

# Measuring Diversity of Research by Extracting Latent Themes from Bipartite Networks of Papers and References

Oliver Mitesser\*    Michael Heinz†    Frank Havemann‡    Jochen Gläser§

June 9, 2008

## Abstract

Inspired by the hypothesis that the diversity of research might decline as a result of new science policy measures we explore the potential of bibliometric measures for analysing the diversity of research at meso- and macro-levels of (national sub-) fields, countries, and organisations. Our aim is to render changes in the diversity of research landscapes measurable and therefore comparable in time series as well as between different countries. We discuss different methodological approaches and some results based on a method that extracts latent themes from bipartite networks of research papers and their cited references by singular value decomposition of the citation matrix.

## 1 Introduction

The diversity of science appears to be moving to the centre stage of science policy discussions. Recent approaches to the governance of science by performance-based block funding for universities have the potential to affect diversity. These attempts to increase the selectivity of research funding reduce the number of funded units and are thus likely to diminish diversity (Adams and Smith 2003). At a more subtle level, diversity is threatened by the adap-

tive behavior of scientists. Whenever science policy increases punishment for failure, e. g. via reduced funding, researchers are likely to choose projects that are safe in that they are approved of by the scientific community and have a high probability of success. Such safe projects follow the mainstream of a field and use approaches that are known to yield results. Research that deviates from the mainstream is increasingly unlikely to be pursued, which reduces the diversity of problem formulations and research strategies in a field (Harley and Lee 1997; Whitley 2007).

These arguments, albeit persuasive, lack empirical foundation. While the micro-mechanisms that make researchers move flock to the mainstream could be identified (Gläser and Laudel 2007; Gläser et al. 2008), no convincing measurement of research diversity at higher levels of aggregation has so far been provided. Opinions of scientists on the subject cannot be considered as reliable evidence for two reasons. Firstly, perceptions of a changing diversity depend on scientists' individual scientific perspectives and their opinions about science policy. They may therefore be biased. Secondly, quality and marginality of a scientific enterprise are often inseparable. Nonconformist approaches might be perceived as bad science by the majority. Conversely, scientists might rationalise insufficient recognition of their work as being due to the specificity rather than quality of their work.

Testing the above-described 'homogenisation thesis' requires measures of diversity that do not depend on scientists' perceptions of that diversity. Bibliometric indicators can be used to construct these measures because they are unobtrusive and objective, i. e. they neither affect the behaviour they measure nor depend on scientists' opinions about the attribute that is measured.

\*Technische Universität Darmstadt, Universitäts- und Landesbibliothek Darmstadt, Germany, oliver dot mitesser at gmx dot de

†Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI), Germany, michael dot heinz at ibi dot hu-berlin dot de

‡s. footnote before, frank dot havemann at ibi dot hu-berlin dot de

§Freie Universität Berlin, Institut für Soziologie, Germany, Jochen dot Glaser at FU-Berlin dot de

H. Kretschmer & F. Havemann (Eds.): *Proceedings of WIS 2008*, Berlin

*Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*  
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

This is an Open Access document licensed under the Creative Commons License BY

<http://creativecommons.org/licenses/by/2.0/>

The concept of diversity has rarely been used in science studies. A paper by Stirling (2007) deals with conceptual issues of measuring diversity in science and society. To our knowledge the first author who published a bibliometric approach to research diversity was Hariolf Grupp (1990). More recently, diversity measures have been used to gauge the interdisciplinarity (which was conceptualised as thematic diversity) by bibliometric methods (Bordons, Morillo, and Gómez 2004; Rafols and Meyer 2007; Rafols and Meyer 2008).

## 2 Method

### 2.1 Measures of Diversity

If there are more species of trees in a wood  $W$  than in another one,  $W^*$ ,  $W$  is supposed to be more diverse. If  $W$  is dominated by oaks and all other species are very rare, then  $W^*$  can make a more diverse impression despite the lower number of species. Therefore the concept of diversity should not only be based on the number of species but also on the evenness of the distribution of all individuals across all species in a habitat. A measure of diversity that takes into account both species number and evenness is the average information content of the statement that an individual in a habitat belongs to a particular species  $i$ . It results from relative frequencies  $p_i$  of  $n$  species in a habitat by the well-known entropy formula

$$H = - \sum_{i=1}^n p_i \log p_i. \quad (1)$$

$H$  is also named Shannon index.<sup>1</sup> Another measure of diversity that takes species number and evenness into account is the probability that two randomly selected individuals belong to different species:<sup>2</sup>

$$S = \sum_{i,j=1}^n p_i(1 - \delta_{ij})p_j = 1 - \sum_{i=1}^n p_i^2. \quad (2)$$

<sup>1</sup>If dual logarithm is used in equation 1 then  $H$  is given in *bits*.

<sup>2</sup>Kronecker's  $\delta_{ij}$  is 1 if  $i = j$  and 0 otherwise. After drawing the first individual it is put back to the population.

This measure was proposed by Simpson (1949);  $\sum p_i^2$  is called Simpson index.

A further aspect of diversity is the disparity of species. If only conifers grow in wood  $W$  it will give a less diverse impression than the mixed forest  $W^*$  even if numbers of species and balance in both woods are equal. To account for species disparity we replace the binary relation of species (*equal* or *unequal*) with a gradually varying disparity. In equation 2 we have to substitute  $1 - \delta_{ij}$  by a measure  $D_{ij}$  of disparity that varies between 0 and 1 (Shimatani 2001). With

$$\langle D \rangle = \sum_{i,j=1}^n p_i D_{ij} p_j \quad (3)$$

the average disparity of a random pair of individuals is measured.  $\langle D \rangle$  is also called Rao index.<sup>3</sup>

The matrix  $D_{ij}$  of disparity ( $0 \leq D_{ij} \leq 1$ ) can be replaced by a distance matrix  $d_{ij}$  that can have elements  $> 1$ . In this case, biodiversity is measured by the average of a taxonomically or genetically defined distance  $\langle d \rangle$  between all individuals in the biotope. If this concept is applied to a single species, the genetic diversity of that species can be defined by the average genetic distance between its  $n$  individuals, which in general are all genetically different. In this case in equation 3 all relative frequencies become  $p_i = 1/n$  and we arrive at<sup>4</sup>

$$\langle d \rangle = \frac{1}{n^2} \sum_{i,j=1}^n d_{ij}. \quad (4)$$

### 2.2 Co-citation Analysis

Thematic structures of the scientific literature are often analysed and visualised by co-citation analysis, a method that was introduced independently by Marshakova (1973) and Small (1973). In a volume of a set of journals frequently cited sources are considered as *concept symbols*. If two sources are often co-cited then citing authors associate them with each other. Small and Sweeney (1985) constructed co-citation clusters

<sup>3</sup>cf. the paper by Ricotta and Szeidl (2006)

<sup>4</sup>Average distance is normally computed by division by  $n(n-1)$  and not by  $n^2$ . This would be adequate to a random drawing without putting selected items back into the population (cf. footnote 2), but  $n$  is often big enough to set  $(n-1)/n \approx 1$ .

of concept symbols using the Salton index of co-citation as a measure of association and applying single-linkage clustering. These clusters are projected back onto the volume of citing articles to establish clusters of papers called *research fronts*.

Schmidt, Gläser, Havemann, and Heinz (2006) tried to use co-citation clusters and research fronts in electrochemistry for an analysis of the latter's diversity as measured by entropy. In this experiment, a field (electrochemistry) was treated as the equivalent of a biotope, themes within a field as represented by co-citation clusters as the equivalent of species, and papers as the individuals belonging to a species. The experiment failed because measuring diversity cannot be restricted to hot research fronts but has to include many less visible research themes, too. Therefore not only concept symbols but also less cited sources had to be analysed. This amplified a negative feature of single-linkage clustering: *chaining* (clustering of items in a long chain whose ends are not thematically related anymore). In our case, the need to include most publications led to one big cluster of nearly all cited sources below some threshold of co-citation strength.

The negative result of this experiment made clear that a classification of literature into disjunct classes—an analogon to disjunct species in a habitat—is problematic. Even though other cluster methods could avoid chaining, it is not adequate to assign only one theme to a paper.

### 2.3 Bibliographic Coupling

Since it is impossible to assign a paper to only one theme, we abandoned the three-level model based on distinct species in a biotope and turned to a two-level model that measures the genetic diversity of one species (Havemann, Heinz, Schmidt, and Gläser 2007).

There are no disjunct classes of individuals in a species. Based on some measure of genetic similarity the average genetic distance between all individuals in a population can be considered as a measure of its genetic diversity. The genetic information carried by an individual points to its ancestors. In scientific literature, some of the direct intellectual ancestors of a publication can be found in the list of cited sources.

Contrary to the genetic information carried by an individual, the bibliometric information on intellectual ancestors is grossly incomplete. References are included and excluded for a variety of reasons, not all of which are linked to acknowledging intellectual ancestors. More importantly, the information about ancestors of the cited papers is located in these papers' reference lists. Thus, a genealogical tree could be obtained only by recursively drawing ancestors from the total citation graph of the scientific literature, i. e. by including references of references ad infinitum. Since extracting all ancestors is very time-consuming and still does not provide the complete ancestry of a paper, we restricted our analysis to direct ancestors. The similarity of two papers was determined by the number of sources that appear in both papers' lists of references. This relationship—complementary to co-citation—is called *bibliographic coupling*. It has been introduced by Kessler (1963).

Due to the incompleteness of bibliographic information the network of bibliographically coupled articles of a volume is very sparse, however. Its density is low, i. e. only a few of all  $n(n-1)/2$  possible couplings between  $n$  articles are real. An average distance computed on this basis cannot be a useful indicator of research diversity.

However, almost all articles in our sample were elements of the main component of the network, i. e. they were at least indirectly coupled. We took advantage of this feature—often found in information networks—by defining the distance between two papers in the main component by the length of the shortest path between them (Havemann et al. 2007). This approach resembles those of Botafogo, Rivlin, and Shneiderman (1992) and of Egghe and Rousseau (2003), who constructed a measure of compactness of networks based on average length of shortest paths.<sup>5</sup>

Using the same sample of electrochemistry articles as in the co-citation analysis, we have calculated the length of the shortest paths in the main component for a time series of the eleven volumes 1995–2005.<sup>6</sup> The distance between two directly bibliographically coupled papers  $i$  and

<sup>5</sup>s. a. the paper by Rafols and Meyer (2007)

<sup>6</sup>Only records of document type *Article* and *Letter* were downloaded from *Web of Science* (WoS).

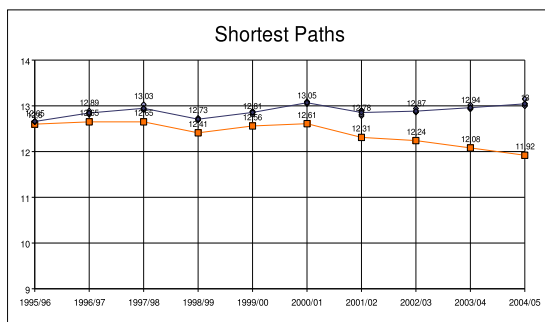


Figure 1: Time series of average shortest paths in a network of electrochemical papers. Red: empirical values, black: model computation.

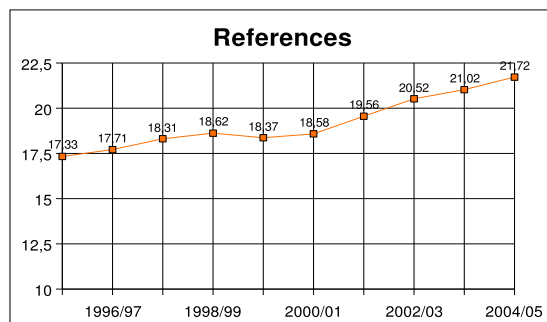


Figure 2: Time series of geometric mean of number of references per paper in 13 journals in electrochemistry.

$j$  was defined as  $d_{ij} = -\log(J_{ij})$ , where  $J_{ij} \leq 1$  is the Jaccard-Index of bibliographic coupling, defined as the ratio between the size of the intersection and of the union of the reference lists  $R_i$  and  $R_j$  of the articles:

$$J_{ij} = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}. \quad (5)$$

For the first six periods (1995/1996 – 2000/2001) the average distance  $\langle d \rangle$  fluctuates around 12.6. Thereafter, it decreases significantly, namely to 11.9 in 2004/2005 (see Fig. 1).

We searched for possible reasons of decreasing distances other than a decreasing diversity of electrochemistry and found that the number of references per paper shows an opposite tendency. It has increased in last years covered by our sample. Since 2000/2001 the (geometric) mean of length of references lists of articles  $|R|$  rises from 18.6 to 21.7 (see Fig. 2).<sup>7</sup> A greater number of cited references per paper leads to more links in the bibliographic coupling network. These additional links act as short cuts, thus shortening the shortest paths between citing papers. We also found that numbers of papers in the 13 electrochemical journals under study have rapidly increased in the last years of the investigated period. This can cause longer paths in the network but also shorten shortest paths by making the network more dense. To check whether decreasing average distances can totally be explained by increasing numbers of links we con-

<sup>7</sup>We use the geometric instead arithmetic mean because the distribution we study is skew.

structed model graphs from our empirical networks by randomly omitting cited sources in reference lists of papers until we got equal means of reference numbers for all periods. In order to exclude any influence of rising numbers of papers, we drew equally sized samples of articles from all periods. We took five samples from each period for an assessment of the scattering of results. Indeed, this procedure let the decreasing trend of  $\langle d \rangle$  disappear (see Fig. 1). This is a strong hint that the decreasing diversity is an artifact produced by changes in the sample over time. However, we cannot be entirely sure. While random samples of individuals can always be used for measuring diversity of populations, randomly omitting references transforms samples of papers into constructed models from which we cannot draw secure conclusions about the real world.

Even though some doubts remain, our chosen approach to measure research diversity as average shortest distance in a network of bibliographically coupled journal papers fails because this indicator is too sensitive to changes in citation behaviour which have nothing to do with changes of diversity.

## 2.4 Latent Themes

A set of articles and their cited sources can be seen as a bipartite network where only links between vertices of different kind are allowed.<sup>8</sup> Co-citation analysis and bibliographic coupling

<sup>8</sup>If articles published in a short period of time are used, the rare cases of articles which are also cited sources of articles in the same volume can be neglected.

are complementary insofar as the former links sources and the latter links citing articles. A method that equally takes into account both these modes of the bipartite network of articles and sources is based on the *singular value decomposition* (SVD) of the rectangular affiliation matrix that describes the network.<sup>9</sup> SVD can be used to extract latent themes of a bibliography. SVD is well suited to our purpose because it returns more than one theme for each article and cited source. This is much more adequate to a subfield's bibliography—as argued above—than hard clustering (Janssens, Glänzel, and De Moor 2007).

The bipartite network of  $m$  papers in a volume of journals in a research field and of the  $n$  sources cited in these papers can be represented by its affiliation matrix  $A$  with  $m$  rows and  $n$  columns. In empirical studies we find nearly always that  $m < n$ . Element  $a_{ij}$  of  $A$  equals one if paper  $i$  cites source  $j$  and zero otherwise. Let  $r \leq m < n$  be the rank of matrix  $A$ . The indices  $i$ ,  $j$ , and  $k$  will run in following ranges:  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ , and  $k = 1, \dots, r$ .

The singular value decomposition of  $A$  is given by the formular

$$A = U\Lambda^{1/2}V^T. \quad (6)$$

The  $r$  columns of  $U$  are normalised eigenvectors of  $B = AA^T$  associated with eigenvalues  $> 0$ . Matrix  $B$  contains the numbers of bibliographic couplings between the  $m$  papers. The  $r$  columns of  $V$  are normalised eigenvectors of  $C = A^T A$  associated with eigenvalues  $> 0$ . Matrix  $C$  contains the numbers of co-citations of the  $n$  sources. The diagonal matrix  $\Lambda^{1/2}$  contains the  $r$  eigenvalues  $\lambda_k > 0$  which both matrices,  $B$  and  $C$ , have in common (as easily can be shown).

It is assumed that  $r$  latent themes can be extracted from the network in a linear ansatz. For this purpose the column vectors of  $A$  (the cited sources), named  $\vec{a}_j$ , are represented by their coordinates  $x_{jk}$  with respect to the  $r$ -dimensional orthonormal basis  $U = \{\vec{u}_k\}$ :

$$\vec{a}_j = \sum_{k=1}^r \vec{u}_k x_{jk}. \quad (7)$$

<sup>9</sup>The method is called *latent semantic analysis* (LSA) if not cited sources but terms are used to describe documents (Deerwester, Dumais, Furnas, Landauer, and Harshman 1990).

Eq. 7 can compactly be written as  $A = UX^T$ . The  $n$  columns of  $X^T$  contain the coordinates of the  $n$  source vectors with respect to the new basis  $U$ .

Analogously, the  $m$  row vectors of  $A$  (the papers) are represented by their coordinates  $y_{ik}$  with respect to the  $r$ -dimensional orthonormal basis  $V = \{\vec{v}_k\}$  resulting in  $A^T = VY^T$  or  $A = YV^T$ . Here the columns of  $Y^T$  contain the coordinates of the  $m$  papers. Comparison with eq. 6 gives  $Y = U\Lambda^{1/2}$  and  $X^T = \Lambda^{1/2}V^T$  or  $X = V\Lambda^{1/2}$ . Equation  $Y = U\Lambda^{1/2}$  translated into coordinates gives

$$y_{ik} = \sum_{l=1}^r u_{il} \delta_{lk} \lambda_k^{1/2} = u_{ik} \lambda_k^{1/2}. \quad (8)$$

The  $r$  coordinates in each of the  $m$  rows of  $Y$  equal the components of paper  $i$  in the directions of the  $r$  themes. In general  $y_{ik}$  can also be negative. Its sign changes if the eigenvector  $\vec{v}_k$  changes its sense of direction (which is not determined). Therefore, the size of theme  $k$  in paper  $i$  cannot be  $y_{ik}$  but is defined as  $y_{ik}^2$ .<sup>10</sup> Then its sign is no longer relevant. The sum of theme sizes  $y_{ik}^2$  in paper  $i$  equals the Euclidian norm of the paper vector  $\vec{a}_i$  (a row in  $A$ ) which does not change if the basis is changed to  $U$ :

$$\sum_{k=1}^r y_{ik}^2 = |\vec{a}_i|^2 = \sum_{j=1}^n a_{ij}^2. \quad (9)$$

Because  $A$  is binary we get

$$\sum_{j=1}^n a_{ij}^2 = \sum_{j=1}^n a_{ij} = |R_i|. \quad (10)$$

Thus, the theme sizes in paper  $i$  sum up to the length of its reference list  $|R_i|$ . With  $p_k = y_{ik}^2/|R_i|$  we can immediately estimate its thematic diversity by calculating its entropy or its Simpson index based on  $p_k$  (Eqs. 1 and 2, p. 2).

The sum of the contributions of two papers,  $i = 1$  and  $i = 2$ , to a theme  $k$  should then be  $y_{1k}^2 + y_{2k}^2$ . The relative joint contribution of both papers together is  $p_k = (a_{1k}^2 + a_{2k}^2)/(|R_1| + |R_2|)$ . Summing sizes of one theme  $k$  in all  $m$  papers and using eq. 8 results in

$$\sum_{i=1}^m y_{ik}^2 = \lambda_k \sum_{i=1}^m u_{ik}^2 = \lambda_k, \quad (11)$$

<sup>10</sup>cf. equation 2 in the paper by Alter et al. (2000)

because the Euclidian norm of the columns of  $U$  equals one. We get that the size of theme  $k$  in the whole bibliography equals its eigenvalue  $\lambda_k$ . The sum of all eigenvalues equals the squared Frobenius norm of matrix  $A$  and thus also the number of links in the network:

$$|A|_F^2 = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 = \sum_{i=1}^m \sum_{j=1}^n a_{ij}. \quad (12)$$

The diversity of the whole volume of papers can then be calculated with relative theme sizes  $p_k = \lambda_k/|A|_F^2$ . The analogue calculation for the theme sizes in cited sources ends up with the same result.

In many SVD-based methods the number of dimensions can be further reduced below  $r$  by omitting eigenvectors which belong to small eigenvalues. This results in a lower number of extracted latent themes, which is desirable in information retrieval. However, for the purpose of measuring research diversity we cannot neglect small themes, as has already been discussed in the context of co-citation clustering.

### 3 Data

We tested the SVD-based extraction of latent themes for two research fields, for electrochemistry—with the set of 13 journals described by Schmidt et al. (2006)—and for the informetric and scientometric part of information science represented by papers in the following five journals:

1. *Information Processing & Management*,
2. *Journal of the American Society for Information Science (and Technology)*,
3. *Journal of Documentation*,
4. *Journal of Information Science*,
5. *Scientometrics*.

For both sets of journals 21 volumes (1986–2006) were downloaded from *Web of Science* (WoS). Details of the datasets are described by Oliver Mitesser (2008).

### 4 Results

We have repeatedly drawn equally sized random samples from each one-year period in both journal sets and calculated a time series of average

entropy and its standard deviation (with respect to sampling). For both research fields we obtained a clear tendency towards higher entropy values (cf. Fig. 3 & 4, graphs on the left side, p. 7). Also for both fields we observe a strong tendency towards higher average numbers of references per paper (cf. Fig. 3 & 4, graphs on the right side).

Next we tested whether the SVD extraction of latent themes is as sensitive to the observed changes of reference numbers per paper as the bibliographic-coupling method described above in subsection 2.3. For this purpose we constructed model networks for each volume of both journal sets by randomly omitting citation links between papers and sources until we reached an average number of 15 references per paper. This reduction changes the entropy for each period, but the tendency towards higher values of entropy is not affected, as Figures 5 & 6 show.<sup>11</sup> The differences between successive periods are not dramatically changed, we only get bigger standard errors.

In 2006 the spectrum of eigenvalues in information science is more even than in 1986 as shown in Figures 7 & 8. In 2006 the biggest front-runner themes differ not so much from the peloton of medium-sized themes as 20 years earlier. In electrochemistry we observe a similar change of eigenvalue spectra (Figures 9 & 10, p. 8).

### 5 Discussion

We cannot yet explain why in both fields entropies of latent themes approach their theoretical maximum during the time span under consideration. In information science we start with 94 per cent of maximum and end up with about 96 per cent. In electrochemistry we have an increase from about 97 to 98 per cent. Obviously this trend cannot continue, it has to slow down in next years.

If the increase of research diversity of both fields is not an artifact—how could it be explained? It is of course possible that all research fields have an inherent tendency to diversify. If this is the case, the homogenisation

<sup>11</sup>The trivial graphs on the right hand sides are generated and displayed to control the random procedure of omitting references.

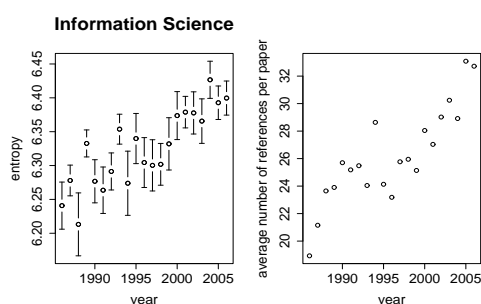


Figure 3: Time series 1986–2006 of average entropy and average number of references per paper in five information-science journals. Entropy averages (maximum  $\log_2 100 \approx 6.64$ ) and standard errors (as error bars) are calculated for 50 samples of 100 papers randomly drawn from each volume.

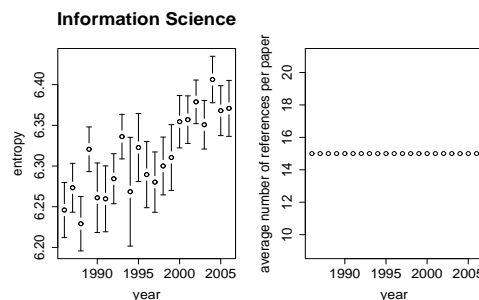


Figure 5: Time series 1986–2006 of average entropy in a model build from five information-science journals with mean number of references per paper randomly reduced to 15. Entropy averages (maximum  $\log_2 100 \approx 6.64$ ) and standard errors (as error bars) are calculated for 50 samples of 100 papers randomly drawn from each model volume.

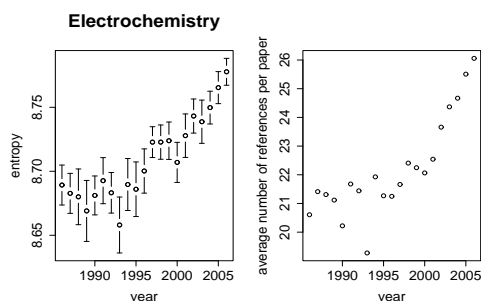


Figure 4: Time series 1986–2006 of average entropy and average number of references per paper in 13 electrochemistry journals. Entropy averages (maximum  $\log_2 500 \approx 8.97$ ) and standard errors (as error bars) are calculated for 50 samples of 500 papers randomly drawn from each volume.

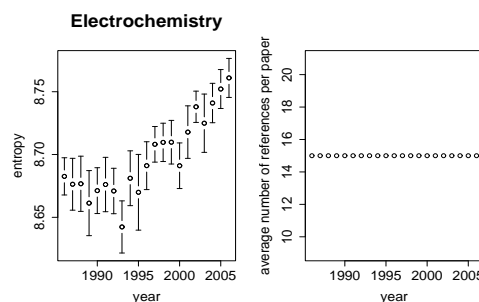


Figure 6: Time series 1986–2006 of average entropy in a model build from 13 electrochemistry journals with mean number of references per paper randomly reduced to 15. Entropy averages (maximum  $\log_2 500 \approx 8.97$ ) and standard errors (as error bars) are calculated for 50 samples of 500 papers randomly drawn from each model volume.

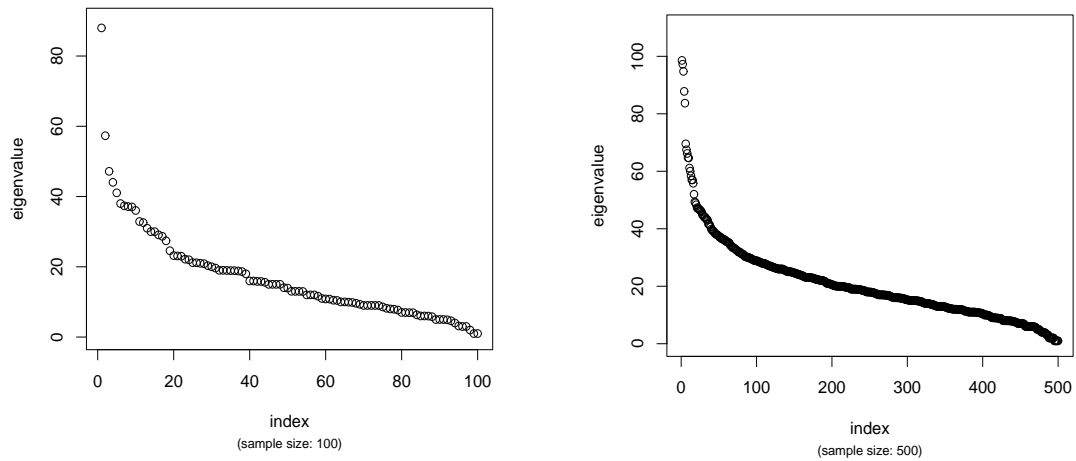


Figure 7: Eigenvalues in information science, 1986. Figure 9: Eigenvalues in electrochemistry, 1986.

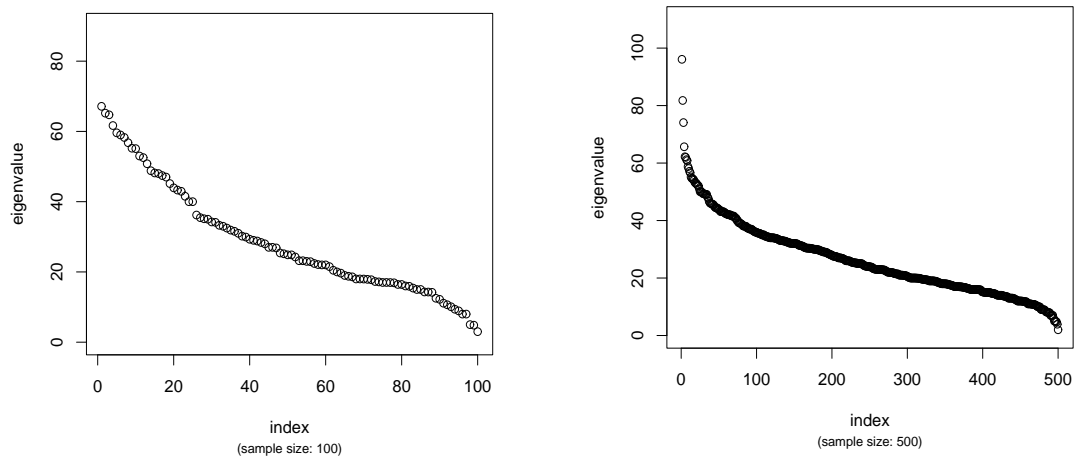


Figure 8: Eigenvalues in information science, 2006. Figure 10: Eigenvalues in electrochemistry, 2006.

thesis could only be tested by comparing trends and their changes in different fields and countries. An SVD based analysis of thematic diversity of a country's research is possible within the approach outlined above and should be the next point on the agenda.

The significant lengthening of reference lists that has occurred in both information science and electrochemistry over the last ten years merits further investigation. In order to make sure that changing citation behaviour does not affect the SVD-based entropy measure of latent themes, we will have to establish in detail the kinds of sources whose citation frequency has increased.

The strong tendency towards higher entropies we found for the bipartite networks of papers and cited sources has to be confirmed by ordinary latent semantic analysis of the bipartite networks of papers and terms. Our results could also be tested by using some variant of probabilistic latent analysis.<sup>12</sup> Furthermore, diversity measures other than entropy should also be tested. We plan to define distances between themes in order to be able to apply the Rao index discussed above.

## Acknowledgement

We are grateful to the developers of the free statistics and graphics software **R**.<sup>13</sup>

We thank Marion Schmidt for discussions about the results of our bibliographic-coupling experiment.

Earlier this year we presented some of our results at a workshop of the Gesellschaft für Wissenschaftsforschung (Society for Science Studies, Berlin).<sup>14</sup> We thank all participants who took part in the discussion after the talk.

## References

Adams, J. and D. Smith (2003). Funding research diversity. *A report from Evidence Ltd to Universities UK*. 1(84036), 102.

<sup>12</sup>cf. the recent paper by Griffiths and Steyvers (2004)

<sup>13</sup>cf. <http://www.r-project.org>

<sup>14</sup><http://www.wissenschaftsforschung.de/asstag.html>

Alter, O., P. Brown, and D. Botstein (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* 97(18), 10101–10106.

Bordons, M., F. Morillo, and I. Gómez (2004). Analysis of cross-disciplinary research through bibliometric tools. In H. Moed, W. Glänzel, and U. Schmoch (Eds.), *Handbook of quantitative science and technology research*, Chapter 21, pp. 437–456. Kluwer, Dordrecht.

Botafogo, R., E. Rivlin, and B. Shneiderman (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems (TOIS)* 10(2), 142–180.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407.

Egghe, L. and R. Rousseau (2003). BRS-compactness in networks: Theoretical considerations related to cohesion in citation graphs, collaboration networks and the internet. *Mathematical and Computer Modelling* 37(7-8), 879–899.

Gläser, J., S. Lange, G. Laudel, and U. Schimank (2008). Evaluationsbasierte Forschungsfinanzierung und ihre Folgen. In F. Neidhardt, R. Mayntz, P. Weingart, and U. Wengenroth (Eds.), *Wissen für Entscheidungsprozesse*, pp. 145–170. Bielefeld: transcript.

Gläser, J. and G. Laudel (2007). Evaluation without Evaluators: The impact of funding formulae on Australian University Research. In R. Whitley and J. Gläser (Eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, pp. 127–151. Dordrecht: Springer.

Griffiths, T. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl. 1), 5228–5235.

H. Kretschmer & F. Havemann (Eds.): *Proceedings of WIS 2008*, Berlin

*Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*  
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

This is an Open Access document licensed under the Creative Commons License BY

<http://creativecommons.org/licenses/by/2.0/>

- Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics* 18(3–4), 219–239.
- Harley, S. and F. S. Lee (1997). Research selectivity, managerialism, and the academic labor process: The future of nonmainstream economics in UK universities. *Human Relations* 50, 1427–1460.
- Havemann, F., M. Heinz, M. Schmidt, and J. Gläser (2007). Measuring Diversity of Research in Bibliographic-Coupling Networks. In D. Torres-Salinas and H. F. Moed (Eds.), *Proceedings of ISSI 2007*, Volume 2, Madrid, pp. 860–861. Poster abstract.
- Janssens, F., W. Glänzel, and B. De Moor (2007). A Hybrid Mapping of Information Science. In D. Torres-Salinas and H. F. Moed (Eds.), *Proceedings of ISSI 2007*, Volume 1, Madrid, pp. 408–420.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation* 14, 10–25.
- Marshakova, I. V. (1973). Sistema svazey meshdu dokumentami, postroyennaya no osnove ssylok. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protssesy i Sistemy* 6, 3–8. (in Russian).
- Mitesser, O. (2008). Latente semantische Analyse zur Messung der Diversität von Forschungsgebieten – Methodendiskussion und Anwendungsbeispiel. Master’s thesis, Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft.
- Rafols, I. and M. Meyer (2007). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. In D. Torres-Salinas and H. F. Moed (Eds.), *Proceedings of ISSI 2007*, Volume 2, Madrid, pp. 631–637.
- Rafols, I. and M. Meyer (2008). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. To appear, cf. [www.sussex.ac.uk / spru / documents / rafols-meyer-diversity2008.pdf](http://www.sussex.ac.uk/spru/documents/rafols-meyer-diversity2008.pdf).
- Ricotta, C. and L. Szeidl (2006). Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao’s quadratic index. *Theoretical Population Biology* 70(3), 237–243.
- Schmidt, M., J. Gläser, F. Havemann, and M. Heinz (2006). A Methodological Study for Measuring the Diversity of Science. In *International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, 10-12 May*, Nancy, pp. 129–137. SRDI – INIST-CNRS.
- Shimatani, K. (2001). On the measurement of species diversity incorporating species differences. *Oikos* 93(1), 135–147.
- Simpson, E. (1949). Measurement of diversity. *Nature* 163(4148), 688.
- Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science* 24, 265–269.
- Small, H. and E. Sweeney (1985). Clustering the Science Citation index® using co-citations. *Scientometrics* 7(3), 391–409.
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface* 4(15), 707–719.
- Whitley, R. (2007). Evaluation without Evaluators: The Consequences of Establishing Research Evaluation Systems for Knowledge Production in Different Countries and Scientific Fields. In R. Whitley and J. Gläser (Eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, pp. 3–27. Dordrecht: Springer.