LES EMERGENCES TECHNOLOGIQUES DANS LE DOMAINE DES DISPOSITIFS OPTOELECTRONIQUES : IDENTIFICATION ET CARACTERISATION

Dominique BESAGNI (1), Claire FRANCOIS (1), Marianne HÖRLESBERGER (2), Ivana ROCHE (1), Edgar SCHIEBEL (2)

dominique.besagni@inist.fr, claire.francois@inist.fr, ivana.roche@inist.fr, edgar.schiebel@arcs.ac.at, marianne.hoerlesberger@arcs.ac.at

(1) INIST - CNRS, 2 allée du Parc de Brabois, CS 10310, 54519 Vandoeuvre-les-Nancy cedex, France

(2) Austrian Research Centers GmbH, Tech Gate Vienna, Donau-City-Straße 1, 1220 Vienne, Autriche

---

Mots-clés : émergences technologiques, bibliométrie, approche diachronique, classification automatique, modèle diffusion, dispositifs optoélectroniques

Keywords : technological emergences, bibliometrics, diachronic approach, clustering, diffusion model, optoelectronic devices

Palabras clave : emergencias tecnológicas, bibliometría, perspectiva diacrónica, clasificación automática, modelo difusión, dispositivos electrónicos

Résumé : Partant d'un ensemble de données issues de la littérature scientifique sur les dispositifs optoélectroniques, nous nous sommes intéressés à l'identification et à la caractérisation de thématiques émergeantes dans ce domaine technologique. Ce travail s'inscrit dans la suite d'un projet européen dont le principal résultat a été de produire une méthodologie permettant d'identifier des technologies émergeantes et prometteuses à partir de l'analyse de la littérature scientifique internationale. Dans cet article, nous avons combiné deux approches analytiques : la première met en œuvre une analyse diachronique des résultats de classifications et la seconde utilise une modélisation du processus d'évolution terminologique d'un domaine fondée sur des indicateurs bibliométriques : le modèle de diffusion. Notre objectif est d'apporter des réponses aux questions suivantes : Dans le domaine technologique considéré, quelles thématiques pouvons-nous détecter ? Parmi ces thématiques, lesquelles sont déjà consolidées et lesquelles sont nouvelles ? Peut-on déceler des liens entre les thématiques repérées ?

1 Introduction

Les technologies émergeantes jouent un rôle essentiel dans les avancées scientifiques, industrielles et sociétales. Plusieurs analyses économiques montrent que l'innovation technologique contribue considérablement à la croissance économique et cet effet n'a cessé d'augmenter ces dernières décennies. Incontestablement, la connaissance précoce de produits et procédés novateurs et alternatifs est une nécessité stratégique qui joue un rôle prépondérant dans leur processus d'évaluation et contribue à une prise de décision à bon escient répondant à de réels besoins. La détection de technologies émergeantes demeure néanmoins un vrai problème, et donc fait l'objet d'études dans un large spectre de domaines allant du marketing à la bibliométrie.

L'arbre de sélection proposée par Armstrong & Green [1] donne une bonne image de l'ensemble de méthodes de prévision qui peuvent être appliquées, en particulier, pour la détection de ces technologies émergeantes. Elle illustre très bien la dichotomie entre les méthodes quantitatives et celles fondées sur l'expertise et montre la grande diversité des approches existantes : des méthodes Delphi ou Nominal Group Technique, fondées sur la confrontation d'avis d'experts, des méthodes de scénarios qui visent à balayer les différents futurs possibles, jusqu'aux méthodes combinant la connaissance des experts sur le domaine et des techniques statistiques permettant l'identification de facteurs de causalité agissant sur les tendances.

Lorsque le volume des données collectées est suffisant pour justifier l'application de méthodes quantitatives, une question importante demeure concernant le type de données employées dans le processus de prévision. Les bases de données de littérature scientifique et de brevets sont les sources les plus souvent utilisées dans la détection de nouvelles thématiques par des méthodes bibliométriques, appliquant des techniques statistiques relativement simples comme les courbes de croissance ou plus sophistiquées comme la classification automatique ou l'analyse de réseaux ([2] ; [3] ; [4] ; [5]). L'introduction d'un troisième type de données relatives aux financements nationaux de projets est très intéressante et permet, si les trois sources d'information sont analysées ensemble et non séparément, de mieux comprendre les interfaces de type triple-hélice engendrées par les relations qui se tissent naturellement entre l'université, l'industrie et les gouvernements ([6] ; [7]).

Ce travail fait suite à un projet financé par la Commission Européenne et dont le principal résultat a été de produire une méthodologie pour l'identification de technologies émergeantes en partant d'une hypothèse de base, fondée sur un constat : les technologies émergentes actuelles trouvent de plus en plus leur fondement dans les connaissances développées en Physique. L'influence grandissante de la Physique dans les domaines technologiques peut s'expliquer par l'augmentation considérable de la complexité des technologies d'aujourd'hui. Il n'est, en effet, pas difficile de trouver des exemples montrant les nombreux liens existants entre les plus récentes découvertes de la recherche en Physique et les technologies de pointe dans les Sciences Appliquées et les Sciences de la Vie.

Dans ce projet, l'étude de l'intersection entre, d'une part, la Physique et, d'autre part, les Sciences Appliquées et les Sciences de la Vie, a produit 45 domaines technologiques candidats, potentiellement porteurs d'innovation. Leur validation, par des panels d'experts scientifiques, a permis d'aboutir à une sélection finale des dix domaines technologiques les plus prometteurs.

Cet article présente le développement de méthodologies pour la détection de nouvelles thématiques et leur application dans un de ces domaines technologiques. Nous avons choisi celui des Dispositifs optoélectroniques car il est un des plus prometteurs de la dernière décennie. Et il faudra certainement s'attendre à ce que les développements réalisés dans ce domaine technologie aient, dans un futur proche, des impacts sociétaux forts, tout particulièrement sur la qualité de la vie, la sécurité, l'énergie, la santé ou encore les transports. Les applications des LED dans l'éclairage urbain et d'habitation ou dans l'industrie automobile sont en effet chaque fois plus nombreuses. De même, les diodes électroluminiscentes organiques, les diodes électroluminiscentes dans l'ultraviolet profond et les écrans organiques deviennent courants dans les dispositifs électroniques mais aussi dans des appareils à la destination du grand public.

Pour identifier les thématiques émergeantes ainsi que leur degré de diffusion, nous réalisons une analyse diachronique en divisant les données collectées en deux corpus, correspondant à deux périodes de temps successives, et en y appliquant deux méthodologies :

- le « modèle de diffusion », correspondant à une approche TF IDF étendue [8], qui procède à la distribution des mots-clés présents dans les corpus par degré de diffusion et produit une modélisation de l'évolution terminologique du domaine technologique et
- l'analyse des résultats de classifications, en utilisant les outils mis en œuvre dans la station d'analyse de l'information Stanalyst®[9].

Après la description de l'acquisition de données, nous présenterons les méthodologies développées, suivies des résultats de leur application dans le domaine des Dispositifs optoélectroniques. Finalement, nous commenterons les résultats obtenus et, en nous fondant sur eux, nous essayerons d'établir des convergences entre les deux approches analytiques mises en œuvre.

2 Acquisition des données

L'information nécessaire à notre étude se retrouve soit dans les publications scientifiques, soit dans les brevets. Pour appliquer une méthodologie de type bibliométrique, il est en effet nécessaire d'utiliser une base de données documentaire où l'information est structurée et organisée par l'intermédiaire d'une indexation ou d'un plan de classement de l'ensemble des textes. De plus, cette base doit être suffisamment multidisciplinaire pour couvrir les domaines de la Physique, des Sciences Appliquées et des Sciences de la Vie.

La base PASCAL de l'INIST réunit à ce jour environ 18 millions de notices bibliographiques obtenues par l'analyse de la littérature scientifique et technique internationale publiée notamment dans les périodiques et les actes de congrès. Cette source d'information s'avère être tout particulièrement adaptée aux objectifs de notre travail car :

- sa multidisciplinarité permet d'accéder à la fois aux domaines de la Physique et à ceux des applications technologiques ;
- la finesse de son plan de classement offre la possibilité d'analyser des domaines très spécialisés ;
- ses codes de classement multiples permettent de calculer des passerelles entre la Physique et ses applications technologiques. En effet, dans PASCAL, chaque notice bibliographique est affectée d'un ou plusieurs codes. Ces codes de classement sont organisés dans un plan de classement qui se présente comme une taxonomie de chaque domaine et sous-domaine pour l'ensemble des disciplines couvertes par la base. Après analyse du plan de classement de PASCAL, nous avons décidé d'employer une stratégie de recherche simple, consistant en la sélection des notices bibliographiques ayant à la fois un code en Physique, quelque soit la spécialité, et un code correspondant à un domaine d'application technologique ;
- son indexation rend possible la détermination d'une terminologie à partir d'un corpus où chaque notice bibliographique bénéficie d'une indexation par mots-clés manuelle ou automatique. Après validation par un expert scientifique, cette terminologie peut être employée dans notre analyse.

3 Méthodologie

Dans cette section, nous introduisons les deux méthodes employées pour le suivi de l'évolution du domaine technologique, décrites plus en détail dans un précédent article [10].

Le modèle de diffusion est utilisé pour évaluer le statut de chaque terme dans le domaine technologique étudié, en le comparant à son statut dans l'ensemble des domaines technologiques considérés où il est également présent. Un degré d'émergence est alors calculé.

L'analyse diachronique des résultats de classification s'intéresse au contenu du domaine technologique étudié en appliquant une méthode de classification qui va permettre d'organiser les données en thématiques et d'analyser les liens entre ces dernières.

3.1 Le modèle de diffusion

Le modèle de diffusion se fonde sur l'hypothèse que les nouveaux résultats dans un domaine de recherche sont publiés dans des articles. Les mots-clés décrivant ces découvertes plus ou moins fortuites des chercheurs commencent à apparaître dans le domaine, dans un premier stade, sous une forme inhabituelle. Dans un deuxième stade, la recherche sur le sujet s'intensifie et les mots-clés prennent de l'importance en devenant plus fréquents dans le domaine. Finalement, dans un troisième stade, les résultats des recherches conduites sur le sujet vont apparaître dans d'autres domaines de recherche. Par conséquent, la détermination du stade de diffusion se décline au niveau des mots-clés par le calcul, pour chacun, d'un degré de diffusion, respectivement, inhabituel, établi ou transfert. Dans un précédant article [8], un filtrage bibliométrique unique avait été utilisé pour procéder à la distribution des termes de la totalité du corpus par degré de diffusion. Dans cette étude, nous y introduisons le facteur temps en divisant le corpus en deux périodes et en formalisant les chemins migratoires possibles d'un mot-clé, de la première vers la seconde période, comme montré dans le tableau 1.

Tableau 1. Définition des chemins de passage des mots-clés d'une période à l'autre selon leur degré de diffusion à chaque période

Périodes	**Modèle de diffusion**		
	Stade 1 Termes Inhabituels	Stade 2 Termes Etablis	Stade 3 Termes de Transfert
période 1 (P1)	1	3 4 5	6
période 2 (P2)	2		

Dans cet article, nous proposons deux approches de sélection des mots-clés sur lesquels nous appliquons les indicateurs bibliométriques décrits dans [8] et [10] en vue du calcul du degré de diffusion des mots-clés associés à un domaine technologique. L'approche HT (Home Technology) introduit la notion de termes « Home Technology » où chaque mot-clé du domaine étudié est positionné par rapport à l'univers constitué par l'ensemble des termes des domaines technologiques considérés. Il s'agit d'obtenir l'affectation de chaque mot-clé à un seul domaine : on dira alors qu'il est un terme « Home Technology » de ce domaine. L'approche TFIDF (Text Frequency Inverse Document Frequency) se focalise sur le domaine technologique étudié et procède à un classement par

rang de l'ensemble des mots-clés du domaine. Chaque mot-clé est positionné en considérant uniquement son voisinage terminologique dans le domaine étudié.

Après partage du corpus initial en deux périodes, P1 et P2, nous avons, dans chacun des corpus obtenus, identifié deux ensembles de mots-clés du domaine technologique. Une procédure de sélection, décrite ci-dessous, est alors opérée sur ces ensembles de termes. Après une étape permettant de singulariser les termes les plus fréquents (étape 1), et avant la sélection des mots-clés de plus forte fréquence relative par le calcul du TFIDF (étape 3), on applique ou non l'étape 2 suivant que l'on veuille ou non sélectionner les mots-clés « Home Technology » :

1. Sélection des mots-clés avec une fréquence supérieure à un seuil fixé

2. Sélection des mots-clés « Home Technology » du domaine technologique en appliquant l'approche HT :

   a- Soit j l'index des domaines technologiques déterminés dans une étape préalable, $j \in [1,J]$ avec J = nombre total de domaines considérés

   b- Soit i l'index des mots-clés d'un domaine, $i = 1,...,k_j$ avec k_j = nombre de mots-clés différents dans le domaine j

   c- Soient a_j = nombre total de références bibliographiques dans le corpus associé au domaine j

   et a_{ij} = nombre de références bibliographiques dans le corpus du domaine j avec présence du mot-clé i, avec : $1 \le a_{ij} \le a_j$

   d- Alors $p_{ij} = \frac{a_{ij}}{a_j}$ est la probabilité avec laquelle le mot-clé i apparait dans le domaine j

   e- Si, pour chaque mot-clé du domaine étudié, on calcule p_{ij} en considérant $\forall j = 1, J$ nous obtenons J valeurs de la probabilité p_{ij}

   f- Sélection des termes dont la probabilité p_{ij} est maximale dans le domaine étudié : $\max_i \left(p_{ij} \right), j = 1, J$

3. Sélection des mots-clés du domaine technologique après leur classement par rang en appliquant l'approche TFIDF :

   a- Soit j l'index des domaines technologiques, $j \in [1,J]$ avec J = nombre total de domaines considérés

   b- Soit i l'index des mots-clés d'un domaine, $i = 1,...,kj$ avec kj = nombre de mots-clés différents dans le domaine j

   c- Soient g_{ij} le poids local associé au mot-clé i dans le domaine j

   et t_i = nombre de domaines où le mot-clé i apparait dans au moins une référence bibliographique, avec : $1 \le t_i \le J$

   d- Alors $TFIDFTech_{ij} = \frac{g_{ij}}{t_i}$ est la valeur associée au mot-clé i dans le domaine j

   e- Sélection des mots-clés dont $TFIDFTech_{ij} > seuil_{TFIDF}$, ce seuil étant fixé pour produire une liste d'environ 300 mots-clés

Enfin, une étape finale de catégorisation des mots-clés par degré de diffusion est réalisée dans chacune des deux listes de mots-clés obtenues en considérant l'indice Gini et l'indicateur RTF. Aussi :

1. Un terme avec un indice Gini < $seuil_{Gini}$ est un terme de Transfert
2. Un terme avec un indice Gini ≥ $seuil_{Gini}$ et
   - une fréquence relative ≥ $seuil_{RTF}$ est un terme Etabli
   - une fréquence relative < $seuil_{RTF}$ est un terme Inhabituel

3.2 L'analyse diachronique par classification

L'analyse diachronique est fondée sur l'application d'une méthode de classification automatique sur des données associées à deux, ou plus, périodes de temps successives et sur l'étude de l'évolution des cartes et des contenus des classes obtenues. Dans cet article, nous avons procédé selon les étapes suivantes :

1. Partage du corpus initial en deux périodes, P1 et P2

2. Application d'une méthode de classification automatique sur les corpus obtenus pour chaque période, dans lesquels les documents sont représentés par les mots-clés présents dans les références bibliographiques. L'outil de classification et de cartographie utilisé ([11], [12]) emploie un algorithme de classification non hiérarchique, non supervisée mettant en œuvre la méthode des K-means axiales suivie d'une analyse en composantes principales

3. Analyse de l'évolution des deux cartes des classes et des deux ensembles de classes en examinant le vocabulaire associé aux classes de chaque période, en utilisant une matrice de comparaison

4. Validation par des experts scientifiques des hypothèses émises à partir de cette analyse.

4 Résultats

L'interrogation de la base PASCAL, sur la période 1996-2003, a produit 3871 références bibliographiques ayant trait simultanément aux domaines des Dispositifs optoélectroniques, d'une part, et de la Physique, d'autre part. Son découpage en deux périodes, 1996-1999 et 2000-2003, a permis d'obtenir deux corpus constitués par, respectivement, 1797 et 2074 notices bibliographiques qui serviront de base à notre analyse.

L'identification de l'ensemble des mots-clés du domaine technologique à partir de ces deux corpus a produit, respectivement, 2345 et 2738 termes.

4.1 Le modèle de diffusion

Les filtrages successifs, mis en œuvre dans le modèle de diffusion, ont permis, pour chaque période, de sélectionner deux listes de termes selon les approches décrites dans la section 3.1 et de catégoriser les termes obtenus par degré de diffusion. L'approche HT a produit pour P1 et P2, respectivement, 92 et 96 termes. L'approche TFIDF a sélectionné globalement trois fois plus de termes, respectivement, 289 pour P1 et 273 pour P2. Les résultats des distributions des termes par degré de diffusion sont présentés dans les tableaux 2 et 3. Leur lecture permet d'observer que, pour les deux périodes, les catégorisations opérées

ont produit une distribution équivalente du nombre de mots-clés dans les trois degrés de diffusion. Leur lecture permet d'observer que, pour les deux périodes, les catégorisations opérées ont produit une distribution équivalente du nombre de mots-clés dans les trois stades de diffusion. Cela est dû au fait que les seuils introduits par l'étape de

Tableau 2. Distribution par période et par degré de diffusion des termes sélectionnés par l'approche HT

Périodes	**Termes inhabituels**		**Termes établis**		**Termes de transfert**		**Nombre de termes HT sélectionnés**	**Nombre total de termes du domaine**
	Nombre	%	Nombre	%	Nombre	%		
P1 (1996-1999)	35	38%	28	30%	29	32%	92	2345
P2 (2000-2003)	30	31%	37	39%	29	30%	96	2738

catégorisation ont été fixés d'une manière pragmatique et ne permettent pas de situer avec certitude les frontières entre les stades de diffusion. En effet, il s'agit d'une approche floue où l'on admet qu'un ensemble de N mots-clés se partage en trois tiers entre Inhabituels, Etablis et de Transfert. Avec cette seule information, il n'est pas possible de déterminer, d'un point de vue méthodologique, si la diffusion des termes se fait vraiment. L'identification des changements de degré de diffusion de mots-clés entre les périodes P1 et P2 peut apporter une réponse à cette question. Les stades de diffusion définis dans les deux périodes créent un espace où il est possible d'identifier ces mots-clés migrants. La prochaine étape montrera la diffusion de mots-clés selon les chemins migratoires possibles présentés dans le tableau 1.

Tableau 3. Distribution par période et par degré de diffusion des termes sélectionnés par l'approche TFIDF

Périodes	**Termes inhabituels**		**Termes établis**		**Termes de transfert**		**Nombre de termes TFIDF sélectionnés**	**Nombre total de termes du domaine**
	Nombre	%	Nombre	%	Nombre	%		
P1 (1996-1999)	101	35%	96	33%	92	32%	289	2345
P2 (2000-2003)	96	35%	82	30%	95	35%	273	2738

Les taux de diffusion des mots-clés entre les deux périodes de temps (cf. tableau 1) sont présentés dans le tableau 4 pour les deux approches de sélection. La dynamique de la diffusion des termes dans le temps a été examinée, d'une part, dans le cadre strict du domaine technologique étudié (sélection HT) et, d'autre part, en considérant les autres domaines technologiques (sélection TFIDF).

Les chemins 1, 4 et 6 recensent les termes demeurant dans le même stade, ce qui traduit leur faible dynamique. Concernant le stade 1 (Inhabituel), le fait que seulement moins de 10% de ses termes dans la période P1 se retrouvent dans la période P2 dans ce même stade montre qu'il y a une très forte dynamique dans

la terminologie de ce premier stade. Cela était attendu car le premier stade est celui où l'on trouve des termes reflétant des fortes fluctuations dans les découvertes scientifiques et encore peu enclins à migrer vers les autres stades. Il faut remarquer une proportion légèrement plus importante de termes suivant le chemin migratoire 1 dans la sélection TFIDF.

Si l'on considère le chemin de passage 2 nous remarquons, dans la sélection HT, un taux de 20% correspondant au double de la valeur enregistrée dans la sélection TFIDF. Il s'agit d'un résultat intéressant car il reflète une plus forte dynamique dans les découvertes technologiques spécifiques au domaine étudié que dans les découvertes empruntées aux autres domaines technologiques. De plus, on peut remarquer que ce taux de 20% associé à la consolidation de nouvelles connaissances en une période de seulement quatre ans représente une fort honnête réussite.

Tableau 4. Chemins de diffusion entre les périodes P1 et P2 des termes sélectionnés par les approches HT et TFIDF

Chemins de diffusion P1 → P2	**Sélection HT**		**Sélection TFIDF**	
	Mots diffusés	*% de Nombre Termes P1*	*Mots diffusés*	*% de Nombre Termes P1*
Inhabituel→ Inhabituel	3	9%	13	13%
Inhabituel→ Etabli	7	20%	11	11%
Inhabituel→ Transfer	0	0%	2	2%
Etabli→ Etabli	14	50%	44	46%
Etabli→ Transfer	2	7%	14	15%
Transfer→ Transfer	15	52%	50	54%

Des découvertes très importantes qui se diffusent très rapidement dans le domaine technologique étudié mais qui également migrent vers d'autres domaines technologiques peuvent ne pas se trouver dans la sélection HT.

Près de la moitié des termes Etablis et Transfert dans la période P1 restent dans ce stade dans la seconde période. Ces termes reflètent le courant dominant des recherches menées dans le domaine technologique étudié. Le nombre de termes diffusant du stade 2 (Etabli) vers le stade 3 (Transfert) dans la sélection HT est petit (seulement 2 termes), mais il est autrement plus important (14 termes soit 15%) dans la sélection TFIDF. Méthodologiquement parlant cela signifie que dans cette dernière il y a plus de termes avec une probabilité d'occurrence dans une autre technologie plus grande que la probabilité d'occurrence dans le domaine technologique étudié. D'un point de vue technologique, ce sont les découvertes très importantes dans les autres technologies qui vont venir contribuer à la consolidation du statut de domaine de recherche fortement dynamique que l'on associe au domaine des dispositifs optoélectroniques.

Dans le domaine technologique étudié, la liste de mots-clés obtenue par l'approche HT est un sous-ensemble de la liste produite par l'approche TFIDF. Aussi est-il intéressant de connaître, pour chaque période de temps, la distribution par degré de diffusion des termes de la liste HT dans la liste TFIDF. Ces résultats sont présentés dans le tableau 5 et permettent de vérifier une relative stabilité sauf pour les termes de Transfert de l'approche HT qui, surtout en la période P1, se partagent quasi équitablement entre les termes Etablis et de Transfert de l'approche TFIDF.

Tableau 5. Distribution par degré de diffusion des termes de la sélection HT présents dans la sélection TFIDF

	P1			P2		
% présence termes HT dans TFIDF par degré de diffusion	Inhabituel	Etabli	Transfert	Inhabituel	Etabli	Transfert
Inhabituel	94%	6%	-	100%	-	-
Etabli	-	100%	-	-	100%	-
Transfert	-	45%	55%	-	17%	83%

D'un point méthodologique, les résultats présentés dans le tableau 5 s'expliquent par le choix pragmatique des valeurs des seuils utilisés dans l'étape de catégorisation. En effet, ces choix dans les deux approches de sélection (HT et TFIDF) ont été opérés indépendamment pour produire, dans les deux cas, une distribution de termes par degré de diffusion équilibrée. La sélection HT recense les termes avec les plus grandes probabilités d'occurrence dans les documents associés au domaine des dispositifs optoélectroniques. Les mots-clés de cette sélection apparaissant relativement souvent dans les autres domaines technologiques sont catégorisés dans le stade 3. La sélection TFIDF réunit tous les termes présents dans les documents associés au domaine des dispositifs optoélectroniques. Ceci signifie que parmi ces termes se trouvent aussi des termes ayant une probabilité d'occurrence dans d'autres domaines technologiques plus haute que la probabilité d'occurrence dans le domaine technologique étudié.

D'autre part, dans le tableau 6 sont présentés, pour chaque période de temps, les résultats de la distribution par degré de diffusion des termes de la liste TFIDF présents dans la liste obtenue par l'approche HT ainsi que les termes qui en sont absents. Cette absence signifie simplement que ces termes sont considérés des termes « Home Technologie » d'un autre domaine technologique que celui que nous étudions, parmi l'ensemble des domaines technologiques sélectionnés dans une étape exploratoire préalable. La perte de termes entre les deux approches de sélection est, tous degrés de diffusion confondus, très importante dans les deux périodes de temps : respectivement, 68% et 65%. La lecture du tableau 6 permet également d'observer que, pour les deux périodes, cette perte est distribuée de manière assez homogène sur tous les degrés de diffusion.

Tableau 6. Distribution par degré de diffusion des termes de la sélection TFIDF présents dans la sélection HT ou absents de cette dernière

	P1			P2		
Nombre de termes TFIDF dans HT par degré de diffusion	Inhabituel	Etabli	Transfert	Inhabituel	Etabli	Transfert
Inhabituel	33	0	0	30	0	0
Etabli	2	28	13	0	37	5
Transfert	0	0	16	0	0	24
Termes d'une autre Home Technologie	*65 (65%)*	*68 (71%)*	*63 (69%)*	*66 (69%)*	*45 (55%)*	*66 (70%)*

4.2 L'analyse diachronique des résultats de classification

L'approche diachronique appliquée a consisté en la réalisation d'une classification, pour chaque période de temps, à partir de chacune des sélections de termes opérées dans l'approche par modèle de diffusion (cf. section 4.1). Nous avons ainsi obtenu deux jeux de résultats que nous allons ensuite comparer. Dans le tableau 7, nous présentons les résultats des classifications en les positionnant par rapport aux données initiales en nombre de termes et en nombre de documents.

Nous pouvons remarquer que malgré le petit nombre de termes conservés dans la classification, surtout pour la sélection HT, le nombre de documents réunis dans les classes reste élevé, formant un corpus toujours représentatif du domaine technologique étudié.

Tableau 7. Résultats des classifications par période et par processus de sélection des termes (HT ou TFIDF)

	P1		**Nombre total**	**P2**		**Nombre total**
	HT	TFIDF		HT	TFIDF	
Nombre de termes dans les classes (fréquence > 2)	39 (5%)	130 (17%)	**771 termes**	44 (5%)	99 (11%)	**932 termes**
Nombre de documents dans les classes	1403 (78%)	1574 (88%)	**1797 documents**	1899 (91%)	2058 (99%)	**2074 documents**

L'analyse des résultats de classification se fait à partir des cartes obtenues selon la méthodologie décrite en section 3.2. Pour comparer ces cartes, on va commencer par la description de la carte obtenue avec la sélection de mots-clés la plus large, à savoir celle de l'approche TFIDF, et pour la période P1 (cf. figure 1). Sur cette carte, on note un arc important allant d'un sous-réseau traitant de l'optique (*Electroluminescent devices* et *Ligth emitting diodes*) vers un sous-réseau plus important, axé sur la photodétection (*Infrared detectors* et *Photodetectors*) liés par un ensemble de classes ayant trait aux matériaux (*Semiconductor materials* et *Elemental semiconductors*). On note également la relative proximité de la classe *Infrared imaging* avec le sous-réseau sur la photodétection. Pour cette même approche et pour la seconde période (cf. figure 2), on retrouve la même organisation générale avec les mêmes sous-réseaux. Cependant, celui des matériaux gagne en densité et celui de l'optique prend de l'importance. Dans ce dernier nous observons l'apparition de composés organiques (classes *Organic compounds* et *Organic light emitting diodes*). De plus, on voit se confirmer la proximité entre les aspects liés à l'imagerie et le sous-réseau des photodétecteurs.

Figure 1. Carte des classes pour la période P1 avec la sélection des termes par l'approche TFIDF

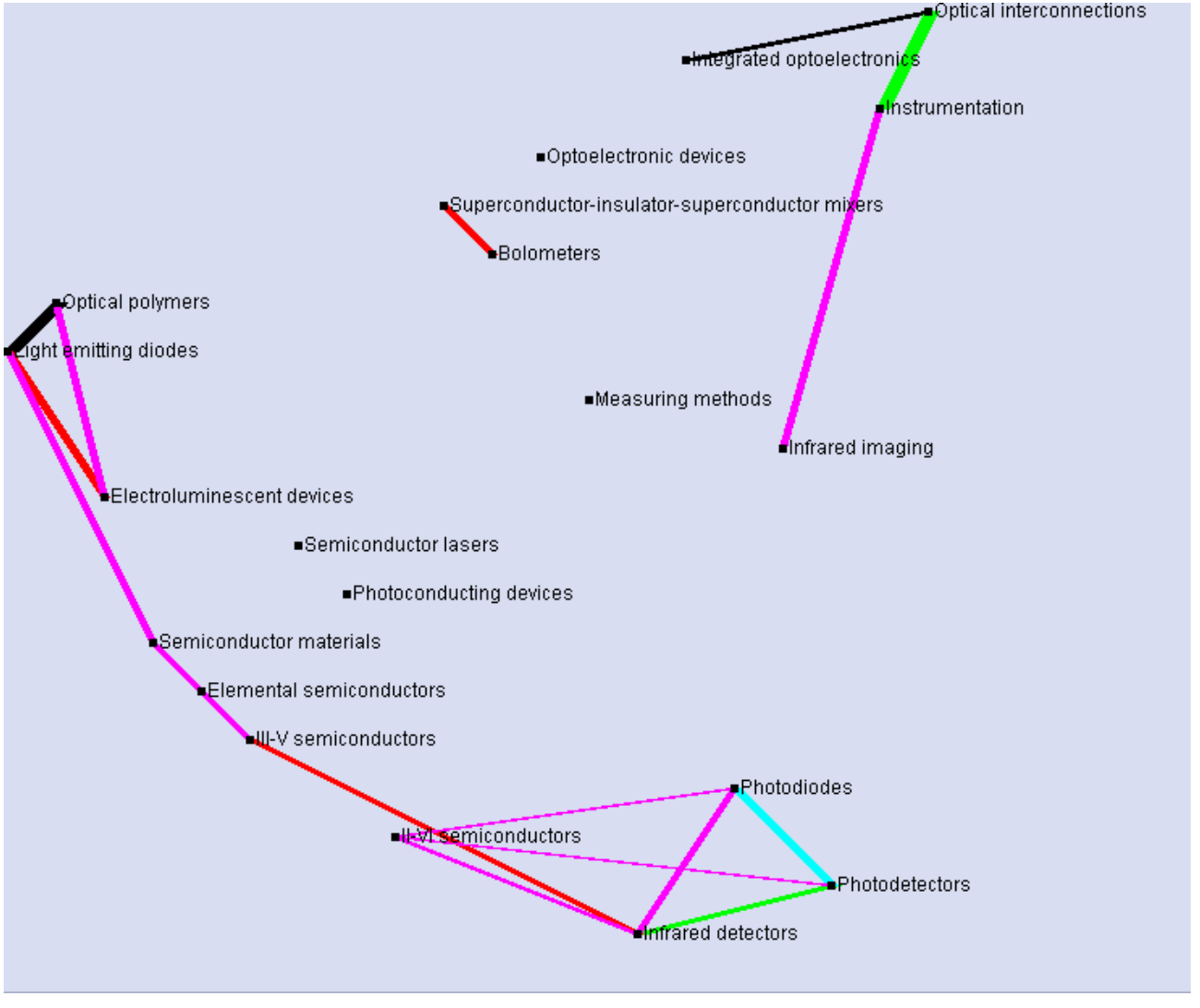


Figure 2. Carte des classes pour la période P2 avec la sélection des termes par l'approche TFIDF

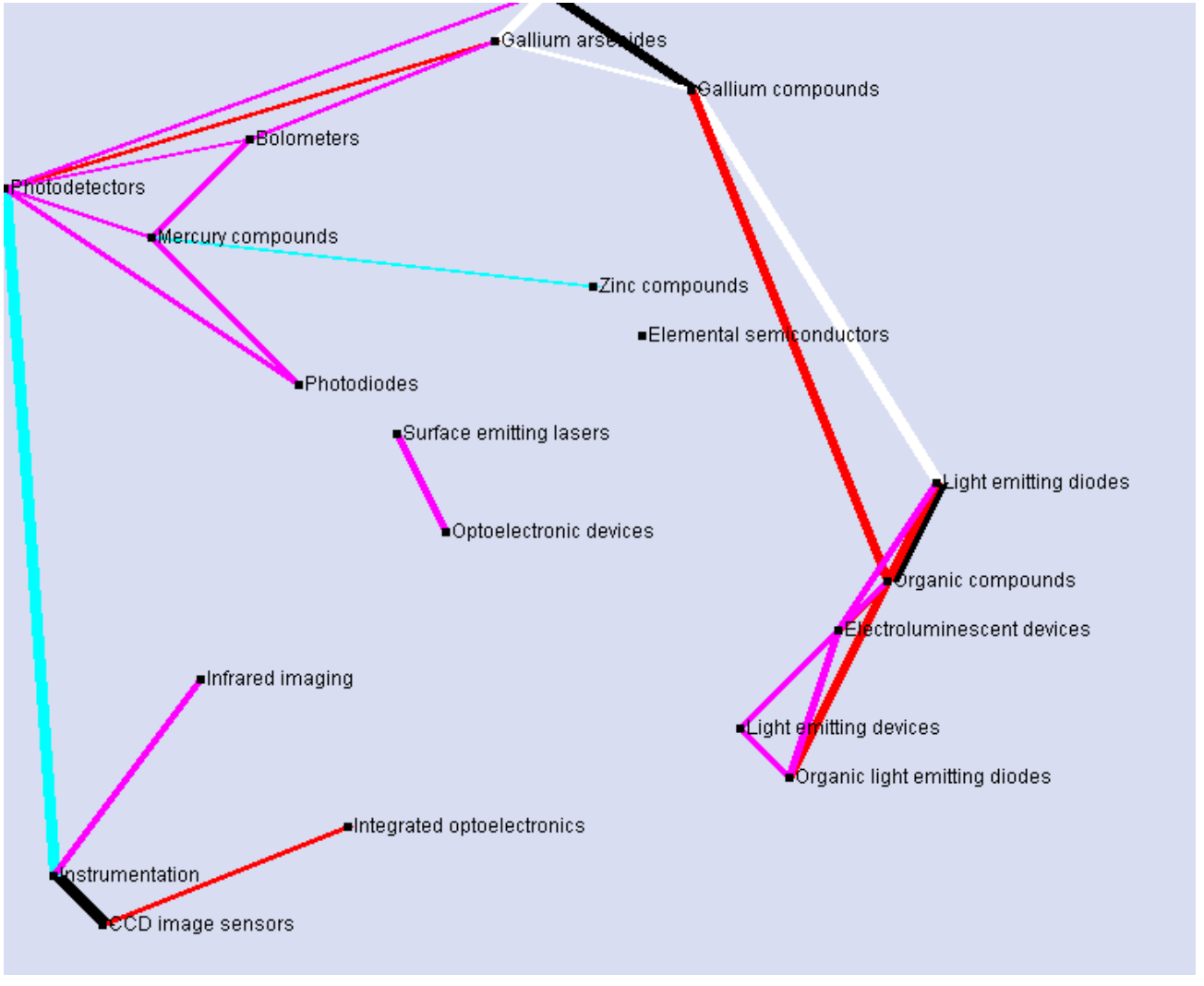


Figure 3. Carte des classes pour la période P1 avec la sélection des termes par l'approche HT

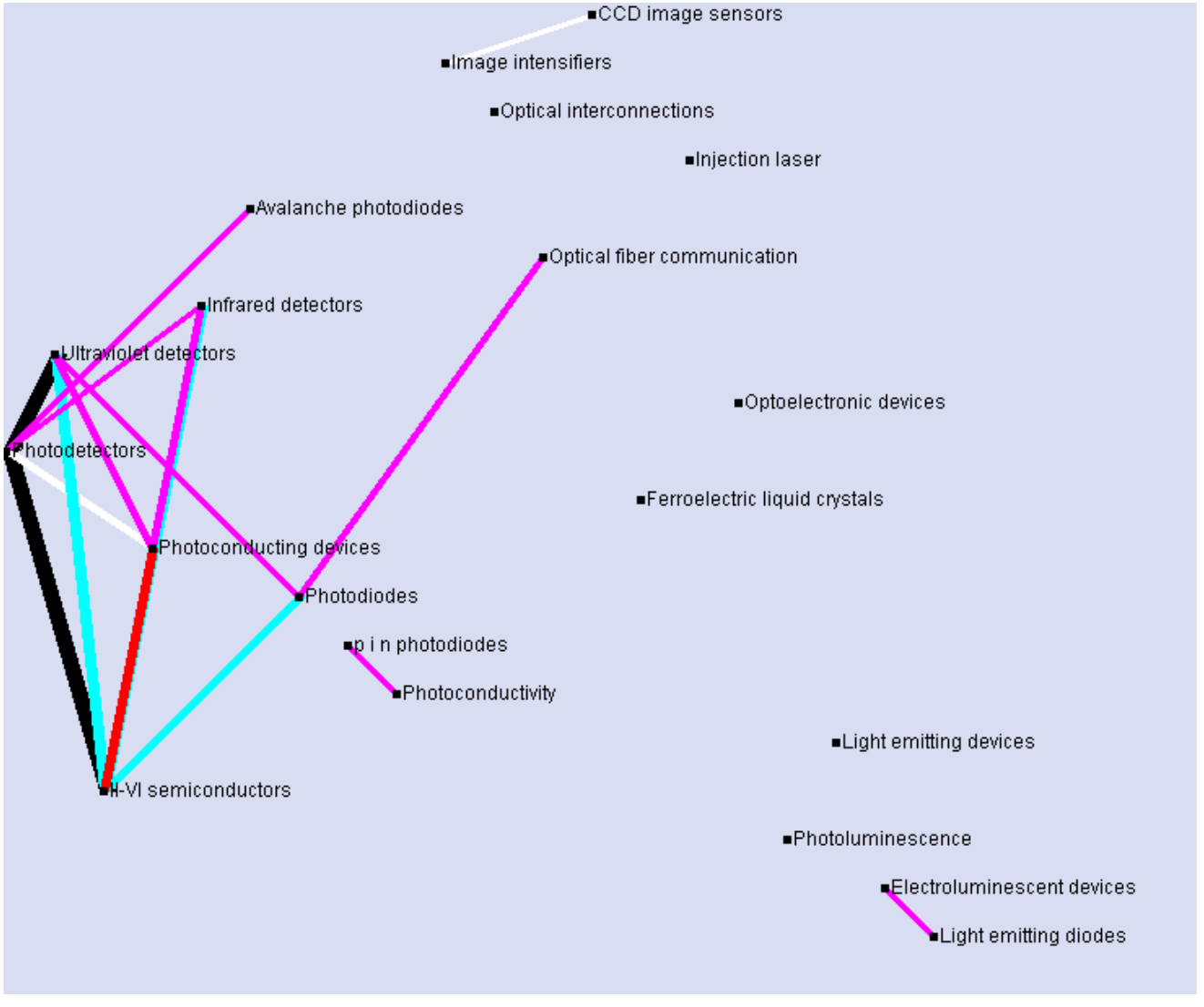
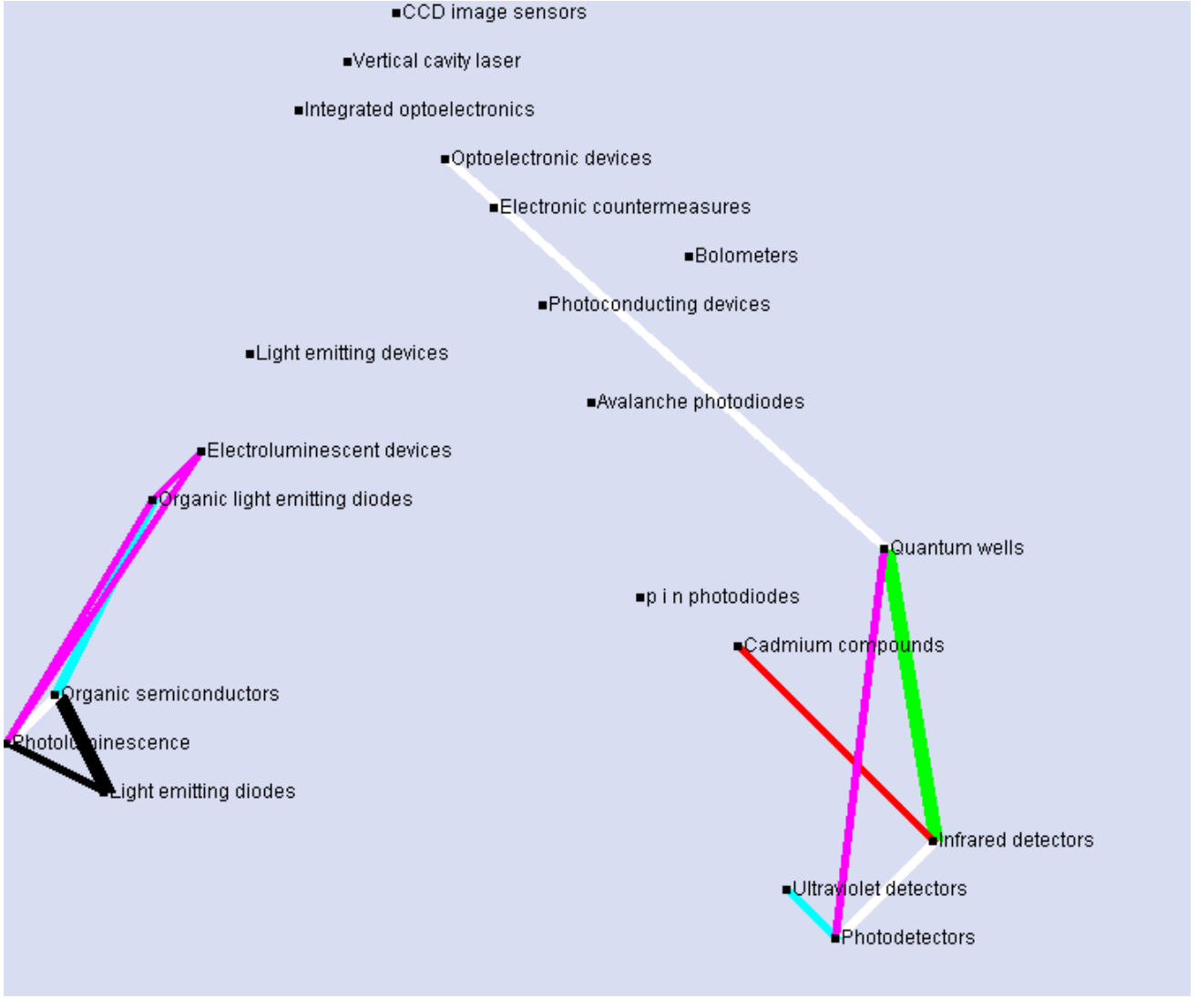


Figure 4. Carte des classes pour la période P2 avec la sélection des termes par l'approche HT



Après les cartes obtenues par l'approche TFIDF, voyons celles obtenues par l'approche HT. Pour la première période (cf. figure 3), on retrouve le sous-réseau des photodétecteurs qui a pris une importance notable par rapport à ce que l'on a vu sur la figure 1, ainsi que le sous-réseau optique qui est à la fois plus petit et très isolé en bas à droite. Par contre, ces deux réseaux ne sont plus liés. En fait, le filtrage opéré par la sélection HT (section 3.1) a supprimé en grande partie le vocabulaire décrivant les matériaux, vocabulaire plus spécifique d'autres domaines technologiques. Nous constatons également que le réseau de l'imagerie reste isolé en haut de la carte, tout en étant proche du réseau des photodétecteurs. Pour la seconde période, la comparaison des cartes obtenues par les sélections HT (cf. figure 4) et TFIDF (cf. figure 2) nous permet d'observer la même scission de l'arc principal en deux réseaux isolés portant sur les photodétecteurs d'une part, et, l'optique d'autre part. La raison en est, là aussi, l'élimination des termes décrivant les matériaux par la sélection HT, avec la notable exception des composés organiques (classes *Organic semiconductors* et *Organic light emitting diodes*) qui apparaissent dans P2. Bien que ce domaine technologique soit centré sur les dispositifs et non sur les matériaux, le fait que ces classes soient présentes montre l'importance des matériaux organiques dans l'évolution de ces dispositifs.

Finalement, en comparant les cartes obtenues par la sélection HT pour les deux périodes (cf. figures 3 et 4), nous observons la même évolution que celle décrite avec la sélection TFIDF (cf. figures 1 et 2), mais de façon plus accentuée. En effet, le réseau des photodétecteurs, très important pour la période P1, est plus restreint dans la période P2. De même, le réseau de l'optique, petit et isolé dans la période P1, prend de l'ampleur dans la seconde période.

5 Discussion

Le modèle de diffusion a pour objectif de caractériser l'évolution d'un domaine technologique par une catégorisation de son vocabulaire suivant son rôle décliné en trois stades : Inhabituel, Etabli ou Transfert. Ce modèle introduit également la notion de « Home Technology » permettant de sélectionner les termes représentatifs du domaine technologique étudié par rapport aux domaines voisins.

L'analyse diachronique est fondée sur l'application d'une méthode de classification automatique organisant les données en thématiques et d'analyser leur proximité relative. La comparaison de deux périodes de temps successives permet d'étudier l'évolution des cartes et le contenu des classes obtenues.

La convergence entre ces deux approches analytiques a déjà été étudiée auparavant dans [10] où nous avons positionné les termes des différentes catégories du modèle de diffusion dans les classes des cartes de l'analyse diachronique pour le domaine de la Biologie moléculaire.

Dans le présent article, nous étudions l'influence de la sélection des termes selon le modèle de diffusion sur l'analyse diachronique par classification.

L'analyse diachronique à partir de l'ensemble du vocabulaire déterminé par l'approche TFIDF a permis de mettre en évidence une inversion de l'importance relative entre les thématiques « Photodétection » et « Optique » avec pour cette dernière l'apparition, en seconde période, des composés organiques.

La sélection des termes par l'approche HT a éliminé une grande part du vocabulaire. Malgré cette perte d'information du point de vue terminologique, l'ensemble des documents obtenu forme toujours un corpus représentatif du domaine. Ceci est confirmé par les observations de l'expert scientifique.

Par exemple, parmi le vocabulaire supprimé se trouvent les termes décrivant les matériaux. L'expert nous confirme que ceci est cohérent avec le fait que le domaine étudié soit centré sur les dispositifs optoélectroniques et non pas sur les matériaux utilisés pour ces dispositifs. En conséquence, la lecture des cartes s'en trouve facilitée par la séparation plus nette des thématiques en des réseaux indépendants décrivant plus clairement l'organisation du domaine.

Cependant, des termes décrivant les composés organiques sont conservés. L'expert explique ceci par le fait que l'utilisation de ces composés dans les dispositifs optoélectroniques a modifié de façon notable leurs caractéristiques. En effet, un intérêt croissant de la recherche pour les semi-conducteurs organiques est justifié entre autres par leur souplesse qui permet d'envisager la réalisation de dispositifs électroniques miniaturisés et flexibles.

Ce travail correspond à une première expérience d'imbrication des deux approches analytiques. Les résultats déjà obtenus devront être confirmés par de nouvelles applications sur différents domaines technologiques.

6 Remerciements

Ce travail fait suite au projet européen PROMTECH - PROMisingTECHnologies [13], réalisé dans le cadre de l'Action Spécifique NEST (New and Emerging Science and Technology) du 6ème Programme-Cadre de l'Union Européenne. Le consortium était constitué par l'ARC System Research GmbH (Vienne, Autriche), le Fraunhofer Institut für Systemtechnik und Innovationsforschung (Karlsruhe, Allemagne) et l'INIST-CNRS (Nancy, France).

Nous remercions également notre collègue Nathalie Vedovotto, ingénieur documentaliste de l'INIST-CNRS, qui nous a apporté son expertise scientifique dans le domaine des Dispositifs optoélectroniques.

7 Bibliographie

[1] **ARMSTRONG J.S., GREEN K.C.,** *http://www.forecastingprinciples.com/selection_tree.html*, 2007

[2] **NOYONS E.,** *Science maps within a science policy context*. Dans : Handbook of Quantitative Science and Technology Research, Eds. Moed H.F., Glänzel W., Schmoch U., Kluwer Academic Publishers, London, pp. 237-255, 2004

[3] **DAIM T.U., RUEDA G., MARTIN H., GERDSRI P.,** *Forecasting emerging technologies: Use of bibliometrics and patent analysis*. Technological Forecasting & Social Change, 73, pp. 981-1012, 2006

[4] **MOGOUTOV A., KAHANE B**., *Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking*. Research Policy, 36, pp. 893-903, 2007

[5] **KAJIKAWA Y., YOSHIKAWA J., TAKEDA Y., MATUSHIMA K.,** *Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy*. Technological Forecasting & Social Change, 75, pp.771-782, 2008

[6] **SALERNO M., LANDONI P., VERGANTI R.,** *The role of funded projects content analysis in early stage disciplines exploration: The case of nanotechnology*. Dans : Proceedings of SPRU 40th anniversary conference – The future of science, technology and innovation policy, 2006

[7] **MOGOUTOV A., CAMBROSIO A., KEATING P., MUSTAR P.,** *Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays*. Journal of Informetrics, 2, pp. 341-353, 2008

[8] **SCHIEBEL E., HÖRLESBERGER M.,** *About the identification of technology specific keywords in emerging technologies: The case of "Magnetoelectronics"*. Dans : Proceedings of ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics, Torres-Salinas D., Moed H. F. (Eds.), Madrid, June 25th -27th, pp. 691-695, 2007

[9] **BESAGNI D., FRANCOIS C., POLANCO X., ROCHE I.,** *Stanalyst® : Une station pour l'analyse de l'information*. Dans : Actes de Veille Stratégique Scientifique et Technologique VSST2004, Toulouse, 25-29 octobre 2004, pp. 319-320

[10] **ROCHE I., BESAGNI D., FRANCOIS C., HÖRLESBERGER M., SCHIEBEL E.,** *Identification and characterisation of technological topics in the field of Molecular Biology*. A paraître

[11] **LELU A.,** *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. Thèse de l'Université de Paris 6, 1993

[12] **LELU A., FRANCOIS C.,** *Hypertext paradigm in the field of information retrieval: A neural approach*. 4th ACM Conference on Hypertext, Milano, November 30th–December 4th, 1992

[13] **BESAGNI D., FRANCOIS C., FRIETSCH R., HÖRLESBERGER M., von OERTZEN J., PRETSCHUH J., ROCHE I., SCHIEBEL E., SCHMOCH U.,** *Final Report (Deliverable 06) for project PROMTECH - Contract N° 15615*. 89 pages + 3 Appendixes, http://promtech.arcs.ac.at/index.php?id=393, 2007