Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 71, No. 2 (2007) 239–269
DOI: 10.1007/s11192-007-1666-3

Do material transfer agreements affect the choice of research agendas? The case of biotechnology in Belgium

VICTOR RODRIGUEZ,[a] FRIZO JANSSENS,[b] KOENRAAD DEBACKERE,[a] BART DE MOOR[b]

[a] *Department of Managerial Economics, Strategy and Innovation, Katholieke Universiteit Leuven, Leuven (Belgium)*
[b] *Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven (Belgium)*

In this paper we examine whether and to what extent material transfer agreements influence research agenda setting in biotechnology. Research agendas are mapped through patents, articles, letters, reviews, and notes. Three groups are sampled: (1) documents published by government and industry which used research materials received through those agreements, (2) documents published by government and industry which used in-house materials, (3) documents published by academia. Methodologically, a co-word analysis is performed to detect if there is a difference in underlying scientific structure between the first two groups of documents. Secondly, interviews with practitioners of industry and government are intended to capture their opinion regarding the impact of the signed agreements on their own research agenda choices. The existence of synchronic and diachronic common terms between co-word clusters, stemming from the first two groups of publications, suggests cognitive linkage. Moreover, interviewees generally do not consider themselves constrained in research agenda setting when signing agreements for receiving research materials. Finally, after applying a co-word analysis to detect if the first group of documents overlaps with the third group we cannot conclude that agreements signed by industry and government affect research agenda setting in academia.

Introduction

Although material transfer agreements (MTAs) may be necessary to exchange research materials between laboratories, academic researchers as well as policymakers

Received June 2, 2006

Address for correspondence:
VICTOR RODRIGUEZ
Department of Managerial Economics, Strategy and Innovation, Katholieke Universiteit Leuven
Naamsestraat 69, B-3000 Leuven, Belgium
E-mail: victor.rodriguez@econ.kuleuven.be

0138–9130/US $ 20.00

have suggested that the trend towards the standardization of MTAs might impede the progress of science and technology by constraining the choice of research agendas. This limitation might be caused by the lack of research materials. Another restriction might be the absence of recipients' freedom to continue a line of research because they no longer own their inventions made through the use of the material. Finally, delays or denials to publish research results that have used the received material might hamper research agendas.

Since knowledge production requires disclosure behaviours (DASGUPTA & DAVID, 1994), the essential variable in the winner-take-all contest among researchers for priority becomes their ability to choose relevant research agendas (CARAYOL, 2003). Thus, scientists and technologists are confronted to what ZIMAN (1987) calls the problem of 'problem choice' in order to determine their research agendas. Generally though, interviewees from industry and government laboratories do not consider themselves constrained in their choice of research agendas when signing MTAs for receiving research materials. On the contrary, MTAs offer important leverage for advancement of their lines of research due to access of materials to carry out the research project.

Our challenge is to detect the effect of MTAs on research agenda setting. The 1990s were the most productive period in biotechnological research: the birth of Dolly, the first successful derivation of human embryonic stem cells, and the completion of the Human Genome Project. Although there has been no formal agreement on a format when a for-profit entity is providing research material to a non-profit organization, a draft text was compromised in 1992 in the United States of America. That is why we chose for our research purposes a time span which runs from 1992 to 2000.

MTAs, pioneered by industry, are increasingly being used by government and academia. Technology transfer offices processed nearly 500 MTAs at the University of Pennsylvania in 1999 (ENSERINK, 1999) and 2000 MTAs at the University of California in 2002 (STREITZ et al., 2003). HANSEN et al. (2005) found that the greatest overall proportion of their survey respondents acquired their last patented technology through MTAs. As agreements, MTAs may take a variety of forms – from letter statements accompanying a shipment of materials to detailed and formally negotiated contracts signed by both parties before a transfer is made. MTAs are executed when proprietary materials are transferred in or out of laboratories. In this paper, we are only focusing on articles, letters, notes, reviews, and patents – further on mentioned as documents – which have used research materials received through MTAs or not. Receivers in MTAs might suffer erosion of research freedom imposed by providers to protect their scientific or technological lead, to slow the dissemination of undesired results, to allow time to negotiate a patent, or resolve disputes over ownership of intellectual property.

MTAs are not only worthwhile studying in their own right, but they might provide insight into wider knowledge flow dynamics between research units. So, we would like

to address the following questions: Is the research agenda choice modified because of MTAs? Can MTAs encroach on the flow of scientific information and distort the content of research programs? Can a research line be eroded or changed because research laboratories sign MTAs? Particularly, in this paper we want to examine whether and to what extent MTAs influence research agenda setting in the sample. We represent research agendas through co-word clusters from titles and abstracts of the sampled documents.

If MTAs signed in industry and government have an effect on the research agenda choice in the same sector, does it mean that terms should differ between co-word clusters stemming from documents which used MTAs and those which did not? So, our first experiment was to detect divergence of research topics between the two groups of documents in industry and government. As we found some common terms, does it mean absence of deviation of research topics and no effect of MTAs on research agenda setting in industry and government? For validating the results we used two steps. First, we asked practitioners from industry and government to judge the impact of MTAs on defining choices in their research agendas. Secondly, as they generally did not suffer from MTAs for setting research agendas, we searched for divergence with academia using co-word analysis.

Moreover, if MTAs signed in industry and government have an effect on the research agenda choice in academia, then does it mean that terms should differ between co-word clusters stemming from documents which used MTAs and those from academia? Again, we found that almost the whole vocabulary of documents which used MTAs consists of terms used in academia. Does this convergence of research topics mean that there is no effect of MTAs signed in industry and government on research agenda setting in academia?

Theoretical background

In this section we resort to well-accepted and researched principles that back our study. Firstly, competitive markets provide poor incentives for the production of knowledge. SCHUMPETER (1934) has argued that scientific and technological progress is brought about through exclusion due to the need of industry to protect itself from the risks associated with innovation. Secondly, scientific and technological productivity is characterized by extreme inequality. MERTON (1968) developed his explanation for inequality in science and technology by defining the Matthew Effect.

We recall that exclusion in the market is achieved through patents. Thus, we were able to find the first principle among government and industry, i.e., they mainly disclose research through patents. The second principle was found in the whole Belgian biotech sample, i.e., only a few assignees of patents or institutional authors of articles, letters, notes, or reviews, disclose the larger part. While we have yet to identify possible

obstacles to the progress of science and technology, we can translate the Schumpeterian and Mertonian principles into an oligopoly of scientific and technological producers.

The question is what sort of scientific and technological contest is socially most desirable. If each of those oligopolistic producers finds it necessary to compete in terms of the quality of their scientific and technological products and research, or by means of producing better science and technology, the performance of the scientific and technological domain may well be satisfactory. Scientists and technologists can minimize the threat of being scooped in a winner-take-all contest by seeking ways to monopolize a line of research (STEPHAN, 1996).

Hence, MTAs have appeared in science and technology to help minimize that risk. If some provisions are not followed, the contract is breached and the wronged party has the right to bring action against the other, such as suing for damages. The transferred material may also be protected against theft or trickery, making a third party liable for damages. Unlike patents or copyrights, MTAs do not rest upon codified legal statutes defining specific rights and obligations (RODRIGUEZ, 2005).

MURRAY (2006) studies the history of the Oncomouse from the institutional theory perspective. Science, as institution, has changed and its logic has adapted to the new economic reality which modified relationships within the institution. Scientific relationships as exchange of research materials use formal governance mechanisms as MTAs.

The tragedy of the anticommons (HELLER & EISENBERG, 1998) helps explain why people underuse scarce resources like research materials. MURRAY & STERN (2005) found evidence for a modest anti-common effect in biotechnology applying the concept of dual knowledge disclosure (MURRAY, 2002). The NATIONAL ACADEMIES (2005) showed no substantial evidence for a patent thicket or a patent-blocking problem in genomic and proteomic. WALSH et al. (2005) have evaluated the impact of MTAs on reach-through claims, publication restrictions, and research material access.

If MTAS hinders research agenda choice, then documents that used materials received through MTAs will not have overlapping terms with documents that used in-house materials. In theory, in a world of costless transactions, the anticommons tragedy could always be avoided by trading proprietary rights (COASE, 1960). In practice, however, avoiding the tragedy requires overcoming transaction costs, strategic behaviours, and cognitive biases of owners.

In a *quid pro quo* approach, two models may facilitate access to patented research materials: patent pools (VERBEURE et al., 2006) and clearing houses (VAN ZIMMEREN et al., 2006). The former concept is used to describe mechanisms whereby providers and users are matched (KRATTIGER, 2004). Patent pools are agreements between patent holders to license each other or to third party their patents.

If research institutions lack of margin of manoeuvre for obtaining patented research materials, there exist three alternatives available: research or experimental use

exemption, conventional one-to-one licensing, and compulsory licenses (VAN OVERWALLE et al., 2006). But problem arises when the material transfer occurs before the provider files a patent application of it. MTAs with confidentiality provisions or trade secret contracts may be the solution.

Empirical framework

As we want to clarify a challenge that MTAs pose for the conduct of science and technology, we have to capture the effect of MTAs on research agenda setting. As literature has long been assumed to represent scientific or technological activity (MERTON, 1942), a map based on publications within a domain can be considered to represent its underlying structure. A method for identifying scientific and technological concept networks and studying their evolution on the basis of documents is co-word analysis. It is about the use of frequencies with which possible pairs of words co-occur in single documents as a means to the elucidation of structures of problems embodied in them (WHITTAKER, 1989).

LEYDESDORFF (1997) considers that the subsumption of textual signals under keywords assumes stability in the meanings, but COURTIAL (1998) replies that words, in co-word analysis, are not used as linguistic items to mean something, but as indicators of links between texts. Signal words are used in papers and patents to guide the reader, as a funnel of interest (WILLIAMS & LAW, 1980). By applying content analysis to a document, our study works in the other direction. Titles and abstracts of articles, letters, notes, reviews, and patents are transformed into a set of co-word clusters in order to capture their cognitive interest structure. Thus, co-word analysis offers a flexible way to enter into and to unravel the content structure of a scientific or technological domain. We assume that cognitive aspects can to some extent be treated quantitatively. Operationalisation always has the potential to introduce bias and distortion. There is however no cognitive message that could be perfectly mapped.

Potential biases stem from the indexer effect and the audience effect. Firstly, the representations of scientific and technological fields were influenced by the ways in which indexers who chose the keywords conceptualized the scientific fields with which they were dealing, so that representations which emerged were more akin to their conceptualizations than to those of their authors whose work it was intended to study. Secondly, authors might choose their title words deliberately in order to address a particular readership.

The definition of a research problem has been shown to be a highly strategic and controversial activity. For coping with the dichotomy between cognitive and social factors influencing knowledge production, CALLON et al. (1983) use the notion of translation. Given a problem Θ_1, its solution is made to depend upon the solution to problems Θ_2, Θ_3, Θ_4, …, Θ_i. What is crucial for all translations is the identification of

problems and the establishing of relationships between them. Consequently, the decision to undertake a particular investigation as resulting from the choice of a research problem can be influenced by the availability of research materials, research publications, or new research lines.

How should the findings of co-word analysis be used? It is up to the researcher to explain the significance of the various indices used, to suggest possible alternatives, and to offer possible interpretations of the data. As science and technology have their own internal landscape which evolves and reshapes itself continually, we should not feel that science and technology can be treated as a black box with inputs and outputs. Since tools such as co-word analysis may open the black box and explore the topography of science and technology, we can articulate cognitive links (LAW et al., 1988).

This particular application of science maps seemed very promising in the 1970s and 1980s. Since mid 1990s it experienced a revival due to information technology breakthroughs. The applicability of new analytical software and the availability of hypertext and graphical interfaces provided new impulses for science mapping based on co-word analysis. Particularly, if we are able to identify themes in a research area by clustering terms from titles and abstracts in publications, then we can create maps on the basis of the cognitive relations between themes (NOYONS, 2001).

Data

As we want to examine whether MTA-use influences the choice of research agendas, we can represent research agendas through co-word clusters from titles and abstract of sampled documents. Another approach could be stemming words from patent claims (VERBEURE et al., 2006) or findings from articles. For our co-word analysis, the documents retrieved were disclosed between 1992 and 2000 by industry, government, and academia in Belgium (Table 1). A document producer is an assignee of patents or institutional author of articles, letters, notes, or reviews.

Table 1. Sampled production of biotech documents in Belgium between 1992 and 2000

1992–2000	Patents	Articles, letters, notes, and reviews	Total documents	Document producers
Industry and government	241	255	496	20
Academia	88	6952	7040	17

Note: in our sample industry is formed by for-profit corporations; government is composed by public research institutes; and academia is constituted by universities and colleges.

We have selected those documents from the database created by GLÄNZEL et al. (2003). Particularly, we have only focused on the core set of disclosures. Papers were articles, letters, notes or reviews published in the journals listed in the Appendix and were retrieved from the following subject categories of ISI Web of Science:

biochemical research methods, biochemistry and molecular biology, biophysics, biotechnology and applied microbiology, cell biology, developmental biology, genetics and heredity, microbiology, and plant sciences.

The retrieved European patents were applied in the following patent classes of the International Patent Classification: C12M (apparatus for enzymology or microbiology); C12N (micro-organisms or enzymes; propagating, preserving, or maintaining microorganisms; mutation or genetic engineering; culture media); C12P (fermentation or enzyme-using processes to synthesize a desired chemical compound or composition or to separate optical isomers from a racemic mixture); C12Q (measuring or testing processes involving enzymes or micro-organisms; compositions or test papers therefore; processes of preparing such compositions; condition-responsive control in microbiological or enzymological processes); C12S (processes using enzymes or micro-organisms to liberate, separate or purify a pre-existing compound or composition; processes using enzymes or micro-organisms to treat textiles or to clean solid surfaces of materials); C07G (compounds of unknown constitution); and C12R (indexing scheme related to subclasses C12C to C12Q or C12S, related to micro-organisms).

In order to study the effect of MTAs on research agenda setting, we have to detect if there is a difference between documents which used materials through MTAs and documents which used in-house materials. We decided to skip university disclosures because of three reasons. Firstly, technology transfer offices were set up in Belgian universities in late 1990s. Secondly, MTAs formalizing material reception by university laboratories were decentralized among researchers, i.e., they kept in their drawers the contract without notifying to the university authorities the signature of MTA. Thirdly, personnel rotation, co-authorship, and time constraints were decisive factors. Consequently, if we would have decided to distinguish which documents were related to MTAs in academia, we should have interviewed each scientist who had published in that period, what was beyond data collection scope.

Hence to have a group of documents which used materials received through MTAs, we asked to industry and government representatives to distinguish whether their documents (Table 1) disclosed between 1992 and 2000 were related to MTAs. So, the first group (F_1) are documents which used research materials received through MTAs by industry and government. The second group (F_2) is made up of documents which used in-house research materials in industry and government. The third group (F_3) is constituted by documents published by academia. Table 2 compiles the number of documents corresponding to three different time periods for each group in the sample.

Table 2. Distribution of documents between 1992 and 2000

Period	F_1	F_2	F_3
1992–1994	11	113	2070
1995–1997	11	135	2547
1998–2000	20	206	2423

Words were retrieved from both titles and abstracts of documents disclosed between 1992 and 2000 (Table 3). The F_1, F_2, and F_3 documents contained respectively 645, 2102, and 15019 distinct non-trivial stems or stemmed words, further on mentioned as terms. There were 291 terms which occurred at least twice in F_1, 2015 in F_2, and 15012 in F_3. There were 150 which occurred at least thrice in F_1, 1385 in F_2, and 10685 in F_3. The number of co-occurrences was different: 1750 with frequencies of three or more among the F_1 documents, compared to 40384 among F_2 documents and to 1406536 among F_3 documents. The total number of terms was 1386 in F_1, 15020 in F_2, and 321074 in F_3. There were 35 terms per document in F_1 and F_2, and 46 in F_3.

Table 3. Summary data

Stems or stemmed words	F_1	F_2	F_3
Total used	1386	15020	321074
Number per document	35	35	46
Number of different stems	645	2102	15019
Number occurring 3 or more times	150	1385	10685
Number occurring 2 or more times	291	2015	15012
Co-occurrences ($F_{min} = 3$)	1750	40384	1406536

Despite the heterogeneity of the research subjects – for instance, plant cell division, staphylokinase derivatives, etc. – treated in the sample, we were able to establish relationships between terms from different clusters. Terms were not only cognitive content to sketch out research programs, but also they helped to identify different translation strategies. A research program is defined as a series of hierarchised problems whose resolution is considered to be crucial to the future of the domain (BASTIDE et al., 1989).

The problem of relationships between input and output data in order to explain the role of MTAs in the development of research programs is a difficult one. In this study we have considered co-word analysis as a means of resolving it. A research system dynamically evolves as a result of the decisions taken by its parts to engage their activity in a given direction. The co-word analysis technique was developed to detect the degree of convergence of these decisions through the analysis of a publications database. The research areas themselves can be classified as lying in the mainstream of the work underway in a research system or, on the contrary, as being of secondary importance because they are isolated, peripheral, or unstructured in the network (TURNER & ROJOUAN, 1991).

Methodology

Indexing

We indexed with the software Jakarta Lucene, using the vector space model, titles and abstracts of 496 documents in industry and government, and 7040 in academia. The grammatical structure of the text and common stopwords – words with little or no semantic value – were neglected. Furthermore, Porter's stemmer was applied to all remaining words encountered in the titles and abstracts, leading to an initial vocabulary or thesaurus of 5012 terms in industry and government, and 31786 in academia. To maintain the most important terms for analysis, only terms occurring in noun phrases were kept. This list was compiled by using part-of-speech tagging functionality of the software LT POS and LT Chunk. Moreover, the log-likelihood method of DUNNING (1993) for detection of bigrams was followed to detect bigrams, trigrams, and tetragrams within those noun phrases (MANNING & SCHÜTZE, 2000). After manual editing, the final list contained 66 phrases in industry, government, and academia. Besides, by cutting off Zipf's curve, terms or phrases that only occurred in one document or in more than 50% of all documents were neglected. Finally, 11 synonym rules were written in industry, government, and academia, e.g., to map 'severe acute respiratory syndrome' onto SARS. Hence, the vocabulary of the final index contained 2125 stems or stemmed phrases in industry and government, and 15019 in academia. The result of indexing was a 2125×496 term-by-document matrix in industry and government, and 15019×7040 in academia.

Term co-occurrence analysis

As it is necessary to have a minimal number of documents to execute the statistical analysis, we opted to group documents in 6 sub-sets as shown in Table 1. Each sub-set passed through the following filters. From the indexed term-by-document matrix, a sub-matrix was constructed for each sub-set by only keeping those documents (columns) that belong to the sub-set and only those terms (rows) that appear in at least two documents. Another filter for term significance was applied by imposing a threshold equal to 5 for the largest term frequency–inverse document frequency (TF-IDF) value of a term in the complete set, further explanations of this weighting scheme can be found in the Appendix. A last term filter was applied by requiring that the largest equivalence index of a term in a sub-set be higher than 0.2 in order to drop terms that have no strong association with others in the sub-set. After applying those filters per group and period we finally obtained the sub-sets of terms shown in Table 4.

Table 4. Terms per sub-set

Period	1992–1994	1995–1997	1998–2000
F_1	46	84	132
F_2	683	800	505
F_3	1311	1502	1384

Clustering

We applied Ward's hierarchical clustering (JAIN & DUBES, 1988) by considering as input the distance matrix derived from the equivalence index matrix for each sub-set. To determine the optimal number of clusters in each sub-set, we inspected four diagrams: dendrograms, stability diagrams, mean silhouette curves and silhouette plots.

A first judgment is offered by the dendrogram. For instance, Figure 1 depicts a dendrogram for F_1 in 1995–1997, cut-off at 11 clusters. The vertical line illustrates the candidate cut-off point of 3 clusters with best terms 'enhanc', 'gene', and 'staphylokinas' for clusters c1 to c3, respectively. The horizontal lines connect clusters in a hierarchical tree. The line length represents the distance between two connected terms or clusters. At each leaf node, the term representing the cluster has the highest mean TF-IDF value in the sub-set.

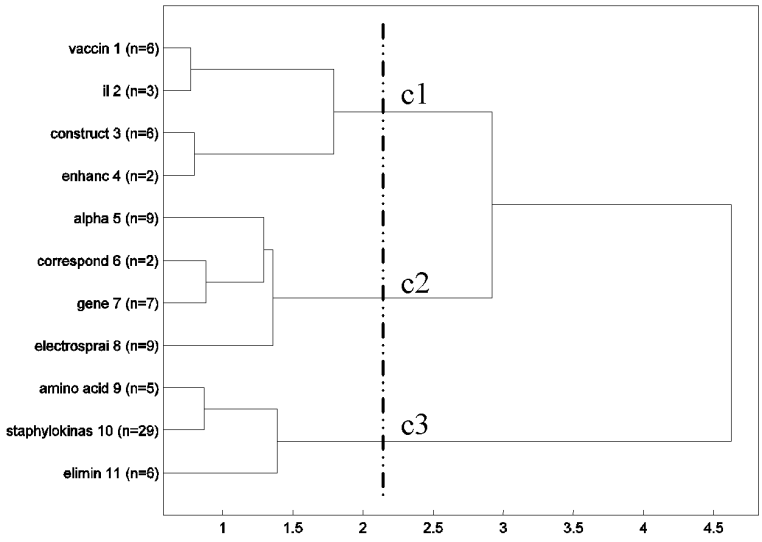

Figure 1. Dendrogram for F_1 in 1995–1997

A second appraise for the optimal number of clusters is given by the stability diagram proposed by BEN-HUR et al. (2002). As an example, Figure 2 shows, for 2 up to 25 clusters, the cumulative distribution of pairwise similarities – quantified by the Jaccard coefficient – between 1000 pairs of clustering solutions for random sub-samples, and each comprising 85% of the terms. High pairwise similarities indicate a stable clustering pattern. Due to the small sub-sample size of this example, for 17 clusters or more there is at least one of them containing only one term and consequently they are not a valid cluster number. A solution with 2, 3, 7 or 8 clusters is a good one; and 4, 5, or 6 clusters are not so stable.

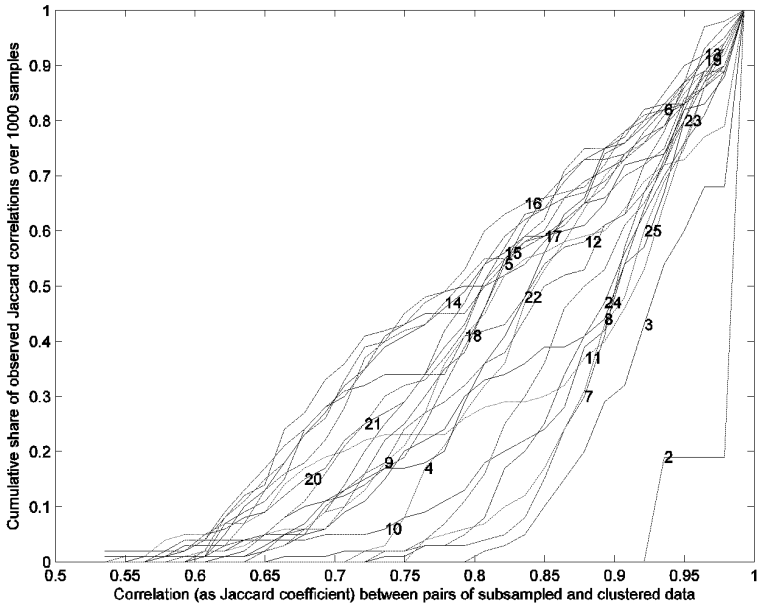

Figure 2. Stability diagram for F_1 in 1995–1997

Thirdly, we observed the mean silhouette curve (ROUSSEEUW, 1987). To illustrate, Figure 3 shows, for 2 up to 25 clusters, a local maximum at 4 clusters – but not as higher than the mean value at 3 clusters – which is a better solution according to the stability diagram in Figure 2. All in all, after examining the three diagrams (Figures 1–3), we established the optimal number of clusters for this sub-set, which is 3. Indeed, Figure 4, for a solution of 3 clusters, depicts a silhouette plot with only a few terms having a negative value – meaning that they would rather be in another cluster.

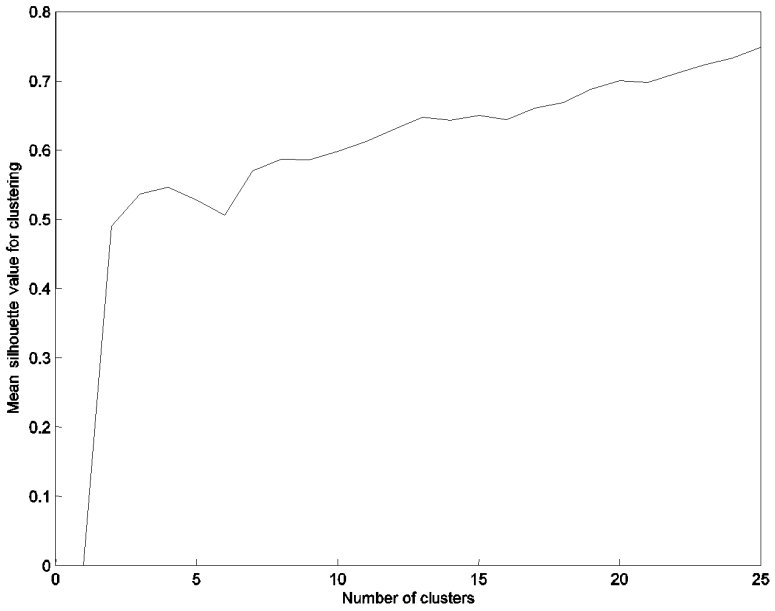

Figure 3. Mean silhouette curve for F_1 in 1995–1997

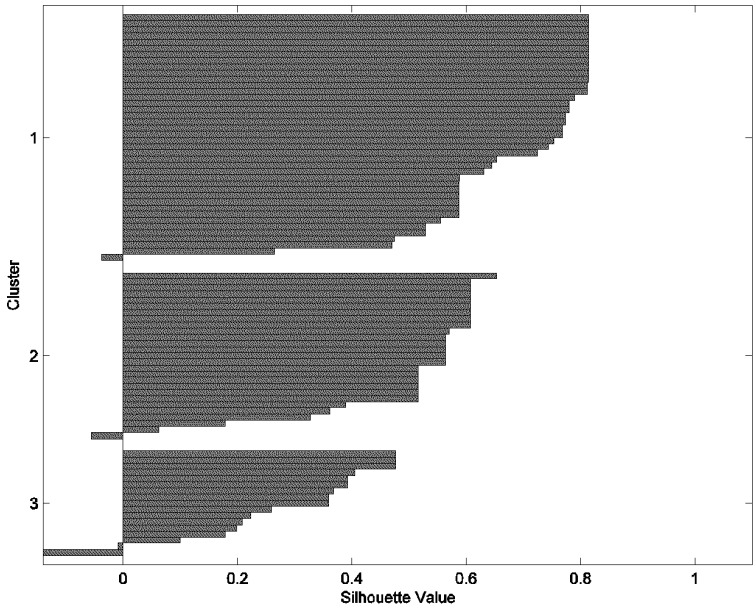

Figure 4. Silhouette plot for 3 clusters of F_1 in 1995–1997

For the remaining F_1, F_2 and F_3 sub-sets, we executed the same procedure and we obtained the optimal numbers of clusters shown in Table 5. A caveat must be stated for F_3, for each period there was a huge rest cluster which had the smallest density. It might be advisable to ignore it, or at least handle it with care.

Table 5. Optimal number of clusters

Sub-set	1992–1994	1995–1997	1998–2000
F_1	4	3	3
F_2	31	30	23
F_3	14	21	19

Strategic diagrams

Once the number of co-word clusters for each sub-set of documents was determined, each cluster was featured by an index of centrality and density, and plotted into a strategic diagram (Figures 5–10) split into four quadrants based on the classification developed by CALLON et al. (1991), viz. 1 is central and visible topics, 2 is isolated topics, 3 is peripheral topics, and 4 is unstructured topics. Clusters were identified in the strategic diagram by the term of the cluster which has the highest mean TF-IDF value in a sub-set.

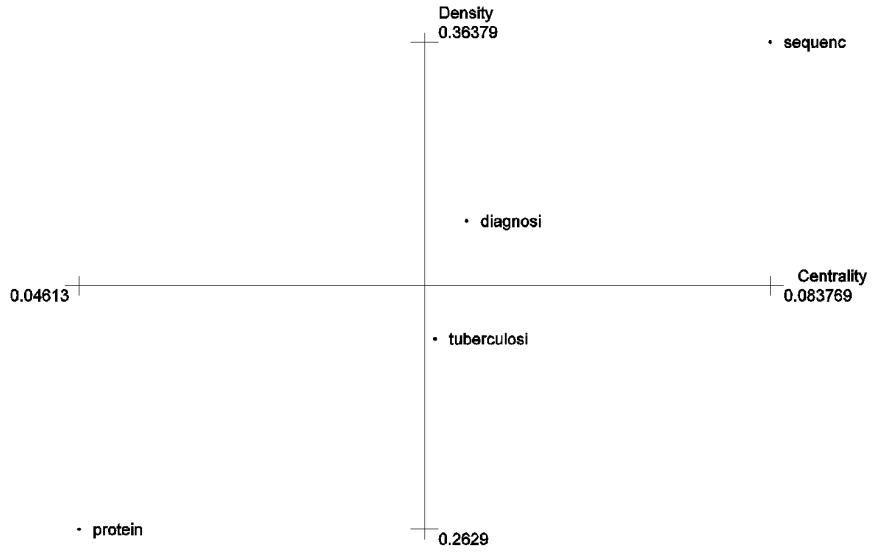

Figure 5. Strategic diagram for F_1 in 1992–1994

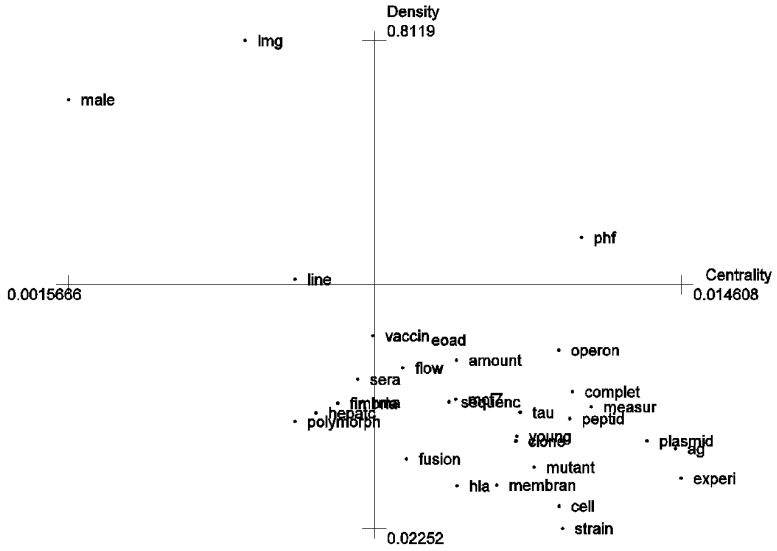

Figure 6. Strategic diagram for F_2 in 1992–1994

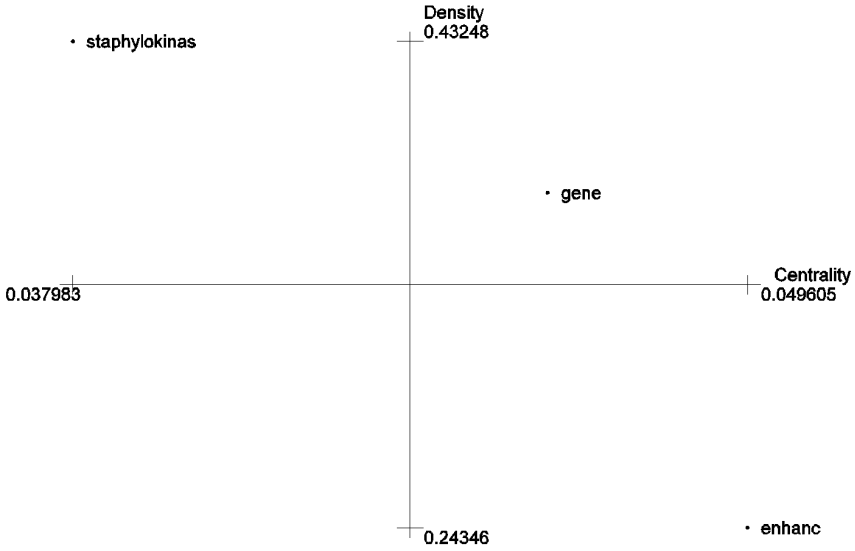

Figure 7. Strategic diagram for F_1 in 1995–1997

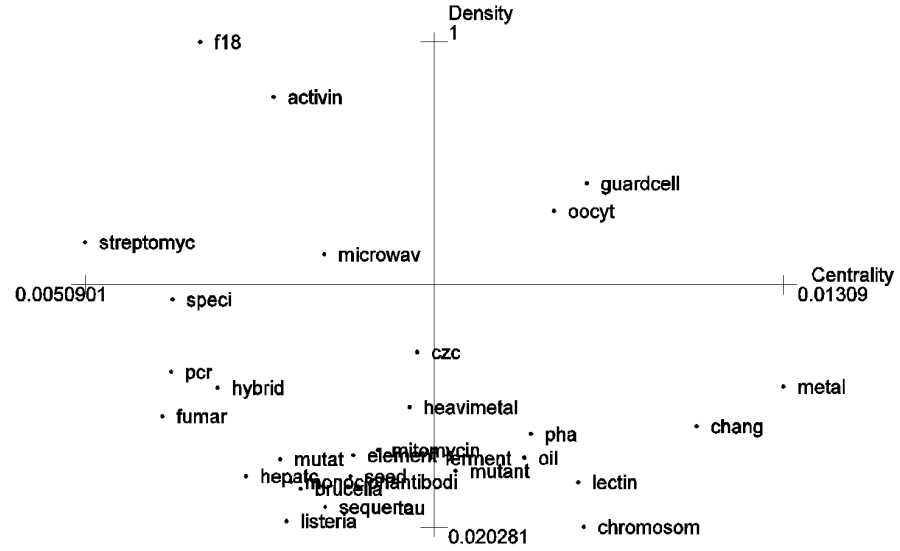

Figure 8. Strategic diagram for F_2 in 1995–1997

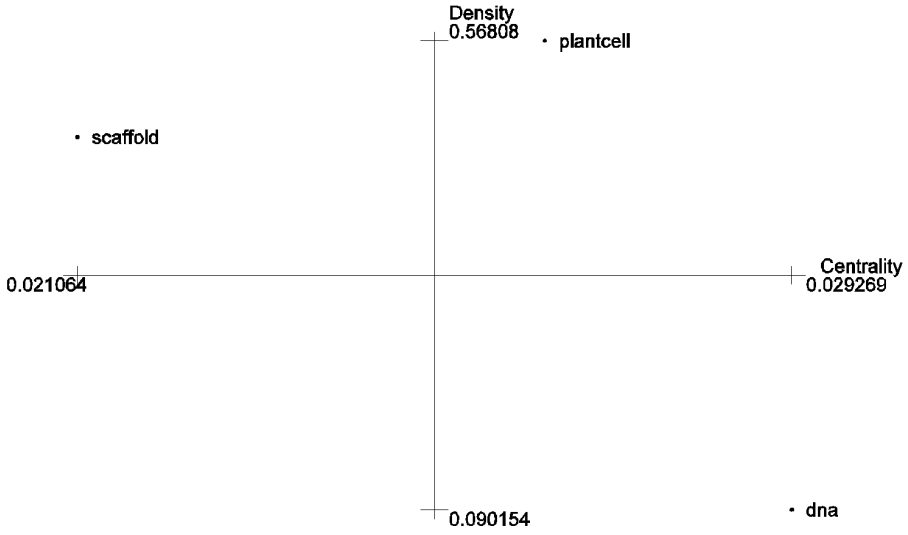

Figure 9. Strategic diagram for F_1 in 1998–2000

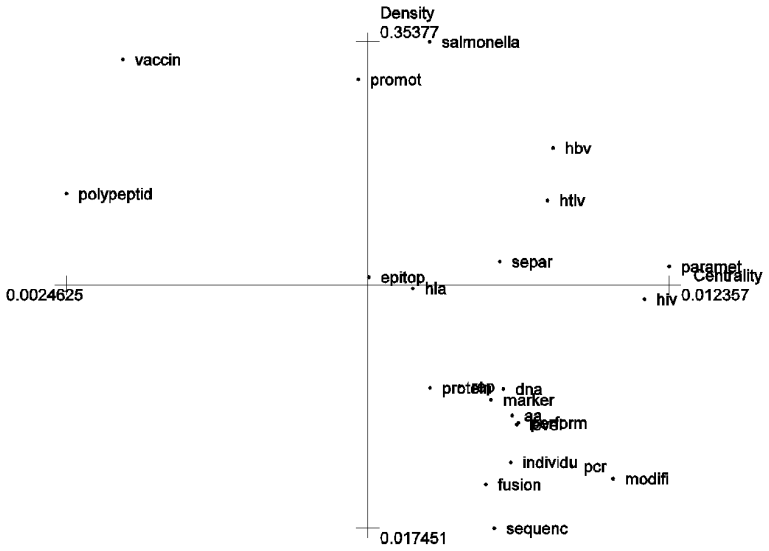

Figure 10. Strategic diagram for F_2 in 1998–2000

Results

As we want to examine whether MTA-use influences the choice of research agendas in industry and government, we represented research agendas through co-word clusters from titles and abstract of sampled documents. Research agendas were analyzed synchronically – the relationship among clusters in the same time period – and diachronically – the evolution of clusters over time. How can divergence of research topics between two groups of documents be grasped? Is it plausible to state that difference in underlying scientific structure between both groups of documents means absence of common terms in those two groups of documents? If we find common terms, does it convey no divergence of research agendas in industry and government? In the affirmative case, how powerful should be common terms to postulate no divergence?

Given co-word clusters, do synchronic and diachronic common terms tend to introduce relations among them? This question suggests that topics located in different strategic diagrams could be cognitively linked to one another despite the fact that, at any given moment in time, these links might not yet be identified. This rather strong assumption could lead to qualitative indications. If these effects really do exist, we would have obtained both a sort of qualitative and quantitative model for describing science and technology in the making. This model would consist of a list of co-word clusters composed of topics interacting through a synchronic and diachronic model (COURTIAL, 1989).

Regarding robustness of common terms, we can use three approaches: the strategic diagram quadrants, theoretical ambitiousness, and mean TF-IDF value. Firstly, if we split the common words into the four categories of topics, then we could consider that central and visible topics are more powerful than the other ones. Secondly, if we order common terms decreasingly according to their mean TF-IDF values, we could consider that higher ones are more powerful than the others (Figure 11). Thirdly, if we rank some of the common terms according to the theoretical ambitiousness level of RIP & COURTIAL (1984), we could consider that those placed at higher levels in theoretical ambitiousness (Table 6) are more powerful than those placed at lower levels.

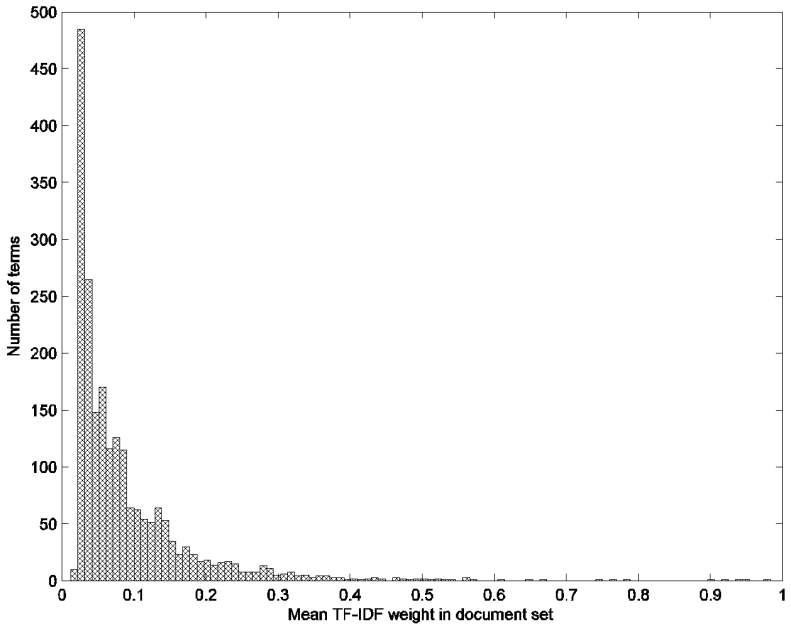

Figure 11. Histogram of mean TF-IDF values for the vocabulary

Table 6. Theoretical ambitiousness of common terms

Level	Rip and Courtial	Weingart and Van den Daele
1	Screening	Measurement, monitoring
2	Costs	
3	Design	Measurement, monitoring Functional explanations, input-output relations
4	Immobilization	
5	Product isolation	
6	Parameter optimization	Functional explanations, input-output relations
7	Mathematical modelling	Functional explanations, input-output relations Causal explanation, mechanisms
8	Physical kinetics	
9	Biokinetics	
10	Biodynamics	Causal explanation, mechanisms

Source: RIP & COURTIAL (1984).

Synchronically, we have found common terms between F_1 and F_2 clusters in the three time periods: 36 in 1992–1994, 57 in 1995–1997, and 86 in 1998–2000. If we use the central and visible approach for robustness (Table 7), we can find only one common term between F_1 and F_2 in the period 1992–1994, none in the period 1995–1997, and 17 common terms in the period 1998–2000. If we use the mean TF-IDF value approach, we can find that the mean TF-IDF value of common terms for 1992–1994 is 0.42242, for 1995–1997 is 0.354434, and for 1998–2000 is 0.33575, which are higher than that of the vocabulary, 0.0927.

Table 7. Common terms between F_1 and F_2 clusters in same periods

Period	Central and visible terms	Isolated terms	Peripheral terms	Unstructured terms	Total common terms	Total terms by period	Common/ Total terms
1992–1994	1	2	11	22	36	729	5%
1995–1997	0	15	17	25	57	848	6%
1998–2000	17	4	0	65	86	637	14%

If we use the theoretical ambitiousness approach, we can find 4 common terms at low level and 3 at middle level in 1992–1994; 7 common terms at low level and 2 at middle level in 1995–1997; 3 common terms at low level, 3 at middle level, and 1 at high level in 1998–2000. In 1992–1994, all low level common terms are unstructured; among the middle level common terms, 2 are unstructured and 1 is peripheral. In 1995–1997, among the low level common terms, 2 are isolated, 1 is peripheral, and the rest is unstructured; among the middle level common terms, 1 is peripheral and 1 is unstructured. In 1998–2000, all the low and middle level common terms are unstructured, the high level common terms turn out to be central and visible. So, newly appearing biotech research themes, mainly measurement and monitoring, are predominant when intersecting term clusters of F_1 and F_2 in the same periods.

Diachronically, we have found common terms between F_1 and F_2. There are 34 common terms between F_1 clusters in the period 1992–1994 and F_2 clusters in the period 1995–1997. There are 27 common terms between F_1 clusters in the period 1992–1994 and F_2 clusters in the period 1998–2000. There are 55 common terms between F_1 clusters in the period 1995–1997 and F_2 clusters in the period 1998–2000. If we use the central and visible approach (Table 8), we can find none central and visible common terms between clusters F_1 1992–1994 and F_2 1995–1997. But we found seven central and visible common terms between clusters F_1 1992–1994 and F_2 1998–2000, 14 between clusters F_1 1995–1997 and F_2 1998–2000.

Table 8. Common terms between F_1 and F_2 clusters in future periods

Period	Central and visible terms	Isolated terms	Peripheral terms	Unstructured terms	Total common terms	Total terms by period	Common/ Total terms
F_1 1992–1994 F_2 1995–1997	0	8	12	14	34	729	5%
F_1 1992–1994 F_2 1998–2000	7	1	0	19	27	884	3%
F_1 1995–1997 F_2 1998–2000	14	0	0	41	55	637	9%

If we use the mean TF-IDF value approach, we can find that the mean TF-IDF value of common terms between clusters F_1 1992–1994 and F_2 1995–1997 is 0.424959, between clusters F_1 1992–1994 and F_2 1998–2000 is 0.444159, between clusters F_1 1995–1997 and F_2 1998–2000 is 0.3381, which are higher than that of the vocabulary, 0.0927. If we use the theoretical ambitiousness approach, we can find 4 common terms at low level and 1 at middle level between clusters F_1 1992–1994 and F_2 1995–1997; 1 common terms at low level and 3 at middle level between clusters F_1 1992–1994 and F_2 1998–2000; 5 common terms at low level and 4 at middle level between clusters F_1 1995–1997 and F_2 1998–2000.

For the cluster intersection between F_1 1992–1994 and F_2 1995–1997, we can find that among the low level terms 1 is isolated, 2 are peripheral, and 1 is unstructured; that the middle level term is peripheral. For the cluster intersection between F_1 1992–1994 and F_2 1998–2000, we can find that the low level term is unstructured, that among the middle level terms 1 is central and visible and 2 are unstructured. For the cluster intersection between F_1 1995–1997 and F_2 1998–2000 we can find that among the low level terms 1 is central and visible and the rest are unstructured, that all middle level terms are unstructured. Again, newly appearing biotech research themes are predominant when intersecting F_1 clusters in a certain period with F_2 clusters in a future period. Nonetheless, functional explanation and input-output relation are almost as frequent as measurement and monitoring.

If we would have not found common terms, might it be due to differences in disciplines of biotechnology? A research trajectory is defined as much by context as it is by content. Context is not so much a function of the disciplines of origin of scientists or technologists, but of the links that they create in their own research between different research problems. Local context is measured by the density or internal cohesion index. By contrast, global context is measured by the centrality index.

As we found some common terms, characterized by high mean TF-IDF, newly appearing research topics, and usually measurement or monitoring, does it mean absence of deviation of research topics and no effect of MTAs on research agenda setting in industry and government? If MTAs signed in industry and government have

an effect on the research agenda choice in the same sector, does it mean that terms should differ between co-word clusters stemming from documents that used MTAs and those which did not?

Discussion

For validating the results we used two steps. Firstly, we asked practitioners from industry and government to judge the impact of MTAs on defining choices in their research agendas. Secondly, we searched for divergence of research agenda between F_1 and F_3. Apart from asking interviewees to discriminate documents (Table 1) for obtaining F_1 and F_2 in order to allow us to perform the co-word analysis, we have also posed other questions to them (Table 9) in order to obtain their opinions on MTA-impact on the choice of research agenda setting.

Table 9. Sampled practitioners' opinion on MTA

Question	Yes	No
Was there in the period 1992-2000 any research output not submitted for publication (article or patent) as a consequence of any MTA signed by your organization?	10%	90%
Was there in the period 1992-2000 any research output delayed for publication (article or patent) as a consequence of any MTA signed by your organization?	40%	60%
Have you given up any research project between 1992 and 2000 because you did not have the research material?	60%	40%
Have your research projects been delayed because you were obliged to create the material (available in the literature) needed for those projects?	90%	10%
Have you ever decided to make the material (available in the literature) after reading the MTA clauses imposed by the provider of the material?	30%	70%
Was the availability of research material at no cost between 1992 and 2000?	40%	60%
Were there conflicting obligations to provider and financial sponsor between 1992 and 2000?	40%	60%
Were the use restrictions (material distribution, limited use to laboratory, return of unused material) difficult to track between 1992 and 2000?	50%	50%
Did material definitions include unmodified derivatives, variants and confidential information between 1992 and 2000?	90%	10%
Was there any determination of commercial research between 1992 and 2000?	40%	60%
Was there any distinction between use and transfer of the material between 1992 and 2000?	70%	30%
Were there royalty-free, nonexclusive license rights to the provider between 1992 and 2000?	70%	30%
Were there permissions of providers to license to third parties between 1992 and 2000?	0%	100%
Were there joint ownerships of research results due to material received between 1992 and 2000?	70%	30%
Did the provider bear the costs of patents prosecution of research results based on the material received between 1992 and 2000?	20%	80%
Did any provider terminate problematic situations between 1992 and 2000?	20%	80%

Note: Opinions based on 20 interviewees from industry and government research biotech labs.

Practitioners answered the questionnaire following a broad perspective without focusing in a particular MTA to avoid breaching the contract and contingent legal consequences. A final caveat is warranted, the patterns observed in Table 9 should rather be considered only as indicators due to the small sample size.

Providers control receivers' publications to determine whether their own confidential information has been improperly disclosed, and whether there are new intellectual property rights. As far this publication screening is concerned, 90% of interviewees said that there was no document not submitted for publication because the provider rejected their disclosure request and 60% of them said that there was no document delayed to be submitted for publication because the provider postponed it.

Regarding access to the research materials that are available in the literature, 70% of interviewees said that it happened to them that they decided not to make a material after reading the MTA clauses imposed by the provider. Apart from that, 60% of them also said that some research projects were delayed because they were obliged to create the material. Nonetheless, 60% of interviewees said that they have given up some research projects, because they did not have the research material.

As there is a perceived limitation to publish research results or access research materials by the minority, will researchers either investigate other topics or a different line of research? The opinion of the majority goes with our findings, i.e., MTAs might not affect research agenda setting. Accession problem in cases where MTAS are no possible might be solved through clearinghouses, patent pools, research or experimental use exemption of patents, conventional one-to-one licensing, and compulsory licenses.

As some materials can be very costly to make, and it can be financially unreasonable to supply them to multiple investigators, 60% of interviewees said that the research material was not free. If this is a deterring factor, the agreement can include the proposal of a one-time fee to allow for cost recovery. Such a fee can reasonably include the cost of materials, the extra labour required to make them, and shipping or other fees (COUNCIL OF GOVERNMENTAL RELATIONS, 1997).

Since the provider of the material is usually not funding the research, the recipient needs to ensure that there are no conflicting obligations between its financial sponsors and the provider. Among 60% of interviewees there were no conflicting obligations but what about the rest? Does a conflicting obligation hamper a research project? We believe that the project can be executed. Nevertheless, what matters here is who will benefit from the research results, the financial or the material provider? Consequently, research agenda setting might not be affected by this factor.

In addition to what has been said, the half of the practitioners found that the use restrictions were difficult to track. Are control of asset distribution, limited use to laboratory, and return of unused material determinant to change the research subject? We do not believe that these administrative tasks imposed by MTAs can affect research agenda setting.

Providers asserted ownership not only of the physical material, but also of unmodified derivatives, variants and confidential information, according to 90% of interviewees. Does this broad definition of material affect research agenda? This not only represents a direct loss, but could also cause indirect damage by limiting the freedom of recipients to continue a line of inquiry because they no longer own their research results.

If pre-emptive MTAs cloud ownership rights, investigators may be restricted in their ability to interact with a future sponsor. Investigators may need a commercial developer to convert an invention into a product, but intellectual property clauses in MTAs may prevent the institution from granting rights to a future developer. When 60% of interviewees said that the material received should not be used in commercial research, we have to think that no sponsor wants to pay for research benefits that it cannot have. In this case, research agenda are affected by the ban of commercial research.

Regarding research collaboration – 70% of interviewees said that MTAs distinguished between use and transfer of the material – we must think of the difficulty to cooperate with other scientists. May this affect research agenda? Are investigators setting aside research lines because they cannot collaborate? On the contrary, 70% of them said that there were royalty-free, nonexclusive licence rights to the provider, which allows cooperation. Further research must be done before stating the relationship between research collaboration and research agenda choice.

The fact that providers do not license to third parties or do not bear the patent prosecution costs does not affect research agenda choice of receivers. However, joint ownerships of research result, between providers and receivers, may increase choices of research lines. Difficulties may arise when a receiver uses two materials from two different providers, or has made one of the materials under company sponsorship. In such a situation, it is quite likely that the MTA-clauses covering the two materials are in conflict. According to 80% of interviewees, providers did not terminate such problematic situations.

If MTAs signed in industry and government might have an effect on the research agenda choice in academia, then would it mean that terms should differ between co-word clusters stemming from F_1 and F_3? Before applying the filters, we compared the vocabulary of industry and government (645 terms) to that of academia (15019 terms) and we obtained 62 non-overlapping terms. So, 90% of F_1 terms were included in F_3 before filtering. If we look at important terms (after filtering) we still find 31 common terms, or more than 10 percent. Does this mean that there is "overlap" of research topics between F_1 and F_3? We think that this modest but existing overlap might indicate no effect of MTAs signed in industry on research agenda setting in academia.

Concluding remarks

As MTAs signed in industry and government did not affect research agenda setting neither in the same sector nor in academia, can we say that MTAs do not hamper the progress of science? However, a study like this one may contradict policymakers and academic researchers' prior opinion. Concretely, the study of research trajectories takes into account density and centrality indexes. These two measures constitute powerful instruments for studying the dynamics of a research network. Those indexes enable us to characterize research themes given: (i) their degree of development, i.e., whether or not topics are solidly constituted; (ii) their positions in the network, i.e., whether or not topics are obligatory passage points in the network (CALLON et al., 1991).

The existence of common terms between clusters suggests that they are cognitively linked to one another. This qualitative analysis of common terms also showed that only a few of them joins visible and central topics, but the majority of them belongs to unstructured topics in an emerging domain in biotechnology. Methodological work undertaken in the present study also suggest that the research themes identified by the co-word technique are relatively stable when using alternative statistical procedures, thereby alleviating the concern that clusters of terms might be nothing more than very unstable statistical artefacts.

Regarding the approach to detect convergence of research agendas, is it sound to decide whether MTAs affect them by just looking at common terms? Before any clustering effort, term selection was performed by implementing a few term filters. Besides the standard procedure of cutting off Zipf's curve by neglecting terms that occur in more than 50% of documents or only in one document, we only considered terms that appear in noun phrases. In addition, other filters were applied to the terms in the sub-sets. For instance, only terms that appear in at least two documents in the sub-set were retained. Also, a term was neglected when its maximal TF-IDF value, in the complete dataset, did not exceed 5 based on visual inspection of the histogram. Furthermore, terms that did not have an association index with any other term in the sub-set higher than 0.2, were also neglected during clustering. These filters do not necessarily pose a problem, but we must keep in mind that they have been applied and that every common term has passed these filters. Some other (common) terms may not have passed them, but then they are, of course, not the best descriptors for the research in a sub-set.

Our interviews with industry and government practitioners, chief scientific officers or similar positions, have corroborated that our findings for industry and government were valid. Generally, interviewees from Belgian biotech said that there was not any research output not submitted or delayed for being published between 1992 and 2000 because the material provider rejected or postponed it, when formalizing the exchange through MTAs. Furthermore, interviewees mostly said that they decided not to make

the material, available in the literature, after reading the MTAs clauses imposed by the provider. Nonetheless, they also said that some research projects were delayed because they were obliged to create the research material, available in the literature, to carry out the research project. In worst cases, they have given up some research projects because they did not have the research material.

Biotechnology has the potential to produce plenty of breakthroughs in existing industries such as agriculture, food-processing, and human health. "This is the first time that science is the actual business" (PISANO, 2002). Hence, scientific research in biotechnology might not be hindered by MTAs because material providers are looking for the next blockbuster. Material providers need research because their products are weak or going to loose profits in the next few years. Receivers will handle research materials internally, when possible, what it would otherwise cost more in terms of research agenda choice. In particular, the WORLD HEALTH ORGANIZATION (2006) recommends that attention should be paid to upstream research that enables and supports the acquisitions of new knowledge and technologies that will facilitate the development of new products.

Despite the plethora of issues related with MTAs, we only focused on research agenda setting in this paper. We set aside a myriad of research questions which need further studies. Firstly, which are the intellectual property strategies of organizations that signed MTAs? Are patent and copyright portfolios modified by MTAs? Are MTAs extending the territory of proprietary information into that of free information? Are MTAs clauses optimized when MTAs rationalizes a number of common contracting and legal features in the organization of research and development (R&D)? Secondly, do MTAs hamper research collaboration? How do MTAs affect the structure, organization, and objectives of scientific or technological partnership? What is the dimension of MTAs impact on technology transfer? What are the MTAs implications in terms of frequency, intensity, benefits, risks, evolution, and variety of scientific collaborations from the standpoint of industry, government, and academia? To what degree might MTAs diversify exposure to corporate R&D? Finally, do MTAs increase visibility of authors or assignees by citations? Do MTAs increase their publication activity? Do MTAs have a positive or negative impact on scientific communication? To what extent might MTAs diversify access to research materials? These questions will be dealt in upcoming publications.

*

We thank the Steunpunt O&O Statistieken for providing the venue where these ideas were initially discussed and much of the work was done. We received very helpful comments from Reinhilde Veugelers, Geertrui Van Overwalle, and Geert Duysters at the Seminar of the International Network for Innovation Research "Topics in Innovation, Science and Technology Research" held in Leuven on May 12, 2006.

We especially appreciate the excellent reviewer of Scientometrics, who made remarkable suggestions on May 25, 2006. We are indebted to Fiona Murray for providing access to her unpublished works. We are grateful for the excellent assistance provided by Mariëtte Du Plessis, Jean Gilbert, and Rebecca Crabbé. We acknowledge very much the support from the KUL Research Council (GOA AMBiorics), the FWO (G.0499.04), the IWT (GBOU-McKnow-E) and the Belgian Federal Science Policy Office (IUAP P5/22).

References

BASTIDE, F., COURTIAL, J. P., CALLON, M. (1989), The use of review articles in the analysis of research area. *Scientometrics*, 15 : 535–562.

BEN-HUR, A., ELISSEEFF, A., GUYON, I. (2002), A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 7 : 6–17.

CALLON, M., COURTIAL, J. P., LAVILLE, F. (1991), Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22 : 155–205.

CALLON, M., COURTIAL, J. P., TURNER, W. A., BAUIN, S. (1983), From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22 : 191–235.

CARAYOL, N. (2003), Objectives, agreements and matching in science-industry collaborations: Reassembling the pieces of the puzzle. *Research Policy*, 32 : 887–908.

COASE, R. H. (1960), The problem of social costs. *Journal of Law and Economics*, 3 : 1–44.

COUNCIL OF GOVERNMENTAL RELATIONS (1997), *Material Transfer in Academia*. Council of Governmental Relations, Washington, DC.

COURTIAL, J. P. (1989), Qualitative models, quantitative tools and network analysis. *Scientometrics*, 15 : 527–534.

COURTIAL, J. P. (1998), Comments on Leydesdorff's article. *Journal of the American Society for Information Science*, 49 : 98.

DASGUPTA, P., DAVID, P. A. (1994), Toward a new economics of science. *Research Policy*, 23 : 487–521.

DUNNING, T. (1993), Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 : 61–74.

ENSERINK, M. (1999), NIH Proposes Rules for Materials Exchange. *Science*, 284 : 1445.

GLÄNZEL, W., MEYER, M., SCHLEMMER, B., DU PLESSIS, M., THIJS, B., MAGERMAN, T., DEBACKERE, K., VEUGELERS, R. (2003), *Biotechnology: An Analysis Based on Publications and Patents*. Steunpunt O&O Statistieken, Leuven.

HANSEN, S., BREWSTER, A., ASHER, J. (2005), *Intellectual Property in the AAAS Scientific Community: A Descriptive Analysis of the Results of a Pilot Survey on the Effects of Patenting in Science*. American Association for the Advancement of Science, Washington, DC.

HELLER, M. A., EISENBERG, R. S. (1998), Can patents deter innovation? The anticommons in biomedical research. *Science*, 280 : 698–701.

JAIN, A., DUBES, R. (1988), *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

KAUFMAN, L., ROUSSEEUW, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York, NY.

KRATTIGER, A. F. (2004), Financing the bioindustry and facilitating biotechnology transfer. *IP Strategy Today*, 1: 1–45.

LAW, J., BAUIN, S., COURTIAL, J. P., WHITTAKER, J. (1988), Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. *Scientometrics*, 14 : 251–264.

LEYDESDORFF, L. (1997), Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, 48 : 418–427.

MANNING, C. D., SCHÜTZE, H. (2000), *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MERTON, R. K. (1942), Science and technology in a democratic order. *Journal of Legal and Political Sociology*, 1 : 115–126.

MERTON, R. K. (1968), The Matthew effect in science. *Science*, 159 : 56–63.
MURRAY, F. (2002), Innovation as overlapping scientific and technological trajectories: exploring tissue engineering. *Research Policy*, 31 : 1389–1403.
MURRAY, F. (2006), The Oncomouse that roared: Resistance and accommodation to patenting in Academic Science. *MIT Sloan School of Management Working Paper*.
MURRAY, F., STERN, S. (2005), Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. *National Bureau of Economic Research Working Paper*, 11465.
NATIONAL ACADEMIES (2005), *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health*. National Academies Press, Washington, DC.
NOYONS, E. (2001), Bibliometric mapping of science in a science policy context. *Scientometrics*, 50 : 83–98.
PISANO, G. (2002), Pharmaceutical biotechnology. In: NELSON, R. R., VICTOR, D. G., STEIL, B. (Eds), *Technological Innovation and Economic Performance*. Princeton University Press, Princeton, NJ.
PORTER, M. F. (1980), An algorithm for suffix stripping. *Program*, 14 : 130–137.
RIP, A., COURTIAL, J. P. (1984), Co-words maps of biotechnology. An example of cognitive scientometrics. *Scientometrics*, 6 : 381–400.
RODRIGUEZ, V. (2005), Material transfer agreements: Open science vs. proprietary claims. *Nature Biotechnology* 23 : 489–491.
ROUSSEEUW, P. J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 : 53–65.
SCHUMPETER, J. (1934), *The Theory of Economic Development*. Harvard University Press, Cambridge, MA.
STEPHAN, P. E. (1996), The economics of science. *Journal of Economic Literature*, 34 : 1199–1235.
STREITZ, W. D., DE BEAR, I., CALMETTES, C. S., REINHART, F. (2003), *Material Transfer Agreements: A Win-Win for Academia and Industry*. Association of University Technology Managers, Northbrook, IL.
TURNER, W. A., ROJOUAN, F. (1991), Evaluating input/output relationships in a regional research network using co-word analysis. *Scientometrics*, 22 : 139–154.
VAN OVERWALLE, G., VAN ZIMMEREN, E., VERBEURE, B., MATTHIJS, G. (2006), Models for facilitating access to patents on genetic inventions. *Nature Reviews Genetics*, 7 : 143–148.
VAN ZIMMEREN, E., VERBEURE, B., MATTHIJS, G., VAN OVERWALLE, G. (2006), A clearing house for diagnostic testing: The solution to ensure access to and use of patented genetic inventions? *Bulletin of the World Health Organization*, 84 : 352–359.
VERBEURE, B., MATTHIJS, G., VAN OVERWALLE, G. (2006), Analysing DNA patents in relation with diagnostic genetic testing. *European Journal of Human Genetics*, 14 : 26–33.
VERBEURE, B., VAN ZIMMEREN, E., MATTHIJS, G., VAN OVERWALLE, G. (2006), Patent pools and diagnostic testing. *Trends in Biotechnology*, 24 : 115–120.
WALSH, J. P., CHO, C., COHEN, W. M. (2005), View from the bench: Patents and material transfers. *Science,* 309 : 2002–2003.
WHITTAKER, J. (1989), Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science*, 19 : 473–496.
WILLIAMS, R., LAW, J. (1980), Beyond the bounds of credibility. *Fundamenta Scientiae*, 1 : 295–315.
WORLD HEALTH ORGANIZATION (2006), *Public Health: Innovation and Intellectual Property Rights*. World Health Organization Press, Geneva.
ZIMAN, J. (1987), The problem of 'problem choice'. *Minerva*, 25 : 92–106.
ZIPF, G. K. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.

Appendix

Journals

Articles, letters, notes and reviews were published in the following journals. The parenthetical figures show the number of documents of our sample that appeared in the indicated journal. Sources that contain less than 10 documents of our dataset are not listed here.

Journal of Biological Chemistry (175); European Journal of Biochemistry (109); Biochemical Journal (104); FEBS Letters (104); American Journal of Medical Genetics (92); Genomics (84); Journal of Chromatography A (73); Journal of Clinical Microbiology (71); Clinical Genetics (69); Biochemical and Biophysical Research Communications (64); Nucleosides & Nucleotides (62); Cancer Genetics and Cytogenetics (57); Applied and Environmental Microbiology (56); International Journal of Systematic Bacteriology (56); Human Molecular Genetics (52); Journal of Medical Genetics (50); Archives of Physiology and Biochemistry (49); Biochemistry (47); Genetic Counseling (44); Journal of Inherited Metabolic Disease (44); Systematic and Applied Microbiology (44); Gene (43); Biochemical Pharmacology (42) Human Genetics (42); Antimicrobial Agents and Chemotherapy (41); Antiviral Chemistry & Chemotherapy (40); EMBO Journal (40); Plant Physiology (38); Nucleic Acids Research (37); Molecular and Cellular Endocrinology (35); American Journal of Human Genetics (34); Biochimica et Biophysica Acta (33); Molecular Microbiology (33); Mammalian Genome (31); British Journal of Pharmacology (30); Genes Chromosomes & Cancer (30); Histopathology (30); Journal of Bacteriology (30); European Journal of Clinical Microbiology & Infectious Diseases (29); Plant Molecular Biology (29); Veterinary Microbiology (29); Journal of Molecular Biology (28); Phytochemistry (28); Belgian Journal of Botany (27); Nature Genetics (27); Applied Microbiology and Biotechnology (26); Cytogenetics and Cell Genetics (26); European Journal of Clinical Chemistry and Clinical Biochemistry (26); Journal of General Virology (26); Plant and Soil (26); Chromatographia (25); Histochemical Journal (24); Analytical Biochemistry (23); Archives Internationales de Physiologie, de Biochimie et de Biophysique (23); Molecular & General Genetics (23); Physiologia Plantarum (21); Plant Journal (21); Annales de Genetique (20); Clinica Chimica Acta (20); Cytokine (20); Cytometry (20); FEMS Microbiology Letters (20); International Journal of Food Microbiology (20); Journal of Applied Bacteriology (20); Journal of Neurochemistry (20); Neurochemistry International (20); Yeast (20); Cell Calcium (19); Microbiology-UK (19); Oncogene (19); Peptides (19); International Journal of Developmental Biology (18); Archives of Biochemistry and Biophysics (17); Plant Cell (17); Proteins-Structure Function and Genetics (17); Biological Trace Element Research (16); Biotechnology and Bioengineering (16); Electrophoresis (16); Human Mutation (16); Cell and Tissue Research (15); Journal of Cell Biology (15); Journal of Virological Methods (15); Molecular Pharmacology (15); Mutagenesis (15); Reproduction Nutrition Development (15); Research in Microbiology (15); Biomedical Chromatography (14); Clinical Infectious Diseases (14); Journal of Histochemistry & Cytochemistry (14); Molecular Plant-Microbe Interactions (14); Biological Mass Spectrometry (13); Journal of Plant Physiology (13); Journal of Steroid Biochemistry and Molecular Biology (13); Plant Science (13); Planta Medica (13); Biochemical and

Molecular Medicine (12); Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology (12); European Journal of Human Genetics (12); Journal of Antimicrobial Chemotherapy (12); Journal of Chromatographic Science (12); Journal of Chromatography B-Biomedical Applications (12); Journal of Experimental Botany (12); Journal of General Microbiology (12); Journal of Labelled Compounds & Radiopharmaceuticals (12); Journal of Liquid Chromatography (12); Protein Engineering (12); Theoretical and Applied Genetics (12); Zentralblatt fur Bakteriologie-International Journal of Medical Microbiology Virology Parasitology and Infectious Diseases (12); Biochimica et Biophysica Acta-Molecular Cell Research (11); Biochimie (11); FASEB Journal (11); International Journal of Peptide and Protein Research (11); Journal of Cell Science (11); Letters in Peptide Science (11); Protein Science (11); Immunogenetics (10); Journal of Natural Products (10); Molecular and Cellular Biology (10); Plant Cell and Environment (10); Plant Cell Reports (10); Plant Cell Tissue and Organ Culture (10); Plant Growth Regulation (10); Planta (10); Progress in Histochemistry and Cytochemistry (10).

Software

The software Jakarta Lucene for indexing, available at <http://lucene.apache.org>, is a high-performance, open source, full-featured text search engine library written entirely in Java. The software LT POS is a part-of-speech tagger that uses a lexicon and a hidden Markov model disambiguation strategy. The software LT Chunk is a syntactic chunker or partial parser that uses the part-of-speech information provided by LT POS and employs mildly context-sensitive grammars to detect boundaries of syntactic groups. Both LT POS and LT Chunk are available at <http://www.ltg.ed.ac.uk/software/pos/index.html>. The software Matlab, available at <http://www.mathworks.com/products/matlab>, was used to carry out other tasks.

Porter's stemmer

Stemming involves the removal of a word affix such as plurals, verb tenses and deflections, and the replacement by the canonized equivalent. The Porter's stemmer uses a simple rule-based scheme to process the most common English words. An advantage of stemming is the equation of different forms of the same word, resulting in a reduced dimensionality of the vector space and thus lessening computational costs and the curse of dimensionality for a clustering task. A disadvantage is the possible loss of morphological information necessary for discerning between different meanings of two similar words (PORTER, 1980).

The Dunning's log-likelihood method for detection of bigrams

The likelihood ratio tests the hypothesis that terms occur independently in a vocabulary. When rejected, the words presumably are correlated. It is a parametric

statistical text analysis based on the binomial or multinomial distribution and may lead to more accurate results than other text analyses that, often unjustifiably, assume normality, what limits the ability to analyse rare events.

Zipf's curve

When all words that occur in a document set are sorted in decreasing order of frequency f, and those number of occurrences are multiplied with the rank r, the result will approximately be a constant C, i.e. $C = rf$. This is formulated in the famous law of ZIPF (1949). The words in the tails of the curve can be considered as bearing less content than terms in the middle of the curve. Hence, "cutting off Zipf's curve" by neglecting terms or phrases that only occur once or in more than 50% of all documents is a pre-processing step to retain the most important words.

Term frequency – inverse document frequency value

The TF-IDF weighting scheme is very popular in information retrieval for determining the most relevant documents to a user's query. It represents the relevance or importance of terms in a document by taking into account all documents in the corpus. Textual information in documents is encoded as k-dimensional vectors, where each component w_{ij} represents the weight of term t_j in document d_i. The set of all terms t_j ($j = 1, ..., k$) is the vocabulary. TF-IDF values are calculated as follows:

$$w_{ij} = f_{ij} \log (N / n_j)$$

where f_{ij} is the term frequency, i.e., the number of occurrences of t_j in d_i, N represents the total number of documents, and n_j is the number of documents containing term t_j in the vocabulary.

The TF-IDF weight of a term in a document is high if the term frequently occurs in that document but only occurs in a few others of the document collection, i.e., a low document frequency, or consequently a high IDF. As a result, terms that occur in a lot of documents are considered common terms and are down-weighted. For a (sub)set of documents one profile vector can be constructed by calculating the mean of all document vectors involved. The ranking of terms according to their resulting mean TF-IDF weights gives information about the most important concepts present in the set.

Equivalence index matrix

Each of the sub-sets composed of documents of a specific group g ($g = 1$ for F_1, $g = 2$ for F_2, and $g = 3$ for F_3) in one of the three periods p, was filtered. From the global indexed term-by-document matrix A_b containing binary values, a sub-matrix $A_{b,g,p}$ was

constructed after filtering terms. From each sub-matrix $A_{b,g,p}$, a term co-occurrence matrix $C_{g,p}$ was constructed by multiplying $A_{b,g,p}$ with its transpose:

$$C_{g,p} = A_{b,g,p} \,.\, A_{b,g,p}'$$

Then, each $C_{g,p}$ was converted into an equivalent index matrix $E_{g,p}$ by transforming the co-occurrence frequency for two terms i and j to their equivalence or association index e_{ij} (CALLON et al., 1991), by applying the following function:

$$e_{ij} = 0, \text{ if } c_i = 0 \text{ or } c_j = 0 \text{ or } c_{ij} = 0$$

$$e_{ij} = c_{ij}^2 / (c_i \,.\, c_j), \text{ otherwise}$$

in which c_i and c_j are the respective document frequencies of terms i and j in the sub-set and c_{ij} is their co-occurrence frequency in that sub-set. Subtracting each equivalence index matrix $E_{g,p}$ from 1, results in a distance matrix that can be used as input for a clustering algorithm.

Ward's agglomerative hierarchical clustering

Hierarchical clustering algorithms group objects in an iterated manner, either from singleton clusters to a cluster containing all objects – agglomerative clustering – or vice versa – divisive clustering. The strategy used to determine which objects or clusters to group in each iteration affects the outcome. In Ward's method, those objects are grouped such that the increase in total error sum of squares over all clusters is minimized. Like any other algorithm, it has its advantages and weaknesses and it is certainly not perfect. One of the disadvantages of agglomerative hierarchical clustering is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired (KAUFMAN & ROUSSEEUW, 1990). What we sometimes observe when using hierarchical clustering is the forming of one very big cluster and a few small very specific clusters (JAIN & DUBES, 1988).

Dendrogram

A dendrogram visualizes the iterative grouping or splitting of clusters in hierarchical clustering. The horizontal lines connect clusters in a hierarchical tree. The line length represents the distance between two connected terms or clusters. An imaginary vertical line would illustrate a cut-off point for a specific number of clusters.

Stability diagram

The stability-based method of BEN-HUR et al. (2002) for determining an optimal number of clusters, exploits measurements of the stability of clustering solutions obtained by perturbing the data set. The result is a stability diagram. Stability is characterized by the distribution of pairwise similarities between clusters obtained from sub-samples of the data. High pairwise similarities indicate a stable clustering pattern. The method can be used with any clustering algorithm; it provides a means of rationally defining an optimum number of clusters, and can also detect the lack of structure in data.

Mean silhouette value

The silhouette value for a term ranges from -1 to +1 and measures how similar it is to terms in its own cluster versus terms in other clusters (ROUSSEEUW, 1987). The mean silhouette value for all terms assesses the overall quality of a clustering solution.

Density and centrality index

Density is defined as the mean of the equivalence indices e_{ij} over all term pairs in a cluster (internal links) and centrality is the mean of e_{ij} for all possible pairs of words of which one is an element of the cluster and the other is not (external links).