Bibliometric mapping as a science policy and research management tool

Noyons, E.C.M.

10 'State of the Art': A Case Study of Scientometrics, Informetrics and Bibliometrics[22]

In this chapter, the results of a mapping study in the field of Scientometrics, Informetrics, and Bibliometrics (SIB) are presented. This field may also be called more generally 'quantitative studies of science'. During the study, we found that the delineation is not as simple as it seemed beforehand. A study published in the same period of time as our study was performed (White & McCain, 1998), showed that SIB researchers may all have their own way of describing the field. Therefore, by allowing the researchers in the field to define the field themselves, we could finally suggest a selection procedure of publications to which they agreed.

By mapping our own field, we have field experts readily at hand. Thus, we were able to validate rather easily the structure as well as the utility of the map interface. Given the fact that the experts were so closely involved, we could explore on the basis of their comments, possible new developments and perspectives for science mapping. We will report about these explorations in this chapter.

10.1 Field delineation, data collection, and methodology

Mapping your 'own' field, has the advantage of experts being directly available (colleague-researchers at CWTS). In addition, it is expected to be easy to attract other experts to evaluate the results (colleague-researchers worldwide in the field of SIB). On top of that, the policy-relatedness of SIB, draws researchers working in political organizations, so that the (policy-related) users are involved as well.

The first step of the study concerned the delineation of the field on the basis of opinions of the researchers in the field. For this purpose, we addressed an Internet discussion list of researchers being member of the International Society for Scientometrics and Informetrics (ISSI). This forum of about 200 members contains researchers in the SIB field. Part of them is working in research policy-related organizations. They were asked to provide names of journals that belong to the core of the field. Secondly, they were asked to list the most important keywords or terms of their own research. About 20 researchers (10%) returned a list. Although the responding rate was not very high[23] most of the supplied information was valuable.

Second, the aggregated list of suggestions was proposed to the forum again and they were asked to give their reactions to the list. This step was built in to check the

[22] An internet version of this project is available at: http://sahara.fsw.leidenuniv.nl.

[23] The main reason for the low response is the fact that the survey was sent to the electronic discussion list. Colleagues could send their suggestions to my personal e-mail address but chose to send them to the discussion list so that all possible respondents could read the contributions by the earlier respondents. Once 'their' suggestions were already proposed by these earlier respondents, they did not feel the urge to contribute as well.

validity of the suggestions and to get rid of journals with too general a scope. Finally, we selected journals fully covered by the Social Science Citation Index (SSCI) only, in view of the planned impact analyses. As a result, eleven journals were selected. We collected the 1991 to 1997 bibliographic data of all publications in these journals, and took that as a starting point for our analyses.

The set contains the following journals:

- Information Processing & Management;
- International Information & Library Review;
- Journal of Documentation;
- Journal of Information Science;
- Journal of the American Society for Information Science;
- Library and Information Science;
- Research Policy;
- Science Technology & Human Values;
- Scientometrics;
- Serials Librarian;
- Social Studies of Science.

As we were able to retrieve the abstract data for the publications of 1992 to 1997, we based our analyses on these years. The basic structure of the field was derived from the 1995/1997 data and the period of 1992/1994 was studied as well.

The titles and abstracts of articles, letters, notes and reviews in the selected journals were subject to a linguistic analysis and the noun phrases were extracted (for details, see Chapter 11). For the most recent period (1995-97), the most frequent noun phrases were identified and used as a list of 'candidate field-specific keywords' representing the core of SIB. On the basis of the expertise at CWTS, a subset of 52 field keywords was selected from this list to be used to structure the field. By calculating the co-occurrences of these keywords, and normalizing the 'raw' co-occurrence matrix with the cosine of co-occurrence vectors, we created a matrix containing the similarity data of the keywords in terms of their cognitive orientation. Thus, keywords with a similar co-occurrence profile (with all other keywords) have a high similarity index (Noyons and Van Raan, 1998a).

$$sim(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{((\sum_i x_i^2)(\sum_i y_i^2))}}$$

where:
x_i = number of co-occurrences of keyword x with any other keyword
y_i = number of co-occurrences of keyword y with any other keyword

Cosine vector of co-occurrences

This similarity matrix was object to a cluster analysis in order to identify clusters of cognitively related topics. The cluster analysis yielded five clusters. This is locally an optimal solution based on the combination of three criterions to determine the 'ideal' number of clusters (c.f., SAS User's Guide, 1989).

The keywords clusters delineate subdomains of SIB. Publications representing the subdomains are retrieved by the keywords. Thus, the keyword clusters denominate subdomains of SIB.

As publications may represent more than one subdomain, we can use the overlap between subdomains (in terms of common publications) as input for multidimensional scaling (MDS). The resulting two dimensions of MDS yield the map of SIB. In the map subdomains with a similar cognitive orientation (many common publications) are in each other's vicinity, and those with a different orientation are distant from each other. In our case, the map (based on the cosine vector co-occurrence data) represents a 'perfect' solution for the cluster co-occurrence data (badness-of-fit: 0.00; distance correlation: 1.00).

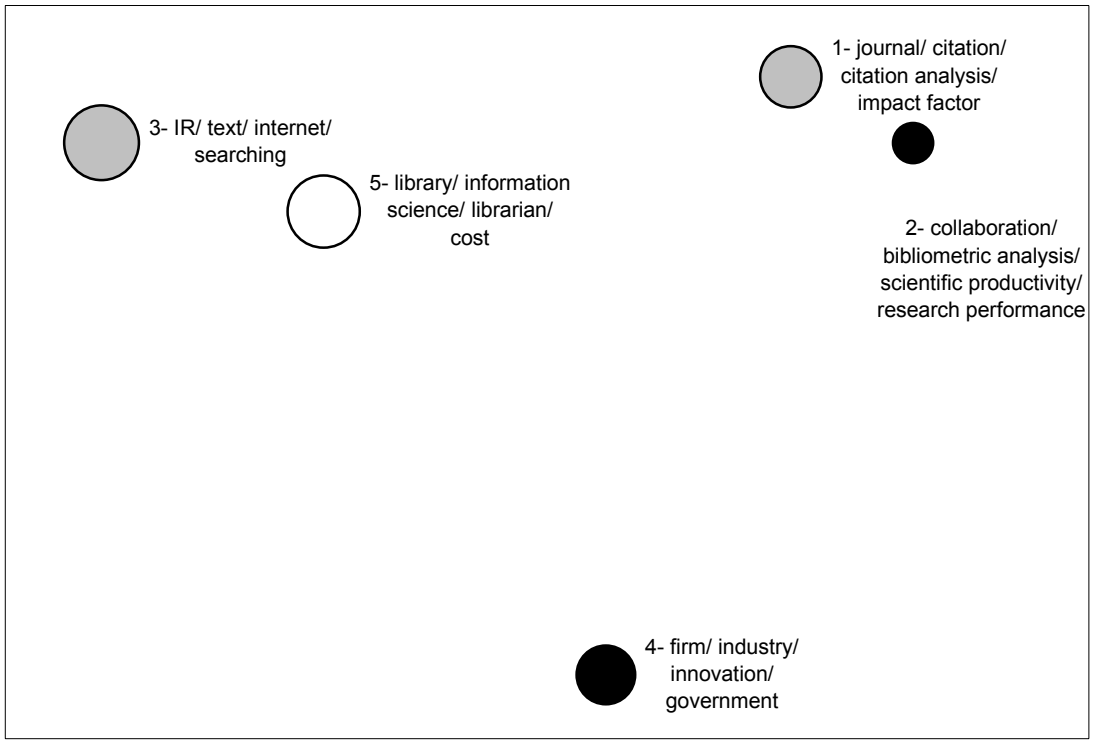
10.2 Main results

As discussed above, our clustering analysis of the 52 keywords yielded five subdomains within SIB. In order to identify the contents, we assigned to each of these subdomains a name based on the four most prominent (i.e., the most frequent) keywords within.

Table 10–1 Five identified subdomains in SIB (1995-97)

Cluster	*Nr Pubs 1992-94*	*Nr Pubs 1995-97*	*Subdomain name*
1	157	172	journal/ citation/ citation analysis/ impact factor
2	48	73	collaboration/ bibliometric analysis/ scientific productivity/ research performance
3	174	245	IR/ text/ internet/ searching
4	71	156	firm/ industry/ innovation/ government
5	244	227	library/ information science/ librarian/ cost

In terms of research areas, we identified these subdomains as: (1) evaluative bibliometrics; (2) research performance, in particular collaboration; (3) information retrieval; (4) science and technology (S&T) policy studies, and (5) library science and management. Four of these five subdomains show an increase of activity in absolute numbers from 1992 to 1997. We present the map of SIB (based on the data of 1995-1997) in Figure 10-1.



The circle surfaces indicate the relative number of publication represented by a subdomains. The colors indicate the activity trend during the period 1992-1997 per subdomain: black indicates a strong increase; white indicates a strong decrease of activity. The calculated explained variance is 1.00.

Figure 10-1 Map of SIB 1995-1997

The map shows the close relatedness of 1 and 2 on the right hand side, and of 3 and 5 on the left hand side. Subdomain 4 (S&T policy studies) is found distant from all other four subdomains at the bottom of the map. The main difference between the latter and the four other seems to be the use of data. As all other four subdomains use publication data for their research, subdomain 4 makes use of other data sources (patent data; OECD statistics; survey data) as the research in this subdomains more society/industry-related. The difference between 1 and 2 on the one hand, and 3 and 5

on the other is also obvious. In the former we are dealing with the evaluative bibliometric research, and in the latter with the research related to libraries.

As research is so significantly different in at least three areas of the map, it is to be expected that the information within the subdomains differs as well. To explore this, we implemented a map interface. This interactive tool enables a user to view by subdomain the general statistics concerning actors (countries, authors, etc.), reference statistics (most cited references, most cited institutes), and internal structure (co-word network map of most frequently used keywords). In Figure 10-2, a computer screen shot of the interface is presented.

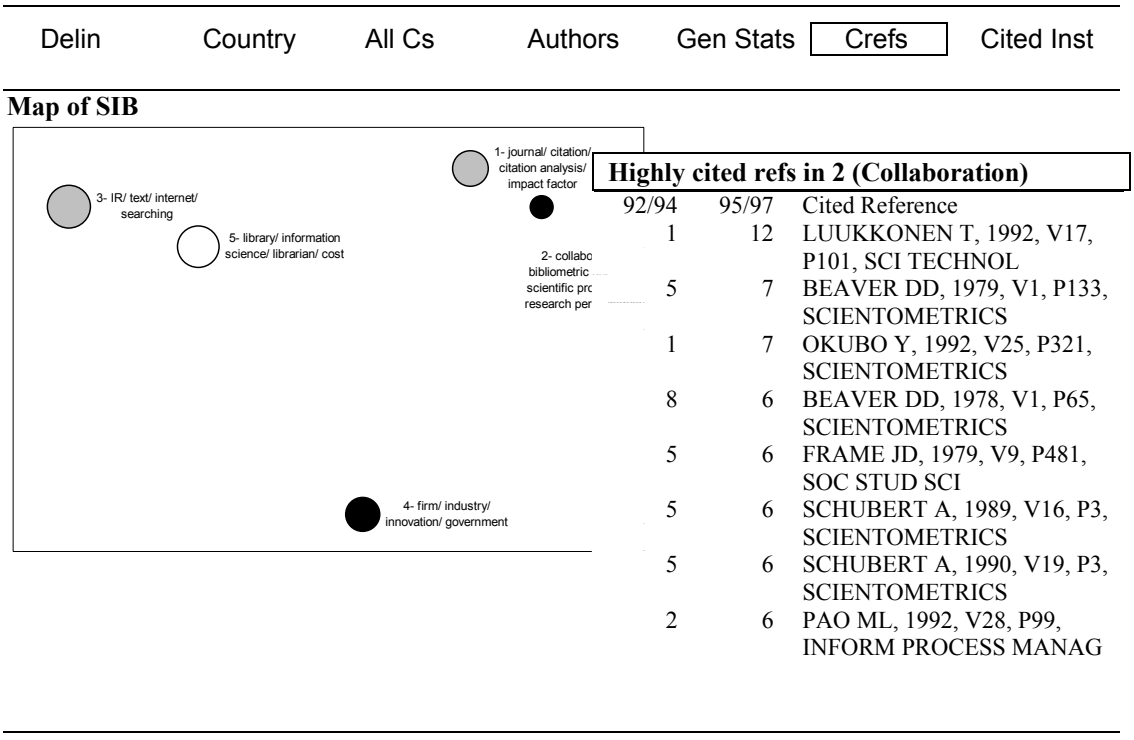


Figure 10-2 Screenshot of mapping interface

This map interface enables a user to evaluate most easily, the internal intrinsic validity (c.f., Figure 3-2) of the generated map. By selecting an information item by clicking one of the top buttons, the top rankings of each subdomain can be retrieved. As the information behind the publications representing each subdomain is directly available, a user does not have to go through piles of papers in order to find the information needed to evaluate the structure. For instance, although subdomain 1 and 2 seem to be covering similar research topics (and therefore they are in each other's vicinity), the lists of most cited references, show significant differences. In subdomain 1 (evaluative

bibliometrics) older work from *Garfield*, *Narin*, and *Cronin* is on top, together with more recent work from *Baird*. In subdomain 2 (collaboration), recent work from *Luukkonen* and *Okubo* and older work from *DeBeaver* is on top.

Furthermore, the aggregation by institution of cited references within a citation window of three years, shows both in subdomain 1 and 2, *Leiden University* and the *Library of the Hungarian Academy of Science* on top, accompanied in subdomain 1 by the *University of Strathclyde* and *Indiana University*. In subdomain 2 however, two Scandinavian (*Inst Studies Research & Higher Education*, Oslo; and *Umea University*) and two French institutes (*Ecole Cent Paris*; and *CNRS Paris*) accompany *Budapest* and *Leiden*.

Finally, the structure can be studied in more detail by the subdomain maps. Following the same procedures as the general overview map, we created detailed maps of each subdomain. Per subdomain we identified the most prominent (subdomain) keywords and normalized their co-occurrence to a matrix of cognitive similarity. On the basis of each subdomain matrix, we generated subdomain (network) maps. For all subdomain keywords, we provided the online version with titles of publications covered. Thus, the user is able to 'descend' to the smallest building block of the map, the publication.

As an example, we present the detail map of subdomain 2 (collaboration) in Appendix A. In this map, the most frequent keywords are positioned in a two dimensional space, where words with a similar cognitive profile (co-occurrences with other words) are in each others vicinity. Moreover, the map is enhanced with the identified cluster structure and with connecting lines indicating a strong co-occurrence relation between two individual words. In a second version of the subdomain map the activity trends around the keywords is indicated.

10.3 Expert input

Although many visitors have browsed through the SIB landscape and its additional information, only a few of them gave comments. Seven SIB researchers took the effort to write comments on the maps and on the additional information through the Internet feedback form (see Appendix B).

The feedback form covered two aspects to which the respondents could give comments. The first refers to the structure as a representation of the field SIB. The second refers to the utility of such maps as a policy-supportive tool. Finally, the respondents could give general comments to the method and results.

The overall opinion of the respondents to the structure was positive. Six of the seven respondents recognized the structure as being a good representation of the field as delineated. The seventh respondent was not sure he recognized the structure, as it seemed too much fragmented to him. Furthermore, six respondents could track down their own research in the map. The seventh commented that his research would be

dispersed over more than one subdomain. This is, however, the case with all respondents. The other six located their work in subdomain 1 (evaluative bibliometrics) and in at least one of the others (four times in subdomain 2, two times in 3, and once in subdomain 5). As a result, we may conclude that the structure appears not appropriate to pinpoint researchers work to exactly one area. We doubt however, in view of the purpose of the map, whether it should. Our maps of science should represent research fields. The subdomain should represent meaningful clusters of topics. The fact that respondents combine research in subdomain 1 with research in three other, seems to justify the fragmentation of the structure. Together with the fact that six respondents acknowledged the structure as being a proper representation of the field, the map seems appropriate for our purposes.

With respect to topics not covered by the maps very few were mentioned. There were no missing topics mentioned by more than one respondent. There were, however, some doubts with respect to the reference of the maps to the 'real world'. Two respondents found the subdomain labeling too 'synthetic'/formal. One of the respondents did not understand all the used keywords. Another regretted that a term like 'information science' was not covered by most subdomains, but rather by one. Of course, the latter observation is a consequence of the used method. The *topic* 'information science' is covered by all subdomains but the *term* is used to delineate one subdomain only. Finally, one of the respondents provided a long list of keywords he would have expected. The list consists of two types of keywords. The first type covers keywords that are much too general (c.f., index, address, utility), the second type covers more specific terms which are probably missing because they have too low a frequency. In the next chapter we will suggest an improved keyword selection procedure.

The question about the policy supportive utility yielded very few comments. Two respondents mentioned the dynamics to be useful. A third respondent mentioned the linkage of subdomains to institutes (actors and cited institutions) to be a useful aspect. One respondent admitted that he did not understand the way the dynamics were generated and therefore could not comment on utility. Two respondents expressed their concern about the ability of policy makers to understand the maps as being representations of scientific research. One of the respondents attributed great value to the maps. As a decision maker himself, he saw the structure and its evolution as something he already suspected. He stated that research policy in his institution would be influenced by the conclusion that could be drawn from our study.

In this chapter we presented the science mapping method, as it is has been applied by CWTS in the past few years, based on our experiences in the Part II studies. We applied the method to our own research field hoping to attract experts more easily to evaluate the results. The comments revealed that, on the whole, the method yielded acceptable results. Moreover, the (interactive) presentation appeared to be a useful

improvement. However, the selection procedure for keywords describing the core of the field needs to be revised.

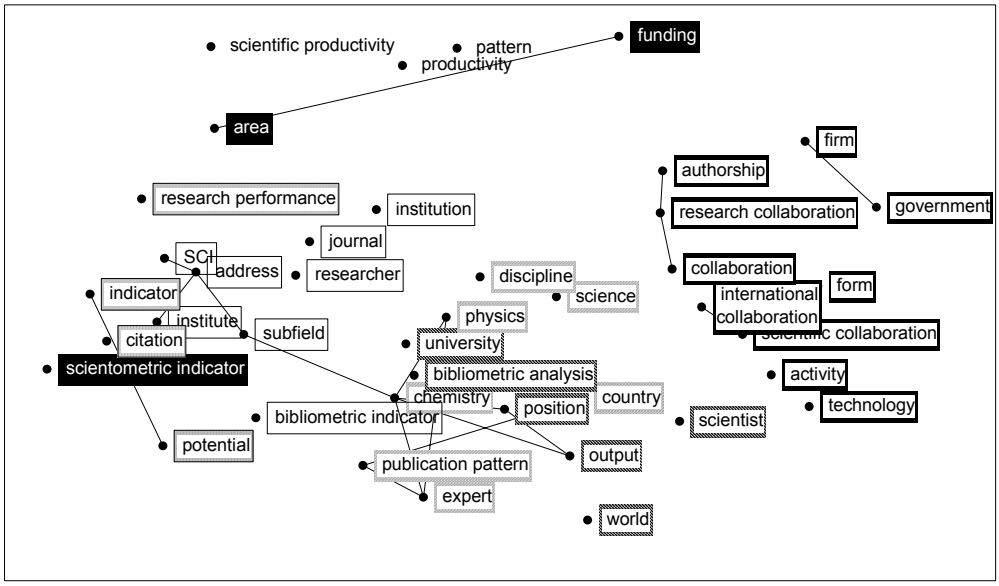
References

Noyons, E.C.M. and A.F.J. van Raan (1998). Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Network Research. *Journal of the American Society for Information Science* 49. 68-81.

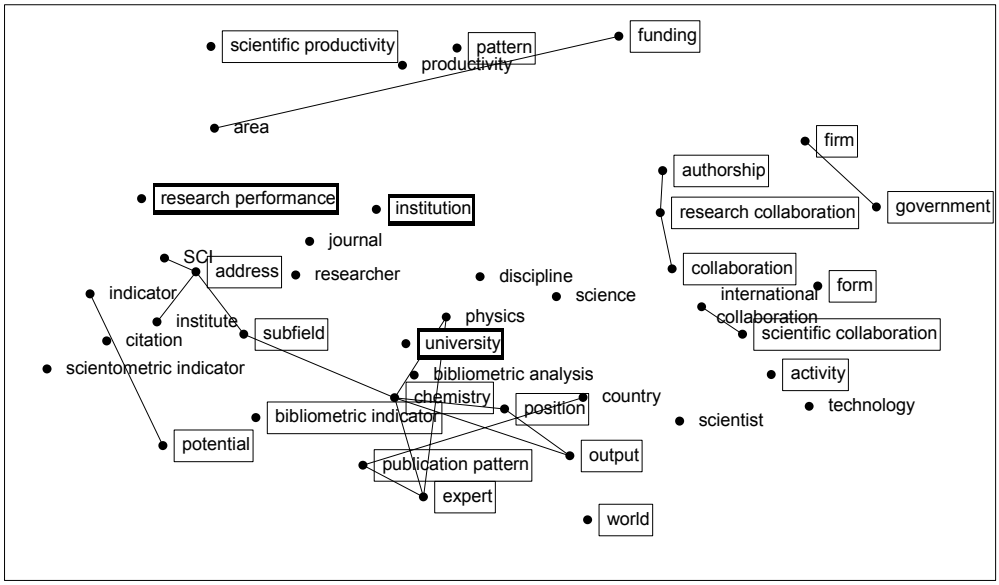
SAS Institute Inc. (1989) *SAS/STAT User's Guide*, Version 6, Volume 1. SAS Institute Inc., Cary, NC, USA, ISBN: 1-55544-376-1

White, H.D. and K.W. McCain (1998). Visualizing a Discipline: An Author Co-citation Analysis of Information Science, 1972-1995. *Journal of the American Society for Information Science* 49. 327-355.

Appendix A



A version



B version

Detail map of subdomain 2 (collaboration) in 1995-1997

Keyword co-occurrence map. Distances based on the cosine co-occurrence vector. The A version of the map is enhanced with the cluster structure (shaded datapoint labels), and with connecting lines between strongly related pairs of individual words. In the B version the keywords with an increasing activity are indicated (label boxes) and those with a decreasing interest (black boxes) during 1992 to 1997. The calculated explained variance is 0.82. Stress: 0.25.

Appendix B

Feedback form of SIB project

Recognizing the landscape

1. Do you recognize the landscape? Does the structure refer to your perception of the field SIB (as defined by the eleven journals)?
   - ○ Yes
   - ○ No
   - ○ Not sure

2. Can you locate your work in one or two sub-domains in the map?
   - ○ No
   - ○ Yes, namely:
   1. ☐ journal/ citation/ citation analysis/ impact factor
   2. ☐ collaboration/ bibliometric analysis/ scientific productivity/ research performance
   3. ☐ IR/ text/ internet/ searching
   4. ☐ firm/ industry/ innovation/ government
   5. ☐ library/ information science/ librarian/ cost

Comments:
……

3. Do you know of areas of interest of the past few years that are represented neither in the overview map nor in any sub-domain map?

……

General Comments

1. Did you come across unexpected structures and/or other findings? And if so: does this refresh your impressions of the field or does it undermine the validity of the maps?

……

2. Did you find any result that could be of importance for policy decisions regarding SIB research? In other words: can you (virtually) think of a situation in which a particular political decision could benefit from the results in these maps that would not have been visualized by a traditional presentation (tables etc).

……

3. Do you have any other comment or question regarding the maps as a tool for policy support?

……