Étude diachronique de la production scientifique de l'Europe centrale francophone en utilisant la base de données PASCAL

Pascal Cuxac, Alain Collignon

ETUDE DIACHRONIQUE DE LA PRODUCTION SCIENTIFIQUE DE L'EUROPE CENTRALE FRANCOPHONE EN UTILISANT LA BASE DE DONNEES PASCAL

Alain Collignon
INIST-CNRS 2 allée du Parc de Brabois, CS 10310 54519 Vandoeuvre les Nancy Cedex France
e-mail : alain.collignon@inist.fr

Pascal Cuxac
INIST-CNRS 2 allée du Parc de Brabois, CS 10310 54519 Vandoeuvre les Nancy Cedex France
e-mail : pascal.cuxac@inist.fr

Résumé *: Les bases de données bibliographiques sont quotidiennement utilisées par les chercheurs car elles représentent un état de l'art, aussi bien des méthodes scientifiques que des résultats et des connaissances actuelles. Pourtant, l'utilisation des bases de données bibliographiques pourrait être compromise. En effet, la concentration des grands acteurs privés du domaine est préoccupante. Depuis plusieurs années, les grands éditeurs se concentrent verticalement et contrôlent toute la chaîne, de l'édition à la diffusion. De plus, l'accès aux résultats de la recherche et à l'information scientifique et technique est en train de changer de paradigmes et ce grâce au « libre accès » et plus particulièrement au phénomène d'auto-archivage des travaux de recherche dans des archives ouvertes.*

Dans ce contexte, l'objet de cet article est de démontrer quel peut être le rôle ou l'influence que peut avoir la base de données bibliographique PASCAL dans la diffusion et la valorisation de la production scientifique de l'Europe centrale francophone et plus particulièrement pour l'Albanie, la Bulgarie, la Macédoine, la Moldavie et la Roumanie. Nous montrerons également comment un outil d'interrogation et d'analyse de l'information comme la plateforme Stanalyst permet d'analyser finement les publications scientifiques et peut être également un outil pour un large public allant du documentaliste au décideur en passant par le veilleur et le scientifique.

Mots clés : *Europe centrale ; Francophonie ; Scientométrie ; Clustering ; Analyse diachronique.*

1. Introduction

L'INIST est une unité de service du Centre National de la Recherche Scientifique (CNRS) qui collecte, analyse et diffuse les résultats de la recherche mondiale en sciences, technologie, médecine ainsi qu'en sciences humaines, sociales et économiques.

Hérité des anciens centres de documentation du CNRS, cet institut, créé en 1988, est localisé dans l'est de la France sur le technopôle de Nancy.

Dès le début de sa création, l'INIST a été imaginé comme une usine de l'information :

l'activité documentaire s'organise selon un processus de production qui permet de traiter d'importants volumes de données et de répondre aux exigences de qualité des utilisateurs (Dusoulier, 1993). Ce processus permet à partir d'un document primaire d'élaborer de nombreux produits et services. L'INIST produit d'une part, des bases de données bibliographiques PASCAL et FRANCIS signalant respectivement 18 millions de référence en sciences, technologie et médecine depuis 1973 et près de 3 millions de références en sciences humaines et sociales depuis 1972 et d'autre part, fournit la copie à la demande des documents recherchés.

Dès 2000, avec l'arrivée de services de diffusion de versions électroniques de revues, l'INIST s'oriente vers l'acquisition et la mise à disposition de ces nouvelles formes de publication à travers des portails documentaires généralistes ou thématiques. Un service d'édition électronique est également créé à cette époque, dans le cadre de la diversification des activités d'offres de service personnalisés et intégrés pour les utilisateurs de l'Institut.

L'évaluation de la recherche scientifique a connu ces premières prémices dès le début du siècle à partir d'études statistiques sur la littérature scientifique. Au début des années 60, ces études ont connus une accélération avec la création du Science citation Index (SCI). L'impact de la scientométrie sur la mesure de la science et son impact sur la vision de la recherche internationale est fondamental (Courtial, 1990). Bien entendu, différentes études ont été menées sur les pays de l'Europe centrale, nous pouvons citer par exemple l'étude menée par l'Observatoire des Sciences et Techniques en 2004 (Thèves et Séchet). Plus récemment, A. Repanovici nous présente l'utilisation d'un outil en libre accès pour mesurer la productivité de la recherche en Roumanie (Repanovici, 2010).

Nous verrons dans cet article quel rôle ou influence peut avoir la base de données bibliographique PASCAL dans la diffusion et la valorisation de la production scientifique de l'Europe centrale francophone et plus particulièrement pour l'Albanie, la Bulgarie, la Macédoine, la Moldavie et la Roumanie. L'objectif de notre étude est de réaliser un constat sous forme de cartographie des domaines scientifiques étudiés.

Nous commencerons par présenter les activités autour de la production de la base PASCAL, puis nous poursuivrons avec une étude infométrique diachronique portant sur la production pour les périodes 2000 – 2004 et 2006 – 2010 relatives aux publications d'Europe centrale francophone. Tout au long de cette étude infométrique, nous étudierons différents paramètres permettant de mettre en lumière le rôle de la base de données PASCAL pour la production scientifique de l'espace francophone de l'Europe centrale.

2. La base de données PASCAL

2.1. Généralités

Pour réaliser ses missions l'INIST s'appuie sur un fonds documentaire unique qui couvre l'ensemble de la littérature cœur en sciences, technologie, médecine et sciences humaines et sociales (Guichard, 1999).

Ce fonds est essentiellement constitué par 19 700 titres de périodiques dont 4 500 abonnements en cours, 125 000 thèses en sciences et technique, 115 000 conférences, 75 000 rapports et 13 000 monographies.

Le stockage est réalisé sous forme papier représentant 25 kilomètres de rayonnage en bibliothèque et au format électronique représentant plus de 7 millions d'articles numérisés en format image (TIFF) depuis 1990.

La production de la base de données PASCAL est placée sous la responsabilité du département d'ingénierie de l'information scientifique. Ce département, divisé en quatre services scientifiques et un service dédié à la gestion et au suivi de production, est composé de 56 ingénieurs documentalistes, spécialistes dans leur domaine d'indexation. Quelques

coopérants, comme par exemple le réseau documentaire d'information en santé publique - BDSP[1], viennent renforcer cette activité. Plus de 3 000 titres de périodiques, majoritairement en langue anglaise, sont analysés annuellement, correspondant à un accroissement de 480 000 références bibliographiques. La base PASCAL, interrogeable en français, anglais et espagnol est distribuée via différents supports comme des serveurs (Questel, Dialog, Datastar, …) ou des fournisseurs d'information tels que Ovid ou EBSCO. Depuis 1973, date de la création de PASCAL, l'INIST possède plus de 18 millions de références réparties dans différents domaines scientifiques (figure 1).

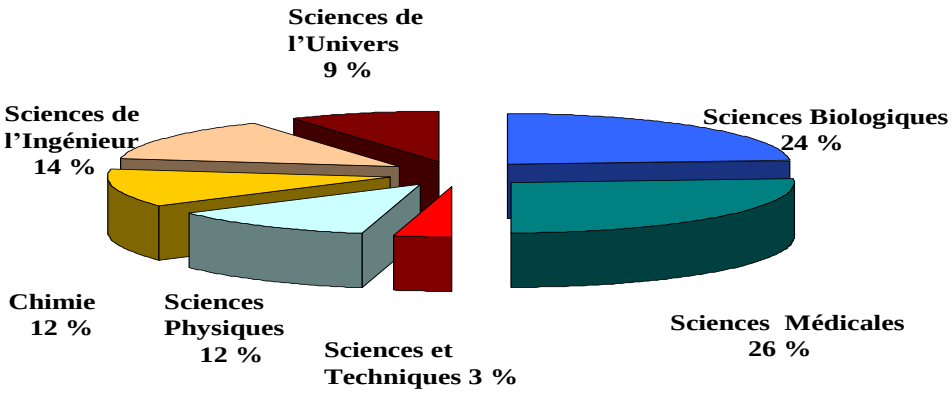

Figure 1. *Répartition des domaines scientifiques dans PASCAL.*

2.2. L'indexation

Le travail d'indexation réalisé par les ingénieurs documentalistes comprend différentes tâches permettant de représenter le contenu des articles analysés par des :

- descripteurs contrôlés choisis dans les vocabulaires spécifiques utilisés à l'INIST ;
- mots-clés non contrôlés (candidats descripteurs) ;
- codes de classement choisis dans les plans de classement élaborés par l'INIST.

L'indexation implique deux tâches intimement liées :

- l'indexation proprement dite ;
- la constitution et la gestion des ressources nécessaires à l'indexation :
  - langages documentaires :
    - vocabulaires de référence ;
    - plans de classement ;
  - référentiels de règles de pré-indexation.

L'indexation peut-être soit manuelle, dans ce cas les descripteurs et les codes de classement sont saisis par le rédacteur au niveau d'une zone de saisie ou directement dans une grille d'indexation (Ménillet, 1992), soit assistée. Pour cette dernière, deux moteurs d'indexation sont utilisés, l'un pour réaliser la pré-indexation (approche symbolique ou linguistique faisant appel à un programme qui utilise des règles de reconnaissance de chaînes de caractères dans les textes analysés pour produire une liste d'éléments d'indexation), et l'autre la collocation lexicale (approche statistique utilisant un dictionnaire d'association entre mots du titre et éléments d'indexation ; ce dictionnaire est constitué automatiquement lors d'un processus d'apprentissage sur un corpus indexé).

[1] BDSP : Banque de Données en Santé Publique ; htpp://www.bdsp.ehesp.fr

3. Etude scientométrique de la production scientifique

3.1. Le corpus étudié

Notre étude s'est limitée à étudier deux corpus de notices bibliographiques signalées pour la période 2000 - 2004 et pour la période 2006 – 20010 dans la base PASCAL tous domaines scientifiques confondus, dont un auteur au moins appartient à un pays francophone de l'Europe centrale.

Pour la période 2000 – 2004, le corpus étudié contient plus de 9 000 notices bibliographiques (pour un total de 1 652 périodiques analysés) tandis que pour la période 2006 – 2010, le corpus étudié contient plus de 10 000 notices bibliographiques (pour un total de 1 472 périodiques analysés).

3.2. La méthode d'analyse scientométrique utilisée

3.2.1. Construction des graphes de relations pays

Nous avons entrepris l'étude des coopérations scientifique entre pays à partir d'une analyse de graphe. Un tel graphe, s'il est construit en prenant en compte quelques paramètres, permet, mieux qu'un tableau, de mettre en évidence non seulement les relations entre pays mais également leur rôle dans le réseau.

Le graphe est construit en prenant en compte les pays d'affiliation des auteurs d'articles dans les corpus étudiés. Nous obtenons alors un graphe non dirigé. Un premier filtrage est réalisé pour ne prendre en compte que les arêtes qui ont au moins une extrémité correspondant à un des pays étudié : en effet le but est d'analyser les relations qu'ont ces pays entre eux et avec le reste du monde.

Pour des raisons de clarté, le graphe est élagué pour supprimer les liens qui apparaissent moins de 10 fois (ce qui donne en moyenne plus de 2 coopérations par ans entre pays pour les périodes de 5 ans que nous avons étudiés).

Les nœuds et arêtes du graphe sont ensuite colorés et dimensionnés en fonction des paramètres suivants :

- Taille du nœud : augmente avec le paramètre de « *betweenness centrality* »
- Couleur du nœud : du plus clair au plus foncé quand le *degré* du nœud augmente
- Epaisseur de l'arête : augmente avec la *fréquence* du lien entre les deux pays reliés
- Couleur de l'arête : du plus clair au plus foncé quand le paramètre de « *edge-betweenness centrality* » augmente

« betweenness centrality » : Cette notion est apparue l'analyse de réseaux sociaux pour déterminer le rôle de chaque acteur dans le réseau. Cette mesure d'intérêt commun est basée sur la notion de plus court chemin. Si deux nœuds sont connectés, il peut exister plusieurs chemins entre eux. Les plus courts chemins sont ceux pour lesquels le nombre de nœuds sur le chemin est minimal. La "betweenness centrality" pour un nœud v est alors le nombre de plus courts chemins entre deux autres nœuds s et t qui passent par v(noté σst(v)), divisé par le nombre total de plus courts chemins allant de s à t (σst).

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} = nombre de plus courts chemins entre s et t ;
$\sigma_{st}(v)$ = nombre de plus courts chemins entre s et t passant par le nœud v

degré d'un nœud : Dans les graphes non orienté, le degré (ou la valence) d'un nœud est le nombre d'arêtes ayant une extrémité en ce nœud. Si le graphe est aléatoire on a alors une distribution normale de ce paramètre (Erdős, 1959).

Fréquence du lien : reflète le nombre de fois ou on constate une coopération entre deux même pays (nœuds du graphe).

« edge betweenness centrality » : Girvan et al. (2002) ont étendu la notion de betweenness centrality aux arêtes. en définissant la "edge-betweenness centrality".

3.2.2. La plateforme d'analyse Stanalyst

Les corpus, issus de la base de données bibliographique PASCAL du CNRS, sont créés et analysés en utilisant la station d'analyse Stanalyst (Grivel et al. 1995) ; http://stanalyst.inist.fr). Cette station regroupe quatre modules (figure 2) permettant de créer la requête d'interrogation des bases de données, de faire une analyse bibliométrique des résultats, de filtrer les termes d'indexation et enfin de faire une classification non supervisée qui sera ensuite représentée sous forme de carte.

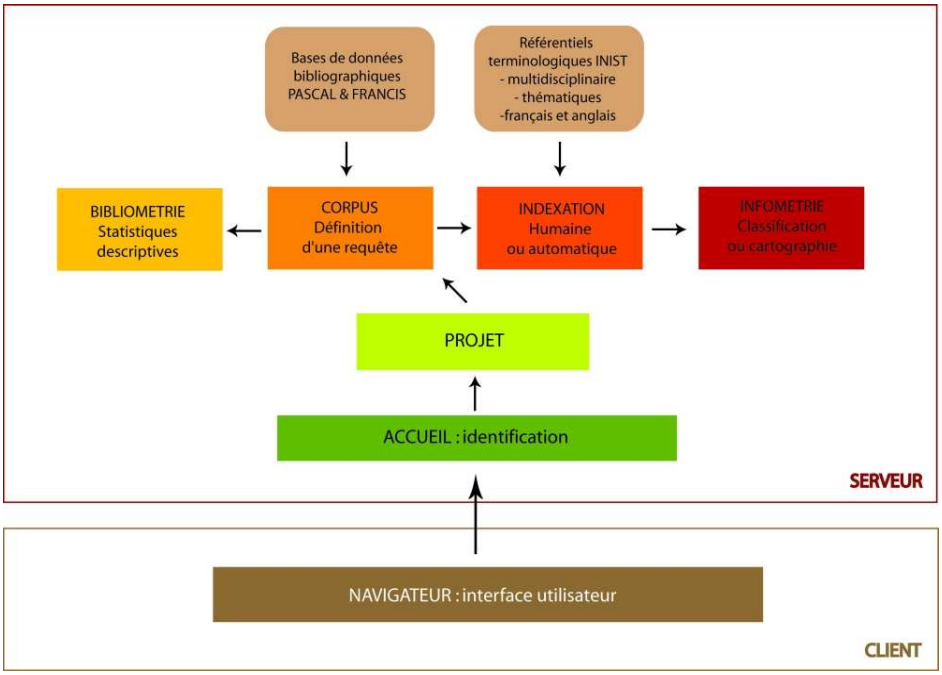

Figure 2. *Architecture de la plateforme Stanalyst*

Le corpus issu de l'interrogation de la base de données PASCAL est constitué de notices bibliographiques structurées en champs (figure 3).

Le module « bibliométrique » de Stanalyst permet de faire des statistiques descriptives sur plusieurs de ces champs (auteurs, pays d'affiliations, pays d'édition, langues, descripteurs, etc.).

FT : Epilepsie par encornement bovin (zébu): à propos d'un cas au Burkina Faso
ET : (Epilepsy by bovine (zebu) goring: a case report in Burkina Faso)
AU : NAPON (C.); DRAVE (A.); KABORE (J.)
AF : Service de neurologie du CHU Yalgado-Ouedraogo, BP 7022 Ouagadougou 03/Burkina Faso (1 aut., 2 aut., 3 aut.)
DT : Publication en série; Niveau analytique
SO : Bulletin de la Société de pathologie exotique; ISSN 0037-9085; Coden BSPEAM; France; Da. 2009; Vol. 102; No. 4; Pp. 217-218; Abs. anglais; Bibl. 4 ref
LA : Français
EA : The post-traumatic epilepsy is responsible for 20% of the symptomatic epilepsies. Accidents on public highway constitute more than 70% of the causes. We report a singular case of fronto-polar post-traumatic epilepsy by zebu goring which appeared two years after the traumatism. The neurological …
CC : 002B01; 002B17A03; 002B16B
FD : Epilepsie; Traumatisme crânien; Infection; Bovin; Etude cas; Burkina; Hôpital; Médecine tropicale; Homme
FG : Artiodactyla; Ungulata; Mammalia; Vertebrata; Afrique; Pathologie de l'encéphale; Pathologie du système nerveux central; Pathologie du système nerveux
ED : Epilepsy; Head trauma; Infection; Bovine; Case study; Burkina Faso; Hospital; Tropical medicine; Human
EG : Artiodactyla; Ungulata; Mammalia; Vertebrata; Africa; Cerebral disorder; Central nervous system disease; Nervous system diseases
SD : Epilepsia; Traumatismo craneoencefálico; Infección; Bovino; Estudio caso; Burkina Faso; Hospital; Medicina tropical; Hombre

Figure 3. *Exemple d'une notice bibliographique PASCAL*

Le module « indexation » permet soit de faire une indexation automatique du corpus si on possède une ressource terminologique, soit d'utiliser l'indexation native des notices. Dans tous les cas le résultat d'indexation doit être révisé afin d'enlever les termes trop génériques ou non porteurs de sens pour l'analyse considérée, et cela dans le but d'effectuer une classification.

Cette classification est obtenue en utilisant le module « infométrie ». Il permet de réaliser deux types de classifications non supervisées (clustering) :

- Neurodoc : méthode inspirée des K-means (MacQueen 1967). Cet algorithme que nous avons utilisé pour analyser les différents corpus et produire des cartes thématiques va être détaillé ci-dessous.
- Sdoc : méthode de classification basée sur les mots associés (Callon et al., 1983)

Algorithme K-moyennes
Entrée : k : le nombre de groupes recherchés
début
Choisir le centre initial des k groupes
répéter
Affecter chaque document au groupe dont il est le plus proche
Recalculer le centre de chaque groupe
jusqu'à ce que (stabilisation des centres)
fin

Figure 4. *Algorithme des K-means*

L'algorithme des K-means (figure 4) consiste à créer K groupes de documents de façon à ce que chaque groupe soit le plus dense possible et que ces groupes soient les plus distants possibles. L'utilisateur choisit le nombre de classes maximum désiré (K) et va initialiser K classes par un point représentant le centre de chaque classe. Les documents sont ensuite affectés à la classe dont ils sont le plus proche et le centre de la classe est recalculé jusqu'à stabilisation.

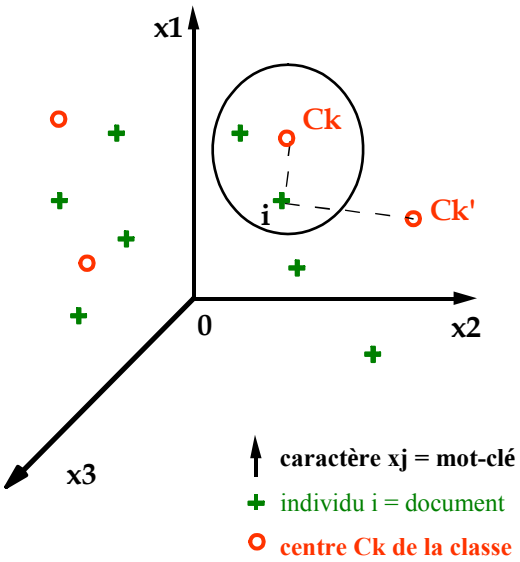

Figure 5a. *Principe des K-means*

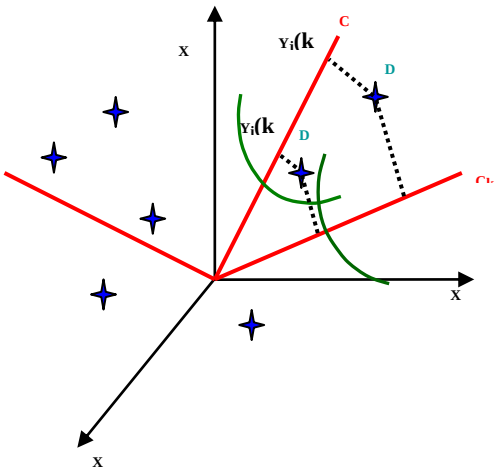

Figure 5b. *Principe des K-means axiales*

La méthode des K-means axiales (Lelu, 1993) diffère en représentant les classes non plus par leur centre mais par un demi-axe (figure 5a et 5b). Les documents se projettent alors plus ou moins loin sur ces axes. Quand la projection se fait au dessus d'un certain seuil fixé par l'utilisateur le document est affecté à la classe. On voit ainsi qu'un document peut appartenir à plusieurs classes (classes recouvrantes). De la même façon on peut projeter les mots-clés sur ces axes. Les documents et mots-clés sont ordonnés selon un degré de ressemblance au type idéal de la classe et sont affectés d'un poids plus ou moins important suivant qu'ils se projettent plus ou moins haut sur ces axes.

A l'issue de cette étape de classification, les documents sont représentés dans l'espace de description, c'est-à-dire dans l'espace des mots-clés (il n'est pas rare d'avoir des corpus représentés dans des espaces à 5 000 dimensions).

Pour obtenir un résultat lisible par l'œil humain nous réalisons une analyse en composantes principales (ACP) afin de représenter le résultat sur un plan. Nous faisons également figurer les valeurs remarquables des cosinus entre classes, ce qui permet d'esquisser un réseau de classes (figure 12).

3.2.3. L'analyse diachronique avec les règles d'association floues

Des travaux précédents (Cuxac et al., 2005) ont montrés qu'on pouvait utiliser des règles d'associations pour mettre en évidence des liens, et donc une évolution, entre classes de classifications différentes.

Une règle d'association A→B extraite d'une base de données, représente un lien entre les deux ensembles de propriétés A et B. Pour mesurer la qualité de cette règle, on dispose de nombreux indices (Kodratoff, 2001) fonctions de ces effectifs, dont les plus courants sont le support, qui est le nombre d'objets vérifiant les propriétés de A et de B, et la confiance, qui est le quotient de ce support et du nombre d'objets vérifiant les propriétés de A, c'est-à-dire du support de A. On peut aisément transposer ce formalisme de la fouille de données à la classification d'un corpus bibliographique, en prenant pour propriétés les classes et pour objets les mots-clés. Habituellement les règles d'association sont utilisées pour des données binaires (0/1). Ici la méthode de clustering utilisée (KMA) permet d'affecter aux mots de chaque classe un poids. Utilisant ce poids, nous avons alors des données qui ne sont plus binaires et les règles d'association habituelles ne sont plus utilisables. On les remplace alors

par des règles d'association floues (Cadot et Napoli, 2004). Du point de vue pratique, on redéfinit l'appartenance aux classes et le support d'une classe de la façon suivante : pour chaque mot-clé, sa valeur d'appartenance à la classe qu'il caractérise correspond à son poids dans la classe. Ainsi, si le mot-clé i a le poids a_i pour la classe A, et b_i pour la classe B, sa valeur d'appartenance aux deux classes est la plus petite des deux valeurs, min (a_i, b_i). Pour plus de détail on se rapportera à l'article de Cuxac et al. (2005).

3.3. Résultats

3.3.1. La publication : quoi, qui et où ?

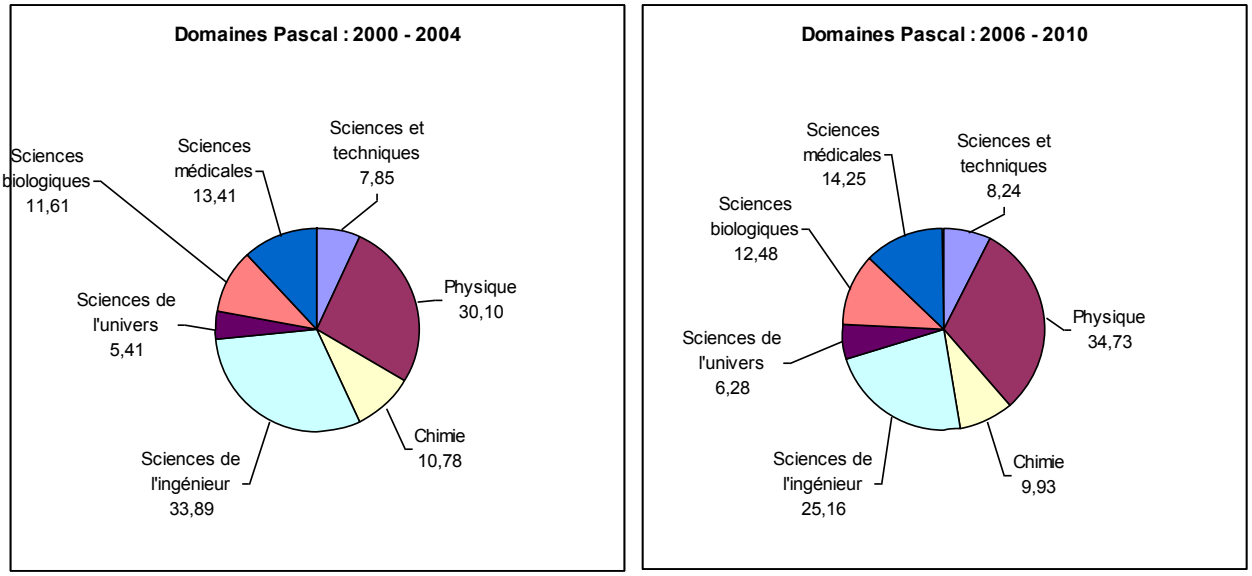

Figure 6. *Répartition des domaines PASCAL analysés.*

La figure 6 représente la ventilation de la production scientifique des pays considérés, telle qu'elle est analysée dans PASCAL pour les deux périodes étudiées. Nous constatons, que la physique (près de 35%) et les sciences appliquées (environ 30%) sont les deux domaines scientifiques qui représentent les champs de connaissance les plus étudiés. Les domaines en relation avec les sciences biologiques (environ 10%) et les sciences médicales (près de 15%) sont moins développés. On constate peu de changements entre les deux périodes.

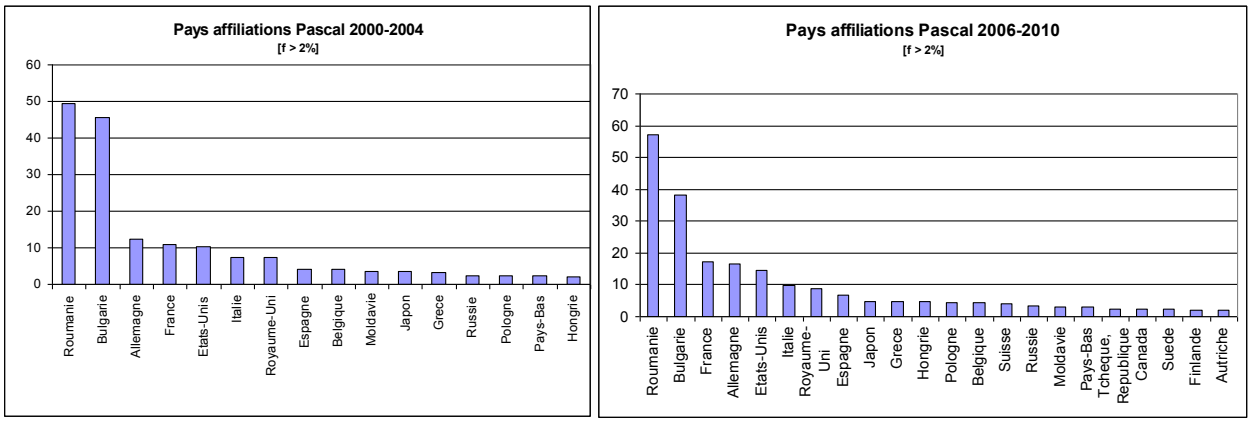

Figure 7. *Ventilation de la production scientifique en fonction des pays d'affiliation.*

Pour la période 2000 – 2004, la répartition des pays d'affiliation (figure 7) s'effectue sur 17 pays, tandis que pour la période 2006 – 2010, 23 pays participent à cette ventilation.

Malgré, cette différence, nous constatons pour les deux périodes considérés, que la Roumanie (pour plus de 50 %) et la Bulgarie (environ 43 %) sont les deux pays moteurs.

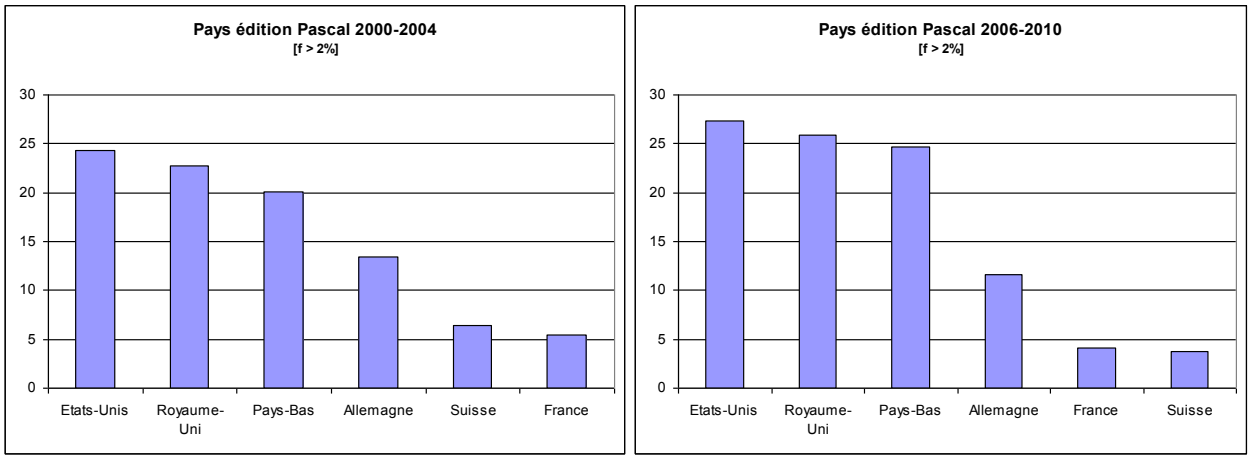

Figure 8. *Ventilation de la production scientifique en fonction des pays d'édition*.

Quelque soit la période, ce sont dans les journaux édités soit par les Etats-Unis, le Royaume-Uni ou les Pays-Bas que l'on retrouve la majorité de la production scientifique de la région étudiée.

3.3.2. La publication : comment ?

Après s'être intéressé à l'étude concernant l'organisation de la publication, notre réflexion va porter maintenant sur le constat de publication par pays.

Dans un premier temps, nous regarderons la distribution de la production par discipline scientifique et par pays (figure 9). Dans un second temps, nous nous attarderons sur la relation inter-pays (figure 10 et 11).

La figure9, nous indique que les périodes d'analyse n'ont pas d'influence sur les thèmes de recherche, mais démontre la présence de pôles de recherche différents par pays.

En effet, en Albanie, les sciences médicales et les sciences biologiques sont au cœur de la recherche scientifique. Par contre, ces deux disciplines sont certes présentes dans les autres pays mais en moindre mesure. En effet, que ce soit en Bulgarie, Moldavie ou en Roumanie, la physique, et en moindre mesure les sciences de l'ingénieur sont les domaines de prédilections des chercheurs des pays considérés.

Néanmoins nous remarquons une constante. En effet, pour les quatre pays étudiés, les sciences de l'univers, la chimie et les sciences et techniques sont absentes du spectre de recherche.

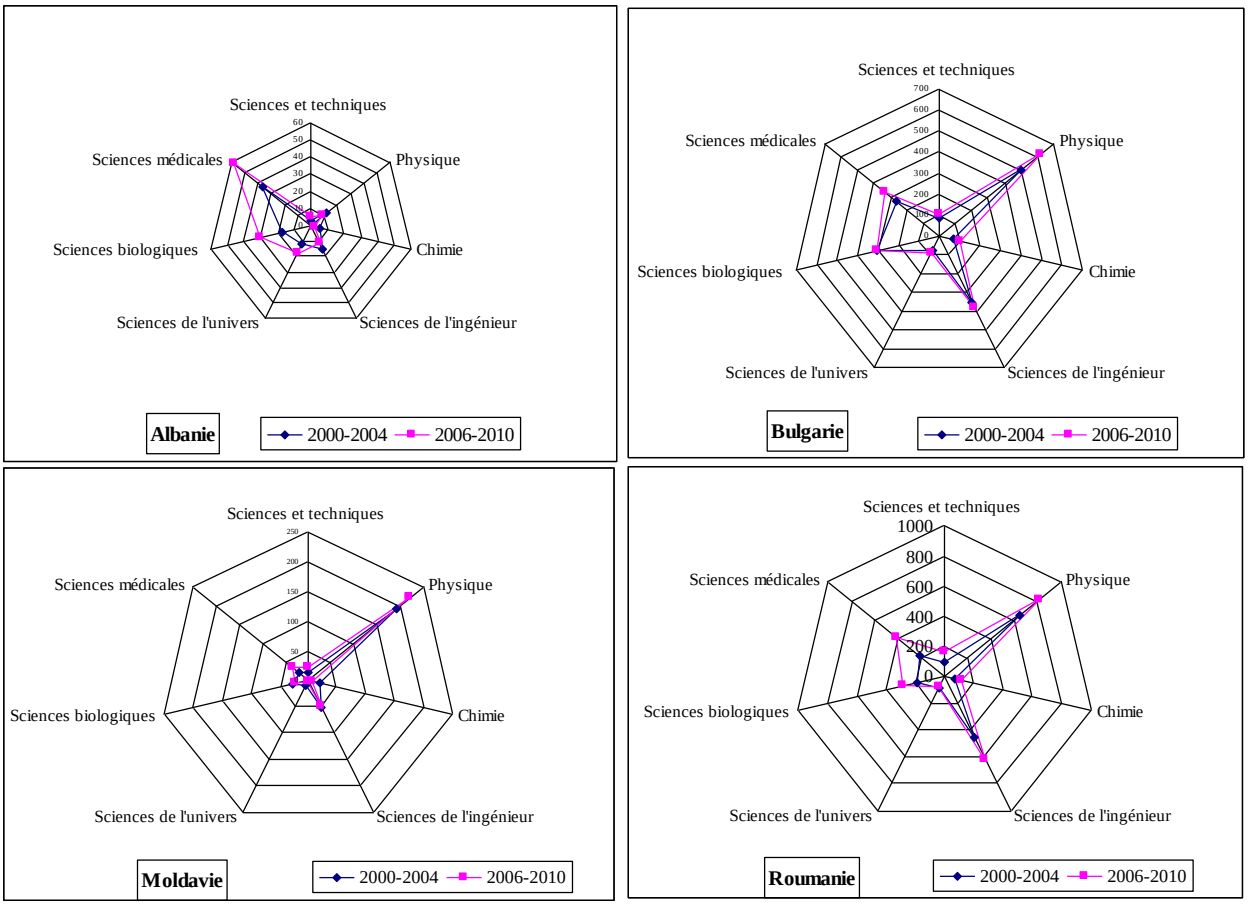

Figure 9. *Distribution de la production par discipline scientifique et par pays.*

Dans les deux graphes ci-dessous sont consignés, pour les deux corpus étudiés (figure 10 : période 2000 – 2004 et figure 11 : période 2006 – 2010) les interactions inter pays dans la co-publication.

La lecture de ces deux graphes, nous impose des constantes majeures :

- La Roumanie et la Bulgarie sont les deux pays moteurs de la recherche scientifique pour la région considérée,
- L'Albanie et la Moldavie ne collaborent peu voir pas du tout avec la Roumanie et la Bulgarie, et que cette coopération se ferait indirectement au travers de pays tiers comme par exemple la Grèce ou l'Australie.
- Bien que nous nous intéressions à une zone géographique dite francophone, nous remarquons que la coopération scientifique avec des pays francophone comme la France, la Belgique est faiblement active,
- De même, nous pouvons nous interroger sur le réseau de chercheurs particulier et personnel que possèdent la Roumanie et la Bulgarie qui a tendance à s'étoffer avec le temps (voir figure 11).

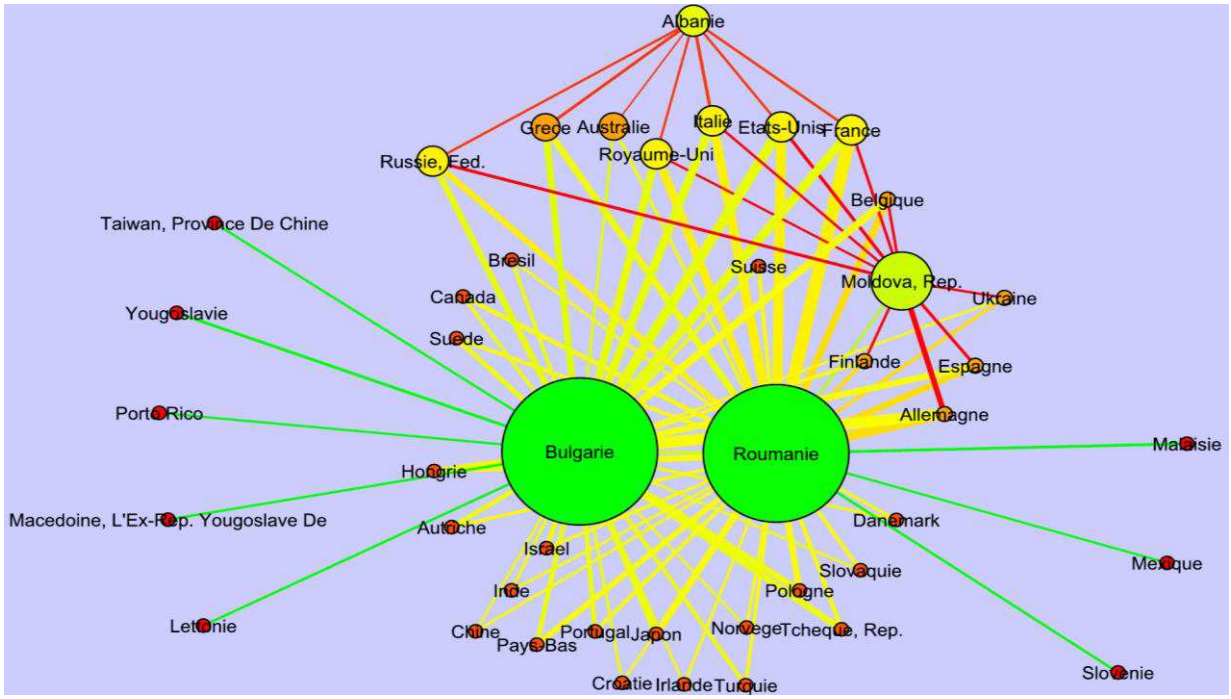

Figure10. *Graphe de la co-publication pour la période 2000 – 2004.*

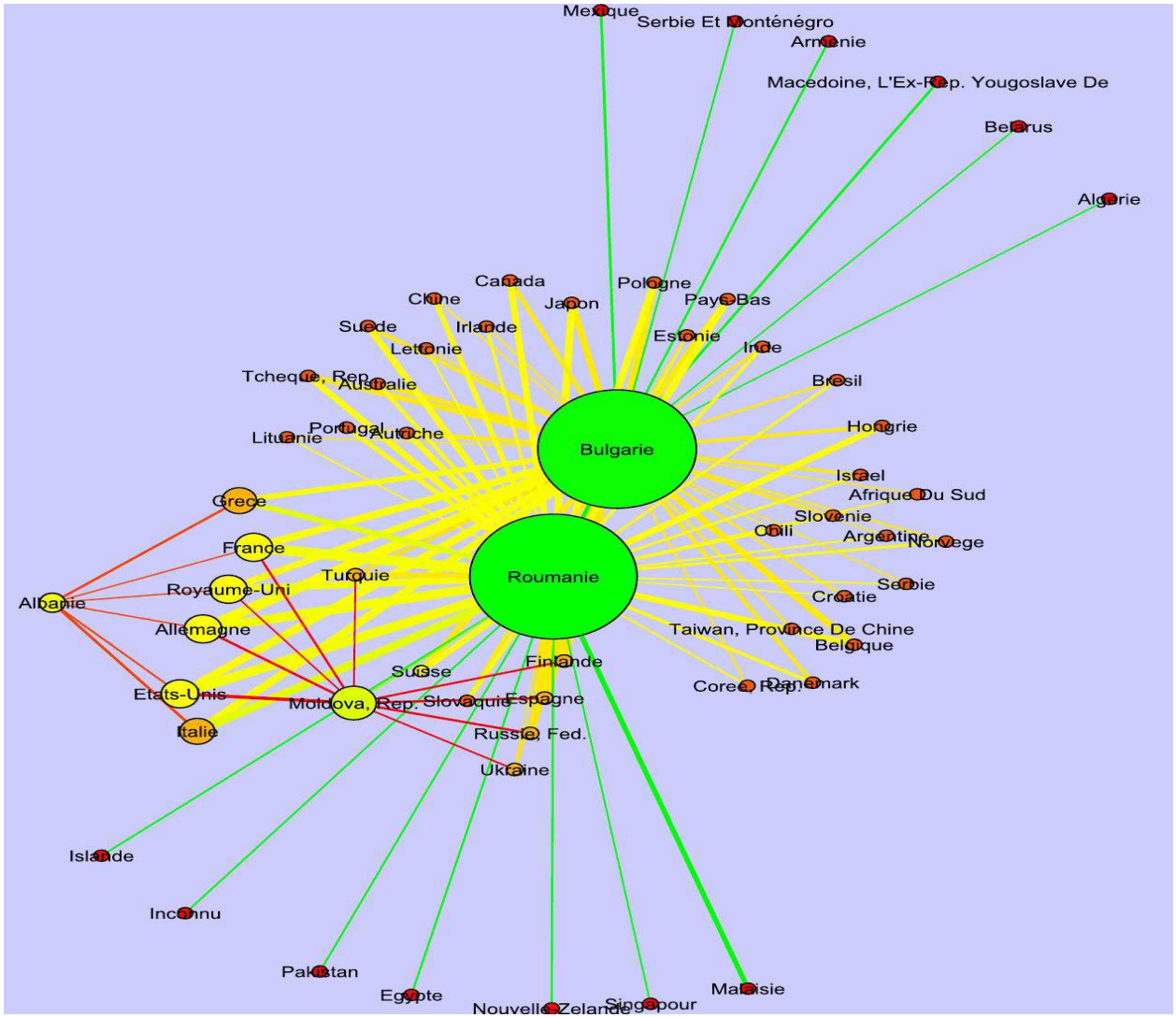

Figure 11. *Graphe de la co-publication pour la période 2006 – 2010.*

Pour conclure cette partie, nous pouvons dire que notre étude bibliométrique réalisée sur deux corpus de près 10 000 notices bibliographiques appartenant aux différents domaines présents dans la base bibliographique PASCAL, nous démontre le rôle essentiel joué par les pays francophones d'Europe centrale dans la production scientifique. Nous pouvons noter l'action centrale jouée par la Bulgarie et la Roumanie.

Les chercheurs de cette région se tournent volontiers vers les Etats-Unis et le Royaume-Uni comme pays d'édition, ainsi que vers les Pays-Bas. Nous pouvons noter que si la coopération scientifique avec le reste du monde n'est pas négligeable, la coopération avec d'autre pays francophones est tangible. Il en est de même pour la langue de publication, ou l'anglais est utilisé majoritairement.

3.3.3. Résultat de classification diachronique

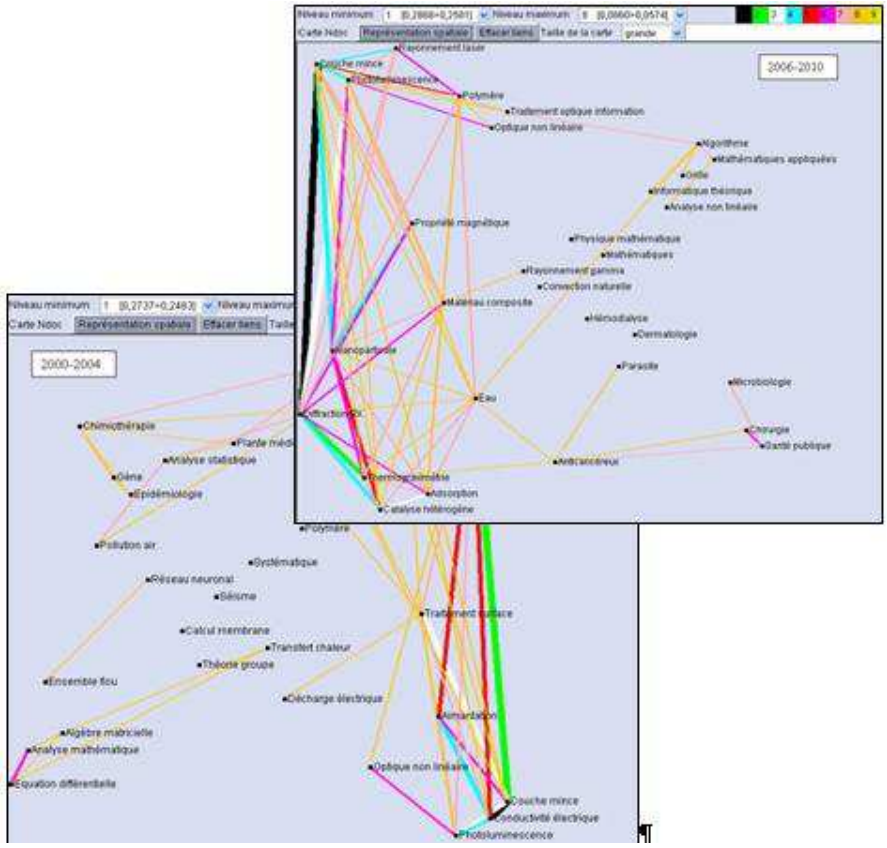

Figure 12. *Représentations cartographiques des clustering obtenus pour les deux périodes étudiées avec 30 classes.*

Les résultats de ce clustering à 30 classes permettent de mettre en évidence un certain nombre de thématiques présentent dans les corpus ainsi que leurs liens plus ou moins important. Toutefois, le corpus étant polythématique, les classes obtenues restent trop généralistes pour que l'analyse diachronique par règles d'associations floues soit pertinente. Pour cela, nous avons fait des classifications à 100 classes (que nous ne présenterons pas ici par soucis de lisibilité).

L'application des règles d'associations floues permet de détecter un nombre non négligeable de modifications. Nous ne présenterons ici que quelques exemples de changements qui nous semblent bien illustrer les évolutions entre les deux périodes.

Parmi les classes de la première période qui disparaissent en seconde période on remarque, entre autres, les classes « Pâte papier » et « Electromyographie ».

Parmi celles qui apparaissent en deuxième période on note « Traitement optique de l'information », « Réacteur nucléaire », « Urologie », « Cosmologie ».

Si on associe les classes découvertes par le clustering à des thématiques de recherches, nous pouvons en conclure que les premières thématiques ne sont plus vraiment étudiées, mais que par contre de nouveaux domaines de recherches sont apparus, montrant ainsi que le tissus scientifique est en plein évolution dans les pays étudiés.

Nous pouvons aller au-delà de ces remarques et mettre en évidence une évolution plus fine comme des regroupements ou des éclatements de thématiques :

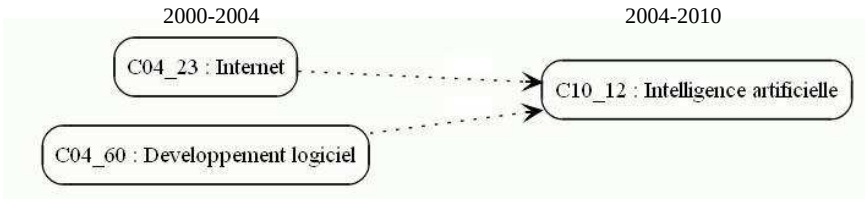

Figure 13. *Un exemple de fusion de classes (thématiques)*

La figure 13 montre les classes de la première période « Internet » et « Développement logiciel » fusionner pour donner la classe « Intelligence artificielle »

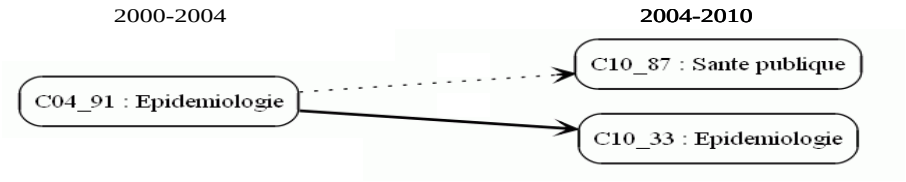

Figure 14. *Un exemple d'éclatement de classes (thématiques)*

La figure 14 est un exemple d'éclatement : la classe de la première période « Epidémiologie » éclate pour donner deux classes plus précises (« Santé Publique » et « Epidémiologie ».

4. Conclusion

Cette étude de la production scientifique de l'Albanie, Bulgarie, Moldavie et Roumanie, correspondant à l'espace francophone de l'Europe centrale, s'appuie sur l'analyse d'indicateurs et de graphes produits à partie de la base de données PASCAL. Ces indicateurs bibliométriques ne mesurent pas l'ensemble de la production scientifique, car tous les domaines scientifiques (comme par exemple les sciences vétérinaires) ne sont pas tous analysés dans cette base bibliographique. Cependant, ces indicateurs reflètent les tendances dans les courants dominants de la science. Nous pouvons tirer quelques conclusions schématiques des paramètres observés à la faveur des deux parties de notre étude, la premières étant consacrée à une étude bibliométrique tandis que la seconde à l'analyse de graphes, sachant que toutes les deux étaient issues d'une analyse scientométrique diachronique.

Les résultats de notre étude n'ont pas l'ambition de donner des éléments de décision pour la recherche dans cette région d'Europe centrale mais plutôt de proposer un constat de celle ci. Par conséquent, nous pouvons noter l'action centrale jouée par la Bulgarie et la Roumanie dans la production scientifique des pays francophones d'Europe centrale. Cette recherche

scientifique bien que orientée vers la physique et les sciences appliquées, n'est pas figée nous montrant ainsi que le tissus scientifique est en pleine évolution dans les pays étudiés.

Nous avons vu que la plateforme Stanalyst est un outil à la fois d'interrogation des bases de données et d'analyse des corpus obtenus (statistiques descriptives et classification). Stanalyst permet l'interrogation des bases de données de l'INIST mais il pourrait également être adapté pour travailler sur d'autres bases de données bibliographiques. Dans le cadre d'une mutualisation des données cet outil permettrait à tout un chacun une interrogation simple mais surtout une analyse fine des résultats pouvant aider à la fois le documentaliste, le chercheur, ou encore le décideur.

Bibliographie

1. Cadot M., Napoli A. (2004), Règles d'association et codage flou des données, 11èmes Rencontres de la Société Francophone de Classification, SFC 2004 (Bordeaux), pp. 130-133.
2. Callon M., Courtial J-P., Turner W. A., Bauin S. (1983) : From Translation to Problematic Networks: An Introduction to Co-Word Analysis, in Social Science Information, vol. 22, pp. 191-235.
3. Cuxac P., Cadot M., Francois C.: Analyse comparative de classifications : apport des règles d'association floues. Revue des Nouvelles Technologies de l'Information RNTI-E-3 Vol. II, Pp.519-530 ; 5èmes journées d'Extraction et Gestion de Connaissances EGC, Paris, Université Paris 5, 19-21 janvier 2005
4. Courtial, J.-P. (1990). Introduction a la Scientometrie, De la bibliometrie a la veille tehnologique. Paris: Anthropos-Economica
5. Dusoulier, N. (1993), « l'INIST, au cœur de la recherche et de l'Europe », Documentaliste- Sciences de l'Information, No 30-1, pp. 19-22.
6. Erdős, P. ; Rényi, A. (1959). Sur les graphiques aléatoires. In Publicationes Mathematicae 6: 290–297.
7. Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
8. Grivel L., François C. (1995) : Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique", Solaris n° 2, 1995, p. 81-112 ; http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2grivel.html
9. Guichard, M. (1999) : L'INIST-CNRS: des services d'information payants et gratuits dans un cadre de service public, Bulletin d'Informations ABF, No.184-5, pp.100-2.
10. Kodratoff Y. (2001), Rating the Interest of Rules Induced from Data and within texts, 12th IEEE -International Conference on Database and Expert Systems Aplications-Dexa 2001, Munich, sept 2001, pp. 265-269.
11. Lelu A. (1993) : Modèles neuronaux pour l'analyse de données documentaires et textuelles, Thèse de doctorat de l'université de Paris VI, 4 mars 1993, 238 pages.
12. Ménillet, D. (1992) : Grilles d'indexation et de préindexation. L'exemple de PASCAL , Documentaliste-Sciences de l'Information, No. 29-4/5, pp.183-190.
13. Repanovici, A. (2010) : Mesure de la visibilité de la production scientifique de l'Université à l'aide de Google Scholar, du logiciel "Publish or Perish" et des methodes de la scientométrie ; www.ifla.org/files/hq/papers/ifla76/155-repanovici-fr.pdf
14. Thèves, J, Séchet P. (2004) : Les systèmes nationaux de recherche et d'innovation du monde et leurs relations avec la France. Eléments de rétrospective, situation actuelle et futurs possibles. Estonie, Lettonie, Lituanie, Slovaquie, Slovénie, Bulgarie, Roumanie ; http://www.obs-ost.fr/fileadmin/medias/tx_ostdocuments/PECOversionOct2005-4_01.pdf