Identification and Characterisation of Technological Topics in the Field of Molecular Biology

Ivana Roche, Dominique Besagni, Claire François, Marianne Hörlesberger, Edgar L. Schiebel

Identification and Characterisation of Technological Topics in the Field of Molecular Biology

Ivana Roche*, Dominique Besagni, Claire François

INIST-CNRS, 2 allée du Parc de Brabois, CS 10310, 54519 Vandoeuvre-lès-Nancy, France, Tel.: +33 (0) 383504600, Fax: +33 (0) 383504733, ivana.roche@inist.fr, dominique.besagni@inist.fr , claire.francois@inist.fr

Marianne Hörlesberger, Edgar Schiebel

Austrian Research Centers GmbH, Tech Gate Vienna, Donau-City-Straße 1, 1220 Wien, Austria, Tel.: +43 (0) 50 550-4524, Fax : +43 (0) 50 550-4500, marianne.hoerlesberger@arcs.ac.at, edgar.schiebel@arcs.ac.at

* Corresponding author

Theme: Primary: S&T indicators for the identification of emerging fields; Secondary: Visualisation and science mapping.

Keywords: emerging technologies, bibliometric indicators, terminology evolution, diffusion model, diachronic cluster analysis**.**

Abstract

Following up the European project PromTech the aim of which was to detect emerging technologies by studying the scientific literature, we chose one field, Molecular Biology, to identify and characterize emerging topics within that domain. We combined two analytical approaches: the first one introduces a model of the terminological evolution of the field based on bibliometric indicators and the second one operates a diachronic clustering analysis. Our objective is to bring answers to questions such as: Which technological aspects can be detected? Which of them are already established and which of them are new? How are the topics linked to each other?

1 Introduction

This paper focuses on methodological approaches for characterising specific topics within a technological field based on scientific literature data as used in the framework of the PromTech project. This project aimed to identify promising emerging technology fields on the observation that many new emerging technologies of the last decades have drawn on the technical application of physical knowledge. Indeed, there are many examples showing that advanced technology in Applied Sciences as well as in Life Sciences is increasingly linked to recent outcomes of research in Physics. The study of the intersection of Physics and these two scientific domains with bibliometric methods, validated by expert panels, led to the selection of ten most promising technological fields. In order to characterise them and analyse their potential evolution, we used the diachronic clustering analysis and the diffusion model, based on bibliometric indicators. This model uses a modified TF IDF approach and additionally evaluates the "home Technology" nature of keywords on the basis of probabilities, and distributes them in different diffusion stages and in semantic categories.

The questions we are trying to answer are: how can we identify and characterise important topics in a set of several thousand articles? Which technological aspects can be detected? Which of them are already established and which of them are new? How are the topics linked to each other?

After describing the diffusion model and the diachronic cluster analysis, we present the results obtained for one of the promising technological fields detected in the PromTech project: "Molecular Biology". The results are visualised with the software-tool Stanalyst® [1].

2 Methodology

The high coverage of scientific literature in Physics and Natural Science, the detailed systematic of classification codes with more than 6,000 items as well as the high quality of the keywords where the reason why the authors used the PASCAL database for this investigation. In PromTech project, the data were extracted in function of their classification categories from the PASCAL bibliographic database. PASCAL contains more than 17 million bibliographic records. They are derived from the analysis of the scientific and technical international literature published predominantly in journals, conference proceedings, reports and dissertations but, in this work, we restricted our study to a corpus of publications coming from journals only. Each publication receives at least one but, if necessary, more than one classification categories related to different scientific fields. The queries operated in this work were exclusively based on the classification categories given in the database and assigned to the individual publications, either manually by scientific experts or automatically based on a content analysis. These classification categories belong to the PASCAL classification scheme, that is a taxonomy of every field and subfield of all the disciplines covered in the database. Each individual publication also benefits from a manual or automatic indexing by keywords. After validation by a scientific expert, that terminology is employed in our analysis.

Let us introduce the two methods we employed to follow the evolution of the technological field.

Firstly, the diffusion model was used to evaluate the term status in the considered technological field, comparing it with its status in the other technological fields where the terms are present and measuring their degree of emergence in order to provide a global field characterisation.

Secondly, the diachronic cluster analysis considered the content of the technological field by a clustering approach allowing to organize the data in sub-topics and to analyse the links between them.

2.1 Diffusion model

Our starting hypothesis was that technologies have their roots in an invention which was created by one person or a group of scientists. After a specific invention is published, the respective knowledge is available to other scientists. If scientists are inspired by the new ideas, they make use of the new theories and findings, and publish their own research work. At that stage, a new terminology is created and also the use of methods, materials, natural science effects and

applications are described in few publications. Unusual terms play a specific role in that stage of the development of a new technology. In later stages, many groups use the findings on a similar technology nomenclature. Bit by bit a well defined terminology is established but is still specific for this new field. In yet other stages, findings diffuse to other technologies and many terms show a more cross sectional property.

Following this diffusion model, a framework was constructed to analyse the stages of the selected technologies:

- STAGE 1: New terms created specifically for the technology can be found in few publications. New terms and terms established in other technologies occur seldom and they form a strong exotic cluster.
- STAGE 2: Technology specific terms are established in the technology field and can occur together with established methods, materials, tools, and applications from other technology fields. They begin to diffuse to other technology fields.
- STAGE 3: This is the stage with the highest maturity. Technology specific terms are greatly established. Natural science effects, methods, tools, materials, and applications are highly accepted. They show a broad diffusion in other research or technology fields.

Therefore, the diffusion model identifies three categories of terms: terms unusual in this topic, the established terms, and cross section terms. The allocation of a term in one of these categories is accomplished thanks to the application of a sequence of statistical filtering essentially based on two pragmatic approaches introduced to identify keywords describing a technological specificity.

First, the so called “home Technology” terms were defined. We assume words which are specific for a technology occurring with a higher probability in one technology field than in the others. The probability is defined by the frequency of one term in a technology field divided by the number of articles in this field. Then, this probability is calculated for the same term for other technology fields. The technology field with the highest probability is declared as the home technology for the considered term.

Secondly, we used a measure for the cross section degree of a term in the selected technology. For this purpose, we introduced the Gini index, a measure of statistical dispersion most prominently used as a measure of inequality of income distribution or inequality of wealth distribution (see [2]). The Gini index varies from 0 to 1. In our case, 0 means a completely uniform distribution of all terms and indicates that each term of the considered field occurs in all other fields. A Gini index of 1 tells us that all terms of the considered field occur only in this field, which means that no term of the considered field occurs in another field. Therefore, more general terms push down the Gini index. Particularly, in our context of emerging technologies, a Gini index value lower than 0.87 means that the considered field does not consist of specific terms.

The technology field is examined according to bibliometric statistics that give an overview of the publication activity, the diversity of the field, the specificity of technical terminology, and the cross section property.

Next, the keywords which form the framework of a technology field and the diffusion stages are selected. For this purpose the following procedure is defined:

1) Selection of all keywords of the technology field;
2) Selection of home technology terms of the technology field;
3) Selection of terms with a higher frequency (more than 2 occurrences);
4) Selection of terms with a higher Relative Term Frequency - Inverse Technology Frequency (RTF-ITechF) (see [3]);
5) The remaining terms are categorised in the three stages of our diffusion model (unusual terms, established terms and cross section terms), using the Gini index and the RTF-ITechF (see [3]);

Finally, each term belonging to the three diffusion stages is secondarily categorised, with expert's help, into "Natural science effect-", "Method-", "Material-", or "Application-" related (from now on named semantic categorization).

2.2 Diachronic cluster analysis

The diachronic cluster analysis is realized with the help of a clustering tool applying first a non-hierarchical clustering algorithm and then a principal component analysis to map the obtained clusters. This tool is implemented in the information analysis platform Stanalyst®. Neurodoc ([4]; [5]), the clustering tool we used, applies a non-hierarchical clustering algorithm, the axial K-means method, coming from the neuronal formalism of Kohonen's self-organizing maps, followed by a principal component analysis (PCA) in order to represent the obtained clusters on a 2-D map.

Our approach consists of:

- splitting the corpus in two successive time periods, namely 1996-1999 and 2000-2003;
- applying cluster analysis on the corpus of each period, in which documents are represented by the keywords existing in the bibliographic references;
- analysing the evolution between the two cluster sets and maps by examining the vocabulary related to the clusters of the two periods.

In order to analyse the evolution of the cluster vocabulary between the two considered periods, we build a comparison matrix pointing out the percent of keywords belonging to the second-period clusters and already existing in the first-period clusters. The cumulated percentage is also calculated for each second-period cluster. Using this matrix, we can identify different cluster behaviours: stability, fusion or splitting. Using the cluster maps, we can also detect status change of the clusters in the global network. These detected phenomena need to be validated by experts from the technological domain.

The cluster behaviours are calculated in the following way.
Let M be the comparison matrix:

$M = (m_{i,j})$ with :

$i = 1,\ldots,n_{P2}$ where n_{P2} = number of clusters obtained in the second period

$j = 1,\ldots,n_{P1}$ where n_{P1} = number of clusters obtained in the first period

m_{ij} = percent of keywords belonging to cluster *i* and existing in cluster *j*

and for each line *i*, the marginal value is:

$$m_{i.} = \sum_{j=1}^{j=n_{P1}} m_{ij}$$

The analysis of the matrix M allows us the build the following hypotheses:

- Let *(i,k)* be a pair of homonymous clusters. The cluster *i* is supposed to be stable if:

$$m_{ij} < \frac{m_{ik}}{2}; \forall j \neq k \quad \text{and} \quad j = 1, n_{P1} \qquad (1)$$

- Let *(i,k)* be a pair of homonymous clusters. The cluster *i* is supposed to be unstable if there exists at least one *j* respecting the condition:

$$m_{ij} > \frac{m_{ik}}{2}; \forall j \neq k \quad \text{and} \quad j = 1, n_{P1} \qquad (2)$$

- Clusters having new titles (titles not existing in the first period cluster list) and obeying the condition (1) are supposed to be stable;
- Clusters *i* with the lowest marginal values have less inheritance from first-period clusters and, we suppose they constitute new terminological notions;
- Clusters *i* with the highest marginal values are supposed to aggregate a great number of notions already existing in the first period clusters.

At this point, scientific expertise is needed. We asked scientific experts for the validation of our hypotheses, for the analysis of the two cluster maps taking into account the cluster contents.

3 Results

Molecular Biology is a research field about Biology at the molecular level. There is a strong interaction with Genetics and Biochemistry. Many efforts are done to understand the mechanism of interactions in cells, the working of DNA, RNA and the regulation. Bibliometric analysis reveals that "Molecular Biology" is a field with high research activities with about 5,000 articles produced during the 1996-2003 period.

3.1 Diffusion model

In comparison to other technology fields (which were investigated in the PromTech project), the percentage of home terms (specific technical terminology), the number of terms per article (diversity of terms) and the average relative term frequency are quite low. That indicates a lot of research activities in various topics without a committed specific terminology. It seems that the terminology consists of many terms but is not established and is somewhat highly

fluctuating. Contrary to the lower percentage of home terms, the value of the average Gini index shows however that the terminology is commonly used within the technology field of "Molecular Biology". The results of the selection procedure for the keyword categorisation into diffusion stages are showed in the Table 1.

Table 1. Results of the keyword selection procedure for the field “Molecular Biology”.

By applying the bibliometric indicators of Table 1 for identifying the diffusion stage and by the know-how of the experts for assigning the terms to the 4 semantic categories we get the results presented in Table 2.

Categories	**Unusual terms** *(Not frequent technology specific terms)*	**Established terms** *(Frequent technology specific terms)*	**Cross section terms**
Natural science effect	Multiple diffraction Angular distribution Helix coil transition Nuclear relaxation …	Molecular force constants Isomerism Microgravity Solvation …	Brownian motion Photosynthesis Circular dichroism Intermolecular forces …
Method	Matrix assisted laser desorption ionization Moller Plesset partition Optical tweezer X-ray topography …	Two-dimensional spectroscopy Heavy atom method Liquid-liquid transformation …	Molecular dynamics simulation XRD NMR spectroscopy Lennard Jones model …
Material	Tissues Pinene Leaf oil Lyotropic liquid crystals …	Biological macromolecule …	Polymers Essential oils Biological compound Animal cells …
Application	Instrument for chemical analysis Multilayer perceptrons Diagnostic techniques Drug delivery systems …	Structure resolution Viruses …	Molecular configurations Polymer solutions Biomolecular effects of radiation Oxygen sensors …

Table 2. Keywords sorted by diffusion stage and by semantic category.

This table shows only the most important terms of the field “Molecular Biology”. The constraint of a paper does not allow to present the whole long term list.

We consider now the content of this table. In the category “Natural science effect” we notice the terms “Multiple diffraction”, “Angular distribution”, “Helix coil transition”, and “Nuclear relaxation” for instance as “unusual terms”. Each of these “unusual terms” has a RTF (Relative Term Frequency) less than or equal to 2%. The sample of the keywords in the category “Natural science effect” such as “Molecular force constants”, “Isomerism”, “Microgravity”, and “Solvation” are allocated to diffusion stage “Established terms”. Each of these keywords holds a RTF greater than 2%. The “Cross section terms” in this category such as “Brownian motion”, “Photosynthesis”, “Circular dichroism” and “Intermolecular forces” feature a Gini index less than or equal to 0.97.

Considering the category “Method” the sample of terms “Matrix assisted laser desorption ionization”, “Moller Plesset partition”, “Optical tweezer”, and “X-ray topography” are the results for “unusual terms” with the RTF less than or equal to 2% as we already mentioned for “Natural science”. The examples of keywords for the diffusion stage in this category are “Two-dimensional spectroscopy”, “Heavy atom method”, and “Liquid-liquid transformation”, whereas the methods like “Molecular dynamics simulation”, “XRD”, “NMR spectroscopy”, and “Lennard Jones model” are assigned to the stage “cross section terms” for this category.

The category “Material” shows us the sample of terms “Tissues”, “Pinene”, “Leaf oil”, and “Lyotropic liquid crystals” for the “unusual terms” diffusion stage. “Biological macromolecule” is listed for the “established terms” stage, and “Polymers”, “Essential oils”, “Biological compound”, and “Animal cells” in the stage “cross section terms”.

In the category “application” we find the examples of terms “Instrument for chemical analysis”, “Multilayer perceptrons”, “Diagnostic techniques”, and “drug delivery systems” in the stage “unusual terms”. Terms with a RTF greater than 2% in the stage “established terms” are such as “Structure resolution” and “Viruses”. Examples for terms with a Gini index less than or equal to 0.97 (cross section terms) in the category “Application” are “Molecular configurations”, “Polymer solutions”, “Biomolecular effects of radiation”, and “Oxygen sensors”.

The analysis reveals some special, probably new biophysical research fields like multilayer perceptrons or lyotropic liquid crystals and also some well known characteristics about this research field like molecular interaction, microgravity or important driving forces like diagnostics and drug delivery systems. But expected and established basic areas like gene expressions, micro arrays, polymerase chain reaction, expression cloning or others are not even home technology terms in “Molecular Biology”. The reason could be that they are now more often used in related science fields like Medical Imaging, Biotechnology, Biodeterioration or others.

3.2 Diachronic cluster analysis

The diachronic cluster analysis applied on two periods, namely 1996-1999 and 2000-2003, allows us to determine which topics of the second period have roots in the first one and which topics are new in the second period.

After analysis of the results of classifications obtained with various numbers of classes over the two periods the expert chose that with 20 clusters in the first period and 20 clusters in the second period. In Table 3, we can see, for each period, the lists of cluster titles, their number of keywords (column KW) and the number of documents (column DOC). The 11 homonymous cluster pairs are highlighted.

Analysing the comparison matrix and the Table 3 we can singularize for the second period:

- eleven homonymous clusters, but only six of them seem to be stable (5, 9, 10, 11, 18 and 19);
- two clusters (13 and 15) with new titles but presenting characteristics of stability;
- three clusters (2, 3 and 17) with low marginal values, (new clusters?);
- one cluster (6) with high marginal value. (cluster with a strong inheritance from different clusters of P1).

	2nd period				
	KW	*DOC*	*Cluster title*	*KW*	*DO*
Cluster 1	27	172	Neural networks	9	20
Cluster 2	69	30	Luminescence decay	59	36
Cluster 3	105	89	NMR spectroscopy	13	17
Cluster 4	12	191	Molecular configurations	37	26
Cluster 5	62	227	Fluctuations	54	41
Cluster 6	7	445	Computerized simulation	23	97
Cluster 7	13	71	Intramolecular mechanics	14	12
Cluster 8	94	375	Atomic force microscopy	65	38
Cluster 9	81	94	Crystal growth from solutions	22	10
Cluster 10	48	117	Crystal structure	29	13
Cluster 11	93	172	Laser applications in medicine	7	17
Cluster 12	23	133	Cellular biophysics	33	83
Cluster 13	37	95	Random processes	31	79

Table 3. Characteristics of the 2 sets of clusters.

On the map of both periods, as we can see in the Figures 1 and 2, it is possible to singularize an interesting dichotomous configuration represented by two very strongly connected cluster networks associating about two thirds of the clusters. However, these networks have quite different characteristics. On the one hand, the topics present in the bigger one are related to the modelling and simulation of biological phenomena. On the other hand, the little network is very homogeneous and deals essentially with instrumentation topics. The remaining clusters are scattered in the map with no significant links with the two previously described networks.

FIRST PERIOD MAP

In the first period map (Figure 1), the bigger network is located in the left side and is formed around the cluster "Computerized simulation" that brought together two sub-networks:

- the first one is located in the upper side and constituted by the clusters "Molecular configurations", "Polymers" and "Fluctuations" related to physical characteristics of biological structures, specially macromolecules and membranes;
- the second one, in the lower side, is formed by the clusters "Diffusion", "Stochastic process", "Neural networks", "Physiological models" and "Biological evolution" dealing with biological process analysis. This sub-network focuses on theoretical aspects as neuronal network studies produced in computer science field and on aspects related to the modelling of physiological process.

The smaller network, that is located in the lower right side, deals with applications of instrumentation techniques to therapeutic measuring.

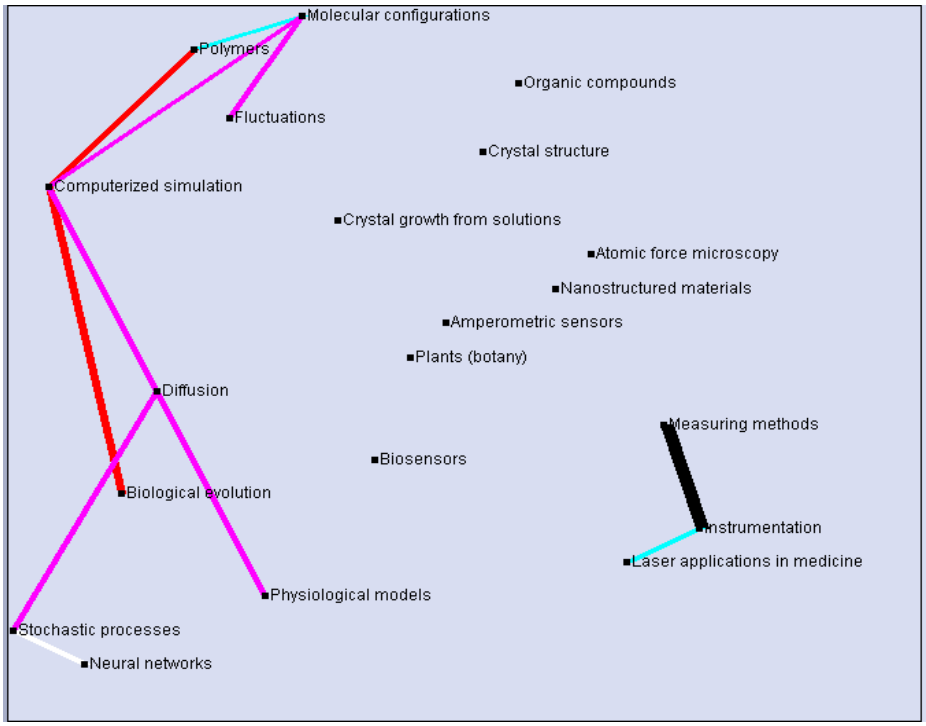

Figure 1. First period cluster map of "Molecular Biology" technological field.

SECOND PERIOD MAP

In the second period (Figure 2), this smaller network is now located in the upper left side. Its evolution shows an important stability in spite of a redistribution of the cluster contents: the clusters "Measuring methods" and "Fluorescence" have strong inheritance from "Instrumentation" cluster.

The bigger network, located in the right side of the map, shows the same global structure with its two sub-networks linked by means of the "Computerized simulation" cluster. As indicated previously, this cluster, number 6 in Table 3,

has a strong inheritance from different clusters which is explained by its position in this network.

The stability of each sub-network structure can be observed:

- in the upper side, the sub-network organized around “Physiological models”;
- in the bottom side, the sub-network centred on “Molecular configurations” cluster.

An in-depth analysis of these sub-networks allows to observe some interesting evolutions. Particularly, we observe a re-centring of the cluster couple “Neural networks” and “Stochastic processes” and a meaningful density increase of this sub-network with the coming out of a new cluster: “Brownian motion”. This cluster is constituted by a majority of studies related to the development of Brownian engines that are devices working in a nanometric scale by which thermally activate processes are controlled and used to generate directed motion in space.

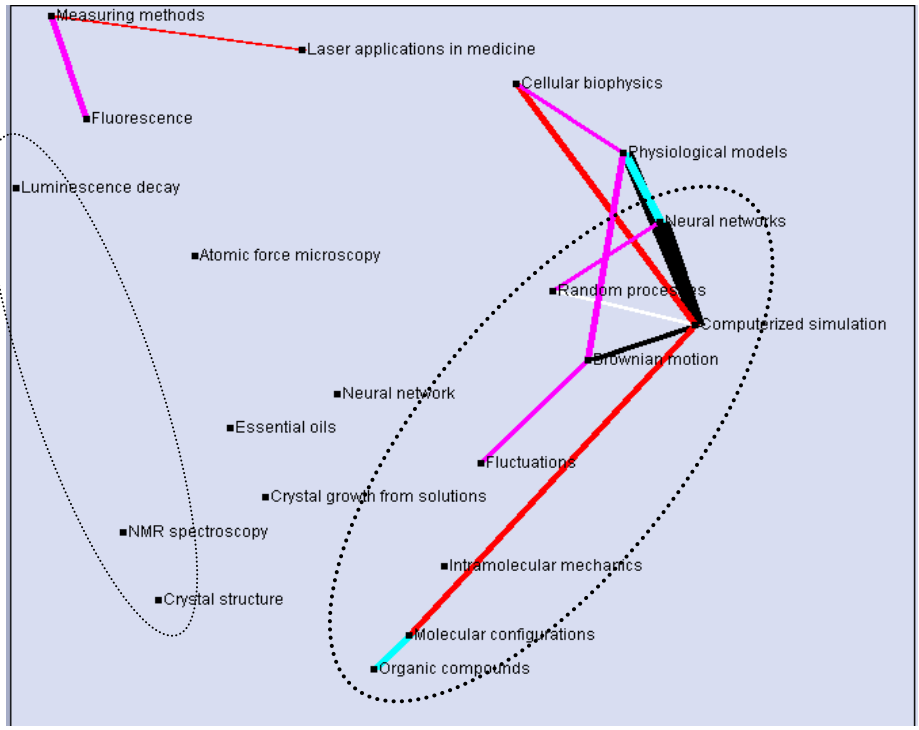

Figure 2. Second period cluster map of “Molecular Biology” technological field.

The first period cluster “Biological evolution” that dealt with the modelling of biological evolution phenomena at genetic and environmental level has a growth of its application field towards principally the development of theoretical models to study processes positioned at the molecular level, in its principal heir in the second period, the cluster “Random process”.

The second period cluster “Cellular biophysics” inherits from 11 first period clusters, principally: “Biosensors”, “Biological evolution”, “Diffusion”, “Stochastic processes” and “Physiological models”. As we can see in the first period map, theses clusters form a sub-network if we consider that the cluster

"Biosensors", that is not linked, is near enough to allow its integration. This sub-network deals with the analysis of biological processes. Finally, in the "Cellular biophysics" cluster we can find studies dealing with the utilisation of tumoral cells as behaviour model, that were absents in the first period clusters.

3.3 Convergences between the two analyses

The comparative analysis of the results of the two methodological approaches operated in this work, namely the diffusion model and the diachronic cluster analysis, reveals some quite interesting convergence points.

Let us consider the 124 home technology terms obtained by the model diffusion selection procedure (Table 1) and their distribution by on the one hand diffusion stage and on the other hand their presence in clusters of first period (P1), second period (P2), both periods or their absence in both periods. The results are showed in Table 4.

	Exclusively in P1 clusters	Exclusively in P2 clusters	Present in clusters of P1 & P2	Absence from clusters of P1 & P2
Unusual	40%	56%	9%	57%
Established	13%	12%	45%	6%
Cross section	47%	32%	45%	37%
Total nb. of terms	15	57	22	30

Table 4. Distribution of "home Technology" terms by diffusion stage, in function of their presence/absence in the clusters of each period.

We can observe, in the last column of Table 4 that more than a quarter of selected terms (30) are absent from the clusters of both periods. This set is formed by a majority of unusual and cross section terms (in total 94%). This absence can be explained if we consider that:

- the unusual terms correspond to fundamental new concepts not frequent enough to participate to a cluster constitution (as "optical trapping", "optical tweezers" or "Moller Plesset partition"), and
- the cross section ones reflect concepts related to applications not yet consolidated and in consequence not frequent enough.

About a half of "home Technology" terms (57) are presented exclusively in P2 clusters. This set is constituted by a majority of unusual and cross section terms (in total 88%). The presence in the second period of a great number of unusual and cross section terms seems to be consistent with the observation in the second period cluster map that:

- in its left side a great number of heterogeneous isolated clusters contains a lot of unusual terms carrying innovation but not yet important enough to be associated to the central clusters, and
- in its right side the clusters forming the network kernel and contain the cross section terms by which the diffusion to other technological fields occurs.

Fifteen terms (that is 12% of the "home Technology" terms) are present exclusively in P1 clusters. This smallest set is constituted by a majority (87%) of unusual and cross section terms. The important presence of unusual terms in P1 is

at first sight unexpected. These terms can be considered as peripheral either representing topics which did not go on in the second period, or representing topics which went in others fields.

Twenty-two terms (that is 18% of the "home Technology" terms) are present in clusters of both periods. This set is formed by a majority of established terms and cross section terms (totally 90%). The presence of a great number of established and cross section terms is not really surprising. Indeed, if we consider that the diffusion to other technological fields begins with established terms, and becomes more marked with cross section terms, their presence in clusters of both periods should be considered consistent. It can also explain the very poor presence of unusual terms in clusters of both periods.

We also consider the cross distribution of the "home Technology" terms by diffusion stage (Unusual, Established or Cross section) and semantic categories (Natural science effect, Method, Material or Application) for both periods (Table 5).

Category / diffusion stage	Natural science	Method	Material	Application
	First period			
Unusual	2	2	2	2
Established	5	4	2	1
Cross section	9	4	7	3
	Second period			
Unusual	15	10	11	1
Established	13	4	2	2
Cross section	10	6	13	6

Table 5. Distribution of "home Technology" terms by semantic category and by diffusion stage for each period.

The number of unusual terms grows significantly in all the categories except in the Application one, with a lot of new terms not yet important enough to be associated to a central cluster and located in a great number of isolated heterogeneous clusters located in the marked region in the left side of P2 cluster map (Figure 2).

Concerning the established terms, we have an equilibrated situation between the two periods for all categories except the Natural science one. This situation is quite normal if we consider that established terms correspond to terminologically consolidated terms in the Molecular Biology field.

In relation to cross section terms, we have the strongest term growths in Material and especially in Application, that is the category facilitating the more the diffusion from "Molecular Biology" field to other technological fields. The clusters involved with application category terms are located in the neighbourhood of the P2 larger cluster network as we can see in the delimited region in the right side of P2 cluster map (Figure 2). In addition, we can suppose that the great number of clusters involved points out a behaviour with intrinsically dispersive tendency in accordance with the definition of these terms that announces their capacity to assure a broad diffusion to other technological fields.

4 Conclusion

It is a challenge to identify emerging technologies. “Molecular Biology” is a broad field and by applying the complementary methods presented here, it can be better characterised and described.

One particularity of our approach is the alternate utilisation of different bibliometric and/or informetric methods and scientific expertise. This expertise is necessary to validate or to complete the results obtained at each step of the work as well as to get the experts personal input on the matter at hand. This could be time consuming, but with our approach, the amount of data submitted to the experts’ appreciation is limited, thus making their task notably easier.

The application of the diffusion model is a novel bibliometric approach giving a more in-depth view of the considered field. The introduced concept of home Technology, associating terms with technologies, allows the development of a new interesting analysis methodology based on the notion of terminology diffusion. The indicators we used such as term frequency, Gini index, relative term frequency, inverse technology frequency, helped working out the different features of a field. Based on these results, we were able to work out the field specifications.

The diachronic approach we adopted consists in splitting the corpus in two periods, applying a content cluster analysis for each period and then detecting the evolution of the topics by examination of the vocabulary related to the respective clusters.

Applying at the same time diachronic cluster analysis and diffusion model allows to confirm the results detected by each method and also to lead to a deeper understanding and characterization of the technological field, The diffusion model approach allows new interpretation of diachronic clustering results introducing external term categorizations.

It is to be noted that the conditions that make that study possible may disappear soon. Indeed, our bibliometric approach is based on the one hand on a multidisciplinary discriminating classification scheme and on the other hand on the possibility of indexing each bibliographic reference with several classification categories. These constraints are very strong and with the assisted indexing procedures currently used on an ever increasing share of the PASCAL database, we could no longer respect these constraints. In the future, if we were to reproduce or to generalize our approach, it would be necessary to develop an automatic categorization method for the bibliographic references along with a protocol to update the classification scheme.

A very interesting next step in this work would be to take explicitly into account the time variable in the diffusion model so as to be able to detect the passage of terms from one diffusion stage to another between time periods. We hope to verify if the results of both methods could be linked as shown in Table 6.

		Diffusion Model		
		Unusual Terms	Established Terms	Cross Section Terms
Diachronic Cluster Analysis	Clusters in the First Period	locally new terms	locally consolidated terms	locally diffusing terms
	Clusters in the Second Period	globally new terms	consolidated terms with roots in the first period	diffusing terms with roots in the first period

Table 6. Diffusion model and diachronic cluster analysis linkage.

Another attractive research axis could be the development of a clustering analysis method using an incremental approach. Such a method would allow to follow more precisely the evolution of a topic in function of the publication date of its related articles.

Acknowledgment: This work has been realized thanks to the European project N° 15615 (NEST) - Sixth Research and Development Framework Plan of the European Union, during 2005-2007. The project acronym is PromTech, and the project full title is "Identification and Assessment of Promising and Emerging Technological Fields in Europe".

5 References

[1] POLANCO X.; FRANÇOIS C.; ROYAUTÉ J.; BESAGNI D.; ROCHE I. (2001). Stanalyst®: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, In: Proceedings of the 8th ISSI, Sydney, July 16-20, 2001.
[2] Gini Index. *http://en.wikipedia.org/wiki/Gini_coefficient*
[3] SCHIEBEL E.; HÖRLESBERGER M. (2007). About the Identification of Technology Specific Keywords in Emerging Technologies: The Case of "Magnetoelectronics". Torres-Salinas, D., Moed, H. F. (Eds.), Proceedings of ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, June 25th-27th, 2007, pp. 691-69.
[4] LELU A. (1993). Modèles neuronaux pour l'analyse de données documentaires et textuelles, PhD Dissertation. Université de Paris 6, 1993.
[5] LELU A.; FRANÇOIS C. (1992). Hypertext Paradigm in the Field of Information Retrieval: A Neural Approach, 4th ACM Conference on Hypertext, Milano, November 30th–December 4th, 1992.
[6] FERBER R. (2003). Information Retrieval, Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web, Heidelberg: dpunkt.
[7] ROBERTSON S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF, Journal of Documentation 60 no. 5, 503-520.
[8] SPÄRCK J. K.; ROBERTSON S. (2006). Inverse Document Frequency - The IDF page. Retrieved November 22, 2006 from: http://www.soi.city.ac.uk/~ser/idf.html
[9] van RIJSBERGEN C.J. (1979). Information Retrieval, London: Butterworths.
[10] WHITE H. D.; McCAIN K. W. (1989). Bibliometrics, Annual Review of Information Science and Technology, vol. 24, pp. 119-186.