. Dousset et B. Gay (2004) ont montré comment analyser les évolutions des alliances dans l'industrie des biotechnologies. La méthode proposée repose sur une représentation graphique de réseaux d'alliances interentreprises à partir de la base de données SDC Platinum V2.3, des sites Internet des entreprises et des nouvelles publiées sur le site Internet de Biospace2. Les graphes construits pour chaque période permettent d'observer les stratégies des entreprises en terme d'alliance et de mettre en évidence des sous réseaux. On observe également les évolutions des réseaux à travers l'apparition de nouveaux acteurs (nouveaux entrants dans le réseau) et la différence de positionnement des entreprises au cours du temps (acteur central du réseau ou non). D'autres études bibliométriques ont porté plutôt sur les relations existantes entre secteur public et secteur privé. Ainsi, Par exemple, G. Toledo et al. (2003) montrent comment les techniques bibliométriques permettent de quantifier les échanges entre les acteurs du secteur public et du secteur privé espagnols au niveau national ainsi que par secteur scientifique ou par Région Autonome. Ce travail a nécessité un important travail de recodification des structures d'affiliations des auteurs classées en douze grandes catégories institutionnelles.

Dans le contexte actuel où les politiques technologiques locales consistent notamment à soutenir les échanges entre la sphère privée et la sphère publique, il est important d'étudier les réseaux d' acteurs (publics et privés) à une échelle territoriale. Il s'agit alors de repérer les acteurs et de mesurer l'intensité des liens notamment

entre la recherche publique et la recherche privée que ce soit au niveau local ou extra local. C'est le but de ce second volet qui consiste à exploiter le corpus de notices bibliographiques à l'aide du logiciel d'analyse de réseaux Tétralogie (S. Karouach et B. Dousset, 2002 ; S. Karouach et B. Dousset, 2003). Dans le cadre de l'outil de veille mis en place pour le Sud Loire, cette analyse a été menée en recodant les structures d'affiliation de façon à identifier les réseaux entre les acteurs du Sud Loire mais aussi avec l'extérieur du Sud Loire (au niveau régional, national ou international). Le tableau 1 résume les classes de structures construites ainsi que l'intitulé de la variable qui apparaît sur les graphiques.

Tableau 1 : Description des variables présentes dans les graphes illustrant les réseaux d'acteurs

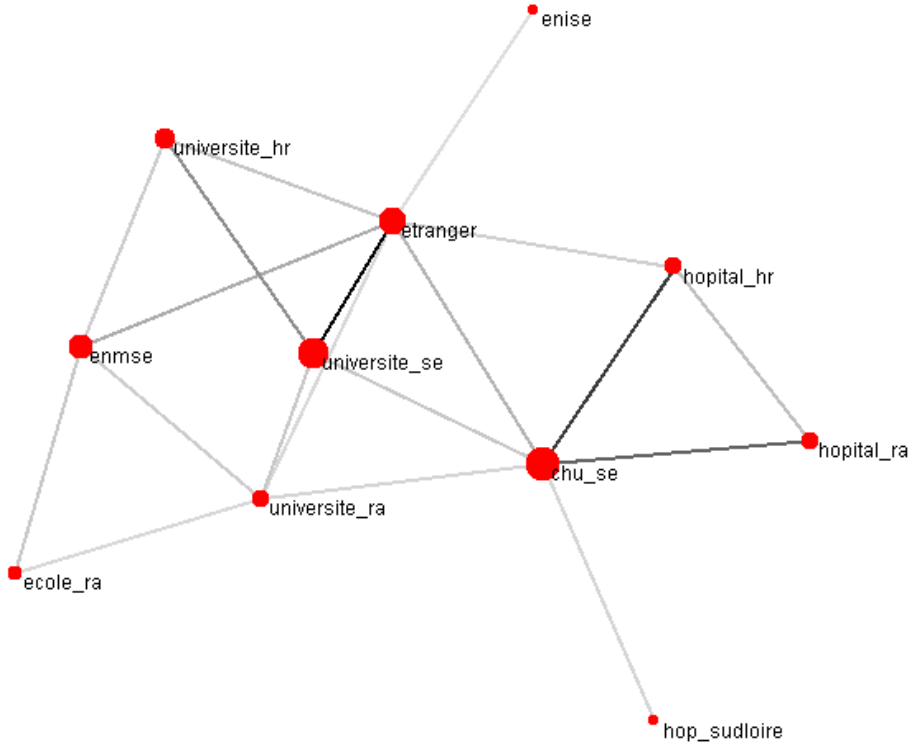
Nom de la variable	Description
universite_se	Université de Saint-Etienne
chu_se	CHU de Saint-Etienne
hop_sudloire	Hôpital n'appartenant pas au CHU et localisé sur le Sud Loire
enmse	Ecole Nationale des Mines de Saint-Etienne
enise	Ecole Nationale d'Ingénieur de Saint-Etienne
entreprise_sudloire	Entreprises de statut privé localisées sur le Sud Loire
universite_ra	Universités localisées dans la région Rhône-Alpes
hopital_ra	Hôpitaux localisés dans la région Rhône Alpes (y compris CHU)
entreprise_ra	Entreprises localisées dans la région Rhône-Alpes
ecole_ra	Grandes Ecoles localisées dans la région Rhône-Alpes
hopital_hr	Hôpitaux localisés en France à l'extérieur de la région Rhône-Alpes (y compris CHU)
entreprise_hr	Entreprises localisées en France à l'extérieur de la région Rhône-Alpes
ecole_hr	Grandes Ecoles localisées en France à l'extérieur de Rhône-Alpes
etranger	Toutes les institutions localisées à l'étranger sans différenciation de statut.
autre_labo_ra	Laboratoires localisés en Rhône-Alpes non rattachés à l'Université (ex : CEA de Grenoble)
autre_labo_hr	Laboratoires localisés hors de la région Rhône-Alpes non rattachés à une université

Un premier graphique construit pour toutes disciplines confondues (figure 3) montre que les principales structures qui publient sont le CHU et l'université de Saint-Etienne (constat visible par la taille des cercles). Au niveau local, l'Ecole des Mines de Saint-Etienne arrive en troisième position. Les liens les plus étroits (traits en foncé) entre les structures sont observables entre :

- le CHU et les hôpitaux régionaux et nationaux
- entre l'université de Saint-Etienne et les institutions à l'étranger. Si toutes les structures locales coopèrent avec des institutions étrangères, nous observons sur le graphique que c'est avec l'Université que les relations sont les plus fortes. Sur la graphique ci-dessous, nous observons que hormis le CHU de Saint-Etienne, les liens entre les acteurs locaux sont si faibles qu'ils n'apparaissent pas. Aucun lien n'apparaît entre l'Ecole Nationale des Mines de Saint-Etienne (ENMSE) et l'Ecole National d'Ingénieurs de Saint-Etienne (ENISE). Les autres liens impliquant l'université de Saint-Etienne sont construits avec les universités régionales ou nationales.

La figure 3 ne fait pas apparaître la catégorie ‘'entreprise" sur le Sud Loire, que ce soit au niveau régional ou national du fait d'un nombre réduit de publications produites par cette catégorie. Mais, d'autres analyses réalisées en excluant le domaine de la santé ont permis de la faire apparaître.

Figure 3 : Les coopérations scientifiques entre acteurs toutes disciplines confondues (filtre : nombre de publications > 33)



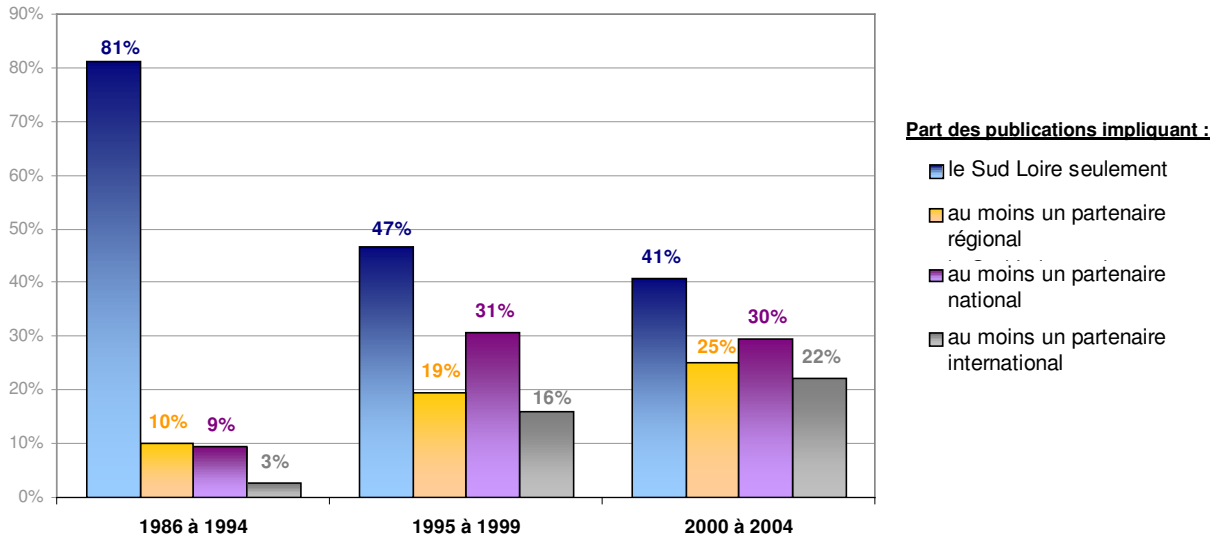
5 Suivi des partenariats territoriaux

Le troisième volet de l'analyse consiste à appréhender la manière dont s'organise la recherche du territoire avec d'autres territoires. Il s'agit donc d'analyser les partenariats territoriaux à travers notamment les coopérations scientifiques. La dimension géographique des collaborations scientifiques a aussi fait l'objet de travaux bibliométriques. C'est le cas de l'étude de J-L Multon et al. (2003) et J-L Multon et al. (2004) réalisée sur l'INRA, dans le but de construire des indicateurs utiles au management de la recherche. Les auteurs étudient notamment l'évolution des collaborations internationales et leur dimension géographique pour connaître les pays avec qui l'INRA collabore le plus. Les analyses croisées entre différents champs comme l'auteur, les journaux, les pays, les villes, les organismes, les universités, les thèmes de recherche permettent pour chacune des entités de l'institut de savoir qui fait quoi, avec qui, et depuis combien de temps. Dans le cadre de notre approche, nous proposons d'utiliser dans un premier temps des méthodes statistiques descriptives pour connaître les principaux territoires partenaires dans les coopérations scientifiques (section 5.1.), puis de faire appel à des techniques de fouille de données pour faire apparaître la complexité des réseaux. Il s'agit alors de mettre en évidence des réseaux à plusieurs territoires et non plus seulement des relations deux à deux (section 5.2.).

5.1 Identification des principaux territoires partenaires du Sud Loire

Appliquée au territoire du Sud Loire, cette approche nous a permis d'observer une forte augmentation des co-publications du Sud Loire avec d'autres territoires, que ce soit au niveau régional, national ou international (figure 4), sur les deux dernières décennies écoulées. La part des publications impliquant uniquement des acteurs scientifiques localisés sur le Sud Loire représentait un peu plus de 80% du nombre total de publications de 1986 à 1994[7]. Cette part ne s'élève plus qu'à 47% et 41% pour les périodes de 1995 à 1999 et de 2000 à 2004. En ce qui concerne les co-publications au niveau régional et national, leur poids a progressé de manière continue sur les trois périodes. Au niveau international, la part des publications a augmenté, tout particulièrement entre les deux premières périodes.

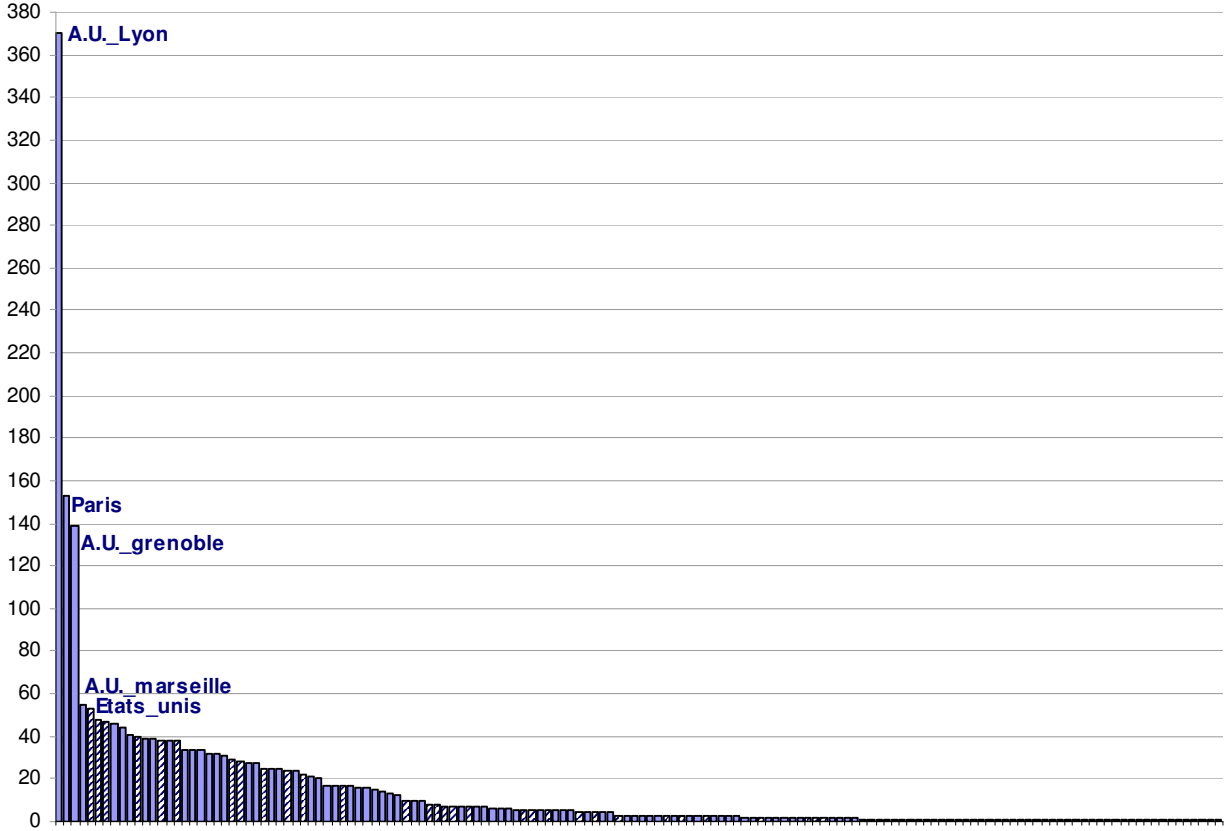
Figure 4 : Répartition des publications en fonction de la localisation des entreprises



Si nous analysons les coopérations à une échelle géographique plus fine (figure 5) pour la période 2000 à 2004, nous remarquons que la distribution est très dissymétrique : le Sud Loire a coopéré de manière intensive avec très peu de territoires. Dans cette distribution, nous observons deux ruptures importantes. On recense 370 notices pour le premier territoire qui est l'aire urbaine de Lyon. Ainsi, 21% des publications du Sud Loire font l'objet d'un partenariat avec Lyon. Ensuite, le nombre de notices chute à environ 150 pour deux territoires (Ville de Paris et aire urbaine de Grenoble). Pour les autres territoires, le nombre de publications est inférieur à 60 et décroît. L'aire urbaine de Marseille arrive en quatrième position avec un peu plus de 50 publications. Le cinquième territoire partenaire est un pays étranger : les Etats-Unis.

[7] Dans le graphique, la modalité ‘’Sud Loire uniquement" fortement sur-représentée pour la période 1986 à 1994 provient notamment du fait qu'avant 1996 tous les organismes collaborateurs n'étaient pas enregistrés dans la base de données Pascal.

Figure 5 : Le nombre de publications associant le Sud Loire avec les autres territoires sur la période 2000 - 2004



5.2 Représentation des réseaux scientifiques territoriaux

L'analyse précédente permet de mettre en évidence les principaux territoires partenaires. L'étape suivante s'attache à rendre compte davantage de la complexité des réseaux de coopérations scientifiques. En effet, il s'agit d'identifier des réseaux à plusieurs territoires et non plus des relations deux à deux. A l'aide de l'algorithme Apriori (R. Agrawal et R. Srikant, 1994), nous proposons de repérer des associations entre les territoires dans le corpus de références bibliographiques des collaborations scientifiques. La matrice étudiée est une matrice booléenne : en ligne apparaît la publication et en colonne le territoire. A l'intersection de la ligne et de la colonne, la valeur 1 s'affiche si le territoire a effectivement contribué à la publication, 0 sinon.

Dans le cadre du Sud Loire, nous avons retenu les règles d'association présentant un taux de support supérieur à 1% (soit 11 notices sur un total de 1063) et un taux de confiance supérieur à 50%. Les 31 règles d'association obtenues ont été réparties en quatre groupes (tableau 2).

Tableau 2 : Règles d'association[8] présentant un taux de support supérieur à 1% et un taux de confiance supérieur à 50% - période 2000 à 2004

Groupe	Règles association	Taux de support P(B,H)	Taux de confiance P(H|B)	Intérêt [B->H]	conviction [B->H]	dépendance [B->H]	nouveauté [B->H]	satisfaction [B->H]
1	au_lyon <= au_grenoble	**7,2%**	**55%**	1,59	2,46	0,21	**0,027**	0,32
	au_lyon <= au_bordeaux	**2,1%**	**56%**	1,62	2,51	0,22	0,008	0,33
	au_lyon <= au_nice	**1,4%**	**56%**	1,60	2,47	0,21	0,005	0,32
	au_lyon <= paris & au_bordeaux	**1,4%**	**75%**	2,15	**4,38**	0,40	0,008	**0,62**
	au_lyon <= paris & au_grenoble	**1,3%**	**68%**	1,97	**3,47**	0,34	0,006	0,52
	au_lyon <= paris & au_marseille	**1,2%**	**62%**	1,78	2,87	0,27	0,005	0,42
	au_lyon <= paris & au_toulouse	**1,2%**	**63%**	1,82	2,98	0,28	0,005	0,44
2	paris <= au_bordeaux	**1,9%**	**51%**	3,56	2,95	0,37	0,014	0,43
	paris <= au_lyon & au_bordeaux	**1,3%**	**68%**	4,74	**4,52**	0,54	0,011	**0,63**
	paris <= au_lyon & au_marseille	**1,2%**	**57%**	3,93	**3,31**	0,42	0,009	0,49
	paris <= au_rennes	**1,3%**	**52%**	3,61	2,99	0,38	0,010	0,44
3	paris <= au_besançon	**2,0%**	**54%**	3,74	**3,11**	0,39	0,014	0,46
	paris <= au_nantes	**2,4%**	**77%**	5,32	**6,12**	**0,62**	**0,020**	**0,73**
	paris <= au_amiens	**1,8%**	**95%**	6,60	**28,77**	**0,81**	**0,015**	**0,94**
	paris <= au_besançon & au_nantes	**1,7%**	**90%**	6,25	**14,38**	**0,76**	0,014	**0,88**
	au_amiens <= au_besançon & au_nantes	**1,3%**	**70%**	**37,21**	**5,50**	**0,68**	0,013	**0,69**
	au_amiens <= paris & au_besancçon	**1,2%**	**62%**	**32,90**	**4,33**	**0,60**	0,012	**0,61**
	au_amiens<=paris & au_besançon & au_nantes	**1,2%**	**72%**	**38,37**	**5,93**	**0,70**	0,012	**0,72**
	au_besançon <= au_amiens	**1,3%**	**70%**	**19,08**	**5,40**	**0,66**	0,012	**0,69**
	au_besançon <= au_nantes	**1,9%**	**59%**	**16,03**	**3,93**	0,55	**0,018**	0,57
	au_besançon <= paris & au_amiens	**1,3%**	**68%**	**18,64**	**5,12**	**0,65**	0,012	**0,67**
	au_besançon <= paris & au_nantes	**1,8%**	**69%**	**18,86**	**5,25**	**0,66**	**0,017**	**0,68**
	au_nantes <= au_amiens	**1,3%**	**70%**	**21,89**	**5,42**	**0,67**	0,013	**0,69**
	au_nantes <= au_besançon	**1,9%**	**51%**	**16,04**	**3,34**	0,48	**0,018**	0,50
	au_nantes <= paris & au_amiens	**1,3%**	**68%**	**21,39**	**5,15**	**0,65**	0,012	**0,67**
	au_nantes <= paris & au_besançon	**1,7%**	**86%**	**26,79**	**11,37**	**0,83**	**0,016**	**0,85**
4	suisse <= au_amiens	**1,0%**	**55%**	12,18	**3,57**	0,50	0,009	0,53
	suisse <= au_besançon & au_nantes	**1,1%**	**60%**	13,29	**4,01**	0,55	0,010	0,58
	suisse <= paris & au_amiens	**1,1%**	**58%**	12,82	**3,81**	0,53	0,010	0,56
	suisse <= paris & au_besançon	**1,1%**	**57%**	12,65	**3,74**	0,53	0,010	0,55
	suisse <= paris & au_besançon & au_nantes	**1,1%**	**67%**	14,77	**4,82**	**0,62**	0,011	**0,65**

Exemple de lecture du tableau 2 : Pour la règle ''au_lyon <= au_grenoble", l'interprétation est la suivante. Les coopérations entre le Sud Loire, l'aire urbaine de Grenoble et l'aire urbaine de Lyon représentent 7,2% du total des co-publications du Sud Loire. Si le Sud Loire co-publie avec l'aire urbaine de Grenoble alors l'aire urbaine de Lyon intervient également dans cette coopération dans 55,4% des cas.

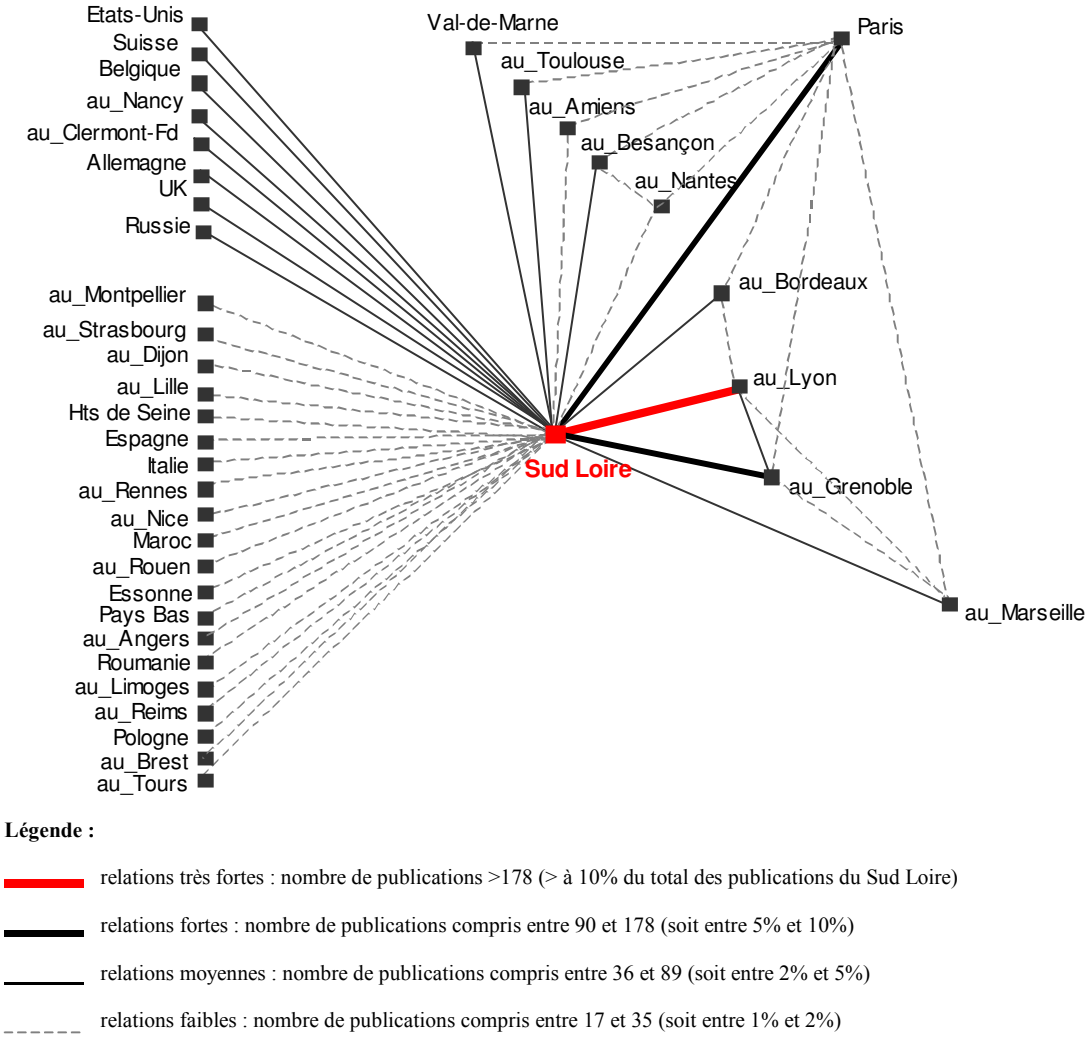
[8] Les règles d'association présentées dans le tableau ne font pas apparaître dans la prémisse de la règle le Sud Loire. Celui ci est implicitement inséré dans la prémisse de la règle puisque le Sud Loire apparaît dans toutes les publications. Lorsque le Sud Loire est intégré dans la condition, les taux de support et de confiance sont donc les mêmes. Celui n'est pas indiqué dans un souci de lisibilité. Les chiffres qui apparaissent en gras dans le tableau sont supérieurs au seuil fixé pour chacun des indicateurs : 1% pour le taux de support ; 50% pour le taux de confiance ; 15 pour l'intérêt ; 3 pour la conviction ; 0,6 pour la dépendance ; 0,015 pour la nouveauté (les valeurs obtenues pour cet indice sont relativement faibles pour toutes les règles compte tenu des taux de support peu élevé) ; 0,6 pour la satisfaction. Pour la définition de ces indicateurs, voir CHERFI H. et al. (2003).

Pour les sept règles d'association du premier groupe, nous observons que l'aire urbaine de Lyon apparaît toujours dans le résultat. Dans la prémisse de la règle, apparaissent les grands centres universitaires tels que Grenoble, Bordeaux, Nice, Paris, Marseille et Toulouse. Ainsi, lorsque le Sud Loire co-publie avec l'un de ces territoires, ce travail s'accompagne au moins une fois sur deux d'une coopération avec l'aire urbaine de Lyon (le taux de confiance variant entre 55% et 75%). Ce résultat montre la place relativement centrale de l'aire urbaine de Lyon dans les échanges entre le Sud Loire et les autres centres universitaires. Pour les autres groupes, on peut remarquer la place importante de la ville de Paris dans les coopérations scientifiques que ce soit avec des grands centres universitaires (groupe 2) ou avec des centres universitaires de taille plus modeste (groupe 3 et 4). Dans le troisième groupe, les valeurs relativement élevées de la conviction, de la dépendance, de la satisfaction et de l'intérêt montrent que les aires urbaines de Besançon, d'Amiens et de Nantes sont fortement dépendantes de la ville de Paris. Cette forte dépendance peut s'expliquer par le niveau de ressources scientifiques des trois aires urbaines qui nécessite une coopération plus systématique avec un grand centre universitaire.

Les réseaux peuvent être également représentés sous la forme d'un graphe d'association. Les différences d'intensité des relations sont mises en évidence grâce à l'épaisseur des traits. Dans le cadre du Sud Loire, deux types de réseaux peuvent être observés (Figure 6).

On observe, d'une part, des coopérations ''simples", ''directes" entre le Sud Loire et d'autres territoires (coté gauche du graphique). Dans ce cas, le Sud Loire copublie avec chacun des territoires sans que d'autres territoires soient associés de manière significative. Ce type de coopérations est majoritaire. On constate, d'autre part, des coopérations qui sont intégrées dans un système plus complexe de relations associant plusieurs territoires (coté droit du graphique). On peut noter la position centrale de la ville de Paris qui intervient aussi bien dans des coopérations avec des grands centres universitaires tels que Lyon, Toulouse, Marseille que dans des coopérations qui impliquent des aires urbaines avec une production scientifique plus modeste (Amiens, Besançon, Nantes). L'aire urbaine de Lyon a une position moins centrale. Elle intervient surtout dans les coopérations avec les villes du Sud de la France. Des liens forts sont observés entre les trois principales villes de Rhône-Alpes : Lyon, Grenoble et le Sud Loire illustrant ainsi le poids important des relations régionales.

Figure 6 : Les réseaux scientifiques impliquant le Sud Loire pour la période 2000 - 2004



6 Conclusion et perspectives

La démarche proposée ci-dessus permet à un territoire de mesurer sa production scientifique et ceci dans les différents pôles de compétences identifiés. Elle permet également de procéder à une analyse territoriale des coopérations scientifiques établies entre les acteurs du territoire. De plus, le recours à des techniques de fouille de données a permis d'extraire des connaissances nouvelles notamment sur la centralité de certains territoires dans les relations qu'entretient le Sud Loire avec l'extérieur. Mais ces techniques offrent encore d'autres perspectives en matière de veille territoriale par exemple en s'inspirant de travaux comme ceux de Lent et al. (Lent, 1997) dans lesquels des algorithmes d'extraction de motifs séquentiels fréquents (Agrawal, 1995) ont été utilisés pour déceler de nouvelles tendances dans une base de données de brevets chez IBM. Cette approche, tout comme les techniques de détection et de suivi de thèmes ou encore d'identification d'informations nouvelles pourraient être utilisées pour suivre les mutations technologiques et identifier des thématiques émergentes sur un territoire. Enfin, cette analyse mise en œuvre sur un autre territoire, doté d'un même pôle de compétence, pourrait également s'inscrire dans une veille concurrentielle territoriale.

7 Bibliographie

AGRAWAL R. et SRIKANT R., 1994, *Fast algorithms for mining association rules*, Proceedings Very Large Data Bases, 1994, pages 487-499

AGRAWAL R. et SRIKANT R, 1995, *Mining sequential patterns*, Eleventh International Conference on Data Engineering, IEEE, Taipei Taiwan, p. 3-14, 1995.

AMF & ETD, 2004, *La veille économique, un nouvel outil pour le développement territorial*, Les notes d'ETD (Entreprises Territoire et Développement).

BARRE R., LAVILLE F., TEIXEIRA N., ZITT M., 1995, *L'observatoire des sciences et des techniques : activités - définition - méthodologie*, Solaris, [http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2barre.html]

CARAYON B., (2003), *Intelligence économique, compétitivité et cohésion sociale*, La Documentation française, Paris, 2003.

CHALUS-SAUVANNET M-C, 2004, *Evolution des pratiques managériales des collectivités territoriales françaises* in *Le management face à l'environnement socio-culturel*, Conférence CEMADIMO et CIDEGEF, Beyrouth, 28 et 29 octobre 2004.

CHERFI H., NAPOLI A., TOUSSAINT Y., 2003, *Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association*, Conférence d'apprentissage CAP 2003, Laval, p 61-76

DESVALS H. et DOU H., 1992, *La veille technologique*, Dunod, 1992.

DOUSSET B. et GAY B., 2004, *Analyse par cartographie dynamique de l'effet de l'innovation sur la structure des réseaux d'alliances dans l'industrie des biotechnologies : application au domaine des anticorps thérapeutiques* in Veille stratégique, scientifique et technologique : VSST'04, volume 1, p. 145-154, Toulouse, 25-29 octobre 2004

FORAY D., 2000, *L'économie de la connaissance*, La découverte, Paris, 2000.

GILAD B. et HERRING J., 1996, *The Art and Science of Business Intelligence Analysis*, JAI Press, 1996.

HALLIMAN C., 2001, *Business Intelligence Using Smart Techniques : Environnemental Scanning Using Text Mining and Competitor Analysis Using Scenarios and Manual Simulation*, Information Uncover, 2001.

HERBAUX P. et BERTACCHINI Y., 2003, *Mutualisation et Intelligence Territoriale*, in ISDM, International Journal of Information Sciences for Decision Making, n°9, 2003.

JAKOBIAK F., 1990, *Pratique de la Veille Technologique*, Editions d'organisation, 1990.

JAKOBIAK F., 1994, *Le brevet, source d'information*, Dunod, Paris, 1994.

KAROUACH S., DOUSSET B., 2002, *Visualisation de relations par des graphes interactifs de grande taille*, in *9ème journée sur les systèmes d'information élaborée : bibliométrie - information stratégique - veille technologique*, Ile Rousse (Corse), octobre 2002.

KAROUACH S., DOUSSET B., 2003, *Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grande taille*, Workshop Inforsid Recherche d'information : un nouveau passage à l'échelle, Nancy, 3 juin 2003.

LENT B., AGRAWAL R., SRIKANT R., *Discovering Trends in Text Databases*, Proceedings of the 3rd International Conference on Knowledge Discovery, KDD'97, AAAI Press, p. 227-230, 1997.

MARTRE H., *Intelligence économique et stratégie des entreprises*, Commissariat Général au Plan. Rapport du groupe présidé par Henri Martre. La Documentation française, Paris, 1994.

MOSS L.T. et ATRE S., 2003, *Business Intelligence Roadmap : The Complete Project Lifecycle for Decision-Support Applications*, Addison-Wesley, 2003.

MULTON J-L., BRANCA-LACOMBE G., DOUSSET B., 2003, *Analyse bibliométrique des collaborations internationales de l'INRA*, in ISDM, International Journal of Information Science for Decision Making, n°6, 2003.

MULTON J-L., BRANCA-LACOMBE G., DOUSSET B., DAGUILLANES C., 2004, *Indicateurs utiles au management de la recherche issus d'un corpus multi-sources*,in Veille stratégique, scientifique et technologique : VSST'04, 2, Toulouse, 25-29 octobre 2004, pp 169 - 78

OCDE, 2003, *Science, technologie, industrie : Tableau de bord de l'OCDE*, 2003.

ROSTAING H., LEVEILLE V., 2001, *Etude bibliométrique pour l'évaluation des programmes de recherche nationaux - Difficulté de mise en oeuvre et d'exploitation dans le cas de la recherche scientifique algérienne*, in Veille stratégique, scientifique et technologique : VSST'01, Barcelone, Espagne, 15-19 septembre 2001.

TOLEDO G., ROMAN ROMAN A., ROSTAING H., 2003, *Analyse du transfert de l'information scientifique et technique entre le secteur public et le secteur privé. Etude des co-publications dans les revues scientifiques espagnoles*, in ISDM, International Journal of Information Science for Decision Making, n°6.