Gems from our Digitization Project

In Knowledge Organization *volume 41 number 6 (2014), we reprinted the introductory editorial by Ingetraut Dahlberg "Why This Journal," from the first issue of* International Classification *volume 1 number 1 (1974). In* Knowledge Organization *volume 42 number 1 (2015) we reprinted her report "The Terminology of Subject Fields" from* International Classification *volume 2 number 1 (1975). The editorial set out the publishing program for the new domain of knowledge organization. The paper on subject terminology laid out the methodology and task for the fledgling domain to begin generating exhaustive sets of terminology. In an email response Dr. Dahlberg asked me whether there had been any response. Unfortunately to date there has not. After the second gem appeared I received this email from Dr. Dahlberg:*

Thank you for reprinting that subject-field article of 1975. I reread it and remembered that there was no response whatsoever on the proposals I made to Infoterm or the Unesco in this paper.

Thereafter the work continued and two further projects increased the number of fields found to become finally 12.500. But fortunately in 1976 a grant from the German Research Association allowed to add to all of the collected terms their definitions. Every day I went to the German National Library in Frankfurt to copy (in writing, of course no typewriter was allowed in the reading room) from special dictionaries the definitions for those terms lacking them so far. We had of course started by copying from 4 encyclopedias what we could find there, alone 20 thick volumes of Brockhaus!. The result of the definitions helped to understand the contents of the terms and to realize that about half of the collection were synonyms! But with the definitions we were able to place them at their respective place in the system which I had already thought of in 1971. All of this led finally to an adequate sorting of the subject-field terms into their subject groups. But the scheme which you also reprinted in the last number is out of date in many cases.

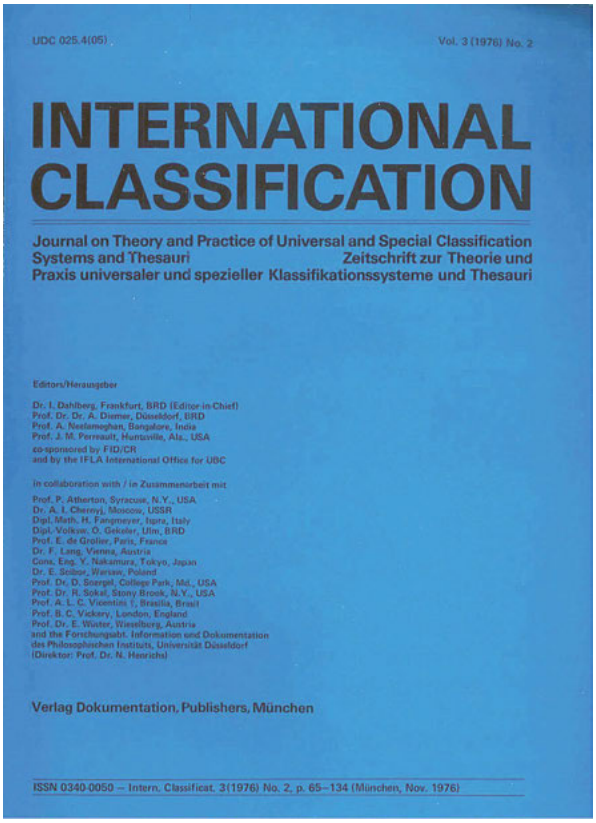
In 1977 I presented the result in a seminar at the DRTC in Bangalore. It was well accepted. And when someone asked, what was the name of the system, I had no answer. Only on my way back from India, I decided it should be called *Information Coding Classification*.

But this is not the end of the story.

We thank Dr. Dahlberg for this rich enhancement and we look forward to the further bits of the story!

In this issue we reprint another classic, an oft-cited but I suspect little-read work by Henry G. Small that sets out the method and potential results of what would become known as bibliometrics for domain analysis. We have seen this research find a home in the information science community, but only recently has it found a resurgence in the knowledge organization community. In some ways this piece by Small is the beginning of the realization of statistical bibliography and bibliometry as set forth by Paul Otlet in Traité du Documentation *(Bruxelles: Editiones Mundaneum, 1934). Notice how Small refers to clustering as classification, and how he names citation an epiphenomenon. I find it fascinating that this piece was published in the earliest pages of our own journal and thus we are excited to add it to our collection of gems.—Ed.*

Reprinted from *International Classification: Journal on theory and practice of universal and special classification systems and thesauri = Zeitschrift zur Theorie und Praxis universaler und spezialler Klassifikationssyteme und Thesauri*. Vol. 3 (1976) No. 2. The masthead identifies the "editors" as: Ingetraut Dahlberg, Alwin Diemer, Arashanipalal Neelameghan and Jean M. Perreault. Emphasis is as in the original.


UDC 025.4(05) Vol. 3 (1976) No. 2

INTERNATIONAL CLASSIFICATION

Journal on Theory and Practice of Universal and Special Classification Systems and Thesauri Zeitschrift zur Theorie und Praxis universaler und spezieller Klassifikationssysteme und Thesauri

Editors/Herausgeber

Dr. I. Dahlberg, Frankfurt, BRD (Editor-in-Chief)
Prof. Dr. Dr. A. Diemer, Düsseldorf, BRD
Prof. A. Neelameghan, Bangalore, India
Prof. J. M. Perreault, Huntsville, Ala., USA

co-sponsored by FID/CR
and by the IFLA International Office for UDC

In collaboration with / in Zusammenarbeit mit

Prof. P. Atherton, Syracuse, N.Y., USA
Dr. A. I. Chernyj, Moscow, USSR
Dipl. Math. H. Fangmeyer, Ispra, Italy
Dipl. Volksw. O. Gekeler, Ulm, BRD
Prof. E. de Grolier, Paris, France
Dr. F. Lang, Vienna, Austria
Cons. Eng. Y. Nakamura, Tokyo, Japan
Dr. E. Solbor, Warsaw, Poland
Prof. Dr. D. Soergel, College Park, Md., USA
Prof. Dr. R. Sokal, Stony Brook, N.Y., USA
Prof. A. L. C. Vicentini F., Brasilia, Brasil
Prof. B. C. Vickery, London, England
Prof. Dr. E. Wüster, Wiesellburg, Austria
and the Forschungsinst. Information und Dokumentation des Philosophischen Instituts, Universität Düsseldorf (Direktor: Prof. Dr. N. Henrichs)

Verlag Dokumentation, Publishers, München

ISSN 0340-0050 — Intern. Classificat. 3(1976) No. 2, p. 65–134 (München, Nov. 1976)

Henry G. Small
Institute for Scientific Information, Philadelphia

Structural Dynamics of Scientific Literature

Small, H. G.: Structural dynamics of scientific literature. In: *Intern. Classificat.* 3 (1976) No.2, p. 67-74.

A methodology is described for the structural analysis of scientific literature using the frequency with which items have been cited together (co-citation) as the measure of association between highly cited documents. A clustering procedure applied to these data results in the formation of groups of documents identifiable as research specialties. The principal concern of the paper is to analyze patterns of change in these structures from year to year, using data from the 1973 and 1974 *Science Citation Index*. The rate of change of clusters and the document turnover in clusters suggest that scientific research fronts are changing very rapidly. Change in the relationships among specialties is also observed by comparing configurations of clusters obtained by multidimensional scaling. By rotation of successive M-D-SCAL plots to a least squares congruence for corresponding clusters, it is possible to show how the clusters have "moved" relative to one another from one year to the next. (Author)

1. Introduction

Science, whether viewed as a social or an intellectual system, is essentially self-organizing. The very process of doing science—creating and communicating new knowledge—generates structure, whether in terms of the interrelation of ideas or of informal communication among scientists. The difficulty is that this structure is an "invisible" one. It is the result of many small and somewhat private thoughts or actions, the collective outlines of which remain obscure.

The problem, then, is how to "discover" or "uncover" this structure and monitor its changes. The journal literature of science and the footnote citation patterns in that literature provide perhaps the most accessible and least obtrusive source of data (1). To be sure, citations are only a by-product or epiphenomenon of scientific activity. Nevertheless, there is reason to believe that citations are an accurate mirror or indicator of this activity. Weinstock (2) provides an exhaustive list of why authors cite. The reasons and motivations for citing appear to be as subtle and as varied as scientific thought itself, but most references do establish valid conceptual links between scientific documents.

The approach to the problem of discovering structure is to design what might be called a "science mapping system" which takes as its input the cumulations of citation data contained in the *Science Citation Index (SCI)* or the *Social Sciences Citation Index (SSCI)*. The basic measure of relationship used is co-citation, the citing of two earlier documents by one later document (3). It is hypothesized that co-citation captures, in most instances, the elementary act of associating ideas or individuals represented by the two cited documents. Hence, the kinds of structures obtained from co-citations are, by and large, interpretable as cognitive or social structures.

The principal finding of the initial experiments (4-7) is that the primary structural unit in science is the scientific specialty. Furthermore, the specialties themselves are linked together in loose networks which eventually connect nearly the entire fabric of science. The most recent work which is reported here concerns the application of this system to the study of how this structure changes

over time. Thus, the aim of this paper is to move from a static, cross-sectional model of the structure of science to a dynamic, longitudinal one.

2. Method

Since the aim of the science mapping system is to measure changes in structure over time as well as structure at any given time, it is necessary to deal with discrete time slices of citation data. The annual cumulations of the *SCI* provide a convenient time slice, not so narrow as to give an unrepresentative picture of most fields and not so broad as to preclude an early detection of significant changes. The *SCI* is a multidisciplinary citation index, covering about 2400 journals in all fields of the natural sciences. In this paper data obtained from two successive *SCI* files, 1973 and 1974, will be discussed. The general strategy is summarized as follows:

1. Select all items in an annual file *(SCI* or *SSCI)* cited more than N times where N is empirically determined (typically N = 15).
2. Determine the number of times pairs of items identified above were co-cited.
3. Use the co-cited pairs of items, appropriately normalized, as input to a clustering algorithm (single-link clustering).
4. Determine the linkages between clusters by summing co-citations among documents in different clusters (called "cluster co-citation").
5. Match sets of clusters obtained independently from successive years to determine new, dropping and continuing clusters.

Setting the threshold for selecting highly cited items is an important step because it determines the level of detail and field coverage of the clusters. Experience has shown that no one threshold of citation frequency is optimal for all fields. However, it is possible to gain some perspective on the interdisciplinary structure of the major fields of science using an overall threshold of about 15 citations per item per year.

The file of co-cited items is created by sorting the selected items file in source (citing) article sequence and forming all pairs (co-citations) of cited items appearing in a given source article. The pairs of cited items are then sorted, and a frequency count is attached to each unique pair indicating the frequency of co-citation. The frequency of co-citation is normalized using the Jaccard formula (in effect determining the percentage of citations to the two items that are co-citations). The pairs are input to a single-link clustering algorithm and clusters of items are formed at a specified threshold of the Jaccard coefficient (8). In order to examine the relationships among clusters, an inter-cluster measure of association is defined at a particular level of the Jaccard coefficient. The form of the cluster-cluster association measure used in this paper is the total number of articles co-citing two different clusters.

Cluster correspondence, the final step in the system, refers to the matching of clustered files derived for successive years to determine what clusters in each year share cited documents and are hence "continuing clusters." This step also determines the clusters that are "new" in the sense of containing documents which did not appear in any cluster the previous year, or "dropped" in the sense that no documents in the clusters appear in the subsequent year's clusters.

3. Comparison of Cluster Statistics for 1973 and 1974

The 1973 *SCI* and the 1974 *SCI* were clustered at identical thresholds (citation frequency threshold of IS and a normalized co-citation threshold of 16 percent). Following the cluster runs the two files were matched to determine the degree of continuity from year to year. In this section, the discussion centers on the statistical comparison of the two cluster runs in terms of changes in overall characteristics, what might be called "external" change. In the following section, focus will be on the change in the make-up or composition of the clustered files, that is, "internal" change.

The magnitude of the *SCI* files for the two years are shown in the "Input File Statistics" section of Table 1. The number of journal articles (source items) in the file declined slightly (-1.5 percent) from 1973 to 1974, but the number of citations from source items increased (+ 4.3 percent), as did the number of cited items (+ 2.3 percent). The mean number of citations per cited item also increased slightly from 1.81 to 1.84, and the mean number of references per source item increased from 12.1 to 12.8. The changes in these numbers probably reflect changes in the *SCI* journal coverage, and to a lesser extent, changes in the scientific literature itself, although the contributions of these two factors would be extremely difficult to trace. Important for the present experiment is that the two files are comparable in size, and no radical differences are evident.

The citation frequency threshold was set at 15 citations per document per year for each file. The characteristics of these files are shown in Table 1 under the heading "Highly Cited File Statistics." With both thresholds 15, the number of cited items selected increased 8 percent from 1973 to 1974. An increase was also observed in the number of source items selected by citing one or more of the selected highly cited items. These increases result from the use of a constant threshold on a file which has increased in size.

The statistics for the creation of pairs of highly cited items which have been cited together are given in Table 1 under the heading "Co-citation File Statistics." The number of distinct pairs of co-cited items, that is, combinations of highly cited items, has increased 25 percent from 1973 to 1974. This is due to the fact that a small increase in the number of highly cited items selected results in a large increase in the number of pairs of such items. In addition, the total network of linkages among the highly cited items has increased in density as shown by the following two statistics: the ratio of the number of all linkages among highly cited items to the number of possible linkages among that number of highly cited items ("percent connected"), and the mean number of co-cited pairs for a single cited item (the number of other cited items with which a given item is linked). These increases in linkage density can be traced to the increase in the number of cited items initially selected, indicating that the additional items selected came from areas of scientific literature where co-citation densities are high. We cannot conclude, in any event, that the natural sciences as a whole are becoming increasingly interrelated, as these statistics might superficially suggest.

The clusters analyzed here were derived at the 16 percent level of the Jaccard coefficient in each year (see "Cluster File Statistics" section of Table 1.) At this level the overall percent connected statistic for the network is very nearly the same for 1973 and 1974, although the mean number of co-citation linkages per document has increased slightly. A total of 1610 clusters containing two or more cited items was generated for 1973 and 1702 clusters for 1974. The mean cluster size increased slightly from 5.0 to 5.2 cited items per cluster.

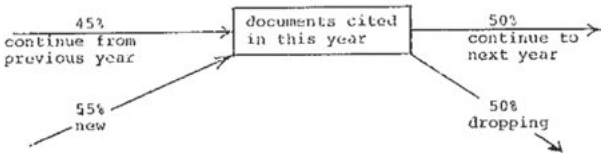
Of interest in gauging the effectiveness of this approach to clustering is the percentage of selected cited items and selected source items which fall into clusters at this level. About 51 percent of the cited items originally selected are included in a cluster at this level. Also, about the same percentage of the source items selected by having cited one or more of the selected cited items are included in one or more of the clusters at this level. Since the selected source items constituted about 36 percent of the total source items contained in the *SCI,* the clusters at level 16 percent are able to "classify" only about 18 percent of the total source items in the annual *SCI* which is, however, roughly 74000 source items. An increase in the number of source items classified could be brought about by a lowering of the initial citation frequency threshold and/or a lowering of the Jaccard level for clustering.

Statistics on the relationships among clusters at level 16 percent are summarized in Table 1 under the heading "Cluster Co-citation File Statistics." First, each source item appears on the average in 1.5 clusters in 1973 and 1.6 clusters in 1974. In other words, while a cited item can be assigned to one and only one cluster, a source (citing) item can be assigned to more than one cluster if it cites items which have been assigned to different clusters. The co-citing of items in different clusters gives rise to nearly identical percent connected networks of all clusters in 1973 and 1974, and a slight increase in the mean number of linkages to other clusters for any given cluster. Again, there is no evidence to suggest a higher level of interdisciplinary research in the natural sciences as a whole.

4. Measurement of Change

The previous discussion has centered on the overall characteristics of the clustered files in the two years. In summary, the two files are similar in structure, the only differences stemming from the fact that the 1974 file is somewhat larger than the 1973 file, giving rise to small increases in average cluster size and the density of the network. But this relative constancy in overall characteristics is misleading in the sense that it masks a great deal of internal change when the cited items making up the clusters are examined from one year to the next. The approach to the measurement of internal change is to determine what cited items are common to the cluster sets in each year.

The statistics on internal change are summarized in Table 1 under the heading "Cluster Correspondence Statistics." There were a total 8042 cited items clustered at level 16 percent in 1973 and 8917 items clustered at this level in 1974. The number of items common to these two sets (identical cited items) is 4046 or 50.3 percent of the 1973 items clustered and 45.4 percent of the 1974 items. From this it can be determined that in 1973, 3996 items (49.7 percent) failed to appear in the 1974 clusters and hence "dropped out," and 4871 items (54.6 percent) in the 1974 clusters were "new," that is, did not appear in the 1973 clusters. If these figures are recast in terms of a single hypothetical year, the results can be diagrammed in the following way:



These statistics are striking in the degree of change they suggest. Even though total numbers of cited documents in clusters remain fairly constant, the make-up of the document set has changed enormously. It is of interest to compare our data on the rate of document turnover in clusters with data obtained by Price and Gursey on the rate of turnover in scientific authors (9). By examining a representative slice of the alphabet and using several con-

secutive years of the *Source Index* of the *SCI,* they found the following pattern for authors publishing in a given year:

In the second part of their study they report "a percent carryover of cited authors from year to year in the *SCI*" which is very similar to the pattern they found for continuing source authors (10). In both their data and our own, the larger percentage of documents continuing to the next year over the percentage from the previous year is the result both of the growth of science and of the *SCI* data base.

Our data, of course, are cited documents rather than cited authors or source authors and furthermore, constitute a very special sample of cited documents, those cited 15 or more times in a year and appearing in clusters. Nevertheless, the pattern of continuing cited documents from year to year is very similar to that found by Price and Gursey for source and cited authors.

The identification of continuing documents leads to the identification of continuing clusters of documents. A "continuing" cluster is defined as one which contains one or more continuing cited items, that is, items which appear in clusters the next year. Statistics on cluster continuation are shown in Table 1 under "Cluster Correspondence." Of the 1610 clusters formed in 1973, 1004 of them (62 percent) "continue" to 1974, and of the 1702 clusters in 1974, 1051 of them (62 percent) "continue"' from 1973. Formulated in the previous manner we have:

This is a lower rate of turnover for clusters than for documents, but nevertheless represents a substantial rate of change of clusters from year to year.

The statistics show that in 1974, 651 clusters were "new" in the sense that they contained documents cited 15 times or more that did not appear in any cluster the previous year. As would be expected, the mean size of continuing clusters is larger than the size of new or dropping clusters (about seven cited documents for continuing clusters compared to about three for new or dropping clusters). Nevertheless, there is a large influx of new clusters, which, if they can be shown to correspond to new areas of knowledge, means that science is changing very rapidly. Clusters which contain all "new" documents may mark the emergence of new areas of research or, at the least, new approaches to existing areas.

Of course, an area need not be totally "new" to mark a significant change in research. For example, an established specialty which corresponds to an ongoing cluster might experience a radical shift which results in a large influx of new cited documents and/or a dropping out of many old documents. The overall percentage of new documents in clusters is 55 percent, hut the overall percentage of new documents in continuing clusters is about 40 percent. A continuing cluster which has a higher percentage of new documents may signal an important change in the specialty. Some preliminary investigations into the stability of clusters over time and the relation of cluster change to social and intellectual change in the specialty will be reported elsewhere (11).

There are, of course, a wide variety of possible patterns of cluster continuation involving the merging or splitting of document groups. The simplest is the passing on of documents from one cluster in year t to one cluster in year t+l. In fact, this is the most frequent pattern. Of the 1004 continuing clusters in 1973, 837 (83 percent) did not diverge, that is, did *not* split into two or more clusters in 1974. Similarly, of the 1051 continuing clusters in 1974, 902 (86 percent) were *not* part of converging patterns from two or more clusters in 1973. The fact that there are more continuing clusters in 1974 than in 1973 indicates a net excess of diverging patterns over converging patterns. The mean number of clusters converging to single clusters in 1974 is 1.25, while the mean number of clusters diverging from single clusters in 1973 is 1.29. It is important to examine these splits and mergers to see if they signal changes in specialty boundaries, i.e., the moving apart of previously connected areas or the coming together of previously unconnected ones.

In summary, the statistics on the rate of change of documents in clusters indicate that a great deal of change is occurring. Apparently, subject and social structure undergo rapid change, at least on the surface or research front of the specialty. New specialties can emerge quite rapidly and also disappear rapidly within a fairly stable overall framework.

5. Changes in Cluster Configurations

In this section, a methodology is outlined for gaining a more detailed view of how systems or networks of clusters change. It was noted earlier that for each pair of clusters in an annual file a measure of inter-cluster association is determined called "cluster co-citation," which is a count of the number of source papers citing documents in *both* clusters. This measure can be used to arrive at a

structure of a set of clusters either in network or spatial terms. The problem is how to measure the structural change which has occurred in going from a configuration of clusters in year t to a configuration of corresponding clusters in year t+l.

As shown in Table 1, the average cluster is linked to about 32 other clusters through cluster co-citation. If the cluster in question continues to the next year, roughly 19 of these links are to other continuing clusters. Of course, the continuing clusters need not be the clusters linked with the previous year. In other words, new links with other continuing clusters may form, or some links with continuing areas may cease to exist. To further complicate the system, any given link between the same two continuing clusters may increase or decrease in value.

The next few paragraphs outline the steps in the analysis of structural change. The first step is to select a starting cluster and its corresponding cluster in the subsequent or previous year. Around the starting clusters, single-link "clusters of clusters" are generated for each year, setting a threshold for cluster co-citation that will produce conveniently sized sets for analysis (usually not exceeding about 30). Each of the clusters in a given year's network is then traced forward to the next year and backward to the previous year to make sure that all the corresponding continuing clusters are included. All clusters for which no predecessors or successors are found are excluded. This confines the analysis to the "continuing" clusters only.

The purpose of the previous steps is to select out of the total sample of clusters a set which exhibits strong inter-cluster linkages, and, further, to insure that continuing clusters have been included for all clusters in the sets.

The next step provides a model for the analysis of change from year to year. The key to this model is the technique of multidimensional scaling as formulated by Kruskal and embodied in the M-D-SCAL program (12), (13). In essence, the technique takes as input a set of relationships among objects (in our case the relationships are cluster co-citations and the objects are clusters) and orders them in a space of N dimensions. The relative positions of objects are obtained when a statistic called "stress," which is a goodness-of-fit measure between input proximities and assigned locations in space, achieves a minimum value. In the present application, the nonmetric option of M-D-SCAL is used, which considers only the rank ordering of the input proximities in calculating stress. As input to M-D-SCAL, a lower-half matrix of cluster co-citations is constructed for the cluster set for each year. (Only a lower-half matrix is needed since co-citation is a symmetric relationship.) Each matrix is scaled independently and configurations are obtained in two dimensions.

In the final step, the centroids of the two M-D-SCAL configurations are superimposed and the 1974 configuration is rotated with respect to the 1973 configuration until the sum of squared distances between corresponding continuing clusters is minimized. The mirror image of the 1974 configuration is also superimposed and rotated to see if the fit can be improved. A joint space of the 1973 and 1974 configurations is constructed using the angle of orientation that minimizes the squared distances.

These steps provide a rough coordination of the two configurations. It is rough in the sense that M-D-SCAL is an approximate procedure where only general relative locations of points are meaningful. Superposition of centroids implies that the center of gravity of equally weighted clusters remains stationary. Rotation to congruence to minimize squared distances between continuing clusters is based on the premise that there is something akin to a principle of least effort involved in the movement of subject areas from one time period to the next. A limitation of the procedure is that the coordinates of the points in M-D-SCAL are standardized such that their mean square distances from the centroid of the configuration is set equal to unity. Therefore, changes from one M-D-SCAL configuration to another cannot reflect changes in scale or size.

Example I: Plate Tectonics

The first example is from the field of plate tectonics. The 1973 cluster # 831 "volcanic activity" and its corresponding cluster in 1974, #1480 "mantle rheology," were selected as starting clusters. Single-link networks were generated around each of the starting clusters in the respective years, using a threshold of 10 cluster co-citations to link clusters in the network. This gave a set of seven clusters in 1973 and a set of five clusters in 1974. When the 1973 clusters were traced forward to 1974, an additional 1974 cluster was added (#1433 "geothermal models") which was not included in the 1974 single-link map. Only corresponding clusters with two or more common documents were considered "continuing."

Lower-half matrices of cluster co-citations were constructed for the seven clusters in 1973 and the six clusters in 1974, and scaled non metrically, obtaining configurations in two dimensions. Superposition of centroids and rotation to a least squares fit of corresponding clusters gave the joint 1973-1974 representation in Figure 1, with arrows connecting the corresponding clusters. The two arrows converging on the 1974 cluster #949 "basalts" indicate that two 1973 clusters (#1014 "trace elements" and #1071 "basalts") merged in 1974.

The joint configuration suggests that considerable movement has occurred for some clusters while others

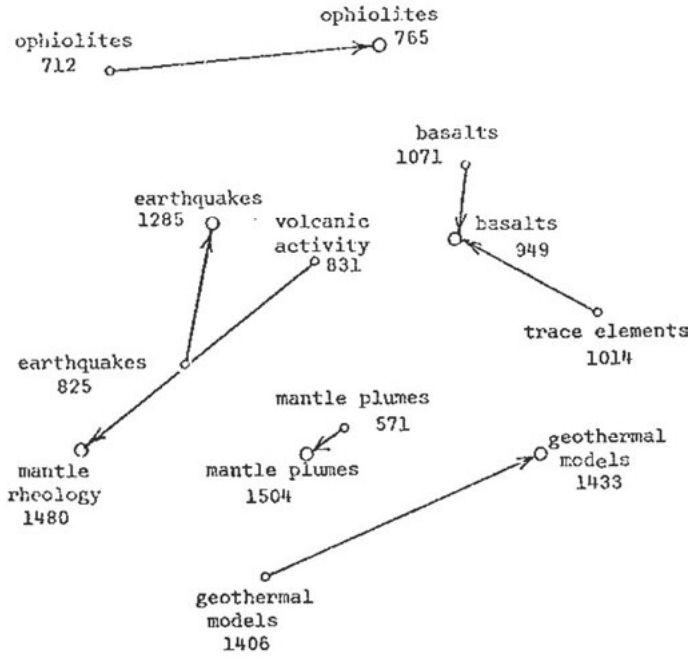


Fig. 1: Plate Tectonics 1973-1974 Joint Configuration

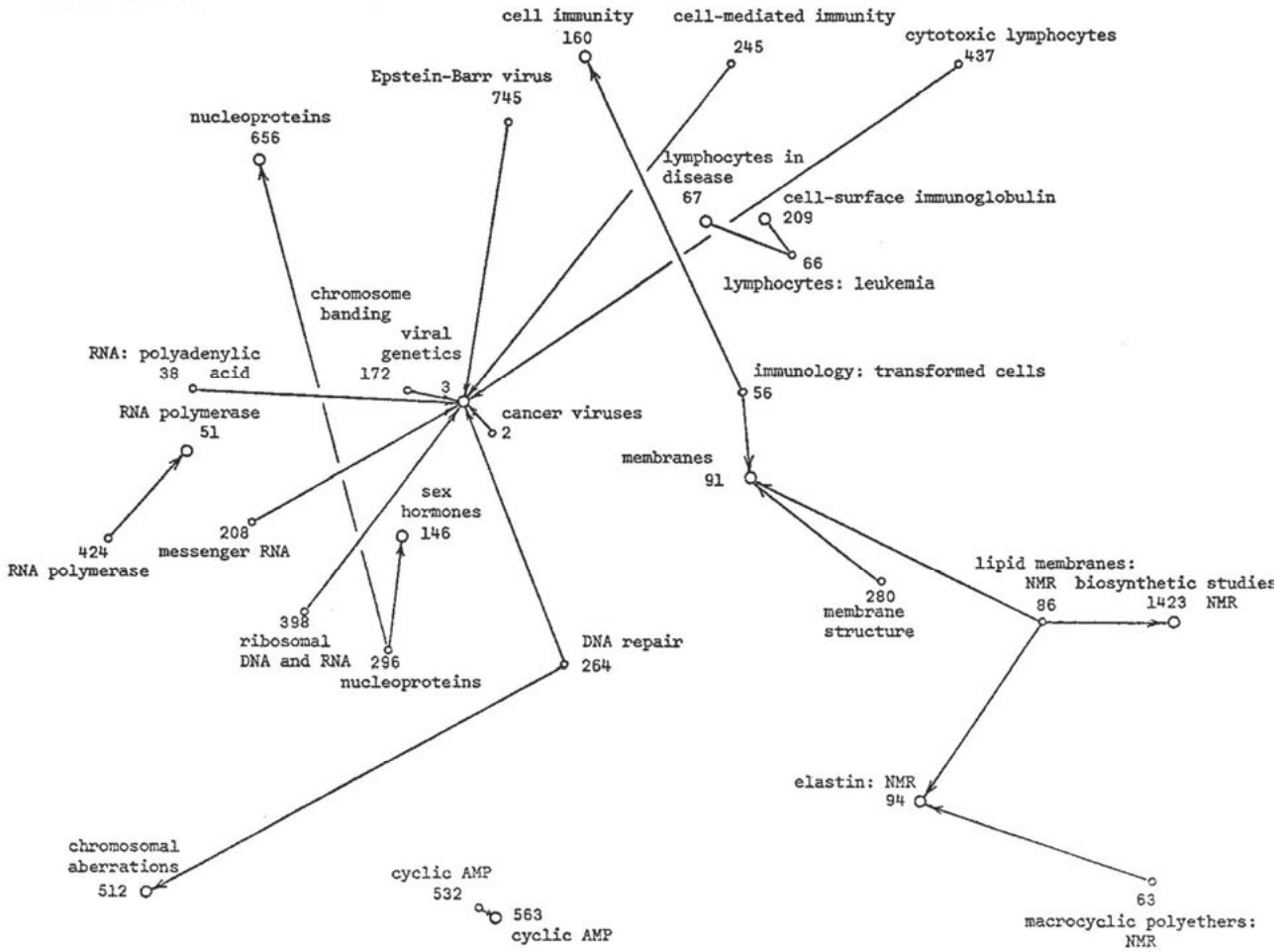


Figure 2. Biomedicine 1973–1974 Joint Configuration

have remained relatively immobile. "Ophiolites" and "geothermal models" are moving roughly parallel to one another towards "basalts" and in the opposite direction from "volcanic activity." "Volcanic activity" is moving away from all areas and "earthquakes" has moved slightly in the direction of "ophiolites."

Some preliminary ideas on the nature of these shifts can be obtained by examination of the titles of the clustered documents. First, the map represents two different approaches to plate tectonics, the geophysical and geochemical. The "ophiolites," "basalts" and "trace elements" clusters in the upper right of the Figure are dominated by geochemists, while the other areas are dominated by geophysicists. Hence, the movement of "ophiolites" toward "basalts" represents a consolidation of interest. The merging of the "basalts" and "trace elements" cluster is also undoubtedly due to their common concern with geochemistry.

The movement of the "volcanic activity" cluster away from the others seems to be due to a drastic decline in this area (the cluster goes from 13 to 2 cited documents) and a shift of concerns away from volcanism to the modeling of the movements of the earth's mantle. Concern with heat flow, which was a part of the "volcanic activity" cluster, has shifted to the "geothermal models" cluster.

Example II: **Biomedicine**

The second example concerns a set of biomedical clusters, which includes some of the largest clusters in the 1973 and 1974 files. This example is more complex than plate tectonics especially in terms of the pattern of merging and splitting of clusters from year to year. The clusters and cluster connections are much larger than for plate tectonics, lending increased statistical significance to the analysis.

The starting cluster for the analysis was, in fact, the largest cluster in the 1973 file in terms of cited documents (#2 cancer viruses) which corresponds most strongly to the largest cluster in the 1974 file (#3 viral genetics). Around these starting clusters, single-link networks of clusters were generated at the level of 20 cluster co-citations. Only clusters consisting of 200 citing documents or more were included in the network. When the 1973 clusters in the network were traced forward to 1974 and the 1974 clusters traced backward to 1973, no additional clusters were added to the networks. As in the case of plate tectonics, only corresponding clusters with two or more common documents were considered "continuing." It is quite remarkable in this case that all continuing clusters in the two years were included in the single-link networks at level 20.

Lower-half matrices of cluster co-citations for the 17 clusters in 1973 and the 12 clusters in 1974 were constructed and then scaled. The smaller number of 1974 clusters indicates that the merging patterns predominate over the splitting patterns. Superimposing the centroids of the M-D-SCAL configurations and rotating to a least squares fit for continuing clusters yield the joint 1973-1974 map shown in Figure 2.

The 1973-1974 map shows considerable movement and also considerable regional organization. The most striking change is the merging of nine clusters to form the "viral genetics" cluster (#3). This represents a confluence of research on genetic aspects of viruses. Especially interesting is the drawing in of two clusters from the immunology region (#245 "cell-mediated immunity" and #437 "cytotoxic lymphocytes"). These merge into cluster #3 along with clusters which are predominantly biochemical in orientation, such as "RNA polyadenylic acid" (#38), "messenger-RNA" (#208), and "ribosomal DNA and RNA" (#398). To the left of "viral genetics" are some areas with strong biochemical orientation, namely "RNA polymerase" and "nucleoproteins." Below it is an important continuing area, "cyclic AMP," shows almost no movement from year to year. The third region is to the upper right of cluster #3 and focuses on immunology. Below the immunology region are areas concerned with cell membranes. Finally, to the lower right of "membranes" we find clusters utilizing the technique of carbon-13 labeling and nuclear magnetic resonance (NMR) in the study of biological problems.

The two way split of "DNA repair" into "viral genetics," on the one hand, and "chromosomal aberrations" (#512) on the other, seems to be due to concern of the latter with the genetic disease xeroderma pigmentosum which is more clinically oriented than viral genetics. The split of "nucleoproteins" (#296) into "sex hormones" (#146) and "nucleoproteins" (#656) and the movement of this latter group to the other side of "viral genetics" is perhaps due to the division of researchers into groups concerned with hormones and those concerned with the nature of histone and chromatin.

Another interesting and marked shift is the splitting of one of the immunology clusters (#56 "immunology: transformed cells") into two parts, one moving to join with "membranes" (#91) and the other "cell immunity" (# 160) moving further away from the membrane region towards immunology. It is also evident that the technique of nuclear magnetic resonance is moving strongly into the biological membrane work and is being applied to other biological systems, as indicated by cluster #86 ("lipid membranes: NMR") splitting three ways into "membranes" (#91), "biosynthetic studies" (#1423) and "elastin" (#94). These areas border on the fields of chemistry and biochemistry and show the infusion of chemical techniques into biomedicine.

6. Conclusions

The goal of this work is to shed some light on the structural dynamics of science and provide some measures of the nature and rapidity of change. The approach taken is to design a clustering system which utilizes the citation connections contained in the journal literature of science, and then to compare the structures obtained from successive cumulations of citation data. The clusters which emerge from such a system, built around the co-citation procedure, have been found to correspond to scientific specialties, and the linkages between specialties have led to the construction of "maps" of science. The work reported here shows how these maps may be animated to indicate changing patterns of relationships among research areas. Much work remains to be done on the validation and explication of these dynamic maps, but they provide highly concrete and detailed pictures to compare with the perceptions of the scientists and open many new avenues of interpretation.

A comparison of the cluster statistics of 1973 and 1974 shows that the overall characteristics of the files have not changed in a major way, and that the changes which are observed are attributable to the growth of the *SCI* and of science. Change is very pronounced, however, when the document make-up of the clusters in the two years is examined. The rate of entry and exit for clusters is somewhat slower than for documents, but it is still surprisingly fast. The annual clusters are probably an indicator of what Derek Price has called "research fronts" (14), and changes in clusters from year to year reflect changes in very recent research emphasis.

Movements of specialties relative to one another have also been observed in the two cases examined, plate tectonics and biomedicine, suggesting that the relationships among specialties are extremely fluid. The general regional organization of the research areas are preserved (e. g. the immunology specialties remain in the same general region) but substantial mobility occurs within regions. These movements are apparently nonsecular in nature and can be traced to short term fluctuations in social or cognitive conditions within the specialties.

It may turn out that absolute rates of change are not as important as large deviations from average or expected rates. Hence, a 40 percent influx of new cited documents into a cluster representing the research front of an existing specialty may not indicate anything more than the successful continuation of research and the incorporation of new results. A new cluster containing five new cited documents may, however, mark an event of some importance. Similarly, some movements of specialty research fronts relative to one another may represent only short lived "Brownian" motions caused by efforts of specialists to incorporate aspects of a neighboring specialty's methods or ideas. Collective or concerted movements of groups of specialties, on the other hand, (such as the merger of several clusters with the viral genetics cluster in 1974), may indicate events of significance.

Acknowledgments

A preliminary version of this paper was presented at the meeting of the Classification Society held in Rochester, N.Y. on May 24, 1976. I am grateful to many of those present for thoughtful comments. The research was supported by National Science Foundation grant SOC73-09096 A02 to the Institute for Scientific Information.

References

(1) Garfield, E.: Citation Indexes for Science. In: Science, 122 (1955) p.108-lll.
(2) Weinstock, M.: Citation Indexes. In: Encyclopedia of Library and Information Science, New York: Marcel Dekker, 5 (1971) p. 16-40.
(3) Small, H. G.: Co-citation in the Scientific Literature: A New Measure of the Relationship between two Documents. In: J. Amer. Soc. Inform. Sci., 24 (1973) p. 265-269.
(4) Small, H. G., Griffith, B. C.: The Structure of Scientific Literatures 1: Identifying and Graphing Specialties. In: Science Studies, 4 (1974) p. 17-40.
(5) Griffith, B. C., SmalL H. G., Stonehill, J. A., Dey, S.: The Structure of Scientific Literatures II: Toward a Macro-and Microstructure for Science. In: Science Studies, 4 (1974) p. 339-365.
(6) Garfield, E., Malin M. V., Small, H.: A System for Automatic Classification of Scientific Literature. In: Indian Inst. of Sci., 57 (1975) No.2, p. 61-74.
(7) Small, H.G., Griffith, B. C.: Automatic Classification of Scientific Literature using Co-citation Clustering. In: Proc, Twelfth Annual Allerton Conference on Circuit and System Theory, (1974) p. 512-521.
(8) Sneath, P. H. A., Sakal, R. R.: Numerical Taxonomy. San Francisco: W. H. Freeman 1973. p. 131.
(9) Price, D. J. D.,Gursey, S.: Studies in Scientometrics. I. Transience and Continuance in Scientific Authorship. In: Intern. Forum on Inform. and Doc., l (1976) No. 2, p. 17-24.
(10) Price, D. J.D., Gursey,S.: Studies in Scientometrics. II, The Relation Between Source Author and Cited Author Populations. In: Intern. Forum on Inform. and Doc., 1 (1976) No. 3, p. 19-22.
(11) Small, H. *G.:* A Co-citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research.

In: Social Studies of Science, to be published May 1977.

(12) Kruskal, J. B.: Multidimensional Scaling by Optimizing Goodness-of-fit to a Non-metric Hypothesis. In: Psychometrika, 29 (1964) p. 1-37.

(13) Green, P. E., Carmone, F, J,: Multidimensional Scaling and Related Techniques in Marketing Analysis. Boston: Allyn and Bacon 1970, p. 144.

(14) Price, D. *1*. D.: Networks of Scientific Papers. In: Science, 149 (1965) p. 510-555.