Scientometrics Vol. 6. No. 6. (1984) 381–400

CO-WORD MAPS OF BIOTECHNOLOGY: AN EXAMPLE OF COGNITIVE SCIENTOMETRICS

A. RIP, J.–P. COURTIAL*

Chemistry and Society Programme/Science and Technology Studies Programme, University of Leiden, P.O. Box 9502, 2300 RA Leiden (The Netherlands)

**Centre de Sociologie de l'Innovation, Ecole des Mines, 62 Boulevard Saint-Michel, 75006 Paris (France)*

(Received October 14, 1983)

To analyse developments of scientific fields, scientometrics provides useful tools, provided one is prepared to take the content of scientific articles into account. Such cognitive scientometrics is illustrated by using as data a ten-year period of articles from a biotechnology core journal. After coding with key-words, the relations between articles are brought out by co-word analysis. Maps of the field are given, showing connections between areas and their change over time, and with respect to the institutions in which research is performed. In addition, other approaches are explored, including an indicator of 'theoretical level' of bodies of articles.

Introduction

Although some scientometricians appear to be making too imperialistic a claim for the role of bibliometric and other metric data about science, there is no reason to ignore such a possibly useful tool. Such a reminder is especially important for science dynamics studies – as they are called nowadays in the Netherlands[1] – because the issue how to describe (or "measure") the dependent variable, that is the nature and substance of the development of sciences and research areas, has been neglected or is passed over too easily.

To give a recent example: a recent issue of *Social Studies of Science* contains two interesting articles. One by *Abir-Am* on the aims and activities of Warren Weaver as the director of the Rockefeller Foundation who gave a push to the development of molecular biology;[2] the other by *Gilbert* and *Mulkay* on the accounts scientists produce of their research practices.[3] *Abir-Am* even complains about the lack of attention in science studies for the study of the *impact* of science policy measures on the development of scientific fields. But she, as well as *Mulkay* and *Gilbert*, limit themselves to analyzing the nature of the interventions and/or scientific practices. The outcome of the intervention or the scientists' actions is suggested, but not described systematically.

To "measure" the dependent variable, or *explanandum* if one prefers that term, two methods are available. The first, cognitive analysis, has a long tradition in the history

of science, but it is only recently that more attention has been given to recurrent features of cognitive structures of science, so as to be able to catalogue them and describe scientific developments in these terms. One may think of *Holton*'s thematic analysis of science, the concept of regulatives introduced by the Starnberg group, and the attempt to catalogue important cognitive elements, as sketched in the science dynamics programme of the University of Leiden.[4]

The second method to "measure" developments of scientific fields is scientometrics, a variety of approaches sharing the general idea that quantifiable aspects of sciences should be extracted and used to "measure" whatever it is that can be measured with them. Publications and citations counts are well known by now,[5] and the most interesting, as well as the most ambitious approach is co-citation analysis, pioneered by *Small* and others at the Institute for Scientific Information.[6] *Lenoir* has claimed that co-citation analysis is *the* solution to the problem of studying developments of scientific fields: it produces maps of the fields, while block-modeling gives the social structure of its practitioners.[7] Such a claim, however, overlooks that co-citation links are sociometric ties, or indicators of the set of accepted authorities in the field at best. It is an additional assumption to take co-citation clusters as reflecting cognitive structure,[8] and *Small*'s attempts at citation context analysis can be taken to show that co-citation clusters are no more than a way to define a body of articles to do content analysis on.[9]

Another, and well-known, criticism is that we lack a comprehensive theory of citing practices in scientific fields, and that what we do know about citing indicates that many other factors besides the acknowledgment of cognitive debts play a role.[10] Citation analysis and co-citation analysis thus appear to measure developments of scientific fields only through an intervening social institution, i.e. citing practices. The implication then is that (co-)citation analysis can only be applied in cases where such a social institution occurs and has a certain stability – that is, only in cases of academic scientific fields oriented toward publication in international scientific or scholarly journals.

To overcome such criticisms, one may try to improve the tools of (co-)citation analysis. Another possibility is to start anew. Realizing that co-citation links are, in the end, a route to the *content* of the articles, one may proceed in a more direct way and use content analysis and/or the coding that is done by documentation services as the basis for scientometric analysis of a state-of-the-field. Such an approach has been pioneered by *Callon, Courtial* and *Turner,* and although their so-called *co-word analysis* is not yet routinely available, it is already clear that an approach based on key-words or signal-words can overcome some of the limitations of (co-)citation analysis.[11] The method is applicable in domains where citing is irregular or absent, and can be used even for reports and internal documents of science policy com-

mittees.[12] The coding procedures (content analysis or from existing documentation services) will introduce problems of interpretation; when co-word analysis is based on existing documentation service data bases alone, one could again speak of a separate social institution that intervenes. An advantage of co-word analysis over co-citation analysis is its relation to recent theories of scientific practice in which the power of words is emphasized.[13] All in all, it is too early to decide whether co-word analysis is "better" than co-citation analysis or not. But it will be clear that co-word analysis complements co-citation analysis, and that its potential should be explored further.

A number of scientometric studies of developments in biotechnology and in the scientific disciplines germane to biotechnology have been carried out by the Technology Policy Unit (University of Aston in Birmingham, UK) and the Chemistry and Society Programme (University of Leiden), under contract with the FAST-Biosociety Program of the Commission of the European Communities.[14] The project aimed to explore different techniques for monitoring developments in the relevant scientific and technological fields. Co-word analysis was used – and amended – to re-analyze the data describing the contents of articles in a biotechnology core journal (*Biotechnology and Bioengineering*) over a period of ten years. This paper will describe the main results of the analysis and, in the concluding section, come back to the general issue of scientometric analysis of developments in scientific fields.

Co-word analysis: the instrument

Starting point for the theory behind the use of co-word analysis is the notion that authors of documents use *signal-words* to guide the reader in the direction they wish him to go. The author enrolls the reader in a "funnel of interest", makes him a captive of the transformation of the field that the author wants to realize.[15] To do so, however, the author has to use signal-words that are accepted in the field, that is, have a latent power that he can mobilize for his own purposes.[16]

In content analysis of the document, or in abstracting it to summarize the interest it may have for other readers, the analyst works in the other direction. The document is transformed into a string of signal-words to capture its cognitive-interest structure. Clearly, the intervention of the analyst may introduce "distortions" – but distortion is a relative concept. There is no "true" cognitive message of the document that could, ideally, be summarized by the perfect abstracter. The impact of the document on the field is realized through the readings of it by others,

through the way they are enrolled by it and interpret its cognitive-interest structure. So the analyst's intervention will transform the document, but often in a way that is comparable to its reception by other readers in the field.

Assuming a competent analyst, a set of documents can be transformed into a data base consisting of strings of signal-words, and for each string the usual bibliometric data (source, date, authors, institution, country, etc.). The number of occurrences of signal-word i is counted, and the number of co-occurrences of signal-words i and j within the strings of signal-words. In this way, a co-occurrence matrix is contructed, with c_{ij} denoting the number of co-occurrences between i and j, and $c_{ii} = c_i$ giving the occurrence of signal-word i (for convenience of data presentation; no matrix algebra will be performed). An example of such a co-occurrence matrix for the data base *Biotechnology and Bioengineering* 1970–1974, is given in Table 1.

The next step in the analysis is to bring out the interesting features of the co-occurrence matrix. One important feature, already visible in the co-occurrence

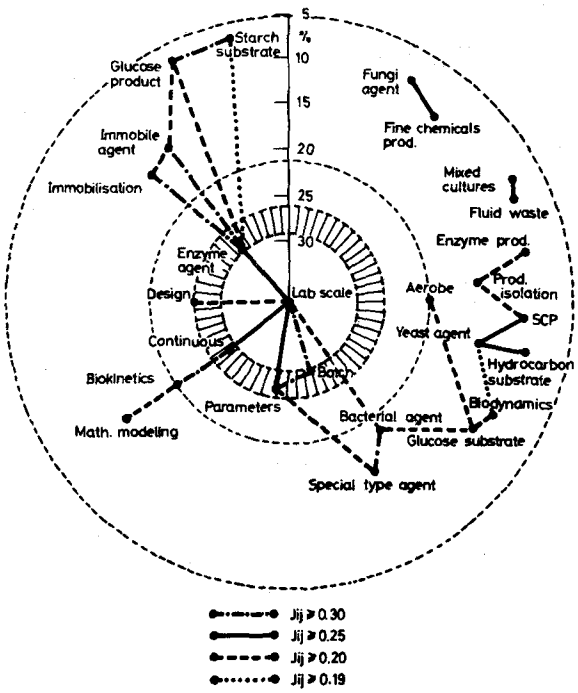


Fig. 1. Circular map, Jaccard, 1970–74
Jaccard links with keywords below 5% frequency have been deleted

matrix (Table 1), is the rank-ordering of key-words according to their frequency of occurrence. The other important feature is the incidence of co-occurrences between key-words and the intensities of the co-occurrences. To distinguish between co-occurrences that are interesting and co-occurrences that are not interesting, an index is introduced that measures the strength of the co-occurrence linkage according to some formula, and a threshold is determined below which co-occurrence linkages are considered to be not interesting anymore.

To give an example, consider the circular "map" of keywords produced by calculating all Jaccard links in the co-occurrence matrix (Table 1) and deleting all links with intensities below 0.19 (the map is given in Fig. 1. and discussed in detail below). The Jaccard index, is often used to bring out linkages, for

$$J_{ij} = \frac{c_{ij}}{c_i + c_j - c_{ij}},$$

example also in co-citation analysis. A glance at the circular map shows that not only remain many key-words invisible (22 of the 49 key-words in this map), but also far fewer linkages between key-words appear than are theoretically possible. A further lowering of the threshold would increase the number of linkages, and one may, in fact, set the threshold at successively lower values to see the fine-structure of the linkage patterns. In many cases, there is a range of threshold values in which the overall pattern of linkages does not change drastically (for instance, because fine-structure linkages appear only within groups of linked key-words and not between groups).

The significance of the linkages that are brought out by the index-threshold operation, can be viewed in two ways. In general, linkages can be seen as *junctions* between bodies of documents, each characterized by the occurrence of the key-word that partakes in the linkage. Researchers may group such key-words together (this happens for instance in the upper lefthand area of the circular "map" (Fig. 1), where immobilized enzymes to produce glucose-type products from starchy substrates are linked together on the map, as well as in the practice of biotechnology). Alternatively, linkages may offer shifts between different research areas (for instance, at the righthand side, Single Cell Protein production is related to product isolation, but the latter is also related to enzyme production; the two areas of production share an interest in product isolation research, but lead separate lives otherwise). A "map" of co-word linkages is thus like a topological map of a city (e.g. when showing bus or underground routes): the lines between the keywords represent possibilities of traveling from one place (key-word) to another, but do not indicate distances, time to be travelled or other metric measures. The arrangement in a two-dimensional plane is arbitrary, and chosen for ease of reading the "map" only.[17]

Table 1
Co-occurrence matrix, 1970–74 (N=458)

rank-ordered keywords	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	etc.
1. lab scale	285	90	97	91	69	87	71	58	69	54	50	31	40	39	46	42	33	27	31	36	34	33	32	28	19	3	29	24	21	
2. enzyme agent	131	10	33	32	22	30	-	5	4	70	10	-	15	59	2	18	-	3	3	6	-	4	14	3	2	9	5	26		
3. batch	128	22	23	58	24	36	44	4	14	23	17	6	26	17	11	22	24	26	8	12	19	9	4							
4. continuous	126	36	32	51	34	25	11	10	24	27	30	23	17	16	24	11	12	1	9	15	8	8	21	14						
5. design	117	16	18	30	18	11	9	3	9	9	26	12	22	23	1	3	7	16	5	3	11	12	5	3	17					
6. parameters	116	15	33	36	33	10	7	18	9	18	19	3	6	18	15	18	15	18	10	12	18	12	7							
7. biokinetics	97	19	18	16	4	5	16	31	23	15	20	11	8	10	6	7	9	6	8	5	9	14	5							
8. aerobe	95	24	22	-	2	16	18	-	26	20	18	13	19	6	14	11	3	22	12	4	17	1								
9. bacterial agent	87	48	3	13	4	8	2	24	5	1	10	13	18	13	6	13	8	10	19	8	3									
10. special type agent	74	2	15	16	4	2	20	6	-	8	12	17	13	7	18	2	15	15	4											
11. immobilisation	74	2	-	3	23	1	2	-	-	-	1	-	2	7	4	1	5	2	13											
12. product isolation	72	23	1	-	4	2	5	22	1	13	2	3	25	1	15	1	2	1												
13. yeast agent	67	5	-	10	5	3	26	19	2	23	2	8	-	4	5	1	2													
14. matemathical modeling	62	12	7	12	18	4	8	5	5	7	3	8	1	1	7	1														
15. immobile agent	61	-	16	-	1	-	1	4	-	-	2	3	1	3	2	15														
16. glucose substrate	56	6	5	3	18	13	2	7	6	4	7	-	4	-																
17. physical kinetics	53	9	6	2	3	9	6	1	1	4	1	3	4																	
18. general micro-organism	53	6	6	-	8	-	3	8	2	1	3	-																		
19. Single Cell Protein	51	5	-	15	7	3	4	5	4	3	-																			
20. biodynamics	50	6	9	6	3	4	1	7	5	-																				
etc.																														

Another view of the significance of the linkages available in the co-occurrence matrix is possible when the universe of documents and key-words is clearly limited – as is the case in our data base, containing all articles in a given journal, published during 1970–1979, and coded with a set of carefully chosen keywords. In such a case, the incidence of co-occurrences can be compared with the expected value, if a set of key-words with given occurrences would be distributed randomly over all articles. The probability of finding a certain value for the co-occurrence between i and j is given by the hypergeometric distribution function for N (the number of articles), c_i and c_j. For example, if one takes the most frequently occurring key-word in the matrix (Table 1), lab scale, with an occurrence frequency just above 60%, it is clear that a co-occurrence value of 69 with key-word 5, design, is too close to 60% of its occurrence of 117 to be interesting.

Thus, if the documents and key-words are given beforehand, it will be useful to consider another index, the *statistical index* S_{ij}, which is the normalized deviation from the expected value of the co-occurrence:

$$S_{ij} = \frac{1}{\sigma}\left(c_{ij} - \frac{c_i c_j}{N}\right)$$

where σ is the standard deviation of the hypergeometric distribution function and $c_i c_j/N$ its mean (or expected value). Calculations of the statistical index show that a threshold $S_{ij} > 2$ (confidence 75% or (much) better) produces the same linkages as appear when using the Jaccard index with a threshold of 0.19, or somewhat more.[18] The Jaccard "maps" can therefore be taken as a conservative picture of the linkages between key-words. They will be used here because they are easy to calculate.

When documents and key-words are *not* given beforehand, for instance in the common case that a body of articles is collected by taking all articles in a data bank that are coded by a chosen keyword, say recombinant DNA (or a few related keywords), there is no clear meaning attached to the universe of documents and key-words occurring in them. In such a case, all junctions between key-words are interesting, and another type of index produces better results. The *inclusion index*, $I_{ij} = c_{ij}/c_i$ (with $c_i < c_j$), measures the extent to which a less frequently occurring key-word is joined to a more frequently occurring key-word. Setting a threshold, say at 0.5, an overall picture is obtained of the "master key-words" dominating a tree of less frequently occurring key-words. (See for example the diagram of Fig. 6.) The notion of "master key-words" can be related to the idea that some signal-words *have* to be used by authors to frame their own attempt at transforming the field.

Before interpreting diagrams with inclusion linkages, spurious inclusions should be deleted. If three key-words i, j, and k (increasing frequency of occurrence) are linked in a triangle, the co-occurrence between i and k may be caused to a large

extent by the joint co-occurrences of i and j, and j and k, and thus be considered spurious. Therefore, the co-word analysis programs contain an algorithm to calculate the expected value of c_{ik}, based on the co-occurrences c_{ij} and c_{jk} which are taken to be given. If the actual value of c_{ik} exceeds the expected value to an amount exceeding a threshold value, the link between i and k is kept as significant. (In the diagram of Fig. 6, this has happened only once.)[19]

Finally, we note that a fourth index may be used, the *proximity index* P_{ij}, given by the formula $P_{ij} = N\, c_{ij} / c_i\, c_j$. The proximity index is a composite index, resulting from the division of the inclusion idex I_{ij} by the frequency of occurrence of the more frequent key-word j. Such a procedure enhances the inclusion links between less frequently occurring key-words, at the expense of the links with the "master key-words". The linkages emphasized by the proximity index may therefore be taken to represent new developments and/or minor areas of research. It is used only because it is easy to calculate; its statistical interpretation is not straightforward.[20]

Co-word maps of biotechnology

Figures 1–8 present the maps produced by co-word analysis of the data for ten years' articles in the journal *Biotechnology and Bioengineering*. In our comments, we shall point at a few interesting features only.

The rank-ordering of the key-words according to their frequency of occurrence is taken into account by having a radial scale (in percentages, with lab scale drawn at about 38% instead of its actual values at about 60%) in the circular maps, and a vertical scale (in percentages) in the diagrams for the inclusion linkages. Looking at the frequencies of occurrence only, and starting with the 1970–1974 map (Fig. 1), it turns out that there are three regions:

– a central region (the hatched circle) with four "methods" key-words and one for an important biological agent, enzymes;

– a broad intermediary region, containing key-words for methods and research approaches, as well as key-words for specific products and processes;

– a peripheral region (not shown on the map), with keywords for products and processes only and very few linkages (or none at all).

Such a structure may well be characteristic of scientific technologies like biotechnology, where there is a wide range of special technologies for specific product/processes in different industrial sectors or public utilities; and a set of general methods and research approaches available to all, and quite probably featuring heavily in a core scientific journal of the field.

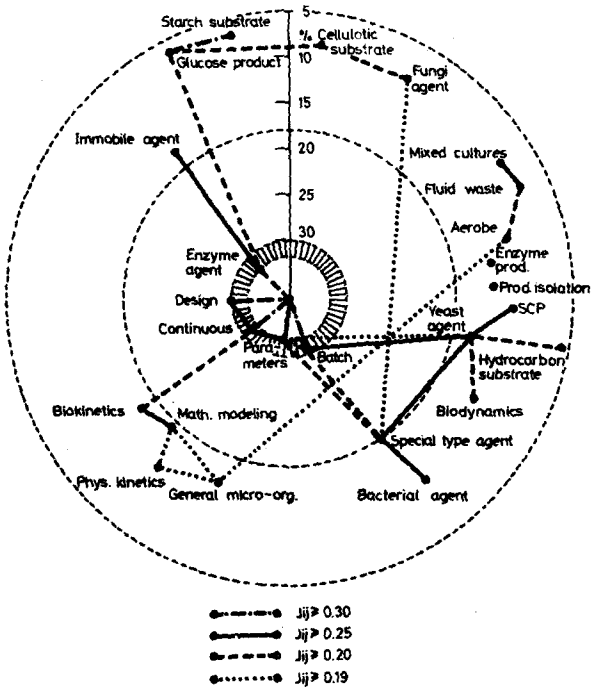


Fig. 2. Circular map, Jaccard, 1975–79
Jaccard links with keywords below 5% frequency have been deleted

Within the central region and the intermediary region, the lines drawn on the map show pathways or junctions between key-words, not similarities. The position of the key-words on the map is not arbitrary, however. An attempt was made to fill up the available space so as to facilitate reading the map, also when comparing it with the subsequent maps. In addition, closeness of key-words is related to (1) intensity of the Jaccard link, and (2) similarity in patterns of linkages above and slightly below the threshold. Thus, in spite of our conviction that the maps are topological, not metric, some sense of distances seems to be unavoidable.

Now consider the map for the next five years (Fig. 2). The central region becomes tighter: the frequency of occurrence of the same five key-words as in the earlier period has become higher and more similar, the mutual linkages have increased, and the distance to the intermediary region has increased (there is a full 12% difference between the two regions). The natural interpretation of this effect seems to be that the "scientific approach" in biotechnology becomes more visible and available, at least to the authors publishing in this core journal. In the intermediary region, shifts occur, for instance in the upper part, where the enzyme-agent group has be-

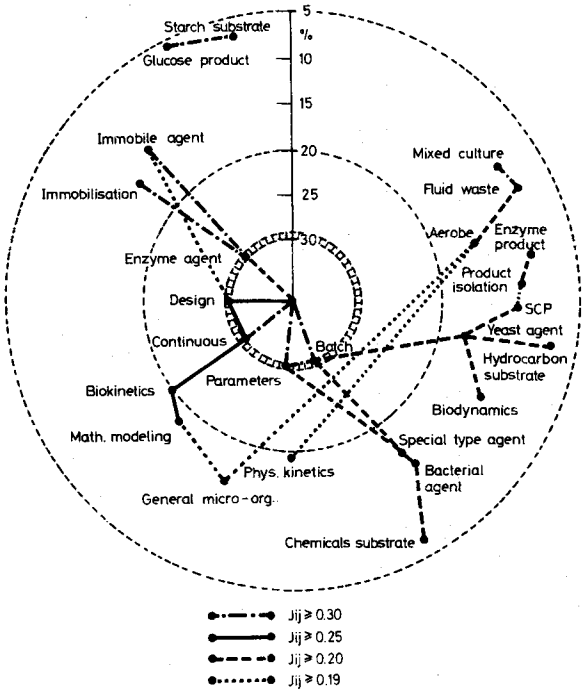


Fig. 3. Circular map, Jaccard, University (N = 1028)
Jaccard links with keywords below 5% frequency have been deleted

come less forceful and glucose is now produced on cellulotic substrates also (which itself is now linked to fungi agent). Other shifts are the increased detail in the group of research methods "by themselves" (to the left) and the more central position of yeast agent (to the right).

To increase our understanding of the maps and the changes in them, co-word analyses may be performed on specific subsets of the data base. This was already apparent above, when changes over time were discussed with the help of maps for two five-year periods. The next three maps (Figs 3–5) give the results of co-word analysis for research produced in different kinds of research institutions: universities, industrial research laboratories and government research institutions. Some striking differences appear. In the universities map (for the whole period of ten years), the inner circle is quite tight, and general and methods key-words preponderate. As one would expect, the "scientific approach" is dominant. The contrast with the industries map is striking: the central circle has disappeared, enzyme agents have become dominant, while continuous has dropped out of the central region and stands rather isolated. (Continous processes, being more sensitive to disturbances, are rarely of

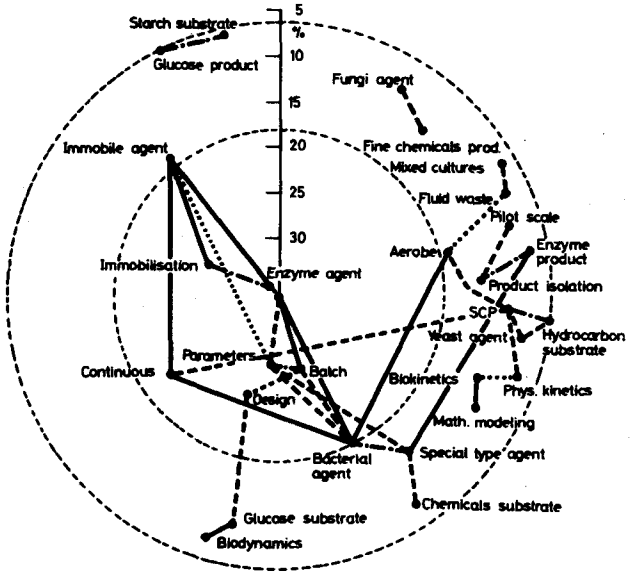


Fig. 4. Circular map, Jaccard, Industry (N = 163)

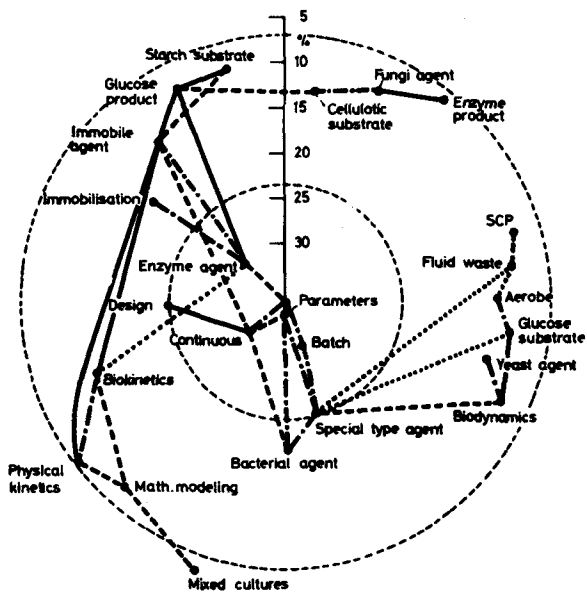


Fig. 5. Circular map, Jaccard, Government (N = 120)

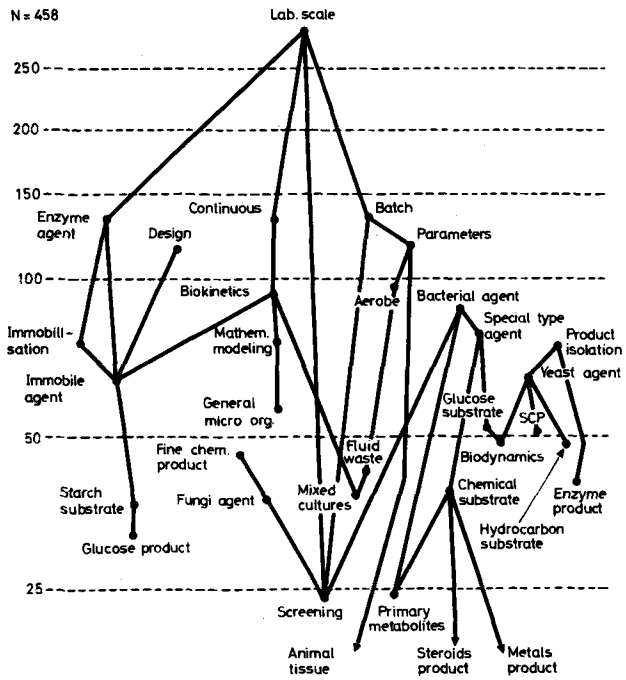


Fig. 6. Vertical map, inclusion index, 1970–1974
threshold 0.333 "Spurious" inclusions deleted

interest for industry, even if they produce higher yields.) Somewhat more linkages appear than in the universities map, because the total number of articles is (much) less, so the Jaccard index will take higher values.

The map for government laboratories, even if based on still fewer articles, shows some interesting features. The "inner circle" remains visible, with parameters becoming the dominant key-word. Cellulotic substrates make their appearance here: this type of research gets its visibility from government laboratories, presumably because it is of long-term policy relevance, especially for agricultural and forestry policy. Another feature is that the more ambitious research approaches are strongly connected to the immobilization-group of key-words. With some knowledge of the field, this can be explained: immobilized-enzyme research fell sharply after the first half of our period, industry focusing on a few practical processes, and government laboratories presumably keeping up the research and becoming more "academic".

The field of biotechnology seems well-suited to mapping with the help of the Jaccard index. The inclusion maps (Figs 6 and 7) do not, in this case, add very

much to our understanding. In fact, they are more difficult to understand because they have too many cross-linkages. The reason for this is, again, the nature of the field, where products/processes form one way of looking at it, and the "scientific approach" a different way, that cuts across the former. This cross-cutting effect has been enhanced by choosing a set of only 49 key-words for the data-base.[21] For other fields, and with key-words drawn from existing documentation services, the inclusion maps appear to produce useful insights.[22]

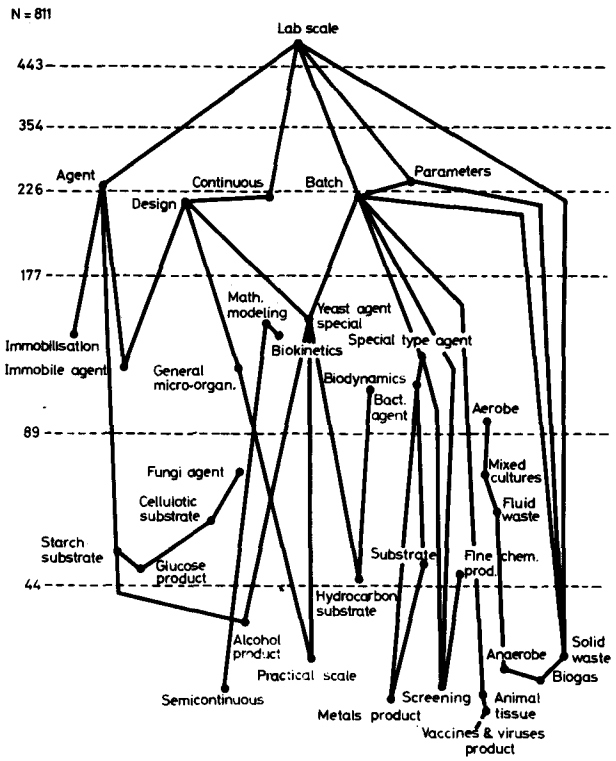


Fig. 7. Vertical map, inclusion index, 1975–1979 threshold 0.333 "Spurious" inclusions deleted

For completeness, we also give the map based on the proximity index, for the period 1970–1974 (Fig. 8). Minor areas of research become visible, while the major structure of the field (as seen through Jaccard or inclusion indexing) disappears almost completely. Most of the areas of research and the junctions between them on the proximity map can also be recognized in the bottom part of the inclusion map (Fig. 6). The interest of the proximity map lies in its potential to indicate

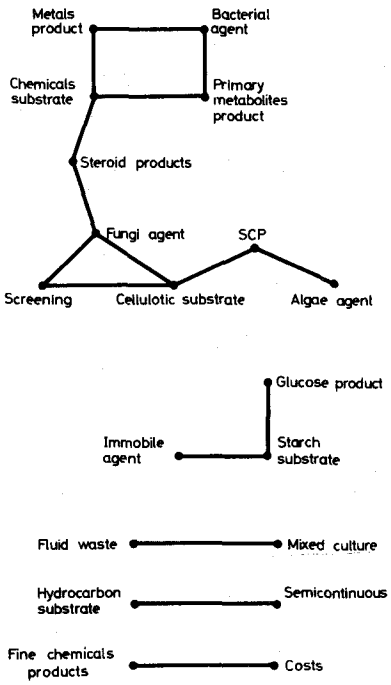


Fig. 8. Proximity map, 1970–1974

new developments, "early promises" that may subsequently transform the field. Only the fungi agent – cellulotic substrate – SCP – algae linkages can be considered a potential transformation. It may well have such potential, but only in the long-term. In the second part of our period, it is still a minor area, although the fungi agent – cellulotic substrate part of the chain has become more visible through the effort of government laboratories.[2 3]

Scientometrics and science dynamics

To assess the value of co-word analysis as an instrument to measure the developments of scientific fields, clearly more fields will have to be covered and validations of the interpretations of the maps have to be produced. Work in progress is doing just that. From the little exercise discussed above, already some conclusions can be drawn.

A centre – periphery model seems to be implicit in co-word analysis, and in our case of biotechnology strikingly so. Co-citation analysis has the same bias, but the

advantage of co-word analysis may be that instruments are available to study the periphery. It is not clear yet if the proximity index discussed above is the most suitable instrument for doing this.

Some features of the development of a field become visible, and sometimes they are surprising, also to people knowing the field. But there seems to be no systematic way of going about the interpretation of the maps. Each map has to be studied by itself, the analyst has to get a feel for the overall picture and the fine-structure. The computer programs facilitate his experimenting with different indices and thresholds, but interpretation remains intuitive.

It should be realized, however, that no scientometric method is able, by itself, to produce interpretations. The intuition of the analyst, his knowledge of the field and his assimilation of experts' knowledge remain necessary. Does this imply that we had better forget about scientometrics and do cognitive analysis, as well as we are able to do so? Yes and no. Cognitive analysis should never be neglected. But the danger of limiting oneself to cognitive analysis (in the case of contemporary sciences) is that one has to become an expert oneself. The views on the state-of-the-field developed in this way will become the view of an actor, another participant in the development of the field, pressing to get a hearing. (Or presenting this view to sociologists and theoreticians of science only, and thus becoming irrelevant to the field as well as difficult to be subjected to quality control.) The development of scientometric instruments will never be a solution to the analyst-actor dilemma. But it will introduce some distance, and quasi-objectified procedures for analyzing scientific fields that can be checked by actors as well as analysts.

What this argument is leading to, is the view that cognitive analysis and scientometric analysis should be *combined* rather than contrasted. From the side of scientometrics, the work on citation contexts seems to be a step, although still a small one, in this direction (see note 9). Co-word analysis will offer the analyst more flexible ways to enter into the content of a science, and does not require a large input from experts in the field studied. From the side of cognitive analysis, there appears to be some reluctance to enlist quantitative methods – in spite of the trend in general history towards an integration of quantitative and qualitative methods.[24] We think that there are clear possibilities for treating cognitive aspects quantitatively. From our work on biotechnology we may draw an example of what we have in mind. In our set of keywords, there were ten key-words denoting research approaches. With the help of Weingart and Van den Daele's general distinction between different levels of scientific capacity in relation to more or less ambitious political goals (see Table 2),[25] we can rank-order the key-words for research approaches according to their theoretical ambitiousness (see Table 3). To do so requires some adaptation of the Weingart–Van den Daele levels. For instance, the

Table 2

Objective		Scientific capacity
description, statistics, assessment	a.	measurement, monitoring
control of systems	b.	functional explanations, input-output relations
construction of systems	c.	causal explanations, mechanisms
provision of goals and of means to reach them	d.	integrated science, theory of complex systems and their behaviour in new circumstances

lowest level is taken to include exploratory, trial-and-error research and "looking for effects to exploit". Then, key-words screening and costs can be fitted to the lowest level. Design incorporates both monitoring and control and is classified mixed a/b. Immobilization and product isolation are often approached with trial-and-error strategies, but may include more systematic input-output studies. Parameter optimalization is clearly a functionalistic research strategy, the bio-reactor remaining a black box. The other key-words imply some attempt at making the black box translucid, by elucidating causal mechanisms (although mathematical modeling may be limited to the improvement of input-output relationships).

As indicated in Table 3, research approaches wholly or partly at level c are considered to be theoretically ambitious. It should be noted that theoretical ambition is always relative to a certain goal, in this case a certain view of the cognitive structure

Table 3

Rank order	Research approach	Weingart–Van den Daele level
1/2	screening	a
	costs	a
3	design	mixed a, b
4/5	immobilization	a or b
	product isolation	a or b
6	parameter optimalization	b
7/9	mathematical modeling	b or c
	physical kinetics	b or c (often)
	biokinetics	b or c
10	biodynamics	c

of biotechnology and its development. The physical and biological processes in the bioreactor and their interrelationships are considered to be the focus of biotechnology. Thus, product isolation will never be very theoretically ambitious. If one would analyze biotechnology in terms of unit operations, however, a different rank ordering would result, and product isolation could become classified as theoretically ambitious.

So far, we have been looking at cognitive aspects. A scientometric indicator of the "theoretical level" of a given set of articles, for example all articles having Single Cell Protein as a key-word, can now be constructed by summing all occurrences of theoretically ambitious research approaches and dividing by the sum of the occurrences of keywords for all research approaches. The theoretical level for the whole ten-year period is 41%, which is a distinct increase with respect to the theoretical level of 7% in 1960. Other results of a general kind are that product-oriented articles have a lower theoretical level (30%) than non-product-oriented articles (51%). For each key-word, the theoretical level can be determined, which leads to expected, as well as unexpected results. Research using hydrocarbon substrates (often for Single Cell Protein production), for example, turns out to have a high theoretical level. The co-word maps (Figs 1, 2) show that this is because biodynamics, the most ambitious research approach, has been pioneered for this area.

The few results of the indicator for theoretical level quoted here (see further the report mentioned in note 14), will indicate the uses that can be made of such "cognitive scientometrics", especially when combined with other analyses, as for instance co-word analysis. There are many other possibilities to be found as soon as the documents of science are seen as persuasive literary products (cf. note 13) and therefore amenable to both cognitive and scientometric analysis. Simple measures as the length of the introductory sections in articles may be explored as indicators for the extent to which the repertoire of a field has become articulated and stabilized. More complicated analysis is necessary if one wants to distinguish between internally generated problems and externally generated problems (with respect to the field studied) on the basis of the interest-funneling as recognizable in the articles.

The indicators derived from such a "cognitive scientometrics" approach may be rough at first, and will always appear somewhat artificial. This, we argued, is necessary to put some distance between the analyst and the field he studies. What other ground would there be for a science dynamics scholar to make a contribution to the debates going on between the scientists (in our case, biotechnologists), policy-makers and other relevant societal groups?

The scientometric studies of biotechnology presented in this paper have still to stand the test of being discussed by biotechnologists and policy-makers. In this respect, our results are preliminary and not intended to convince without doubt. Rather, we hope to stimulate further exploration of the potential of "cognitive scientometrics".

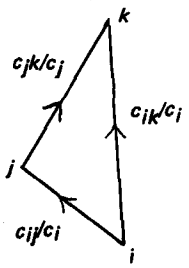
*

The original data base was adapted for the co-word analysis by Walter *van der Es*, with the assistance of Fred *Brouwer*. Their advice and comments, and those of Michel *Callon*, Bill *Turner* and Serge *Bauin* have been very helpful when we were exploring the meaning of the co-word maps.

Notes and references

1. Science dynamics studies are empirical investigation of the factors that determine developments in the sciences, not (only) out of general historical and intellectual interest, but (also) with possible, although long-term science-policy interest in mind. For further information on the way science dynamics studies are being stimulated in the Netherlands, see the report by A. RIP, *EASST Newsletter* No. 2 (May 1982) 16–18.
2. Pnina ABIR-AM, The Discourse of Physical Power and Biological Knowledge in the 1930s: A Reappraisal of the Rockefeller Foundation's 'Policy' in Molecular Biology, *Social Studies of Science* 12 (1982) 341–382.
3. G. GILBERT, M. MULKAY, Warranting Scientific Belief, *Social Studies of Science* 12 (1982) 383–408.
4. G. HOLTON, *The Scientific Imagination. Case Studies,* Cambridge University Press, Cambridge, 1978; G. BÖHME, W. van den DAELE, W. KROHN, *Experimentelle Philosophie. Ursprünge autonomer Wissenschaftsentwicklung* Suhrkamp, Frankfurt a/Main, 1977 and G. BÖHME e.a., *Die gesellschaftliche Orientierung des wissenschaftliches Fortschritts* Suhrkamp, Frankfurt a/Main, 1978; *A Research Programme in Science Dynamics for the University of Leiden: Oriented Science in Contemporary Society* University of Leiden, Leiden, March 1982.
5. *Science Indicators,* the bi-annual review of the health of American science, has spawned all sorts of science and/or technology indicator efforts. For science dynamics, the more interesting attempts are those of D. de SOLLA PRICE, e.g. his quantified model of the citation cycle; see *Current Contents* Nr. 39 (Sept. 29, 1980) 8–20.
6. See for example E. GARFIELD, M. V. MALIN, H. SMALL, Citation Data as Science Indicators, in Y. ELKANA e.a. (Eds), *Toward a Metric of Science* John Wiley & Sons, New York, 1978, p. 179–207.
7. T. LENOIR, 'Quantitative Foundations for the Sociology of Science: On Linking Blockmodeling with Co-Citation Analysis', *Social Studies of Science* 9 (1979) 455–480.
8. The same point is made by M. de MEY, Piaget's Notion of "Cognitive Structure" and Current Cognitive Analyses of the Structure of Scientific Knowledge and Puzzle-Solving in Science, paper presented to the EASST conference, Deutschlandsberg, 24–26 September 1982.
9. H. SMALL, E. GREENLEE, Citation Context Analysis of a Co-Citation Cluster: Recombinant-DNA, *Scientometrics* 2 (1980) 277–301. Also Susan E. COZZENS, Life History of a Knowledge Claim: The Opiate-Receptor Case, paper, 4S Annual Meeting, Atlanta (GA), 5–7 November 1981.
10. See G. N. GILBERT, Referencing as Persuasion, *Social Studies of Science* 7 (1977) 113–122, and D. EDGE, Quantitative Measures of Communication in Science, *History of Science* 17 (1979) 102–134.
11. M. CALLON, J.-P. COURTIAL, W. A. TURNER, S. BAUIN, From Translations to Problematic Networks: An introduction to Co-Word Analysis, *Social Science Information* 22 (1983) 191–235. The group is based in the Ecole des Mines (Centre de Sociologie de l'Innovation) and the CNRS-Centre de Documentation Scientifique et Technique (Service d'Etude et de Réalisation des Produits d'Information Avancés), Paris.

12. Science policy documents have been used in this way in the first co-word analysis: M. CALLON, J.–P. COURTIAL, W. A. TURNER, *L'Action Concertée Chimie Macromoléculaire: Socio-logique d'une Agence de Traduction* Centre Sociologie de l'Innovation Paris/Strasbourg, Groupe d'Etude sur la Recherche Scientifique Université Louis Pasteur, 1979.
13. See especially the work of J. LAW, B. LATOUR, M. CALLON, e.g. R. WILLIAMS, J. LAW, Beyond the Bounds of Credibility, *Fundamenta Scientiae* 1 (1980) 295–315; M. CALLON, B. LATOUR, Unscrewing Leviathan: How do actors macro-structure reality?, in: K. KNORR-CETINA, A. V. CICOUREL, *Advances in Social Theory and Methodology. Toward an Integration of Micro- and Macro-Sociologies* Routledge & Kegan Paul, London 1981, 277–303; M. CALLON, J. LAW, On Interests and Their Transformation: Enrolment and Counter–Enrolment, *Social Studies of Science,* 12 (1982) 615–625.
14. The project, directed by Harry ROTHMAN (Birmingham), has produced a number of reports, some of them to be published in book form. Some titles are: *Biotechnology: A Review and Annotated Bibliography; Fuel Alcohol: The Brazilian Experience; Monoclonal Antibodies: Another Success for Biotechnology; Biotechnology World Patents Analysis; Enzyme Technology Patents; Microbiology: A Scientometric Analysis; Portrait of a Biotechnology Core Journal.* The last report, by W. VAN DER ES and A. RIP, provides the data for the co-word analysis discussed in the present paper. The work for this report has been supported by the Dutch Ministry for Science Policy.
15. The notion of "interest funneling" was introduced by J. LAW and illustrated with the help of detailed analyses of the structure of discourse of the introductions to scientific papers. See also WILLIAMS and LAW, *op. cit.*, note 13.
16. G. LEMAINE, Social Differentiation in the Scientific Community, paper presented to the EASST conference, Deutschlandsberg, 24-26 Sept. 1982, discusses the re-definition of orthodoxy attempted in scientific work, and shows that the limits of re-definition also depend on the risk strategies followed by the authors. See id., Z. f. Wissenschaftsforschung 3 (1) (April 1984) 9–27.
17. It is possible to transform linkages of different intensities into distances in a two-dimensional plane with the help, e.g., of *Kruskal* scaling techniques. This is occasionally done in co-citation clusters (see notes 6 and 9), and was also attempted in the first co-word analysis (note 12). The problem, however, is that the projection onto the plane introduces a heavy "stress", and key-words or co-cited articles that are close to each other in the plane may well be far apart in terms of linkage intensity. No overall metricization should be attempted, therefore, even though local metricization may well be useful to clarify complex sets of linkages.
18. When the hypergeometrical distribution function can be approximated by a normal distribution, the threshold of twice the standard deviation provides for a confidence limit of about 5%. If it cannot, Tschebyscheff's theorem sets an upper limit of 25%. The statistical index favours co-occurrences between low-frequency key-words, since a random distribution of such key-words over all articles makes the expected value of the co-occurrence vanishingly small. In this region, the statistical index does not produce interesting links, and other indices, e.g. the proximity index, have to be used.
19. When probabilities are calculated on the basis of the actual frequencies of occurrence and co-occurrence, the argument runs as follows. The chance to find j, given i, is c_{ij}/c_i, and the



chance to find k, given j, is in the same way, c_{jk}/c_j. The product of these two conditional probabilities is the chance to find k, given i and through the intervention of j.

If the actual value c_{ik}/c_i exceeds the expected value, that is the product of the two conditional probabilities, by an amount larger than the chosen threshold, the link between k and i is considered to be significant in its own right. Since the same criterion is applied to decide whether to keep bi-lateral links, the map produced in this way is homogeneous.

20. The proximity index is the quotient of the actual and the expected value of the co-occurrence, a quantity which has no immediate statistical interpretation. A probability interpretation based on actual frequencies of (co-)occurrence, as in note 19, is also difficult. The index is important, however, to make those linkages visible that would otherwise be dominated by master keywords. CALLON et al. (note 11) give examples of this use of the proximity index for the field of dietary fibre studies. In Fig. 8, the proximity index pattern of our data file (1970–1974) is shown. Since our coding procedure, limited to 49 keywords, has destroyed most of the information about low-frequency word linkages, the pattern is not very informative.
21. The key-words were chosen to emphasize research approaches and disciplinary contributions; see further the report mentioned in note 14. Work is in progress to analyze the same set of articles on the basis of key-word coding by the French CNRS Pascal system, and to compare the outcomes with the present results.
22. The only fully worked-out example is the case of dietary fibres. For detailed information, see also M. CALLON, J.–P. COURTIAL, W. A. TURNER, *L'Analyse des Mots Associés dans la Littérature Scientifique et Technique – Le Cas des Fibres Alimentaires* Ecole des Mines, Paris, Juillet 1981.
23. The proximity index map for the second period, 1975–1979, shows the fungi agent – cellulotic substrate link, but without the SCP – algae connection. A particularly strong group in this map is formed by solid waste – anaerobe – mixed culture – biogas (with fluid waste attached to it), which may reflect the rising interest in environmental biotechnology and the promise offered by anaerobic processes. In view of the difficulty in interpreting proximity links, we shall not present further details.
24. See for example the International Symposium on Quantitative Measures in the History of Science (1976), from which we have quoted D. EDGE's paper (note 10). Institutional history (e.g. of American chemistry) and collective biographies are producing time series and other quantitative measures and follow the trend of general history more rapidly.