



Universiteit
Leiden
The Netherlands

Bibliometric mapping as a science policy and research management tool

Noyons, E.C.M.

Citation

Noyons, E. C. M. (1999, December 9). *Bibliometric mapping as a science policy and research management tool*. DSWO Press, Leiden. Retrieved from <https://hdl.handle.net/1887/38308>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/38308>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/38308> holds various files of this Leiden University dissertation

Author: Noyons, Ed C.M.

Title: Bibliometric mapping as a science policy and research management tool

Issue Date: 1999-12-09

2 Principles of Science Maps

In this chapter the principles of a science map are discussed. These principles account for a trustworthy and useful process and procedure to build a science map which can be used as a policy supportive tool in terms of evaluative bibliometrics.

2.1 What do maps show?

The central question of this section is an important issue in science (and technology) mapping: *what do the maps show?* By discussing the most important principles underlying maps of science (listed below), this question will be addressed.

1. Maps of science as a tool for science policy should represent the scientific knowledge. Scientific knowledge is represented per se by *research output*;
2. Bibliometric science maps are constructed on the basis of *publication data*;
3. Provided that the research output of a field is well covered in a *bibliographic database*, this field can be represented by (a selection of data from) this database;
4. By using *content describing elements* (CDE, the building blocks of a publication description), each publication can be characterized;
5. With help of co-occurrence data of the most frequently used CDEs within a bibliographic database, the *structure of the database* can be unraveled;
6. Under the assumption of principle 3 and 5, a structured bibliographic database of publications in field A represents the *structure of field A*;
7. The *dynamics of the structure* based on the changing co-occurrences, represent the dynamics of the field, as related to the structure of the field.

Each principle will be discussed from the perspective of the matter addressed in this book. We do not claim that this list is exhaustive. Other applications of science maps (e.g., information retrieval) may have other principles.

Research output

Maps to be used as policy-supportive tools should represent scientific knowledge. Policy-related users want to know the structure of this knowledge *and* its evolution in order to validate their activities or explore future developments.

Whatever scientists think or say individually, their discoveries cannot be regarded as belonging to scientific knowledge until they have been reported to the world and put on permanent record.

(Ziman 1984, p. 58)

Following the argumentation of Ziman (1984), we should conclude that the map particularly is a suitable representation of scientific knowledge. He states that:

A mature body of scientific knowledge is like a map. The structure of some region is represented by the relative positions of various conventional symbols, each standing for some selected category or aspect of the real world. (...) The map metaphor also suggests that scientific knowledge is a multiply connected network of concepts, where the validity of any particular proposition does not depend solely on one or two other theoretical propositions or empirical observations.

(Ziman 1984, p. 49)

However nice this metaphor looks, a *map* is 'just' a virtual representation, possibly with no reference to the 'real, physical world'. From the objectivity viewpoint, the data itself should create a structure: the self-organizing maps (Kohonen, 1990) of science. This may cause the map to become incomprehensible and unpredictable if it does not refer to the perception of the field structure by an expert. For this particular reason, the interpretation and validation by experts is of vital importance for the *utility*. If a map is not interpretable for a field expert, it means that the map is not useful for policy supportive means. The map has no reference to the world according to the policy-related user and thus the map cannot contribute to a policy or management discussion.

Publication data

A map presents the structure of a field in a particular period of time (T). The selected publications were published during that period of time. The map based on these publications therefore represent the structure of the field in period T . The *publications* in T , however, represent the *research* performed in a preceding period. It is very difficult to determine the time lag between (completion of) *research* and *publication*. For instance, if we look at publications in journals, in each stage from *research* to *publication*, several factors can be identified that affect the time-lag. At the stage from completion of the research to the first version of a publication, the available time of the researcher to write the article is involved, and sometimes in applied research, an embargo on the results set by the body that commissioned the study are just a few of them. In the process from *first submitted* to *final accepted* of a publication, factors like available time of the author to correct earlier versions, and time and availability of the referees, are at stake. In the process from the *acceptance* to *actual publication*, the backlog of a journal is not to be neglected. Unfortunately, not every journal has the same backlog. There are journals with a publication section for highly significant recent developments (in scientific or social sense) which have a small backlog, in most journals, however, it takes between six and eighteen months before the accepted version of an article is published. This may, but not necessarily, be field dependent.

Therefore, a map based on journal publications in T may represent the research performed in year $T-i$, where i is for instance 2 years. As presentations at conferences seem to be better updated with the present work of researchers, a map based on proceeding papers in T , may represent the research performed in $T-i$, where i is between 1 and 'zero' (publication in the year of research). Again, it will depend on the objective of a study whether this is a problem. A mapping project aimed at unraveling the main structure of a field in a period of time longer than one or two years, will probably not be affected by a relatively short time lag. Moreover, clever selections (based on sources, document types, journal sets, or even on the output of excellent performing research groups) assure consistency of data, and thus reliability of results. Finally, science maps show the structure of the scientific output of researchers, not the research itself. Maps give an indication of how the knowledge is structured, under the assumption that knowledge is represented by the scientific output (Ziman, 1984).

An exploration of the publication delay as defined by the period of time between the date of submission and publication of an article has been reported by Luwel and Moed (1998). They are concerned with this phenomenon in view of the impact of publications (citations received).

Bibliographic database

The availability of reliable data is, beyond doubt, the most important condition for a valid bibliometric study. The choice for a particular database for a particular study does not solely depend on the consensus of bibliometricians. For an important part it depends on the objective of the study. If the required indicators can be extracted from database X and both the users of the indicators and field experts approve of the database X to be used, there is no reason to use database Y , which may be a standard in bibliometrics. For instance, during the evaluation of the project presented in Chapter 9, experts in the evaluated field microelectronics stated that, as far as the most important developments were concerned, the field might as well have been represented by the bibliographic data of just a series of international conferences. On the other hand, a bibliometric study including impact data, 'must' use the ISI citation databases. Not (only) because they form a bibliometric standard thanks to its unique coverage (namely multidisciplinary), but (also) because they are the only databases containing cited references².

Another important consideration is that the scope of the database determines to a large extent the results of the mapping exercise. In Chapter 6 and 7, we report of a study of neural network research, based on data extracted from the INSPEC database. The scope of this database appeared to be relevant for the study, and the funding body (the German Ministry for Science and Technology, BMBF) agreed on that. Nevertheless,

² There is a specific field, high energy physics, which has its own database (SLAC-SPIRES) including cited references

experts in the field concluded afterwards that a considerable part of the field was not represented, being the research conducted within the behavioral sciences (cognitive psychology). As a result, we should refer to the monitored field as being mainly neural network physics and engineering.

Finally, Chapter 5 is referred to as another illustrative example. In this study, we mapped the field of 'optomechatronics' on the science side (publications) and on the technology side (patents). An important subfield observed in the science map appeared to be missing in map on the technology side. It concerned an area of software engineering. The most plausible reason for this is that software as such is difficult to be patented. As a result, the area mainly covered by software developments is hardly covered by a patent database, and thus hardly present in the technology map. It shows up, however, very well in the science map.

In order to answer the question '*what do the maps show?*' one should first answer the question '*what does the database cover?*' The map never shows more than the data discloses. nevertheless, a map is able to reveal hidden structures (within the data); structures which may not be obvious to field experts.

Content Describing Elements

The concept 'Content Describing Element' (CDE) is flexible. Some items in a bibliographic database are beyond doubt CDEs: title, abstract, classification codes, thesaurus terms. They are all able to describe the contents of an article in such a way that it is not easily mixed up with another. In other words, they are completely or to a large extent document-specific. Others, however, may be CDEs as well but are not or less specific for a particular document: author, journal, cited reference. In a search for interesting publications, a researcher often makes a first selection by choosing a particular journal, or a set of journals. Then, he scans titles and authors of the listed articles. In an alternative procedure, he may look up which (new) articles cite a particular publication or author. Thus, the contents are determined by journal, author, title and/or cited reference, or at the least by a combination of these elements.

By nature, CDEs seem to be appropriate elements to build a science field map. In that case the CDEs of publications must become CDEs of a bibliographic database and thus of a science field. For example, publication-specific keywords describe the publication (its main issues) to which they belong, and the field-specific keywords describe the contents (the main issues) of a field. As a result, the keywords of publication X (belonging to field A) are *candidate field keywords* for A, but they belong not necessarily to the most typical keywords for A.

In principle, to build a map we may use any CDE. Ziman (1978) states:

Since science is more than personal knowledge, it can consist only of what can be communicated from person to person. The available media of human communication determine the forms,

and to some extent the contents, of messages that make up scientific knowledge. To start with, as a crude 'zeroth-order approximation', we treat this as a strict limitation; to achieve the ultimate goal of consensuality, science must be capable of expression in an unambiguous public *language*.

(Ziman 1978, p. 11)

This means that in every aspect of a publication a potential communication issue is captured. It will depend on the purpose of the map, which one to use. This dependency is caused both by the data, and by the user involved. For example, a map based on author co-occurrence data, primarily shows the 'social' structure of the field. Researchers working in one and the same institute, and having a good (professional) relationship, are more *likely* to co-author a publication than those who do not. So, from the data 'point of view' - that is, for which purpose should the data best be used - the aim of the map is of great importance. Therefore, should the map be aiming at unraveling the *social* structure, the author co-occurrence data seems most appropriate. On the other hand, if the map should be aiming at unraveling the *cognitive* structure, the author co-occurrence data may appear to be appropriate as well. However, in that case it is likely that the user would object. For an average user, a map based on author co-occurrence data does not *primarily* refer to the *cognitive*, but rather to the *social* representation. He would get confused because the map does not show a representation that refers to his perception of the field concerned. This observation seems trivial as it is illustrated by such opposite examples. If the CDEs are more similar, the discussion of this user dependency becomes more relevant. A cognitive map based on keywords retrieved from *titles* may, for instance, be rejected by an expert who primarily gives 'popularizing' titles to his publications, and may prefer controlled terms. An expert, however, who is most of the time working on new developments (including new topics) may prefer the titles (and abstracts) rather than controlled terms, because they may not cover new topics.

In most cases, the structuring of a field for science policy support is established by *co-word* analysis. The words or terms refer directly to topics and methods, and thus to the field cognition of experts and (other) users. In such analyses, the CDE discussion mainly focuses on the usage of a controlled vocabulary as opposed to free text. A controlled vocabulary relates to keywords taken from an existing list (i.e., thesaurus). This list is maintained and updated by experts in the field to which this list refers. In most specialized bibliographic databases, such keywords are added to publication documents. The usage of the controlled vocabulary has two main disadvantages. The first is best known as the 'indexer effect' (c.f., Healey, Rothman and Hoch, 1986). The indexed terms are added to publication by field experts. It will depend on the expertise (in all its facets) how trustworthy and appropriate this is done. The second disadvantage concerns the rigidity of indexed lists of terms. It will take some time before a new term (topic, theme, subject, method) is introduced into an index. In

information retrieval this is an important point of discussion. The controlled vocabulary (indexed terms, descriptors, et cetera) is more precise (sometimes even more adequate) to be used in bibliographic searches, but lacks the, often important, feature of topicality. A 'free text' search in a bibliographic database returns documents containing up-to-date vocabulary but often omits documents with titles and abstracts in a slightly different jargon.

In policy-supportive studies, it will depend upon the aim of the project, what CDE is to be used. Bibliometric co-word mapping studies aiming at generating an exhaustive historical overview of a science field, will benefit from the usage of controlled terms, whereas studies aiming at exploring recent developments, will benefit from the usage of free text CDEs.

Therefore, in order to answer the question '*what do the maps show?*' first the question '*what do we want the maps to show?*' has to be answered. And in view of that question, it should be determined what kind of data is going to be used to build a map. Furthermore, it should be investigated whether the data and the resulting maps generate a picture of the field that reflects the 'representation' of the user, and is appropriate to answer the raised issue (the aim of a project). To deal with these questions, one should not only be flexible with respect to the information presented in a map, but also with respect to the process of building the maps. The user-bibliometrician interaction is vital for the results, and therefore for the success of bibliometric mapping.

Structure of the field and its dynamics

In Section 2.3, the need of dynamic maps rather than static maps will be discussed. Here, the discussion is focused on the applicability of dynamic maps. In view of the question of '*what does a map show?*' we should also deal with the question '*what do the changes in a dynamic map show?*' Before the field dynamics can be monitored, it should become clear what the starting point is or what the final point is. A dynamic map of a field shows the changing interaction of its elements. In terms of co-occurrences, a dynamic map shows the changing relations between selected elements. In order to use a map for policy-related questions, all questions discussed above, should be answered before the dynamic map can be interpreted. Otherwise, the dynamic map may, for instance, reflect the changing coverage of a database rather than the field dynamics.

In Noyons & Van Raan (1998a, see Chapter 6), it is pointed out that in a bibliometric mapping study aiming at exploring actors' activities as well as field dynamics, a structure should be generated (subdomain delineation) on the basis of one period of time (say one year T) and this structure (delineation) should be used to analyze other years (for instance previous years, $T-i$). In this case the most recent representation of the field becomes the point of reference. A map showing the field dynamics reveals

the dynamics as related to T . The interpretation of the field dynamics is therefore dependent on the situation in the point of reference T . For instance, if the analysis of the field in year T identifies a subdomain X , which seems to be a merger of two specialties (x_1 and x_2), a dynamic map based on the structure of T , does not notify the fusion of x_1 and x_2 into X as such. It does however reveal the dynamics of X as if it existed already in $T-i$. This approach reveals the dynamics of X within the whole field as defined by the 'present' (T) situation. It should be noted that the fusion of x_1 and x_2 into X is already a fact and from a policy point of view it does not seem to make sense to evaluate into detail that this merging has taken place. But it *does* seem to make sense to explore the dynamics of X from the present point of reference: who was responsible for the development of X . By retrieving the actors from X in $T-i$, the founding actors of X are revealed. In other words, this type of approach is essential in studies to 'trace' developments in scientific knowledge.

From an *historical* point of view it may make more sense to monitor the field evolution with a *past* situation as a point of reference. This approach is appropriate to show how developments in the past 'disappeared' or 'exploded' in recent time.

2.2 Co-word analysis as a bibliometric tool

Co-word analysis concerns co-occurrence analysis of specific words. These words are retrieved from publications. Every publication can be described by words. Often it makes sense to use phrases rather than single words. These phrases are (meaningful) groups of words. Together, the meaningful words and phrases are referred to as *keywords*. They describe the main issues of a publication. These keywords are available in documents in bibliographical databases. They may be 'uncontrolled', i.e., extracted from free text fields (titles, abstracts), they may be added by authors (author keywords), and they may be 'controlled', i.e., added to the publications by the database producer (indexed, thesaurus, or controlled terms). We already discussed that each type has its advantages and disadvantages (see Healey, Rothman and Hoch, 1986; Whittaker, 1989, and section 2.1). The non-indexed keywords extracted from titles and abstracts are preferable, as they can be extracted from almost every bibliographic database. This makes them more generally available and thus flexible and better adjustable to the policy issue addressed. If, for example, a field is perfectly covered by a specific database, a mapping study based on co-word analysis can always be performed, whereas in only a limited number of cases cited reference data or controlled terms are available. Moreover, with co-word analysis of 'free text'-extracted (uncontrolled) terms, different bibliographic databases can be combined.

As a result from the discussion in the previous and in the present section, we discern two kinds of keywords. The first is the keyword that describes together the contents of a *publication* and in combination with all other keywords of a publication, it discriminates one publication from the other. We will refer to this kind of keyword as

being a *publication keyword* (PKW). The second type is the one describing the contents of a *publication collection* or database and will be referred to as being a *field keyword* (FKW). Together with all other FKWs it discriminates one science field from the other.

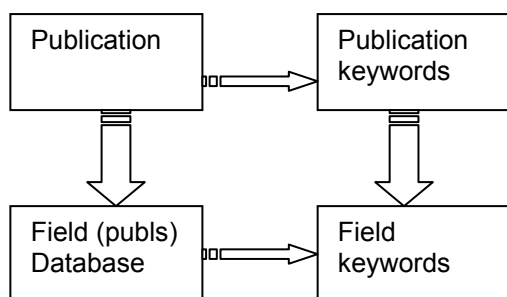


Figure 2-1 Publication keywords (PKW) and field keywords (FKW)

2.3 Mapping as a bibliometric tool

As we discussed earlier, the enthusiasm for bibliometric maps (or co-citation/co-word modeling) in the seventies and eighties has been tempered since the early nineties. Reasons for this might have been the high costs involved, the modest validity according to the experts evaluating the results, and the inaccessibility of the method and results (the maps). If we consider the three parties involved in quantitative policy-oriented studies of science (see Chapter 3), we identify at the same time three aspects to which *objections* to mapping are directed.

1. Evaluated scientists (as objects): the results;
2. Scientometricians (as producers): the data and methods;
3. Policy makers (as users): the utility.

The first objection points at the lack of recognition by researchers in the field. In particular co-word mapping has suffered from this (Healey, Rothman, and Hoch, 1986). Rip (1997) states that co-word maps are sometimes hard to understand. They would show 'pathways' rather than a structure.

A similar kind of aggregation would occur naturally when research group leaders would report on the state of the field and ongoing and future work of their groups in relation to it. Co-word maps are thus suitable to purposes of tracing connections and locating work strategically.

(Rip 1997, p. 17)

This passage particularly points out the utility of co-word maps for research evaluation or monitoring. In that sense, one may wonder whether 'pathways' differ from, (or are inferior to) a structure. Moreover, Rip pleads for the independence of scientometrics where the results are concerned (see also Chapter 3). Once *data* and *method* have been validated, the resulting maps show a *point representation* (see section 3.2) of the field, i.e., a representation generated by the creator (the scientometrician) on the basis of approved data and method, and as such robust. It will depend on the expert, evaluating/validating the results, whether the structure is 'recognized'. It is, however, important to notice that the validation of *data* and *method* often comes down to the validation of *results*, the generated maps. As a result, the first and second objection are closely related. In view of this, we conducted a mapping study of the field in which scientometricians are active, scientometrics, informetrics, and bibliometrics (SIB). In Chapter 10, we report the method and results as well as the comments of field experts.

As to the third objection, we refer to section 3.2. Furthermore, we address the issue of the utility of a map as a representation of a science field. *Why would we create maps?* What does the spatial (positional) information add to the information we already have by distributing publications over identified subdomains. A map puts the subdomains in a two or three dimensional space in such a way that the subdomains that share many publications are in each others vicinity, and those who share few or no publication, are distant from each other. We experienced in several studies that users of our results, focus merely on the *division* into subfields, rather than on the added and typical 'mapping' information of the *positioning* of the identified subdomains. They evaluate the structure first without using the positional information. In such cases, characteristics of each subfield are compared to those of the others. For instance, by comparing the activity of actors (countries, institutes, departments) in the identified subfields, strengths and weaknesses in terms of activity of an actor can be determined. In the study presented in Chapter 9, we visualized the activity patterns of four departments of a research institute within the mapped structure of the field concerned. It appeared that the formal institutional structure with different research departments nicely fitted into the structure of the field as obtained by co-occurrence analysis. We observed that, next to the identification of subfields, the two dimensional positioning accounts to a large extent for the activity profile of each department within the institution. Thus, also the positioning on the map appears to be a valid indicator.

2.4 Science mapping as a policy supportive tool

A map of science represents the (static) situation of a field in a particular period of time, using the publication data in that time span. Often the need of validation of the map is expressed (Moed, 1989; Bauin et al., 1991; Tijssen, 1992; Hinze, 1997). In any

evaluative study, the results should be checked by experts in the field, at the least to preclude accidental errors.

Once experts have expressed their contentedness with a map of their science field, regarding the structure on the basis of keyword clusters, it is still the question what to do with this information. The identification of clusters of words as subdomains (or 'themes', c.f., Callon, Law and Rip, 1986) as such could be sufficient to generate tables in order to evaluate the activity of a specific actor in the field and to compare it to other actors. The positioning of these subdomains in a two or three dimensional space is disputable as to add no valuable information, regarding its utility. In other words: *what can we do with this information?*

An analogous situation exists for weather reports on television. Some years ago, the illustration of the weather of 'today' was not more than a map of the country or region with clouds, sun and indicators for high and low pressure areas. The map showing the situation of today's weather caused the audience (user) to lose interest because most of the information referred to something they already knew. (*I know that there are clouds above the area I live, because I've seen that and it is has been raining the whole day*). Recently, these static maps have been replaced by animated maps. They show how the situation in the sky has evolved from the situation of, say, the day before. Thus, the map showing the 'final' situation is the same as the static map, but we now have more insight in how the situation has evolved to the present, thus allowing us to make, in a way, our own personal view on how things might be in the near future. For instance, with the presumption that the movements of clouds and high/low pressure areas will be continued, we are able to make our own weather forecast. On the other hand, it gives the weather forecast on television more credit, because we see how clouds sometimes move in unexpected directions.

When mapping a science field, we find ourselves in a similar position. The comments to static maps of the present are often similar to the comments to static weather report maps (*I know that these are the main areas within the field, and I know that the area I'm working in is small because ...*). The policy user of such maps may say that the maps looks nice (the expert said so) but what can he do with the spatial information. Subdomain x is in the vicinity of y but what does this tell hem about the relation of x and y besides the cognitive. By showing how the field (map) has evolved to the present situation³, the user can put this *relation* between x and y in perspective of its evolution. The *relation* is evolving in a certain direction, and does this indicate a particular development to be expected in the near future (e.g., merger of x and y or further separation)

Whether an extrapolation of certain trends will become true remains, of course, to be seen.

³ Films of such evolutions are available at <http://sahara.fsw.leidenuniv.nl>.

2.5 From scientific output to science maps

The 'process' from scientific output to science maps has been described along the lines of some basic (bibliometric) principles. Moreover, it has been pointed out how these principles could be implemented in order to create science maps that can be used for certain policy-related issues. The process as far as been discussed in this chapter is depicted in the next Figure.

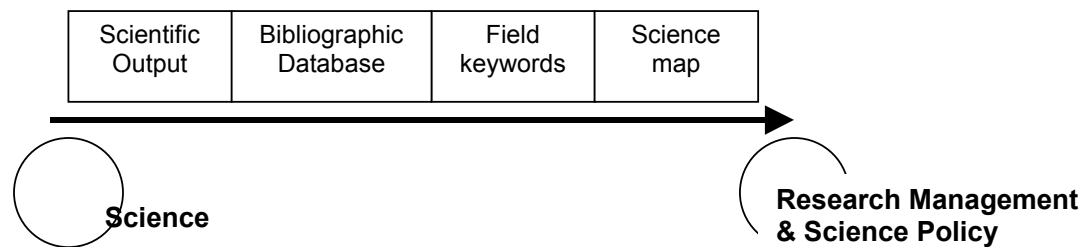


Figure 2-2 From scientific output to science maps

Furthermore, if we take into account the required utility of science map, the 'end product' should not be 'just a map' but rather a *map interface*. The interface discloses by automated procedures (e.g., via graphical internet browsers), all kinds of information 'behind' the map, such as actors, detail maps and field dynamics. Primarily, the policy-related issue raised will determine the contents and design of the map interface.

The process from publication (bibliographic) database representing a science field, to the map interface to be used to address the raised policy-related issue would look like:

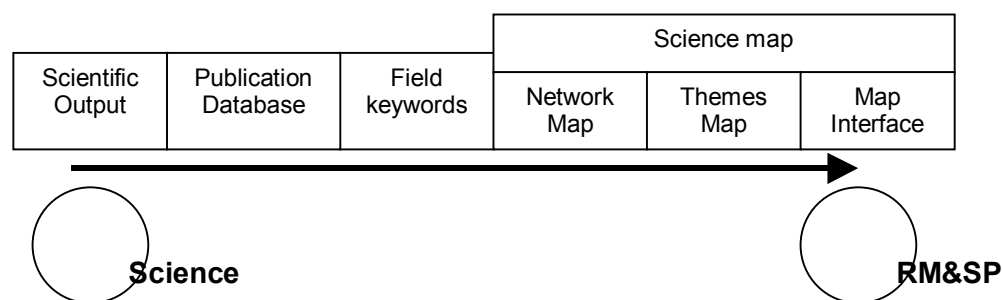


Figure 2-3 From scientific output to map interface

The transition from network map to themes map (see Figure 1-1) is one-on-one. The themes map is a simplification of the network map. As a result the former contains the information of identified subdomains .

In this chapter of principles of science mapping as a policy supportive tool have been discussed. The bottom line is that the issue to be addressed to a great extent determines the data to be used. Furthermore, it has been argued that in particular the dynamics (evolution) adds great value to the utility of science maps.

References

- Bauin, S., B. Michelet, M.G. Schweighoffer, and P. Vermeulin (1991). Using Bibliometrics in Strategic Analysis: "Understanding Chemical Reactions" at CNRS. *Scientometrics* 22. 113-137.
- Callon, M., J. Law, and A. Rip (1986). *Mapping the Dynamics of Science and Technology*. The MacMillan Press Ltd., London, ISBN: 0 333 37223 9
- Healey, P., H. Rothman, and P.K. Hoch (1986). An experiment in Science Mapping for Research Planning. *Research Policy* 15. 233-251.
- Hinze, S. (1997). *Mapping of Structures in Science & Technology: Bibliometric Analyses for Policy Purposes*. Ph.D. Thesis Leiden.
- Kohonen, T. (1990). The Self-Organizing Map. In: *Proceedings of the IEEE*, Vol 78 no. 9, September 1990. 1464-1480.
- Luwel, M. and H.F. Moed (1998). Publication Delays in the Science Field and their Relationship to the Aging of Scientific Literature. *Scientometrics* 41. 29-40.
- Moed, H.F. (1989). *The Use of Bibliometric Indicators for the Assessment of Research Performance in Natural and Life Sciences: Aspects of Data Collection, Reliability, Validity and Applicability*. DSWO Press, Leiden University.
- Rip, A. (1997). Qualitative Conditions of Scientometrics: The New Challenges. *Scientometrics* 38. 7-26.
- Tijssen, R.J.W. (1992). *Cartography of Science: Scientometric Mapping with Multidimensional Scaling Techniques*. DSWO Press, Leiden University.
- Whittaker, J. (1989). Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. *Social Studies of Science* 19. 473-496.

Ziman, J.M. (1978). *Reliable Knowledge*. Cambridge University Press, Cambridge, ISBN: 0-521-40670-6

Ziman, J.M. (1984). *An Introduction to Science Studies: the Philosophical and Social Aspects of Science and Technology*. Cambridge University Press, Cambridge, ISBN: 0-521-34680-0

