

THE KNOWLEDGE BASE OF THE FIRM IN BIOTECHNOLOGY BASED SECTORS: PROPERTIES AND PERFORMANCE.

Pier Paolo Saviotti, INRA-SERD, Université Pierre Mendès-France, BP 47, 38040 Grenoble Cedex 9, France. Tel: +33 4 76825831; Fax: +33 4 76825455; E-mail: saviotti@grenoble.inra.fr;

and

IDEFFI, CNRS-UNSA, Sophia-Antipolis
06560 Valbonne -FRANCE

First draft, not to be quoted without the permission of the author.

This paper is based on work carried out in collaboration with a number of colleagues of both the institutions mentioned above. In particular it relies on work carried out jointly with (in alphabetical order) Marie Angèle de Looze, Marie Antoinette Mauperthuis and Lionel Nesta.

To be presented at the Workshop

Frontiers and Trends

Frontiers of Innovation Research and Policy

**A Workshop between the Instituto de Economia/UFRJ and ESRC-Centre for
Research on Innovation and Competition- The University of Manchester**

Rio de Janeiro September 26-27 2002

THE KNOWLEDGE BASE OF THE FIRM IN BIOTECHNOLOGY BASED SECTORS: PROPERTIES AND PERFORMANCE.

- 1) Introduction
- 2) General treatment of KB
- 3) Mapping and measurement of KB
 - 3.1) The techniques used
 - 3.1.1) Lexicographic analysis
 - 3.1. 2) Measures of KB properties
- 4) Results
- 5) Discussion
- 6) Summary and conclusions

THE KNOWLEDGE BASE OF THE FIRM IN BIOTECHNOLOGY BASED SECTORS: PROPERTIES AND PERFORMANCE.

Pier Paolo Saviotti, INRA-SERD, Université Pierre Mendès-France, BP 47, 38040 Grenoble Cedex 9, France. Tel: +33 4 76825831; Fax: +33 4 76825455; E-mail: saviotti@grenoble.inra.fr; and IDEFI, CNRS-UNSA Sophia-Antipolis, 06560 Valbonne -France.

1) INTRODUCTION

Advanced capitalist economies are expected to be moving towards the so called phase of the knowledge based economy. In such an economy knowledge would become the main competitive asset of firms, as opposed to capital goods in previous periods. In fact the knowledge based economy is not emerging suddenly, but it is the result of more than a hundred years of evolution, during which R&D has been institutionalised (Freeman and Soete, 1997) and it has become a standard component of economic life. In this period there has been an enormous expansion of education at all levels, including higher education. As a consequence intangible capital has gradually become more important than tangible capital (Foray, 2000). Intangible capital is not simply contained or embodied in the R&D department of a firm, but it is determined by the interactions of the different elements of knowledge held by the individual members of a firm or of an organisation. Insofar as a firm is concerned the concept of knowledge base captures the intangible capital that determines a firm ability to compete. The knowledge base (KB) of the firm can be defined as the collective knowledge that a firm can use for its productive purposes (Saviotti, 1996). The collective character of the KB arises from the necessary interactions between the members of the firm. Every firm is characterised by a very advanced form of division of labour. The final output of the firm can only be produced if the very large number of individual stages of production are co-ordinated or combined. The particular form of division of labour adopted, which is specific to each firm, determines the frequency and the types of interactions occurring within the firm. The knowledge base of a firm thus evolves in the course of time, since any set of interactions modifies the KB but the interactions themselves are determined by the KB. However, the KB achieves a certain stability since the firm has a set of routines and decision rules that are modified very infrequently. The KB of a firm can thus be expected to be very specific and to have a considerable degree of path dependence.

In spite of the general recognition of the importance of knowledge in economic activities we know very little about the ways in which firms create and use knowledge. This paper attempts

to give a more analytical and empirical content to the concept of knowledge base and to discuss the implications of the KB for the behaviour and performance of the firm. This paper will be mainly concerned with firms in the biotechnology based sectors.

2) THE KNOWLEDGE BASE OF THE FIRM.

To say that firms use knowledge might seem a trivial statement. Any productive activity at an time has always required some kind of knowledge. However, what is peculiar to modern firms and organizations is their use of knowledge that is produced by institutions whose main goal is to create knowledge. This is of course a modern phenomenon following the institutionalisation of science in German universities in the second half of the XIXth century and the subsequent institutionalisation of industrial R&D (Freeman and Soete, 1997). Thus modern firms increasingly use scientific and technological knowledge created outside the boundaries of the firm and of the sector in which they operate by other organisations whose main goal is to create knowledge. Furthermore, within the same firms there are departments or divisions which are specifically charged with creating knowledge. This contrasts markedly with previous periods during which the creation of new knowledge took place jointly with its utilisation.

In spite of these changes firms are not primarily knowledge generating organisations, but they use knowledge in order to compete. Competition takes place amongst the final products of the firms. Thus we can say that a firm's KB is used to create its revealed technological performance (RTP). KB and RTP follow distinct but related dynamics. To the extent that knowledge is required to create new goods and services the creation of a new KB will have to precede that of new goods and services. In a sense RTP at a given time t will embody the KB at a previous time $t-\Delta t$. The magnitude of the delay Δt is likely to depend on the characteristics of the sector, of the firm and of the technologies concerned. However, although the dynamics of the KB and of the firm's products are inter-related they are not identical, as confirmed by a number of studies (see for example Pavitt, 1998). It is possible for the knowledge base of firms to converge or to remain very similar during periods in which their products become increasingly differentiated, and the reverse. Moreover, firms use more technologies than those that they incorporate in their products (Brusoni, Prencipe, Pavitt, 2000). Thus a subsequent stage of our study should analyse the dynamics of firms' outputs and its relationships to that of the KB.

As previously pointed out, the KB of a firm is the 'collective' knowledge the firm can use to achieve its productive purposes. Although the production of knowledge does not necessarily follow the same rules as the production of goods, the two are equally likely to be produced by means of division of labour. Within teams work is allocated to individuals, within firms to departments, within an economy to sectors etc.. Organisational boundaries exist because particular groups of people collectively perform given functions. That is, division of labour exists at all levels of aggregation in an economy. From this point of view science can be treated as an activity that is itself carried out by means of division of labour. In science this gives rise to disciplines, sub-disciplines, specialities etc..

Of course, such division of labour would be quite ineffective if it were not accompanied by an adequate co-ordination of these activities. Within a firm co-ordination takes place by means of the interactions occurring at different levels, within and between departments and divisions. The division of labour in knowledge creation presents an additional problem in that the partitions in science do not correspond to those in technology and industry. A given component of knowledge can be classified both within a scientific speciality and within a technological and industrial sector. Conversely, the components of the KB of a firm can be classified either based on the scientific discipline from which they derive or on the industrial application to which they can contribute.

In a highly knowledge intensive firm two different and not necessarily easily compatible forms of division of labour are combined: the division of labour in science and that in industry. Thus, even in the R&D departments of firms in a highly science based field such as biotechnology, we can expect the individual components to which the division of labour gives rise to be partly determined by scientific disciplines and partly by the nature of industrial applications. This double influence on the division of labour in industrial R&D is reflected in the classification of patents, the immediate output of R&D, and of the technological classes that are contained in patents. Technological classes are partly based on scientific disciplines (e.g. C07C: acyclic or carbocyclic compounds; C07D heterocyclic compounds) and partly on industrial applications. Before concluding that the classification of patents is too inaccurate to be used we should remind ourselves that the same ambiguity exists in the classification of industries, some of which are defined by the nature of their output and some by the activities carried out. In spite of this ambiguity the concept of industry has been used for a long time and continues to be used. We can thus consider that the R&D activities of a firm are both the

result of the division of labour in knowledge production. We expect these activities to be partly determined by scientific developments and partly by the industrial sector in which they are carried out.

The various elements of knowledge used, be they more science or more applications related, need to be co-ordinated. Thus, the patents of a firm are not independent but linked. We can detect these linkages by several means, for example by the technological classes or by the key words contained in the text of the patents. Patents sharing common key words or technological classes are linked and the frequency of co-occurrence is proportional to the intensity of the linkage. We can thus detect and map the networks of which knowledge is constituted. The techniques used to detect the co-occurrence, and thus the linkages, will be discussed later. For the moment it is important to stress that these techniques can give us an interesting and useful picture of the KB of a firm.

A crucial question in what follows will be 'To what extent are the maps and measures we construct a good representation of the actual KB?'. Although we will have to come back to this question in the discussion of the techniques and of the results, we can begin to address it now. Returning to our definition of the KB, we only know that it is created by the interaction of the various types of knowledge that the firms uses. In a firm in one of the biotechnology based sectors different types of knowledge are present. Clearly, knowledge based on biotechnology has to play a fundamental role, but it is not the only type: for example, knowledge of the regulations required to test and market drugs, agrochemicals or food is almost equally important. Thus, in principle a complete map of the KB should include all the types of knowledge that the firm uses to create its final output, or RTP. Information on all these types of knowledge is rarely available. In what follows we will rely on patent statistics to create a representation of the KB. During this period the challenge for LDFs was to internalise the new biotechnological knowledge. The other types of knowledge required to create final products did not change substantially and we can expect these other competencies not to be those that create competitive advantages for firms. The component of the KB that is likely to have contributed the most to the competitive advantage of firms in the sectors considered is the scientific and technological component. Thus, by relying on patent statistics we capture the most important part of the KB during the period studied.

We can also ask to what extent our maps and measures are a good representation of the division of labour in R&D processes within a firm. The answer to this question depends on the correspondence between the technological classes contained in the map and the allocation of tasks to individuals and teams. Such correspondence is unlikely to be perfect. For example, in some cases researchers in different teams might collaborate in the creation of the knowledge leading to a given patent. However, for a number of reasons we can expect the map to be still related to the division of labour in the firm's R&D activities. First, a large part of the patents are likely to originate from a team. Second, even when a patent originates from several teams it is quite likely that the combined competencies are complementary and that they correspond to the technological classes contained in the patent. Third, we can expect the KB of a firm to change as the firm strategy changes, for example by incorporating new scientific disciplines and new industrial applications. In this case the composition of the R&D personnel is going to change, for example by incorporating a growing percentage of researchers competent in the new disciplines and industrial applications. In the mean time we expect the patents produced and the technological classes they contain to change reflecting the new composition of R&D activities. Thus, the time profiles of the individuals-activities map and of our patents-technological classes map can be expected to be very similar. We expect the map we obtain to be an *approximate* representation of the division of labour in R&D activities, but to be a good enough approximation to deserve further study. Within the limits of this approximation the techniques we used in this paper show us both the division of labour, identified as the nodes of the network of knowledge created by the firm, and coordination, identified by the links between the different technological classes and key words belonging to different patents.

In addition to this organisational correspondence between the structure of the KB and the map we obtained, there is another, more 'epistemological', correspondence between the two. Knowledge can be considered as a correlational structure, since scientific theories correlate the variables corresponding to the observables of a given subset of the external environment of the firm (Saviotti, 1996, 1999). Thus knowledge is in its very essence a network, and the networks we detect are in principle morphologically compatible with the structure of knowledge and adequate to represent it.

The extent to which the representation of the KB that we develop in this paper is applicable to firms in other sectors depends on the sector considered. We chose biotechnology related

sectors both because we have a professional interest in them and because they are the sectors for which the approximation to the KB that we develop here is likely to be better than for other sectors. In general the pharmaceutical and chemical are the sectors for which a closer correlation is found between R&D expenditures and patent production. Even in these sectors the KB of the firm contains other components, such as financial, marketing and legal competencies, but during the period under study such competencies do not confer a competitive advantage to any LDF with respect to its competitors. During the period under study the main task facing firms in biotechnology based sectors was to internalise the new biotechnological knowledge. Other industrial sectors are characterised either by a lower science intensity or by a different mixture of types of knowledge. According to Pavitt (1998) the management of interfaces is a very important aspect of knowledge management in large corporations. Yet the important interfaces are sector dependent. For example, while in pharmaceuticals and chemicals the interface between industrial R&D laboratories and academic research is very important the critical interface in automobiles is that between industrial R&D laboratories and production. Thus we expect the representation of the KB that we develop here to be a better approximation for the biotechnology based than for other sectors. However, we do not expect our representation to be completely inapplicable to other sectors. Rather we think that in the case of other sectors this approach will need to be complemented by information about types of knowledge that are not clearly reflected in patents. To start with the biotechnology based sectors gives us the possibility to test the method in the case in which it is likely to perform best. Adaptations of the method to other sectors will then become possible.

3) THE MAPPING AND MEASUREMENT OF THE KNOWLEDGE BASE.

Starting from the considerations of the previous section we can begin to define the KB by means of its composition, that is the list of technologies used by the firms. This list can be accessed by means of the technological classes assigned to each patent by patent examiners. We can expect that no two firms will have the same list of technological classes. Yet what would still be missing is a measure of the *distribution* of technological classes within the KB and of their *interactions*. In the KB of a firm elements of knowledge are not isolated, but they are used jointly. The structure of the KB of a firm is then a network of interconnected elements of knowledge. Even where the individual elements of knowledge were the same in two KBs, they could lead to very different outcomes both for what concerns knowledge accumulation and economic returns, depending on their patterns of interactions. A list of the

technological classes present in the KBs of the two firms is an interesting piece of information but it leaves us far from fully understanding their structure. The concept of structure involves both the elements of the KB and their interactions. The structure of any piece of knowledge can then be understood as the combination of the constituent elements and of the links between them. Second, the individual elements of knowledge that we find in firms' KBs are defined by the prevalent form of intellectual division of labour in society. Disciplines, specialities etc. are the result of such a process. As it happens with any process of division of labour, final outcomes can only be obtained if there is co-ordination of the individual steps of a 'production' process. Production is here written in inverted commas because it must be understood in a fairly general manner. For example, it might mean the production of knowledge or of any immaterial outcome. Thus the structure of links joining the elements of a KB is the result of the prevailing division of labour and co-ordination in the production of knowledge. Such division of labour is defined outside the firm, in its selection environment, and it can be expected to influence in the same way the KBs of different firms. However, we can also expect each firm to superimpose upon an externally determined structure its own specific contribution. The actual KB of each firm is likely to be the result of the intellectual division of labour in the society in which the firm operates and of specific influences internal to the firm. Such internal influences are likely to be very path dependent, that is to depend on the history of the firm. The study of the interactions between the elements of knowledge of a KB is, therefore, of fundamental importance.

The two methodologies that we will describe in what follows are intended to map and measure these two aspects of the KB.

3.1) THE TECHNIQUES USED

3.1.1) LEXICOGRAPHIC ANALYSIS.

Lexicographic analysis allows us to identify a number of key words in a particular text and to establish the extent of correlation between these key words based on their co-occurrence (de Looze et al, 1999). The more often the key words co-occur, the more closely related they are. In fact, the strength of the links between key words is calculated as the frequency of their co-occurrence. The technique is described in greater detail in Appendix 1. The networks of knowledge thus constructed can be displayed graphically and provide us with a very intuitive image of the structure of firms' KB.

3.1. 2) MEASURES OF THE PROPERTIES OF THE KB.

The knowledge base of a firm can vary in many ways. For example, it can be concentrated on very few topics or be distributed over a very large spectrum of subjects. The former will be a very specialised KB, the second a much more general one. In addition we can expect that the integration of new fields of knowledge introduced by a firm in order to change its KB will disorganise the structure of the KB. Thus new fields of knowledge will not be perfectly coordinated amongst themselves or with pre-existing ones as soon as they are introduced. Their integration within the KB can be expected to improve in the course of time. Thus the KB can have a number of interesting properties that it is worthwhile to measure.

The concept of coherence has been used very often in the literature on the economics of the firm. However, while it is a concept of great intuitive appeal, it is not easily definable in an operational sense. Teece et al (1994) proposed a method to measure the coherence of the firm and the used it to test their hypothesis that, at least within certain types of environment, coherent firms were more likely to survive than incoherent ones. An example of incoherent firm would be a conglomerate producing a set of completely unrelated products. On the other hand, a coherent firm would have been characterised by a set of related products. In the approach by Teece et al (1994) coherence was defined as *relatedness*. They argued that related products are more likely to be produced together than unrelated ones, due for example to economies of scope.

In the case of firms in highly knowledge intensive sectors it is at least equally interesting to calculate the coherence of their KB. To the extent that a KB is a precursor of future products, but also taking into account that the KB does not need to map exactly onto the structure of the firm's products at a subsequent time (Pavitt et al, 2000), it is important to measure not only the coherence of the firms products, but also that of its KB. If the KB represents the crucial resource of the firm, its coherence is likely to be an important determinant of the firm's performance. In the case of the so called life science company, the model followed by most firms in the 1990s, it was considered advantageous to have a common knowledge base in order to produce a set of heterogeneous products. Although the coherence of the KB was not explicitly mentioned, it follows that within the model of the life science company the coherence of the KB was privileged with respect to that of the outputs.

The principle on which the measure is based is that related products are likely to be produced jointly more frequently than unrelated products. Considering that there is a probability that any two products can be produced jointly accidentally, the degree of coherence of any set of products is obtained by comparing the observed frequency of co-occurrence with the calculated probability that the products occur together randomly. In what follows an adaptation of the technique used by Teece et al to measure the coherence of the KB and developed by (Nesta, 2001) is described. Conceptually this involves replacing the firm's products with the elements of knowledge it uses. In our case the elements of knowledge are the technological classes associated to patents. Apart from the actual technique used in its measurement, an important problem arises about the meaning of coherence. Such a problem is more acute in the case of the KB, but it exists also in the case of products. Relatedness is an ambiguous concept in the sense that both similar and complementary activities can be said to be related. In the case of the KB it is quite likely that most of the technological classes included in the KB are complementary rather than similar. To the extent that they are the result of a process of division of labour, most of these technological classes are unlikely to be similar. On the other hand, a certain amount of similarity is required in order for the different technological classes to be coordinated. Thus, we expect the degree of coherence that we measure to be predominantly influenced by the complementarity of the technological classes included and to a smaller extent by their similarity.

As pointed out above, if two activities i and j are related, they are more likely to occur together than completely unrelated activities. Eq (4) measures the probability that any two activities occur together randomly. From that probability we can calculate the expected value μ_{ij} and the standard deviation σ_{ij} of the frequency of co-occurrence of the two activities (Eqs 5 and 6). If C_{ij} represents the measured frequency of co-occurrence i and j , then r_{ij} (Eq 7) measures the degree of relatedness of the two activities or, in the view of Teece et al, their coherence. In the case of firms having many activities the indicator r_{ij} has to be averaged over the activities of the firm. This is obtained by means of Equations (8) and (9), in which a weighted average of r_{ij} is calculated with respect to all possible pairs of activities (Eq 8) and with respect to the activities of firm f (Eq 9).

$$P [X_{ij} = x] = \frac{\binom{N_i}{x} \binom{T - N_i}{N_j - x}}{\binom{T}{N_j}}$$

$$\mu_{ij} = E (X_{ij} = x) = \frac{N_i N_j}{T}$$

$$\sigma_{ij} = \mu_{ij} \left(\frac{T - N_i}{T} \right) \left(\frac{T - N_j}{T - 1} \right)$$

$$COMP_{ij} = r_{ij} = \frac{C_{ij} - \mu_{ij}}{\sigma_{ij}}$$

(4)-(7)

Eq (7) then measures the extent of relatedness, that we interpret as mainly complementarity between technologies i and j. It is obtained by subtracting from the observed frequency of co-occurrence of technologies i and j its expected value and by dividing the result by the standard deviation. The degree of coherence of firm i with respect to all the technologies it uses, that is the coherence of its KB, is then obtained by calculating first what Teece et al called the Weighted Average Relatedness (WAR) (Eq(8)) for the firm and then its weighted average with respect to all the technological classes used by the firm f:

$$WAR_i = \frac{\sum_{j \neq i} r_{ij} P_j}{\sum_{j \neq i} P_j} \quad (8)$$

$$COH_f = \sum_{i=1}^n \left[\frac{WAR_i P_{fi}}{\sum_i P_{fi}} \right] \quad (9)$$

Let us now describe the other measures of the properties of the KB used in Nesta (2001).

THE DEGREE OF SPECIALISATION.

The calculation of the degree of specialisation began with that of the Relative Technological Advantage (RTA) of firm f in technology i :

$$RTA_{if} = \frac{P_{if}}{\sum_i P_{if}} \bigg/ \frac{\sum_f P_{if}}{\sum_{if} P_{if}} \quad (10)$$

The degree of specialisation is then derived from RTA_{if} as the ration of its standard deviation to its mean:

$$SPE_f = CV_{RTA,f} = \frac{\sigma_{RTA,f}}{\mu_{RTA,f}} \quad (11)$$

(2)

SCOPE OF THE KB

The scope of the knowledge base of the firm is measured by the technological portfolio of firm f , defined as the number of technological classes in which the firm has patents.

$$PORTEK_f = \sum_{i=1}^{30} C_{if} \quad (12)$$

(3)

Of course, the scope of the firm also corresponds to the degree of differentiation of its KB.

Eq (7) then measures the extent of relatedness, that we interpret as mainly complementarity between technologies i and j . It is obtained by subtracting from the observed frequency of co-occurrence of technologies i and j its expected value and by dividing the result by the standard deviation. The degree of coherence of firm i with respect to all the technologies it uses, that is the coherence of its KB, is then obtained by calculating first what Teece et al called the Weighted Average Relatedness (WAR) (Eq(8)) for the firm and then its weighted average with respect to all the technological classes used by the firm f :

(8)

$$WAR_i = \frac{\sum_{j \neq i} r_{ij} P_j}{\sum_{j \neq i} P_j}$$

(9)

$$COH_f = \sum_{i=1}^n \left[\frac{WAR_i P_{fi}}{\sum_i P_{fi}} \right]$$

SIMILARITY

This is a relative property of two different KBs:

$$SIM_{ij} = S_{ij} = \frac{\sum_{k=1}^n C_{ik} C_{jk}}{\sqrt{\sum_{k=1}^n C_{ik}^2 \sum_{k=1}^n C_{jk}^2}}$$

where C_{ik} and C_{jk} measure the co-occurrences of technologies i and j with a third common technology k .

The techniques previously described seem considerably different. Lexicographic analysis allows us to obtain a graphic representation of the KB of a firm. Such a representation has the advantage of being very intuitive. Representations of the KB obtained at different times provide us with a picture of how the firm's KB changes in the course of time. However

lexicographic analysis is not well suited to quantitative measurements. On the other hand, the other measures based on patent statistics can be used both at the level of aggregation of the firm and at higher levels of aggregation. Thus, the two techniques are complementary in that one provides a more intuitive representation of the KB and the other one a more accurate representation. In the description of the results it will be explained how the joint use of the two techniques allows us to provide a more complete representation of the KB.

4) RESULTS

The results that are going to be described here are taken from a number of papers and from a PhD thesis. These results have been obtained in some cases by studying a small set of firms, in other cases by means of data bases containing information about a large number of firms.

4.1) LEXICOGRAPHIC ANALYSIS: THE STUDY OF AVENTIS.

Aventis was formed in 1999 by the merger of Rhône Poulenc and of Hoechst. Both companies were previously chemical and, at different times starting from the end of the 1980s, decided to change the nature of their activities and to become life science companies. Like many other firms who had previously undertaken the same strategic reorientation Hoechst and Rhône Poulenc had to change their KB. When this strategic reorientation was already underway on both sides the two firms decided to merge. In a merger we expect the resources of the two companies to be reorganised in order to take advantage of the potential synergisms. In the case of firms in highly knowledge intensive sectors we can expect their KBs to be reorganised. Thus the KBs of the two firms needed to change both to incorporate new knowledge and to adapt to the merger. The results described here are taken from two papers (Saviotti, et al, 2002a, 2002b) that contain a more detailed description of the study.

The study covered the period 1990-1998. As we see from Fig 1 the percentage of biotechnology patents of the two forms had already started rising before the merger. Biotechnology patents started rising earlier for Rhône Poulenc because it undertook the change of strategy towards the life sciences earlier than Hoechst.

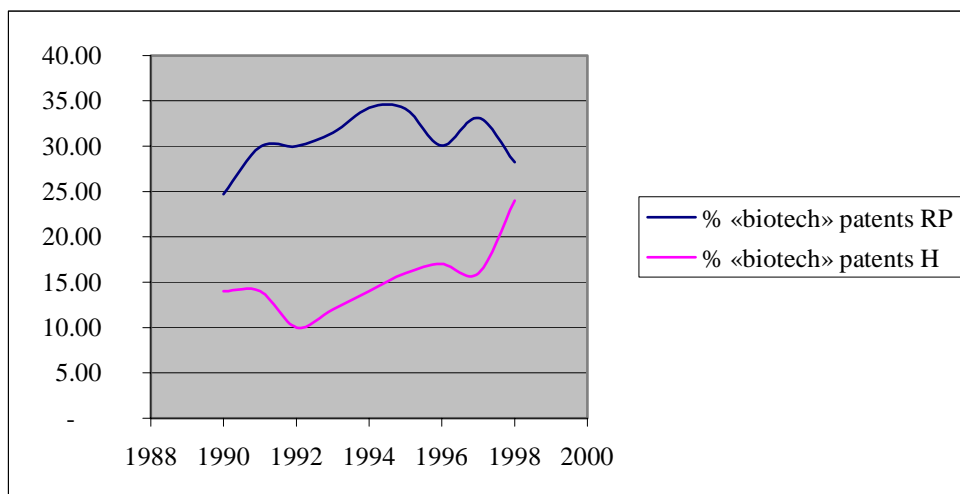


Figure 1: Evolution of the shares of «biotech» patents for Hoechst and Rhône Poulenc, 1990-1998

Fig. 2 shows the graphic representation obtained by means of lexicographic analysis for the KB of Hoechst for the period 1990-1992. The KB is here represented as a network whose nodes are the technological classes used by the firm and whose links indicate which technological classes are used jointly. The numbers attached to each link indicate the frequency of co-occurrence, or the strength of the link. Within the diagram we can see that some technological classes (e.g. A61K, C07C, C07F, C09G, C09D) are much more central than others, because they are linked to a greater number of other classes with links of higher strength. Furthermore, the class A61K, 'Preparations for Medical, Dental, or Toilet Purposes', is the main class containing compounds used for pharmaceutical purposes. A61K separates two subsets of the map in Fig. 2. The classes that are NW of A61K are those mostly related to, biotechnology while the ones that are placed E and SE of A61K are classes mostly related to the technologies that were previously used by Hoechst, that is chemical technologies. It is to be observed that some of these chemical technologies, such as C07C, acyclic or carbocyclic compounds, keep being used in the pharmaceutical industry. Thus, although the KB changes we cannot expect all the old technologies to disappear instantly to be replaced by new ones. The actual process of change is more likely to be a gradual change of composition in which new technologies increase progressively their weight while old technologies decline. Also, some old technologies may always be required in combination with new technologies. Of course, we expect the composition of the KB to change more rapidly during periods of radical than of incremental change.

Going back to Fig. 2 we can see that the NW corner represents the new technologies that Hoechst is trying to acquire and the rest of the map the technologies that the firm was previously using. To the extent that the firm carries out the stated strategy to become a life science company we can expect the weight of the biotechnology based classes to increase.

In Fig. 3, representing the map of the KB of Hoechst during the 1996-1998 period, we can see that the number of biotechnology related classes decreases slightly from 9 to 8 while the number of chemically related classes falls from 23 to 13. Thus, the weight of biotechnology classes increases during the period 1990-1998. A similar though not identical evolution was followed by Rhône Poulenc, as shown in Figs 4 and 5. Here we can observe the same partition in biotechnology related classes West of A61K and of chemically related classes East of A61K. The biotechnology subset of the map seems at the beginning better structured than the chemical subset. Also, the internal connectivity of the biotechnology subset of the map seems to increase in the period observed. However, the proportion of chemical classes increases rather than falling. This probably does not indicate a reversal of strategy, as the growing percentage of biotechnology patents indicates (Fig. 1), but rather a desire to rationalise the surviving chemical activities. On the whole we can conclude that the study of the KBs of the two firms shows that both of them were changing their KB according to their stated strategy of becoming life science companies.

After the merger Aventis was considerably reorganised, taking into account also the changed external environment. At the middle of the 1990s most firms in the biotechnology based sectors were still following the model of the life science company, but such model has now been abandoned by all of them. The rationale of the life science company consisted in the existence of a knowledge base common to a series of activities that had traditionally been carried out separately. After the advent of the new biotechnology it was assumed that to concentrate within the same firm a series of heterogeneous activities could give rise to important economies of scope. It is to be observed that this strategy implicitly meant that to increase the coherence of the KB could compensate for the heterogeneity of the markets that would thus be brought under the same roof. This strategy would seem rational when the relative importance of knowledge with respect to other resources is expected to grow, as we move towards the knowledge based society. The subsequent evolution of the concept of the life science company seems to indicate that either we are still far from the knowledge based society or there have been short term fluctuations that made the same strategy temporarily not viable. We also have to take into account that a new type of knowledge, such as the new biotechnology, will subsequently undergo a process of differentiation that might make it more convenient to specialise in

given subsets of an initially unique KB. Thus, the balance of advantages and disadvantages to be obtained by adopting a common KB to produce a range of heterogeneous products may change in the course of time. The actual strategic evolution of most firms in the biotechnology based sectors consisted in abandoning the concept of the life science company and in specialising either in pharmaceuticals or in agrochemicals. Several firms (Novartis, Astra Zeneca, Pharmacia etc) have completely separated their agrochemical from their pharmaceutical activities. After the merger Aventis did the same by selling Crop Science to Bayer. In what follows the KB of Aventis after the merger and of some of its subsidiaries will be represented.

Fig. 2. Diagram for period 1 (1990-1992) of the co-occurrences between the main IPC classes of Hoechst:

(the most central classes are represented in dark grey)

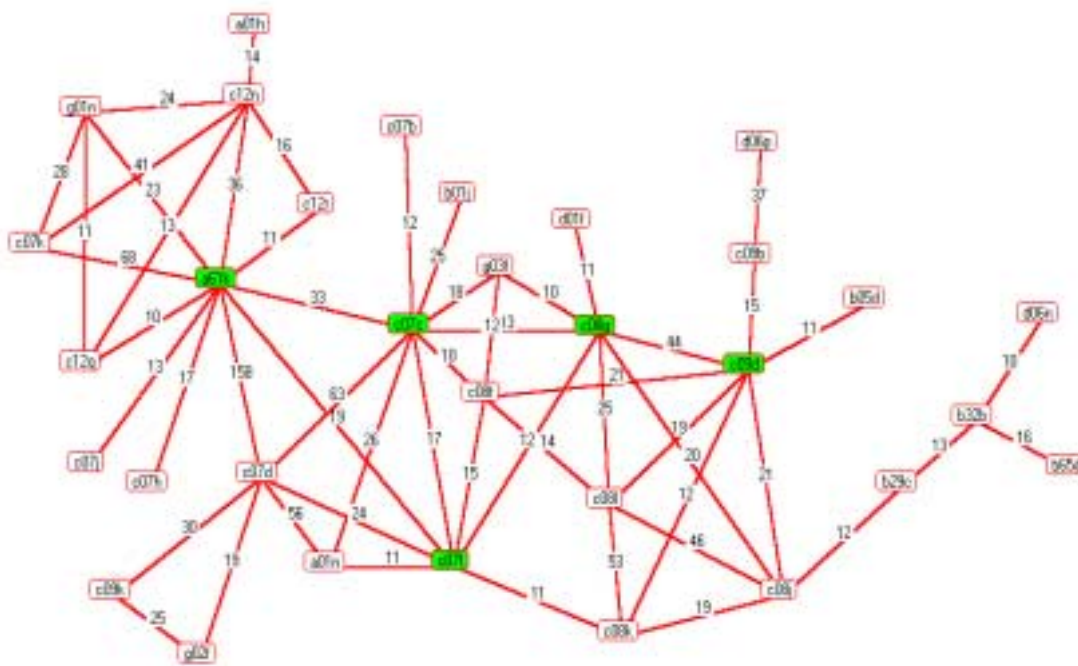


Fig. 4. Diagram for period 1 (1990-1992) of the co-occurrences between the main IPC classes of Rhône-Poulenc (the most central classes are represented in dark grey)

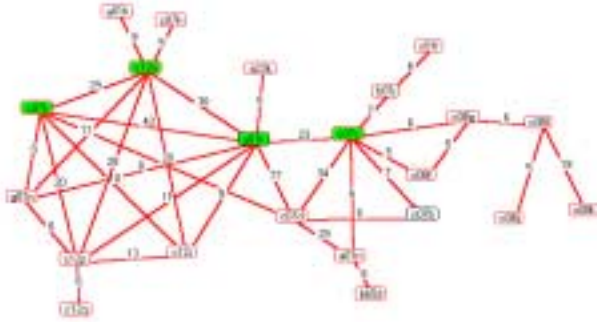
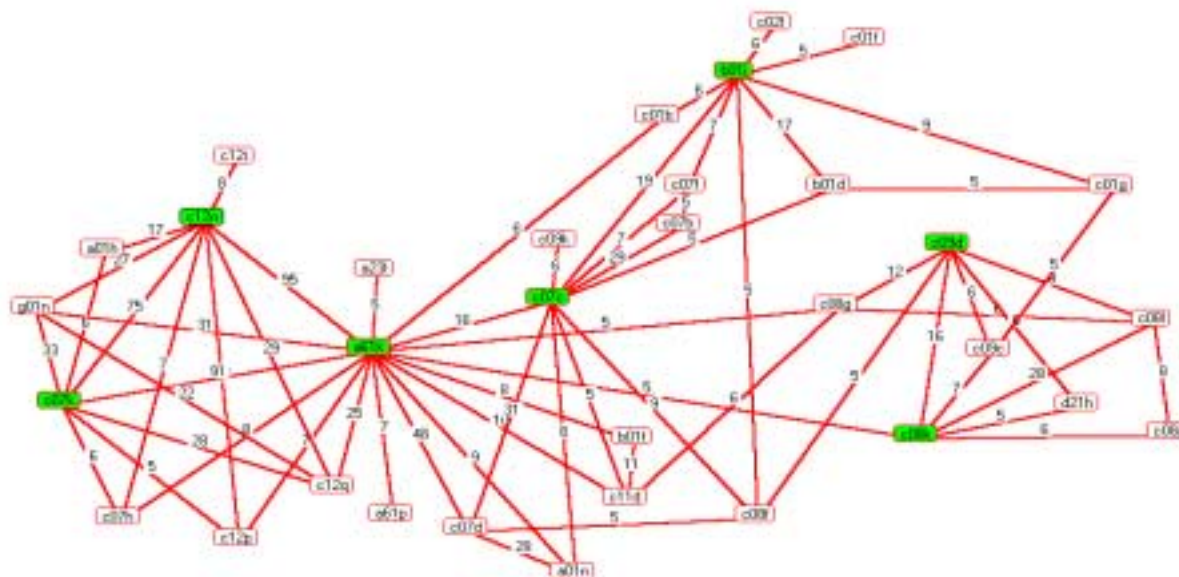


Fig. 5. Diagram for period 3 (1996-1998) of the co-occurrences between the main IPC classes of Rhône-Poulenc



After the merger Aventis underwent a considerable reorganisation. In 1999 Aventis had become a life science company, performing two types of activities, pharmaceuticals and agrochemicals (Fig. 6). However, subsequently, and following the same strategy as most other life science companies, it sold Crop Science to Bayer, thus becoming a pharmaceutical company. Figs 7-12 show the KBs of Aventis, of Aventis Pharma, of Pasteur Merieux, of Aventis Behring and of Crop Science.

The KB of Aventis as a group (Fig. 7) does not show any more the separation between biotechnology classes and chemical classes that was evident in both Rhône Poulenc and Hoechst before the merger. The main technological classes of both types are still present, with A61K being the most central one. However, the weigh of the chemical classes has fallen and the whole KB seems a much more closely integrated network. The map of the KB of Aventis after the merger provides clear evidence that the transition towards the life science company had been achieved.

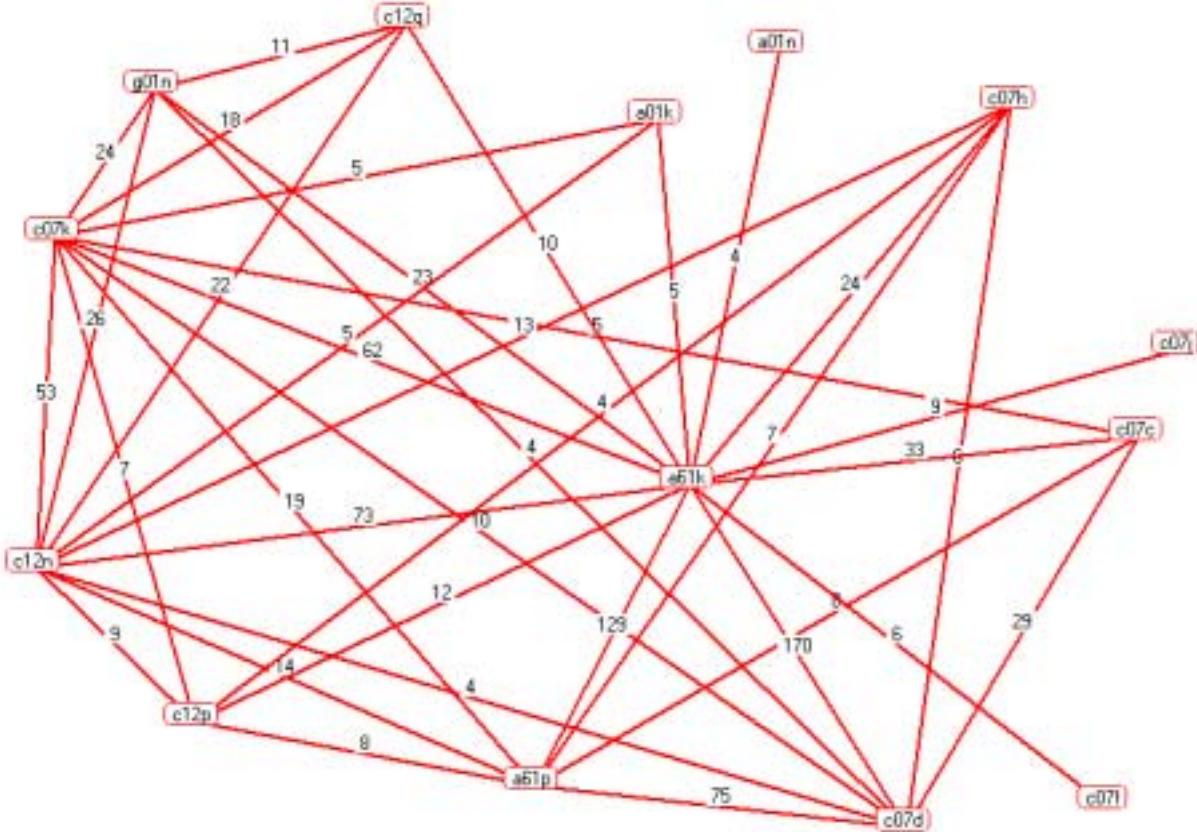


Fig. 8. The knowledge base of Aventis Pharma, as represented by the network of the technological classes of the firm and by their links.

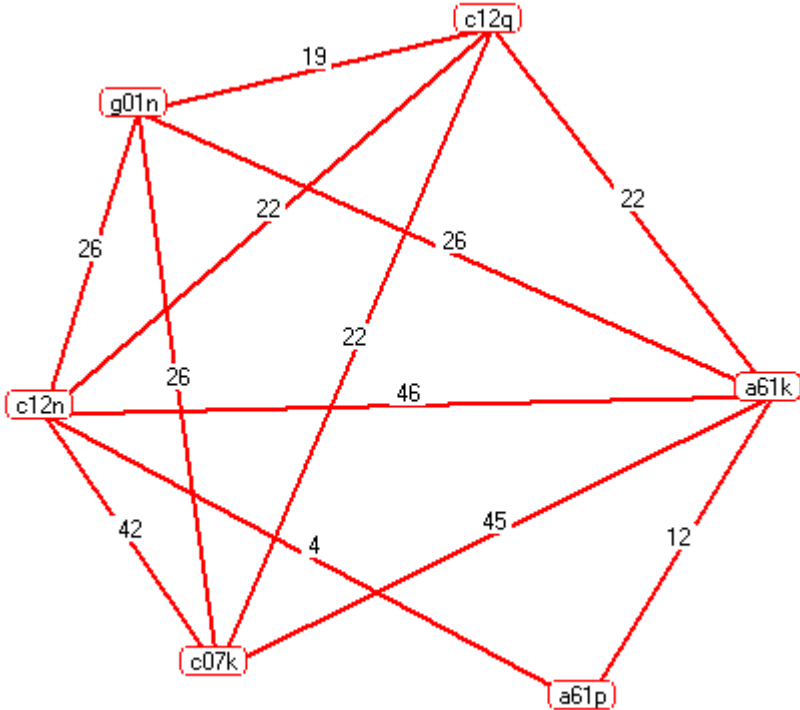


Fig. 9. The knowledge base of Aventis Pasteur, as represented by the technological classes of the firm and by their links.

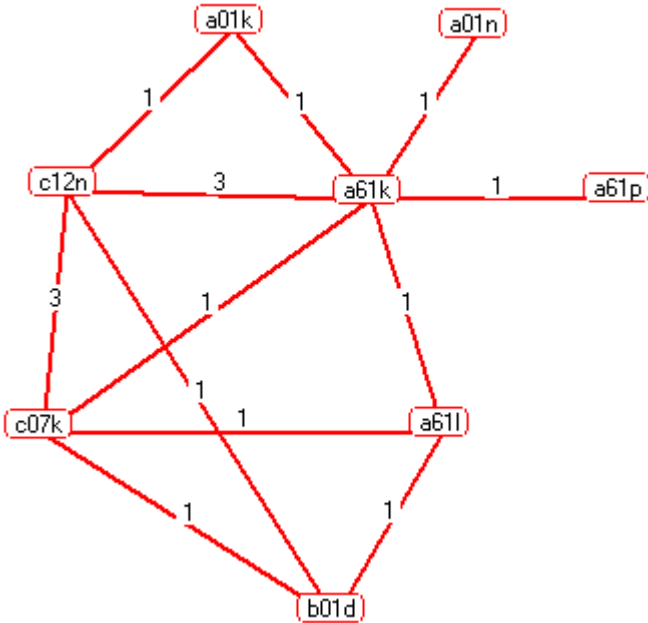


Fig. 10. The knowledge base of Aventis Behring.

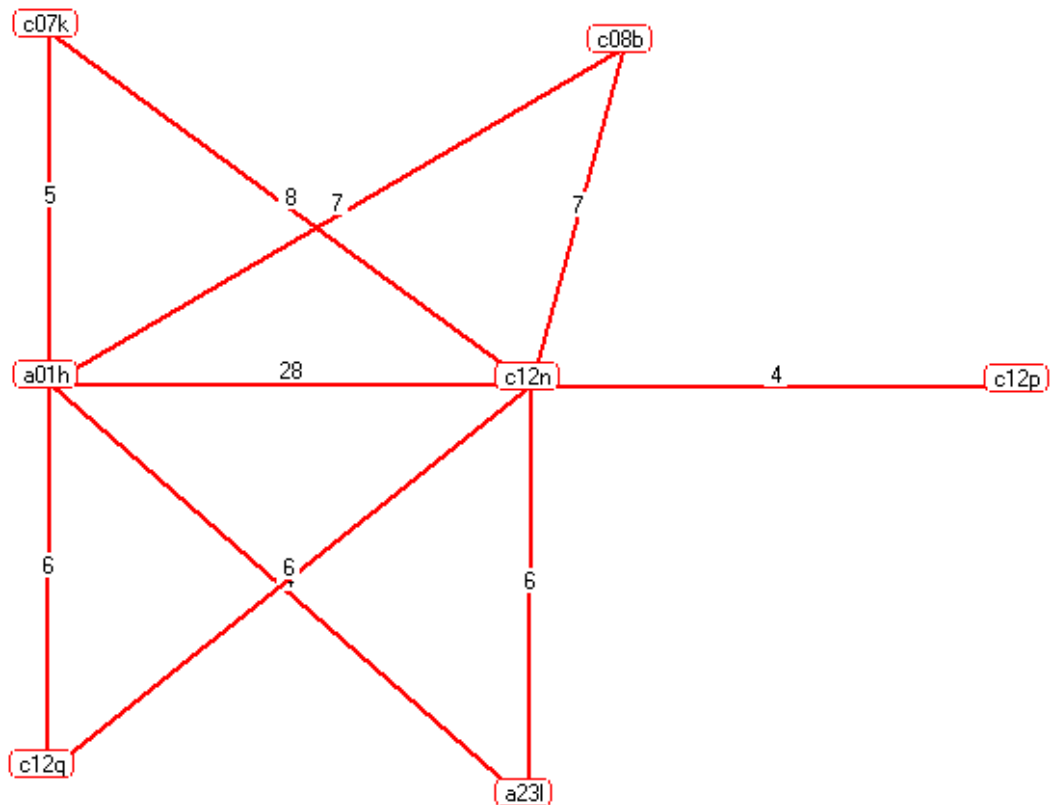


Fig. 11. One of the two networks composing the KB of CropScience, as represented by the technological classes used by the firm and by its links. This network corresponds to a 'plants' pole.

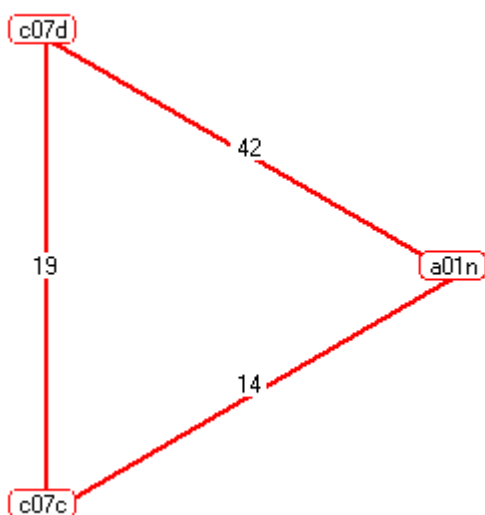


Fig. 12. The second network in the KB of CropScience, as represented by the technological classes used by the firm and by its links. This network corresponds to a 'chemical' pole.

If we compare the KBs of Aventis and of Aventis Pharma we see that they are very similar. In both cases A61K is the most central class. All the other highly central technological classes are common to the two maps. Thus A61P, C07C, C07D, C07H, C12N, C12P, given here in alphabetical order, are the most central classes in both KBs. It is to be observed that C07C and C07D, 'Acyclic or carbocyclic compounds' and 'Heterocyclic compounds' respectively, are the chemical classes that were the components of the KB of pharmaceutical firms before the advent of biotechnology. These technological classes do not disappear, but are integrated within the new KB. In summary, the KB of Aventis Pharma seems a reduced version of that of Aventis with lower levels of connectivity.

The KBs of the other subsidiaries seem considerably different. All of them contain a much more reduced number of classes and have lower levels of connectivity. This can be explained both by the smaller size and by the greater degree of specialisation of the subsidiaries. For example, Aventis Pasteur is specialised in the production of vaccines and thus uses only the technological classes that are relevant for this purpose. The lower level of connectivity follows inevitably from the smaller size, involving a lower number of patents.

The KB of Crop Science has been reported here (Figs 11 and 12), although the firm has now been sold by Aventis to Bayer, in order to find out how integrated its KB was with that of Aventis. As it turns out, the KB of Crop science is split into two apparently non connected parts. The first (Fig. 11) corresponds to a 'plants' pole while the second (Fig. 12) corresponds to a 'chemical' pole. On the one hand this shows that Crop Science managed to integrate modern biotechnology. On the other hand, the existence of the two networks is justified because agrochemical firms need to use both sets of competencies. As it was the case for Aventis Pasteur and for Aventis Behring, no links are found between the KB of Crop Science and those of the other subsidiaries of the group. At this time we cannot tell whether this result indicates that the KB of Aventis is segmented, without communications between subsidiaries, or whether the communications that take place are not revealed by the techniques we use. Furthermore, it is quite possible that the presence and type of links between the KBs of the subsidiaries of a firm depend on the firm organisation: some firms may have a very integrated structure, in which different subsidiaries have a high degree of interaction, while others may keep subsidiaries very independent. This finding raises a number of questions, both methodological and related to the nature of the KB. Clearly, the problem requires further research. For what concerns mergers and acquisitions, the separation of the KB of a

subsidiary should make easier to separate it from the rest of the firm, for example by selling it off, than if the KB of the same subsidiary were closely integrated with the rest of the firm.

4.2) PROPERTIES AND PERFORMANCE OF THE KB.

Firms are not generally knowledge producers. They use knowledge in order to produce products or services by means of which they compete. However, in highly knowledge intensive firms the production of knowledge is the crucial step in firm growth and competition. Thus we expect the properties of the KB that were described previously to have an impact on firm performance. This general hypothesis was tested by defining a number of dimensions of the performance of the firm and by regressing measures of these dimensions against a series of independent variables including the properties of the KB. The detailed description of these calculations can be found in Nesta (2001). In the present paper only a summary of the results will be supplied in order to be able to discuss the implications of the properties of the KB for firm behaviour and performance.

The dimensions of the performance of the firm chosen were the production of knowledge, measured by the number of patents produced, the rate of profit and Tobin's Q. The former is not an indicator of final performance but of intermediate output. Tobin's Q is the ratio of the stock exchange value of a firm to the value of its assets. The set variables used in the regressions is shown in Table 1 and the results are shown in Table 2. The sample of firms used in the initial runs included 99 firms from different countries, whose main activity was distributed over a number of sectors.

The results obtained show that both the differentiation and the coherence of the KB are important determinants of the firm's performance at the three levels considered. Yet this importance varies during the period studied, in particular between the 1980s and the 1990s. The degree of differentiation turns out to be relatively more important during the 1980s while coherence turns out to be relatively more important during the 1990s, both for what concerns the rate of profit and Tobin's Q. These results seem to indicate the presence of structural differences between the two periods considered. In fact the 1980s were a period during which biotechnology was an entirely new technology, radically different from the knowledge base previously used by firms in the pharmaceutical and agrochemical sectors, which was based on organic chemistry. Genetic engineering started being industrially promising only in the 1970s. Incumbent pharmaceutical and agrochemical firms had a very limited absorption capacity for

the new technology, and many of them were not even convinced that genetic engineering was going to be the basis of future industrial developments. The first large diversified firms (LDFs) that tried to internalise the new knowledge had to rely on contracts with dedicated biotechnology firms (DBFs) (Grabowsky, Vernon, 1994; McKelvey, 1996). It took most of the 1980s for LDFs to acquire a sufficient absorption capacity in the new biotechnology. During this period it became evident that genetic engineering was going to be the basis for a large range of new industrial applications, but which new applications were going to be fruitful was not yet clear. The leading strategic imperative for firms was to enter as many new areas of the new biotechnology as possible, that is, to increase the differentiation of their KB. By the end of the of the 1980s a number of new techniques an applications had emerged that seemed to provide more stable avenues for progress. For example, the unexpectedly rapid progress in the human genome project, due essentially to the acceleration of the operation of sequencing, gave rise to the expectation that gene therapy was going to be an important source of future applications. Also, the emergence of technologies that were transversal within biotechnology, such as bio-informatics, changed the dynamics of knowledge generation and provided important new opportunities for the creation of SMEs based on a new specialisation (Saviotti et al. 2000). In the changed circumstances of the 1990s firms started not only to explore new technologies and applications, but began to integrate better the different components of their KBs. In other words, firms started to increase the coherence of their KBs. In fact, this did not entail an end to the differentiation of their KBs, but it meant that differentiation was accompanied by the preoccupation to increase the complementarity of the components of the KB. Furthermore, it is possible to argue that the transition from the predominance of differentiation to the predominance of coherence corresponded to the passage from exploration to exploitation (March, 1991).

That an important change took place at the beginning of the 1990s is confirmed by a study of the dynamics of innovation networks in the pharmaceutical sector over the period 1980-2000 (Orsenigo et al , 2001). The authors of this study found that the roles played by different actors within innovation networks changed markedly at the beginning of the 1990s. They explained these changing roles by the emergence of transversal, or general, technologies within biotechnology. They hypothesize a process of knowledge growth in which a new field is founded by some new hypotheses. The subsequent development of these hypotheses gives rise to more specialised ones, in such a way that the field becomes organised as a hierarchy of hypotheses in which the earlier ones are more general and the subsequent ones more specific.

Table 1. The variables used in the regressions to test the effect of the KB on firm performance (Source Nesta, 2001)

Measures of performance. Dependent variables			
Patents	DBA	Production of knowledge	Firm/year
Profit	WGR	Short term efficiency of the firm	Firm/year
TobinQ	WGR	Stock exchange value of the firm	Firm/year
Cognitive aspects of the firm-Properties of the KB			
COH	DBA	Degree of coherence	Firm/year
PORTEK	DBA	Technological portfolio	Firm/year
SPE	DBA	Degree of specialisation	Firm/year
RD	WGR	Current R&D effort	Firm/year
CA	DBA	Absorption capacity	Firm/year
Knowledge flows (indep. Variables)			
POOL	DBA and WGR	External sources of knowledge	Firm/year
Description of the firm			
GRFIRME	Internet	Size, Country, date of creation	Firm
ACTIVITE	Internet	Main activity of the firm.	Firm
ACTIFS	WGR	Firm 'real' assets	Firm/year
PARMARCHE	WGR and Internet	Market share	Firm/year

Table 2. The effect of the properties of the KB on firm performance. The variables of group 1 are properties of the KB, those of group 2 depend on scale effects, those of group 3 represent external effects, those of group 4 other properties of the firm. (Source Nesta, 2001)

		Production of knowledge				Profit				Tobin's Q			
		1980s		1990s		1980s		1990s		1980s		1990s	
1	Coherence	Pos.	-	Pos.	Sig.	Neg.	-	Pos.	Sig.	Neg.	-	Pos.	Sig.
	Differentiation	Pos.	Sig.	Pos.	Sig.	Pos.	-	Pos.	-	Pos.	Sig.	Pos.	-
	Specialisation	Pos.	Sig.	Pos.	Sig.	Neg.	Sig.	Neg.	Sig.	Neg.	Sig.	-	-
2	R&D expenditures	Pos.	Sig.	Pos.	Sig.								
3	Absorption capacity	Pos.	Sig.	Pos.	Sig.								
	Knowledge flows	Neg.	Sig.	Neg.	Sig.								
4	Market share					Pos.	Sig.	Pos.	Sig.	Neg.	Sig.	Neg.	Sig.
	Assets					Pos.	Sig.	Pos.	Sig.				
	Profit (residual)									Pos.	Sig.	Pos.	Sig.

Thus it seems that an important transition took place in biotechnology at the end of the 1980s, a transition that led to both an enhanced differentiation of biotechnology and to new strategic and organisational priorities for firms.

5) DISCUSSION

This paper stressed the importance of the KB for the behaviour and performance of firms in a knowledge based society. Clearly, in these conditions it is important to be able to represent the KB and to measure its properties. Two different techniques that can be used for this purpose have been described. Lexicographic analysis can give a more intuitive graphic representation of the KB, but its capacity to derive quantitative objective measures is inferior to the other technique based on patent statistics. The two techniques can thus be considered complementary tools for the study of the KB.

In principle, for knowledge intensive firms we expect the KB to influence most aspects of firm behaviour and performance. In particular, four aspects of firm behaviour seem to be of central importance and are being investigated:

1. Firm strategy
2. Firm organisation
3. Mergers and acquisitions, and their converse, demerger or divestiture.
4. Innovation networks

A change in firm strategy can be expected to lead to a change in both the composition and the structure of the KB. The results previously reported of a study of Rhône Poulenc and of Hoechst bring this out very clearly. A decision by the firms to change from chemical to life sciences firms was followed by a change of their KB in which the weight of biological technology classes increased gradually and these classes became better integrated within the KB. Clearly, we cannot expect all firms to adjust equally to the same change in strategy. Thus, we can conceive that different firms attempting to implement the same strategic change will modify their KB at different speeds and by incorporating a different mixture of new technologies. In this sense we can expect changes of strategy to be accompanied by some degree of path dependence. A study of comparable strategic transitions of other firms is underway in order to find out the range of paths that can be followed to implement strategic transitions involving a change of KB.

Several organisational arrangements can in principle be adopted by firms for what concerns the creation and utilisation of knowledge. Thus, when a firm has subsidiaries we can imagine that KBs of all the subsidiaries either to overlap or to be completely separated. Of course,

these two solutions only define the extremes of a range, with many intermediate solutions possible. The results we found for Aventis seem to show that the KBs of the subsidiaries are completely separate. As it was previously pointed out, we do not yet know whether these results are the consequence of the organisational structure of Aventis or of a partial blindness of the method. Studies of other biotechnology based firms will be carried out in order to both improve the method and to find out to what extent the fine structure of the KB is affected by firm organisation.

When firms merge or purchase other firms we expect their organisation and their KB to change. Both the relationships of the firms before the merger and the dynamics of it can affect its outcome. An interesting question arises as to what the best ex-ante combination of KBs is: is it better for the merging KBs to be similar or complementary? And to what extent? Furthermore, we expect a 'good' merger (but the same hold for acquisitions) to give rise to the exploitation of the potential synergisms between the two firms. On the other hand, we also expect the actual process of merging to cause temporary coordination problems, in such a way that the benefits of the merger can be obtained only with a certain delay. The results we have obtained so far lead to a number of provisional conclusions: first, although no measures of the similarity of the KBs Hoechst and Rhône Poulenc have so far been performed, the two firms seemed to have rather similar KBs; second, the KB of Aventis after the merger seems to follow the same path already started by the two firms towards the life sciences and towards a greater integration. Another interesting implication of previous results for mergers is that, to the extent that coherence is an important of firm performance, we expect the final outcome of a successful merger to led to coherent merged KB. Of course, such a coherent KB is unlikely to be obtained immediately after the merger. Indeed, the reorganisation process is likely to reduce the coherence of the KB below that of the firms before the merger. However, if the merged firm is to perform well, we expect the coherence of the merged KB to recover suitable values after a short delay. Studies of these aspects are underway.

Demergers are also likely to be affected by the KB of the firm. For example, we expect a subset of a firm with a tightly integrated KB to be more difficult to separate than a subset of a comparable firm that has a completely segmented KB. In this sense the apparent separation of the KB of Crop Science from that of Aventis could have facilitated its sale to Bayer. Such an hypotheses needs to be tested in the other studies being carried out.

Innovation networks are typically formed by LDFs, DBFs and by public Research Institutes. It is unlikely that these different actors play the same roles. Both empirical and theoretical studies (for example Pyka, Saviotti, 2002) show that a complementary relationship of the partners is more likely to lead to successful innovation networks. It is, however, possible for a minimum extent of similarity to be required in order for the partners to be able to communicate. Work is currently underway to test these propositions.

In summary, we can expect the KB of a firm to have important influences on several aspects of behaviour and performance of a firm. Some of these aspects and have been discussed here and are the object of further research.

6) SUMMARY AND CONCLUSIONS.

The central concern of this paper has been the representation and the measure of properties of the KB of firms in biotechnology based sectors. The KB has been defined as the collective knowledge that firms can use for their productive purposes. Thus the KB depends not only on the elements of knowledge acquired by individual members of the firm, but also on their interactions. In more fundamental terms, the KB depends on the division of labour and on the coordination inside the firm. The information required to construct a detailed map of individual competencies and of their interactions is both extremely costly and difficult to acquire. In this paper two approximate methodologies to represent and measure properties of the KB are described. The approximation is based on the use of information on patents, including both patent statistics and textual information. Of these two methods lexicographic analysis, a part of scientometrics, allows us to detect the technological classes used by firms and their links. The links are detected and their frequencies measured by means of the co-occurrence of key-words present in the text of the patents and associated to particular technological classes. The representation of the KB thus obtained consists of a network whose nodes are the technological classes used by the firm and whose links measures the interaction of different technological classes. In this paper the technique is applied to the study of the formation of Aventis from the merger of Rhône Poulenc and of Hoechst.

The construction of indicators of innovation based on patent statistics has been developed in the last twenty years. In this paper a number of indicators already developed have been used together with a new one. This new indicator measures the coherence of the KB of the firm.

The development of the indicator of coherence of the KB is based on the work of Teece et al (1994). They had developed an indicator of the coherence of the firm based on its output. Their technique has been adapted to the measure of the coherence of the KB (Nesta, 2001). In knowledge intensive industries we can expect the coherence of the KB to be at least as important as the coherence of the firm's output. Firms are now knowledge producers but the production of knowledge is not their main objective. Knowledge is used to produce goods or services by means of which firms compete. We can then expect the KB of the firm to have an influence on firm performance. It turns out that two properties of the KB, its differentiation and its coherence, are particularly important determinants of the firm performance. Furthermore, the ranking of these two properties as determinants of the firm's performance changes during the evolution of the technology. Differentiation is relatively more important during the 1980s while coherence becomes relatively more important during the 1990s.

The two techniques described above are complementary rather substitutes for the analysis of the KB of the firm. In the paper a number of possible applications described. Thus, we can expect the KB to affect and to be affected by firm strategy, by firm organisation, by mergers and divestitures, by the formation of innovation networks. A change in strategy involves changes in the firm's KB. Different types of firm organisation can be expected to lead to different KBs, for example to segmented KBs when the subsidiaries of the firm do not communicate, or to totally connected KBs in the opposite case. A successful merger can be expected to modify the KBs of both firms to exploit the potential synergism inherent in the merger. The process of merger can be expected to reduce temporarily the coherence of the KB below that of the merging firms, but at a short time after the merger the coherence of the KB should increase again. Innovation networks are now a stable form of industrial organisation. Their existence can be partly explained by the need to acquire new types of knowledge created at an increasing speed. However, their dynamics is still to be studied. For example, is it better for partners in networks to have similar or complementary KBs? These questions form part of a research programme underway.

In summary, this paper can be situated in the context of the economics of knowledge. In particular, it attempts to provide tools to represent and measure properties of the KB of firms and to study their effects on firm behaviour and performance.

REFERENCES

Brusoni S., Prencipe A., Pavitt K., Knowledge specialisation and the boundaries of the firm: why do firms know more than they do? Presented at the conference Knowledge Management: Concepts and Controversies, Warwick University, Warwick, (10-11 February 2000).

De Looze M.A., Roy A., Coronini R., Reinert M., Jouve O., Two measures for identifying the perception of risk associated with the introduction of transgenic s, *Scientometrics*, Vol. 44, (1999) 401-426.

Grabowski H., Vernon J. (1994) " Innovation and Structural Change in Pharmaceuticals and Biotechnology " *Industrial and Corporate Change*, vol.3, n°2.

Foray D., (2000) *L'Economie de la Connaissance*, Paris, Repères.

Freeman C., Soete L., (1997) *The Economics of Industrial Innovation*, London, Pinter.

Mc Kelvey M., (1996) *Evolutionary Innovation*, Oxford, Oxford University Press.

March, J. G., 1991, 'Exploration and Exploitation in Organisational Learning', *Organization Science*, 2(1), pp. 71-87.

Nesta, L., (2001), "Cohérence des bases de connaissances et changement technique: une analyse des firmes de biotechnologies de 1981 à 1997," Thèse de Doctorat d'Economie Appliquée. Grenoble: Université Pierre Mendès-France, 328 pages.

Orsenigo L., Pammolli F., Riccaboni M. (2001) Technological change and network dynamics. Lessons from the pharmaceutical industry, *Research Policy*, Vol. 30, pp 485-508.

Pavitt K., (1998) Technologies, products and organisation in the innovating firm: what Adam Smith tells us and Joseph Schumpeter doesn't, *Industrial and Corporate Change*, Vol. 7, N° 3, pp. 433-452.

Pyka A., Saviotti P.P., (2002) Networking in Biotechnology Industries – From Translators to Explorers', working paper, University of Augsburg, January

Saviotti P.P., (1996) *Technological Evolution, Variety and the Economy*, Cheltenham, Edward Elgar,

Saviotti P.P., (1999) Knowledge, information and organizational structures, in Robertson P.L., (Ed) *Authority and Control in Modern Industry*, London, Routledge (1999)

Saviotti P.P., (1999) Knowledge, information and organizational structures, in Robertson P.L., (Ed) *Authority and Control in Modern Industry*, London, Routledge (1999)

Saviotti P.P., de Looze M.A., Michelland S., (2000) –The changing marketplace of bioinformatics, *Nature Biotechnology*, Vol. 18, pp. 1247-1249.

Saviotti P.P., de Looze M.A., Maupertuis M.A. Nesta L. , (2002a) Knowledge dynamics and the mergers of firms in the biotechnology based sectors, to be published, *International Journal of Technology Management*.

Saviotti P.P., de Looze M.A., Maupertuis M.A., (2002b) Knowledge dynamics, mergers and acquisitions in the biotechnology based sectors, submitted to *Economics of Innovation and New Technology*.

Teece, D. J., R. Rumelt, Dosi G. Winter S.G., (1994), Understanding Corporate Coherence: Theory and Evidence, *Journal of Economic Behavior and Organisation*, Vol. 22, pp. 1-30.

APPENDIX 1.

LEXICOGRAPHIC ANALYSIS.

Description of the Sampler Software

Sampler [JOUVE, 1996] is a 'text mining' environment developed by Cisi¹. It calls for the coupling of linguistic technologies with an outstanding implementation of the statistical means of associating words - largely used in France in the 80's. The software works under UNIX and Windows 95.

In order to preserve its 'bottom-up' logic - i.e. navigate with the text within the text - linguistics is uniquely apt to extract units of relevant information. These units are *terminological nominal phrases having the capacity to represent the concepts and objects of the field outside the text* [IBEKWE, 1995]. *The hypothesis is that word extraction will lead to the identification of the documents' themes in return for later statistical treatments (ibid).*

The extractor is made of a lexis, a list of morphological patterns and contextual clarifying rules.

The treatment is performed through a battery-operated algorithm, allowing therefore to process the text in one shot and to reach quasi-linear processing speeds (1 Mocket every two seconds).

The next step consists of putting forward all the nominal phrases which have been collected during the terminological extraction phase. The latter can be enriched with uniwords, proper nouns (which are extracted automatically also).

This grouping (clustering) is enacted through the associated word method.

The corpus is split into homogeneous textual units : the paragraph (which can be used in terms of parameter), the press message,... for the entire text, the instruction, along with its fields (some of which can be masked) for the structured text. One is then to look for the co-occurrences characterized by the appearance of two nominal phrases within one same textual unit. A parameter - alleged to be either equivalent to or associated with **Eij** - is then calculated so as to quantify the associating strength between two words:

$$\mathbf{Eij} = \mathbf{Cij}^2 / (\mathbf{Freqi} * \mathbf{Freqj})$$

where Cij is the co-occurrence between I and j, Freqi is the frequency of I and Freqj is the frequency of j within the corpus.

¹ CISI, 3, rue Le Corbusier Silic 232 DER Génie Informatique 94528 Rungis Cedex France

This algorithm, by measuring the entire relation, makes it possible to map out certain signals which are allegedly weak: two terms appearing once only and to gather, yet producing together an index of equivalence as high as for terms appearing together thousands and thousands of times.

A saturation algorithm is then applied to group in 'clusters' the nominal phrases which are linked together the most. The number of internal, external terms (relations among clusters via a key word), as well as the minimal strength of the link, can be used for parameters, allowing thus to zoom/unzoom in on the corpus.

The terminological extraction of nominal phrases and the method of associated words work towards a production with a semantically homogeneous representation of the texts (which can be qualified as isotopias) - with no initial semantic resources

[POLANCO, 1995].

The watchman's analysis is made easier by the interactive capacity of the Sampler system and the graphic representation of the navigation structures under the ergonomic shape of lexical fields.

This 'bottom-up' approach seems to be the most viable at the moment in carrying out a linear analysis of the texts; it is starting to be generalized on the Internet/Intranet [GREFENSTETTE, 1997]: it makes it possible to unify transversally distributed data.

It is somehow an advantage for it to be coupled with a downgrading hierarchical analysis approach, which makes for a more global dealing for a first approach [REINERT, 1990].

APPENDIX 2: List of main IPC classes

- A01H NEW PLANTS OR PROCESSES FOR OBTAINING THEM; PLANT REPRODUCTION BY TISSUE CULTURE TECHNIQUES** REPRODUCTION DES ES?PLANTES NOTAMMENT PAR GENIE GENETIQUE
- A01N PRESERVATION OF BODIES OF HUMANS OR ANIMALS OR PLANTS OR PARTS THEREOF; BIOCIDES, e.g. AS DISINFECTANTS, AS PESTICIDES, AS HERBICIDES** CONSERVATION DE CORPS HUMAINS OU ANIMAUX OU DE VÉGÉTAUX, OU DE PARTIES DE CEUX-CI; BIOCIDES, p.ex. EN TANT QUE DÉSINFECTANTS, PESTICIDES, HERBICIDES
- A61K PREPARATIONS FOR MEDICAL, DENTAL, OR TOILET PURPOSES** Preparations for medical dental or toilet purposes
- B01D SEPARATION ÉVAPORATION; DISTILLATION; SUBLIMATION**
- B01J Chemical or physical processes, catalysis, colloid chemistry**
- B05D Processes for applying liquids or other fluent materials to surfaces**
- B29C Shaping or joining of plastics**
- B32B Layered products**
- B65D Containers for storage or transport of articles**
- C01B Non metallic elements**
- C04B Lime, magnesia ciments**
- C07C Acyclic or carbocyclic compounds**
- C07D Heterocyclic compounds**
- C07F Acyclic or carbocyclic compounds containing other elements than carbon, hydrogen, oxygen etc**
- C07H SUGARS; DERIVATIVES THEREOF; NUCLEOSIDES; NUCLEOTIDES; NUCLEIC ACIDS** SUCRES; LEURS DÉRIVÉS; NUCLÉOSIDES; NUCLÉOTIDES; ACIDES NUCLÉIQUES
- C07K Peptides**
- C08F Macromolecular compounds obtained by reaction involving only carbon to carbon**
- C08G Macromolecular compounds (polyesters, resins organopolysiloxan, elastomers .)**
- C08J Working up ; general process of compounding (dispersions, gels etc.)**
- C08K USE OF INORGANIC OR NON-MACROMOLECULAR ORGANIC SUBSTANCES AS COMPOUNDING INGREDIENTS** EMPLOI COMME ADJUVANTS DE SUBSTANCES NON-MACROMOLÉCULAIRES INORGANIQUES OU ORGANIQUES
- C08L COMPOSITIONS OF MACROMOLECULAR COMPOUNDS** Composés macromoléculaires (caoutchouc, plastiques, polymères ,,,)
- C09B ORGANIC DYES OR CLOSELY-RELATED COMPOUNDS FOR PRODUCING DYES; MORDANTS; LAKES** COLORANTS ORGANIQUES OU COMPOSÉS ÉTROITEMENT APPARENTÉS POUR PRODUIRE DES COLORANTS
- C09D COATING COMPOSITIONS, e.g. PAINTS, VARNISHES, LACQUERS; FILLING PASTES; CHEMICAL PAINT OR INK REMOVERS; INKS; CORRECTING FLUIDS; WOODSTAINS; PASTES OR SOLIDS FOR COLOURING OR PRINTING; USE OF MATERIALS THEREFOR** COMPOSITIONS DE REVÊTEMENT
- C09K MATERIALS FOR MISCELLANEOUS APPLICATIONS, NOT PROVIDED FOR ELSEWHERE** SUBSTANCES POUR UTILISATIONS DIVERSES, NON PRÉVUES AILLEURS
- C11D DETERGENT COMPOSITIONS** COMPOSITIONS DÉTERGENTES
- C12M APPARATUS FOR ENZYMOLOGY OR MICROBIOLOGY** APPAREILLAGE POUR L'ENZYMOLOGIE OU LA MICROBIOLOGIE
- C12N**

MICRO-ORGANISMS OR ENZYMES; COMPOSITIONS THEREOF

- C12P FERMENTATION OR ENZYME-USING PROCESSES TO SYNTHESISE A DESIRED CHEMICAL COMPOUND OR COMPOSITION OR TO SEPARATE OPTICAL ISOMERS FROM A RACEMIC MIXTURE** MICRO-ORGANISMES OU ENZYMES; COMPOSITIONS LES CONTENANT
PROCÉDÉS DE FERMENTATION OU PROCÉDÉS UTILISANT DES ENZYMES POUR LA SYNTHÈSE D'UN COMPOSÉ CHIMIQUE DONNÉ OU D'UNE COMPOSITION DONNÉE, OU POUR LA SÉPARATION D'ISOMÈRES OPTIQUES À PARTIR D'UN MÉLANGE RA
- C12Q MEASURING OR TESTING PROCESSES INVOLVING ENZYMES OR MICRO-ORGANISMS** PROCÉDÉS DE MESURE, DE RECHERCHE OU D'ANALYSE FAISANT INTERVENIR DES ENZYMES OU DES MICRO-ORGANISMES
- C12S PROCESSES USING ENZYMES OR MICRO-ORGANISMS TO LIBERATE, SEPARATE OR PURIFY A PRE-EXISTING COMPOUND OR COMPOSITION** PROCÉDÉS DE SÉPARATION OU DE NETTOYAGE FAISANT INTERVENIR DES ENZYMES OU DES MICRO-ORGANISMES
- D01F CHEMICAL FEATURES IN THE MANUFACTURE OF ARTIFICIAL FILAMENTS, THREADS, FIBRES, BRISTLES, OR RIBBONS; APPARATUS SPECIALLY ADAPTED FOR THE MANUFACTURE OF CARBON FILAMENTS** PARTIE CHIMIQUE DE LA FABRICATION DES FILAMENTS, FILS, FIBRES, SOIES OU RUBANS ARTIFICIELS; APPAREILS SPÉCIALEMENT ADAPTÉS À LA FABRICATION DE FILAMENTS DE CARBONE
- D06P DYEING OR PRINTING TEXTILES; DYEING LEATHER, FURS, OR SOLID MACROMOLECULAR SUBSTANCES IN ANY FORM** TEINTURE OU IMPRESSION DES TEXTILES; TEINTURE DU CUIR, DES FOURRURES OU DES SUBSTANCES MACROMOLÉCULAIRES SOLIDES DE TOUTES FORMES
- G01N INVESTIGATING OR ANALYSING MATERIALS BY DETERMINING THEIR CHEMICAL OR PHYSICAL PROPERTIES** RECHERCHE OU ANALYSE DES MATÉRIAUX PAR DÉTERMINATION DE LEURS PROPRIÉTÉS CHIMIQUES OU PHYSIQUES
- G02F DEVICES OR ARRANGEMENTS, THE OPTICAL OPERATION OF WHICH IS MODIFIED BY CHANGING THE OPTICAL PROPERTIES OF THE MEDIUM OF THE DEVICES OR ARRANGEMENTS FOR THE CONTROL OF THE INTENSITY, COLOUR, PHASE, POLARISATION OR DIRECTION OF LIGHT, e.g. SWITCHING, GATING, MODULATING OR DEMODULATING; TECHNIQUES OR PROCEDURES FOR THE OPERATION THEREOF; FREQUENCY-CHANGING; NON-LINEAR OPTICS; OPTICAL LOGIC ELEMENTS; OPTICAL ANALOGUE/DIGITAL CONVERTERS** DISPOSITIFS OU SYSTÈMES DONT LE FONCTIONNEMENT OPTIQUE EST MODIFIÉ PAR CHANGEMENT DES PROPRIÉTÉS OPTIQUES
- G03F PHOTOMECHANICAL PRODUCTION OF TEXTURED OR PATTERNED SURFACES, e.g. FOR PRINTING, FOR PROCESSING OF SEMICONDUCTOR DEVICES; MATERIALS THEREFOR; ORIGINALS THEREFOR; APPARATUS SPECIALLY ADAPTED THEREFOR** PRODUCTION PAR VOIE PHOTOMÉCANIQUE DE SURFACES TEXTURÉES