2.2.3 Science mapping from publications

An example in mathematics & computer science [1]

Ed Noyons and Ton van Raan [2]

Introduction

In this Section we discuss the creation of 'maps of science' with help of advanced bibliometric methods. This 'bibliometric cartography' can be seen as a specific type of data mining, applied to large amounts of scientific publications. As an example we describe the mapping of the field mathematics & computer science (MCS). The mapping is based on 'co-word analysis' [Callon, 1983; Noyons, 1998] and applied to CompuMath, the special Citation Index of ISI [3] for computer science and mathematics. The number of publications covered by this database is about 50,000 per year.

This article addresses the main lines of the methodology. We will illustrate the results with a project carried out for the Swiss Science and Technology Council. The aim of the project was to map the field and to assess the 'position' of the Swiss MCS research for the period 1995-1998. Current research is going on to update the mapping for the years 1999-2001.

Each year about a million scientific articles are published. For just one research field, such as MCS, the amount of papers is already about fifty thousand per year. How is it possible to keep track of all these developments? Are there cognitive structures and patterns 'hidden' in this mass of published knowledge, at a 'meta-level'?

Suppose each research field can be characterized by a list of most important, say 200, keywords or, in most cases, phrases i.e. keyword-combinations ('concepts'). For MCS research such a list will cover words like differential equation, optimization, chaos, fuzzy set, parallel computer, Monte Carlo simulation, and so on. Each publication can be characterized by a subset from the total list of keywords. It is, as it were, a DNA fingerprint of a publication. For all publications, keyword-lists are compared pair-wise. In other words, these many thousand publications constitute a gigantic network in which all publications are linked together by one of more common keywords. The more keywords two publications have in common, the more these publications are related (keyword-similarity) and thus belong to the same research area or research specialty. In the biological metaphor: the more DNA two objects have in common, the more they are related. Above a certain similarity threshold, they will belong to a specific species.

Mathematical techniques are used to unravel these publication networks, by word-similarity measurements, clustering of related publications, and finally

1 This article is based on a report for the Centre of Science and Technology Studies (CEST) attached to the Swiss Science and Technology Council, Bern

2 Dr E.C.M. Noyons, noyons@cwts.leidenuniv.nl and Prof Dr A.F.J. van Raan, vanraan@cwts.leidenuniv.nl, Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands, www.cwts.leidenuniv.nl

3 The Institute for Scientific Information in Philadelphia, the publisher of the Science Citation Index and all other related citation indexes.

mapping the ensemble of these clusters in a two-dimensional space. This procedure visualizes an underlying structure. The fascinating point is that these structures can be regarded as the cognitive, or intellectual structure of the scientific field. Clusters can be identified as subfields and research themes.
As discussed above, the procedure is entirely based on the total of relations between all publications. Thus, the structures that are discovered are not the result of any pre-arranged classification system. The structures emerge solely from the internal relations of the whole universe of publications together.
In other words, what is made visible by our mathematical methods, is the self-organized structure of science. A detailed discussion of science maps based on co-word analysis is given in a recent publication [Noyons, 1999].

Methodology

From the above discussion it is clear that keywords from publications play a central role in the methodology. Only noun phrases (NPs) can become field 'keywords'. In order to identify noun phrases in English texts, we use a computer-linguistics based 'noun phrase extractor' (the 'parser'). The identified NPs are divided into two groups: the single word NPs (SWNP) and the multiword NPs (MWNP). From the list of MWNPs, those with too general a meaning are removed.
The selection of field-specific keywords from the list of remaining MWNPs is presently established on the basis of their frequency distribution, and (if possible) the input of field experts. For each MWNP, we count the number of occurrences in titles and abstracts within the field under study, as well as the number of occurrences in titles in science as a whole (i.e. all publications (about a million!) covered by all ISI citation indexes). On the basis of these results the specificity of the NP within the field and its 'centrality' within the field is determined (see [Noyons, 1999] for a detailed discussion).
By using an 'on-line' feedback form, experts are enabled to remove preliminary selected keywords or to add preliminary excluded NPs from the two lists. In order to identify clusters (subfields) within a field, we first construct a matrix composed by co-occurrences of the selected keywords (about 900) in the set of publications for a specific period of time (we start with the most recent period, in the example: 1997-1998). We normalize this 'raw co-occurrence' matrix in such a way that the similarity of keywords is no longer based on the pair-wise co-occurrences, but on the co-occurrence 'profiles' of the two keywords in relation to all other keywords.
This similarity matrix is input for a cluster analysis. In most cases, we use a standard hierarchical cluster algorithm including statistical criteria to find an optimal number of clusters. The identified clusters of keywords represent subfields. These subfields are labeled with a name by the four most frequent keywords in a cluster.

To construct a map of the field, the subfields are positioned in a two-dimensional space. Each subfield represents a subset of publications on the basis of the above discussed keyword similarity profiles. If any of the keywords is in a publication, this publication will be attached to the relevant subfield. Thus, publications may be attached to more than one subfield. The overlap between subfields in terms of joint publications is used to calculate a further co-occurrence matrix based on subfield publication similarity.
The subfields are positioned in two-dimensional space by multidimensional scaling. Thus, subfields with a high similarity are positioned in each other's vicinity, and subfields with low similarity are distant from each other. The size of a subfield (represented by the surface of a circle) indicates the share of publications in relation to the small number in the field as a whole. Particularly strong relations between two individual subfields are indicated by a connecting line. As discussed above, we begin our mapping procedure with the data for a recent time period (here 1997-1998).

The maps can be published on the CWTS web site [4]. Through this browser based interactive interface the maps can be explored and validated. Information 'behind' the map is provided in the same way (actors, and their output and impact indicators).

The map created by our co-word based methodology does not cover 100% of the MCS publications in Compumath. In other mapping projects, for instance neuroscience, we reach a coverage of 80% or more. In this field, we cover only 60% of the publications in 1997-1998. Most probably this relatively low coverage is related to the communication characteristics of the field. Mathematics abstracts contain a lot of 'non-language' expressions such as symbols and formulas. Therefore, less keywords are available for the co-word analysis.

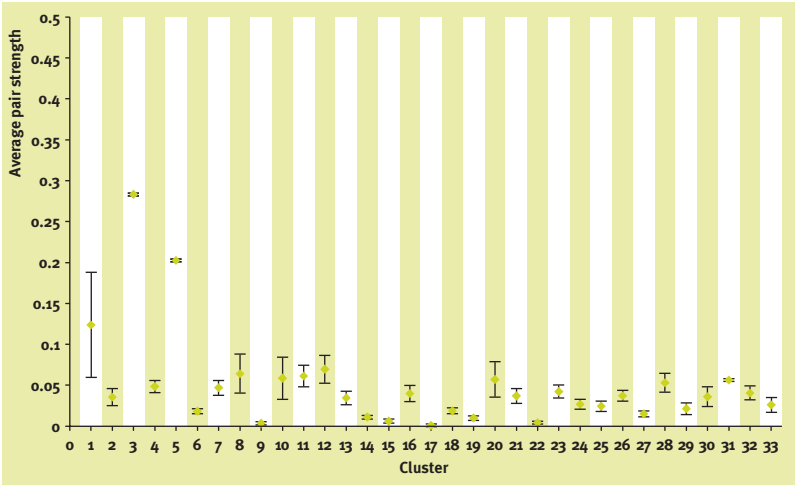


Figure 1
Mathematics Computer Science internal cluster coherence (1997-1998).

4 http://www.cwts.leidenuniv.nl

Analysis mathematics and computer science

The clusters resulting from the mapping procedure have been tested for internal coherence. We calculated the average linkage between all keyword (NP)-pairs within a cluster, and the standard deviation. This internal coherence measure indicates the robustness of the identified subfield. The results are given in Figure 1.

Legend of Figures 1 to 6

subfields	
1	ATM/ performance evaluation
2	Linear model/Bayesian approach/ Monte Carlo method
3	Case based reasoning
4	Fuzzy set/membership function
5	Large number/strong law
6	Simulation study/ maximum likelihood/asymptotic distribution/ distribution function
7	Robotics
8	Scheduling problem/single machine/traveling salesman problem/ job shop
9	Artificial neural network/computer simulation/ computational complexity/ expert system
10	Stability/traveling wave
11	Parallel computer/parallel algorithm
12	Necessary condition/optimal control
13	Optimization
14	Boundary condition/numerical simulation/numerical experiment/partial differential equation
15	Dynamical system/chaos/time series/initial condition
16	Internet/ WWW/website
17	N 1/Finite Group/Monte Carlo simulation
18	Discrete time/continuous time/frequency domain/nonlinear systems
19	Necessary and sufficient condition/Banach Space/Hilbert space/ l 2
20	Vertex/regular graph/chordal graph/distance regular graph
21	Genetic algorithm/objective function/simulated annealing/tabu search
22	Asymptotic behavior/boundary value problem/approximate solution/exact solution
23	Finite element
24	Real time/petri net/ formal method
25	Differential equation/2nd order/runge kutta method/first order
26	Initial data/cauchy problem/global existence/initial boundary value problem
27	Polynomial time/linear time/approximation algorithm/bipartite graph
28	R n/ positive solution/bounded domain/semilinear elliptic equations
29	Complexity/lower bound/upper bound/ branch and bound algorithm
30	Classification/feature extraction/discriminant analysis/texture classification
31	Robustness/disturbance rejection
32	Numerical method/finite difference method
33	Sufficient condition/asymptotic stability/lyapunov function/global stability

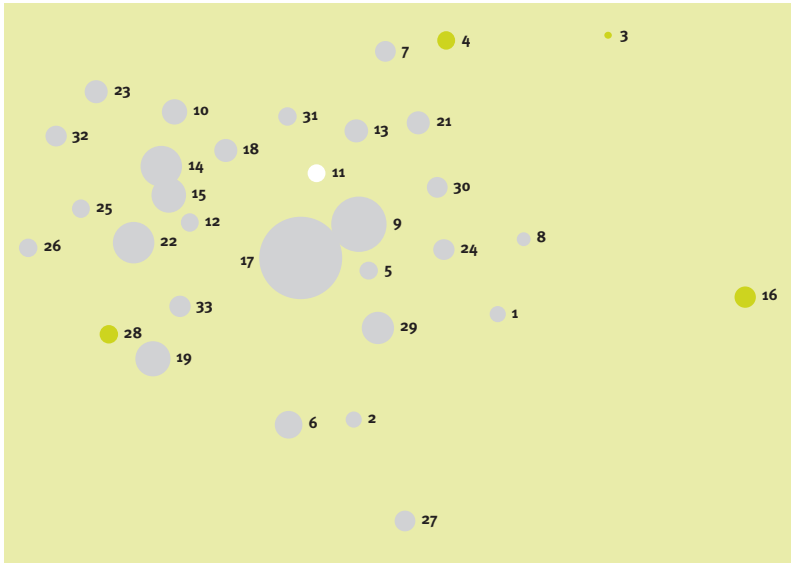


Figure 2
Map of Mathematics & Computer Science (1997-1998). Two-dimensional representation based on the similarities between identified clusters of keywords (subfields). For the list of subfields with corresponding number we refer to the legend of Figure 1. The size of the subfields represents the number of publications in a specific subfield. The color indicates a significant change of publication activity. Green: increase of activity; White: decrease of activity. The badness-of-fit criterion is 0.22, the distance correlation is 0.88.

In Figure 2 we present the structure of MCS by a 2D map of the structural relations between subfields.

As discussed in the methodology section, this map is a two-dimensional representation of a structure resulting from the cognitive, i.e. keyword/concept-based relations of subfields, measured by the co-occurrence of these concepts. These subfields are defined by clusters of related keyword/concepts. By mapping the structural relations between subfields of successive time periods, we create a 'movie' of the evolution of the field within that period. In this movie (see for examples our CWTS web site), we visualize the evolution of individual subfields (growth, in terms of publications), and of their relations with each other (positioning). The information 'behind' the map, can be explored through an interactive browser based interface. Selection of a specific information 'option', enables the user to retrieve data by clicking the relevant subfield circles. The following options are available for each subfield: the most frequently publishing authors, organizations, and countries, as well as the most frequently used journals, the most highly cited publications, authors and organizations. In addition, information is provided to evaluate and validate the structure itself on the basis of word- and citation-linkages between subfields.

In the next steps, we can investigate:
- the relative share of a particular country or institution within the science field.
- The development of activity of this country.
- The impact of the publications of a country or institution compared to the world average within the field.

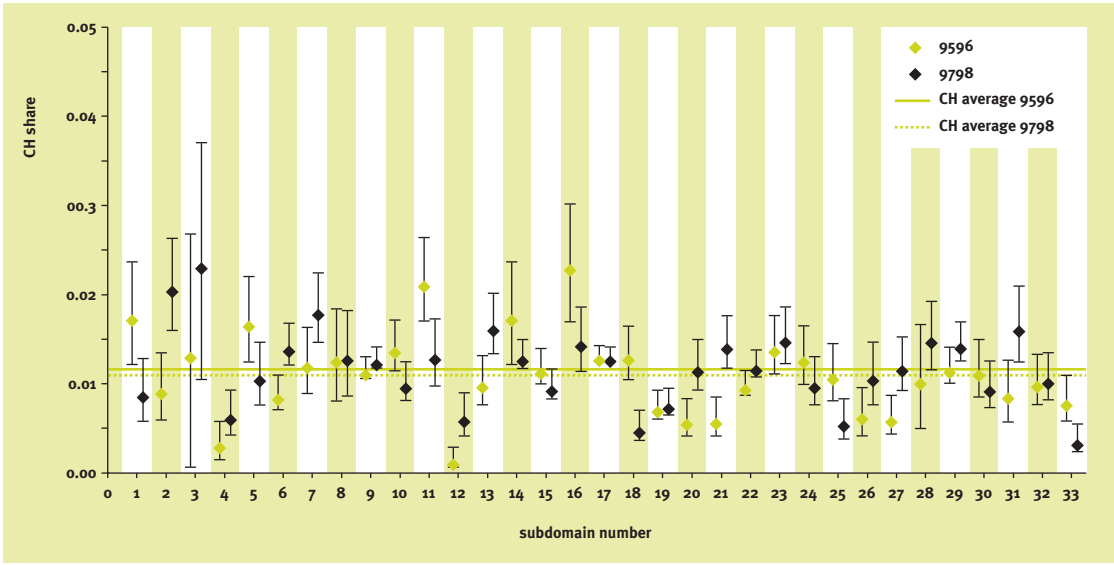


Figure 3
Swiss share in Mathematics & Computer Science subfields (1995-1996 and 1997-1998). For subfields see Figure 1.

These steps are illustrated below with the results of a project carried out for the Swiss Science and Technology Council.

Relative share

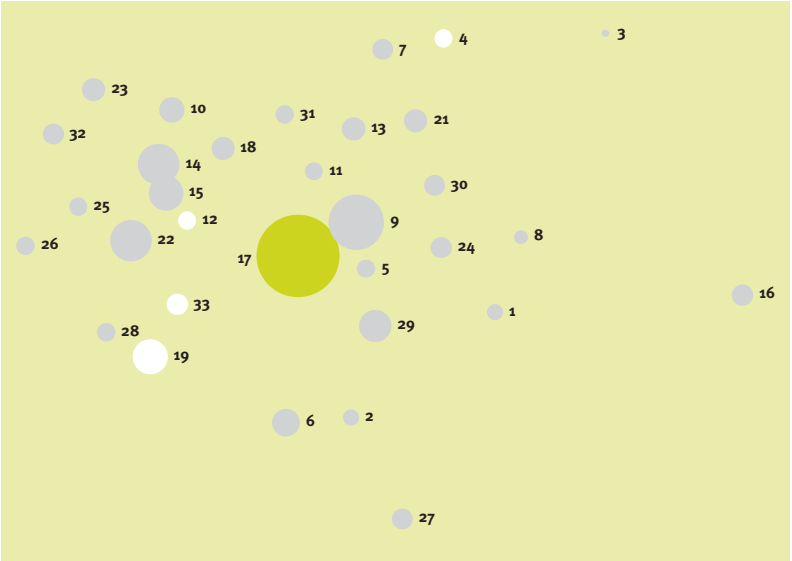
To get an overview of the Swiss activity distribution over the field of MCS, we calculated the share of publications with at least one Swiss address per subfield. This share is a percentage of the total number of publications in a subfield. We determined the Swiss share for the two used periods of time (1995-1996 and 1997-1998) to provide an indication of a trend in the Swiss activity. The error bars added to the data-points indicate the significance of the identified trend. The results are presented in Figures 3 and 4.

We find that the average share of Switzerland in MCS is somewhat more than 1% of the world's total output. In one subfield (17: Finite Group, Monte Carlo simulation), the share of Swiss activity is above this average in the whole period, i.e. both in 1995-1996 and in 1997-1998. This central subfield, in which much general research is covered, Switzerland shows an interest that is significantly above its average in the whole field. In subfields 4, 12, 19 and 33, the Swiss activity share is below its own field average in the whole period.

Development within the subfield

There are two subfields in which the Swiss share decreases significantly in the studied period (18 and 33). There are three subfields (1, 11, 25) in which the decrease remains only just within the calculated error bars. In the case of 11, we are dealing with a subfield with a significantly decreasing world wide interest. In all five subfields with a Swiss share decrease, the absolute Swiss output

Figure 4
Map of Mathematics & Computer Science research with indication of the Swiss share in subfields (1995-1996 and 1997-1998) Colors indicate a Swiss share significantly above (Green) or below (White) its own field average throughout the whole period. For subfields see Figure 1.



decreases as well. There is one subfield (16), in which the Swiss share decreases (though just within the error bars) but in which the absolute number of Swiss output *increases*. In this particular subfield, the world activity increase exceeds the Swiss increase. The conclusion that Switzerland does not keep up with the pace world wide (primarily US publications) is too simple. In the earlier period (1995-1996) Switzerland already showed a relatively high share (more than 2%). In the later period (1997-1998) its activity is lowered to a more average Swiss level (around 1%). The fact that the world-wide activity was increased in 1997-1998, could therefore also be interpreted as a good foresight of the Swiss researchers in this area. We stress, however, to be careful with conclusions based on relatively low absolute numbers.

Furthermore, there are seven subfields in which the Swiss share increases significantly (2, 6, 12, 13, 20, 21, 27). In all these cases the world-wide interest increases as well. Four of these subfields are located at the 'lower' side of the map. Apparently, Swiss MCS research has directed its focus to this area of the field. Swiss activity is also increased in the area above the center. We already mentioned 13 and 20, but also in 3, 4, 7 and 31 an increase of Swiss activity is noted, although not exceeding the error bars. In Figure 5, we summarize the increasing/decreasing share of Switzerland in 1997-1998 in relation to 1995-1996.

Impact

Finally, we indicated in the map those subfields in which Swiss research in MCS reaches an impact that is significantly above or below the world average in 1995-1996. The world average is determined by the average impact of a publication per subfield. Figure 6 shows that the strength of Swiss MCS research is in the core area of the field, i.e. in and around 17 (1, 7, 9, 10, 12, 18, 19, 28, 29, and 33).

Figure 5
Map of Mathematics & Computer Science research with the development of Swiss share in subfields (1995-1996 and 1997-1998).
Color legend:
Green: increase outside error bars.
Dark grey: share increase in 1997-1998 over 20% of 1995-1996.
Black: share decrease in 1997-1998 over 20% of 1995-1996.
White: decrease outside error bars.
For subfields see Figure 1.

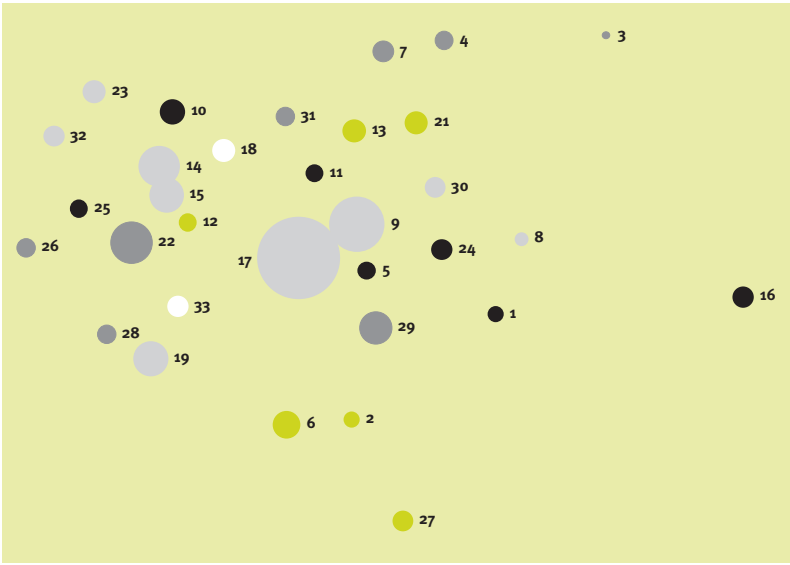
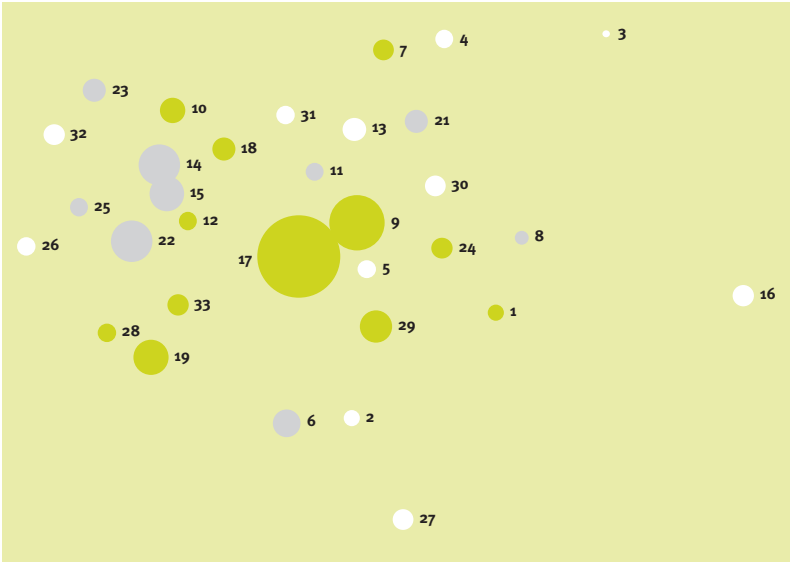


Figure 6
Map of Mathematics & Computer Science research with the Swiss impact as compared to the world average in subfields (1995-1996).
Color legend:
Green: Swiss impact higher than 1.2 times world average.
White: Swiss impact lower than 0.8 times world average.
For subfields see Figure 1.



In general, the Swiss impact in the 'periphery' is lower. In subfields 2, 3 and 4 it is even zero.

Bibliometric analysis of the underlying publication data allows us to identify high-impact institutions. Such specific results can be looked up in more detail via the CWTS map interface.

A detailed discussion of our bibliometric method to measure impact on the basis of citation analysis is given in recent publications [Noyons, 1999; Raan, 1996, Raan, 1999; Raan, 2001].

Conclusion and future vision

With bibliometric mapping we are able to depict the cognitive structure of scientific fields. These cognitive, semantics-based structures act as a 'basic landscape' in visualizing the mutual relations and linkages between subfields and themes within science fields, as well as the *interdisciplinary* relations with other fields. By introducing a time-dimension in the analysis (time-series of maps) we are able to identify *newly emerging themes* ('dynamics of the field'). Thus bibliometric mapping helps to answer crucial questions such as: how does an R&D field look in terms of its cognitive, intellectual structure? How is the field related to its direct 'scientific environment'. Is it possible to explore this 'scientific environment' from the perspective of socio-economic problems? Who and where are the important actors?

Moreover, given the generic character of the methodology, our approach can be extended, if appropriate, immediately to any other data system of (electronically available) documents (e.g. patents, reports, proposals) and databases covering the most recent important international conferences, as well as databases compiled from appropriate sources available via Internet and electronic publishing.

References

- Callon, M., J.-P. Courtial, W.A. Turner, S. Bauin. (1983). From Translations to Problematic Networks: an Introduction to Co-Word Analysis. Social Science Information **22**:191-235
- Noyons, E.C.M., 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 (JASIS) **49**:68-81
- Noyons, E.C.M. (1999). Bibliometric Mapping as a Science Policy and Research Management Tool. Thesis Leiden University. DSWO Press, Leiden
- Noyons, E.C.M., M. Luwel, H.F. Moed. (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes. Journal of the American Society for Information Science (JASIS) **50**:115-131
- Raan, A.F.J. (1996). Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises. Scientometrics **36**:397-420
- Raan, A.F.J. (1999). Scientific Excellence of Research Programs as Pivot of Decision-making. The IPTS Report 40:30-37. Institute for Perspective Technological Studies, Joint Research Institute, European Commission, Seville
- Raan, A.F.J., Th. N. van Leeuwen. (2001). Assessment of the Scientific Basis of Interdisciplinary, Applied Research. Application of Bibliometric Methods in Nutrition and Food Research. Research Policy, to be published