Hybrid Citation-Word Representations in Science Mapping: Portolan Charts of Research Fields?

Michel Zitt

Lereco, INRA, BP 71627, F-44316 Nantes cedex 3, France. E-mail: zitt@nantes.inra.fr

Alain Lelu

Université de Franche-Comté, LASELDI, F-25000 Besançon, France; or LORIA – Campus scientifique – BP 239 – 54506 Vandoeuvre cedex – France. E-mail: alelu@9online.fr

Elise Bassecoulard

INRA, Lereco, F-44316, France. E-mail: bassecou@nantes.inra.fr

The mapping of scientific fields, based on principles established in the seventies, has recently shown a remarkable development and applications are now booming with progress in computing efficiency. We examine here the convergence of two thematic mapping approaches, citation-based and word-based, which rely on quite different sociological backgrounds. A corpus in the nanoscience field was broken down into research themes, using the same clustering technique on the 2 networks separately. The tool for comparison is the table of intersections of the M clusters (here M = 50) built on either side. A classical visual exploitation of such contingency tables is based on correspondence analysis. We investigate a rearrangement of the intersection table (block modeling), resulting in pseudo-map. The interest of this representation for confronting the two breakdowns is discussed. The amount of convergence found is, in our view, a strong argument in favor of the reliability of bibliometric mapping. However, the outcomes are not convergent at the degree where they can be substituted for each other. Differences highlight the complementarity between approaches based on different networks. In contrast with the strong informetric posture found in recent literature, where lexical and citation markers are considered as miscible tokens, the framework proposed here does not mix the two elements at an early stage, in compliance with their contrasted logic.

Introduction

Citation flows, on the one hand, and lexical similarity, on the other hand, are the bases of the main networks used to retrieve information and map knowledge flows, from the Google search engine, which mixes hyperlinks among sites and word retrieval, to mapping techniques used in scientometrics. The mapping of science is the topic of an abundant literature, in which we highlight a few milestones: citation-based mapping was developed especially at Drexel University and Institute for Scientific Information (ISI) in the late sixties and the seventies; citation flows were developed by Garfield (1967); bibliographic coupling was developed by Kessler (1963); document cocitation was developed by Small (1973), Small and Griffith (1974), and Marshakova (1973); author cocitation was developed by White and Griffith (1981) and further developed by McCain (1983), who experimented also on journal cocitation analysis (McCain, 1991); and influence relations were developed by Pinski and Narin (1976). The theoretical framework of citation mapping was developed in Merton's and de Solla Price's wake.

Later, the development of science studies expanded the latitude of interpretations of citation relations, ranging from Merton's inheritors to actor-network theory (ANT). A strong theoretical background to the lexical study of scientific text was a feature of the Anglo-French "sociology of translation" in the eighties, with an influence of French philosophers such as Michel Serres. Forerunner of the ANT, the sociology of translation (see, for example, Callon, Courtial, Turner, and Bauin, 1983) highlighted the intermeshing of cognitive aspects and social stakes of scientific writing and presented coword analysis as especially appropriate for emerging or controversial areas of science.

The formal (partial) analogy between words and citation networks has suggested the use, in parallel, of similar data

Received May 24, 2010; revised August 29, 2010; accepted August 30, 2010

^{© 2010} ASIS&T • Published online 2 November 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21440

analysis techniques: clustering, factor analysis with many variants, multidimensional scaling (MDS), and social network methods. Interestingly, in the domain of factor analysis, Benzecri, promoter of the correspondence analysis, paid great attention to linguistic applications (Benzecri, 1973). Factor analyses and MDS have pervaded the world of bibliometrics and information science, for example, in word mapping (e.g., Noyons et al., 2003). Progress in computing efficiency allows the large-scale application of new factor methods such as latent semantic analysis. Local approaches based on proximity measures (typically Salton and Jaccard) and threshold graphs were privileged by pioneers of both coword and citation analyses.

Some important trends were noticed during the decade: the generalization of applications of the social networks theory, the multi-network approaches, and the need for mapping the Web. These trends were boosted by progress in computer efficiency. The Clever/Google revolution, partly building on earlier bibliometric works (above-mentioned "influence measures"), is, in turn, fed back to bibliometric measures. Science mapping has been a fairly active field over the decade (Noyons & van Raan, 1998) and powerful applications that combine large-scale and document-level have appeared (Börner, Chen, & Boyack, 2003; Chen, 2003; Boyack, Klavans, & Börner, 2005). For the specific application to thematic maps based on dynamic dating of research fronts, we watch the progress from crude circular projections on small fields (e.g., Zitt & Bassecoulard, 1994) to animated spherical projections that encompass all science at the document or the journal level (Boyack, 2004).

Our focus here is on hybridization/comparison of methods that exploit several bibliometric networks such as citation, words, nomenclature classes (e.g., coclassification), and authors (e.g., coauthorship, coactivity publications-patents, affiliation trajectories). We will limit ourselves to the first two, probably the most precise when targeting thematic mapping. Conducting comparisons between the two approaches, on a large-scale at least demands large computing resources, which perhaps explains why the hybridization of citation and words was not studied much in classical bibliometric literature. Achievements in this area included ISI developments on the use of citation linkages to generate Keywords-Plus in citation indexes (Garfield & Sher, 1993). Light forms of hybridization, such as "citations in context" (Small & Greenlee, 1980; Small, 1986) are now common in online databases (a recent contribution is by Callahan, Hockema, and Eysenbach, 2010). On a smaller scale, practitioners of citation mapping mobilize basic text statistics measures to give names to citation-generated clusters. An early example of cluster enrichment that combines both approaches is found in Braam, Moed, and van Raan (1991). Leydesdorff (2004) carried out another interesting test on the relationship between titles of patents and titles of the articles they cite. In the quasi-citation (or "sitation") world of hyperlinks, Googlelike algorithms also mix textual analysis and hyperlinks. Zitt and Bassecoulard (1996) made use of lexical validation for cocitation clusters. Later, these authors proposed a hybrid

process to delineate scientific fields, where citation flows are used to expand/smooth a set of articles built by a lexical query (Zitt & Bassecoulard, 2006; Laurens, Zitt, & Bassecoulard, 2010).

Hybrid approaches are also being applied to clustering processes. van den Besselaar and Heimeriks (2006) have proposed complete hybridization, which builds on formal similarities of citation and vocabulary, by mixing words and reference as pure tokens. Janssens, Glänzel, and De Moor (2008), after efficient neutralization of distributional discrepancies by statistical normalization, also considered citations and lexical tokens as "miscible," allowing the building of proximities, ranging from 0% to 100% citation content (and conversely for words); other clustering applications are announced by the same team (Liu, Yu, Janssens, Glänzel, Moreau, & De Moor, 2010). These proposals express a strong "informetric" posture. Endorsing a perhaps more traditional view, we do not consider words and citation as informetric tokens substitutable in various proportions, but rather as different expressions of the communities' life, which can converge, diverge, or complement each other. We will investigate a comparison of the two approaches that are based on a fairly simple tool, the table of clusters overlaps, revealing a rich basis for generating hybrid pseudo-maps. The focus in this work is on methodological aspects. For that reason, we reemployed a dataset on nanoscience, used in previous steps of the exploration.¹ We first discuss context, then the methods and results, followed by conclusions.

Context: Citations and Terms, Analogies and Differences

Author's choice. The "social act" of choosing references, on the one hand, and choosing terms, on the other hand, is complex, involving individual and collective representations. Many works have studied the scientific rhetoric and the citing behavior, for example, Latour (1987), Cozzens (1989), Luukkonnen (1997), Wouters (1997), Cronin (1984), Leydesdorff (1998), White (2001), and Aksnes and Rip (2009), but information scientists and bibliometricians do not necessarily claim a particular theoretical inscription, a point made, for example, by Garfield (1998). Obvious differences exist between words and cited references at the paper level, in the way they address fundamental aspects such as communication of the research question and results. The research question is clearly more apparent from title words than from a list of references. The combination of knowledge flows for achieving the research is described more directly by references, however, in a mostly imprecise manner as stressed, for example, by Leydesdorff (1998). The insertion of the community and identity claims is described again more directly by references but also possibly by the use of signal terms or combinations of terms.

¹Preliminary results were presented at the 10th Science and Technology Indicators Conference (Zitt, Lelu, & Bassecoulard, 2008).

Language markers describe both the research question and its sociological context. Relativist sociologists have argued how the scientists' rhetoric reflected their interests and positions with the social networks, especially in "hot science" (leading edge and controversial topics). Some terms act as interest markers, in a nonseparable mix of social and cognitive value. They can also act as beacons that are likely to attract attention, persuade the readers, and mark a hot topic and its potential for attracting funds (e.g., today "sustainable growth"; Callon, Courtial, Turner, & Bauin, 1983; Law & Williams, 1982). Such signals in natural language are probably clearer to policy makers than the equivalent bibliographic references to a recent path-breaking work that could be expected to play the same role. One of the main reasons why words look more adapted to "science alive" is not because citations cannot describe recent science but rather because words are universal. Not confined to scientific and kindred databases, linguistic markers are more able to reflect the interweaving of scientific, social, and political contexts, typical of emerging and controversial areas-at the expense of some "translation" effort.

The selection of references and the phrasing of arguments engage different processes of selection. (e.g., White, 2001). On their side, citation linkages ("acknowledged knowledge") appear as tracers of intellectual connections, even though only a part of them can easily be interpreted as a Mertonian acknowledgement of intellectual debts. The list of references, including those with a non-Mertonian slant, usually contains direct information about the community of research or at least the "legitimate repertoire" of the area (Rip 1988). The research background, mixed with identity claims and social dependences, can be directly read from the organization of bibliographic references, while words associations, even in full text, do not usually make the knowledge flows apparent.

Albeit a source of controversy in the early applications of cocitation methods (Mullins, Hargens, Hecht, & Kick, 1977), the capability to reflect scientific micro-communities is usually held as stronger for citation linkages than coword associations. A typical clue of the divergence of two communities that are working on the same theme and using a similar vocabulary is the absence of a common legitimate repertoire of references, which can reveal strong national imprint, sometimes autarkic behavior, adherence to strict schools of thought. In such cases, citation clustering will tend to see two communities, and word clustering a single one. Violations of the homogeneity hypotheses (open space of science) may produce divergences between citation and word mapping. Sometimes, the divergence of communities includes the usage or coining of different terminology, and the two mapping methods will detect the split. The opposite configuration, when the cited repertoire is almost identical but where sub-communities diverge in their terminology, can happen, for example, in applied fields that heavily refer to a common theoretic or methodological substrate (see also "scientific mix" below).

Emulating citations for delineating communities by wordbased analysis demands sophisticated semantic analysis and social studies of signal words and jargons. Another point stressed by the specialists of cocitation is the fact that weaklinkages à la Granovetter (revisited 1983) are far more reliable than their counterparts in word networks, where weak ties are often part of the noise. Last, but not least, instead of a list of references unique at the article level, several texts/lists are available on the lexical side: list of controlled key-words and lists from natural language fields—title, abstract, full text. Moreover, for coword analyses, for example, a "lexical unit" (sentences, paragraphs) must be defined when cooccurrence are calculated on long fields—not mentioning the variants generated by the terms unification-disambiguation processes.

The aggregate level: Combinations and mapping. Formal analogies and differences of citation and words distributions and networks have been sketched in many works (Egghe, 2000; Prime, Bassecoulard, & Zitt, 2002; Zitt & Bassecoulard, 2006). By and large, the underlying distributions are quasi-hyperbolic, with typically more concentration for words. In this respect, citation distributions are less skewed, suggesting a higher degree of complexity (richer "vocabulary"), with parameters also depending on the time frame (citation window). Despite the choice of metrics or similarity measures not favoring high-frequency terms, some lexical groupings are forced by sets of generic terms, where subsets are better discriminated by citation methods. But, a general rule can hardly be proposed: Bandwagon effects in citation behavior may create quasi-cliques of cited items that hinder the discriminating power of citation-based clustering. The inflation of the number of references in authors' practice, which is a long-term trend, may also bring noise to citation clustering. Another phenomenon, typical of the less dense areas in the networks, is the sensitivity of clustering/factor methods in those cases, likely to create (spurious) divergences (mostly) between the two approaches. The discriminating power of the methods can also vary with the scale of observation. A part of the statistical effects can be reduced by a proper normalization of proximity measures proposed by Janssens et al. (op. cit).

Basically, the statistical techniques applicable to measures of token associations (coword, cocitation) and of articles associations (lexical coupling, bibliographic coupling), or combined measures, are similar, although the intellectual habits have sometimes associated preferential data analysis techniques to each network. A complete protocol, say, for citation, may associate a work on structuring items, on the one hand, and citing articles, on the other hand. The classical ISI cocitation framework started with cocitation clusters ("cores") and then assigned citing articles of the current literature ("research fronts"). Similarly, one can start with bibliographic coupling and possibly go backwards towards the cocitation associations, to access the intellectual structure of the theme. Path-breaking works in bibliographic coupling and cocitation mapping have been recalled in introduction. Additional features on cocitation may be found in, for example, Zitt and Bassecoulard (1996) for applications to general clustering rather than leading edge, Chen (1999, 2003) for high-performance methods, and Glänzel and Czerwon (1996) for bibliographic coupling. Mapping by lexical techniques was studied in great detail by sociologists in the "translation" school (Callon et al., op. cit), especially the coword side. In the following, we will limit ourselves to bibliographic and lexical coupling.

"Just-in-time" and other time-features. The time qualification of items and linkages is a basic difference between the two approaches, because articles bear a publication date, while words are not primarily dated. Then, the citation relation between articles (A cites B) is diachronic whether no similar relation exists for words, at least in a direct form (time series of their usage or their combinations can be analyzed). However, the diachronicity of citation disappears as a primary feature in the matrix multiplication, leading from the original matrix [article × cited references] to a bibliographic coupling matrix [article × article] or a cocitation matrix [cited references × cited references], remembering that a complete cocitation run involves a reassignment of citing articles ("research front") to the cited core established from the latter matrix. There is no fundamental discrepancy in the "just-in-time" capability of citation methods compared with a similar word-based technique. The criticism of a lack of immediacy of citation techniques bore on early applications of cocitation at ISI and Drexel University, which deliberately focused on the leading edge of research (strong signals) at the expense of emerging topics. In this particular case, the immediacy is naturally hindered because of the citation lag to get the score of citation or cocitations used as a threshold. Thresholds may also be necessary for processing time constraints, but this is true as well for techniques following the lexical way.

If we now take threshold-free applications of bibliographic coupling technique or even cocitation, the issue of immediacy is alleviated. For example, in a typical application of bibliographic coupling, the matrix of proximity [article \times article] of a particular citing year is known at the same time as it lexical coupling counterpart. The only lag is the delay of awareness by citing authors of a published material in citable form, common to all publication-based techniques.

Just-in-time is then not a distinct advantage of lexical techniques. In classical cocitation or coupling maps, the diachronicity of the underlying relations, if not directly exploited in building the structure of themes, is especially helpful to qualify the time features of clusters (Zitt & Bassecoulard op. cit. 1994), for example, crossed immediacy of the intellectual base with short-term growth indicator. The dynamics of research fronts have often been studied; for example, Morris, Yen, Wu, and Asnake (2003) proposed a dynamic visualization of research fronts and Chen (2003) replaced citation dynamics in the general context of scientific networks visualization. Lexical changes have attracted attention for a long time (among others, Zitt, 1991; Polanco, Grivel, & Royauté, 1995). A variety of techniques is used for dynamic analysis of fields (see, for example, Noyons & van Raan, 1998).

A common feature: Holism. A deep but misleading analogy exists between lexical and bibliographic coupling. Both stress the synthetic vantage point of classical word and citation approaches: A single measure of proximity reflects the kinship of two articles on the basis of shared references or shared terms, with various normalizations. This global and semantic-free rendering of the "scientific mix" dilutes the particular features carried by specific groups of references or terms, for example, their slant towards particular dimensions of research: theory, models, methodology, instruments, observations, tests, applications, or towards particular research objects, products, or scientific areas. Undisputable advantages of the global view are simplicity and independence from a priori semantic classifications. There are also counterparts to this holistic approach: the black-box effect in the way intellectual structure is uncovered, in contrast with in-depth descriptions of semantic networks, and the fact that proximity measures are uneasy to interpret.

When two documents are close to each other on a word coupling index, we do not know on which semantic grounds they are linked: Do they share the theoretical point of view? The instruments? The methodology? The balance between the various components of the "scientific mix" is out of control. Stating that the relative importance of the various dimensions of this scientific mix are reflected in the balance of terms, or cited items, thus contributing to the convergence of lexical and citation methods, is a heroic hypothesis. Yet, to a large extent, similarity of scientific mixes is a condition of proximity. Let us assume that the vocabularies relative to the K semantic dimensions do not overlap, with K = 5. Say the dimensions are as follows: field features, theoretical context and hypotheses, methodology and instruments, empirical results, applications or objects. Take two articles *i* and *j* and suppose, for convenience, lists of equal lengths L for *i* and *j*. The number of words marking the dimension k = 1..K in article *i* is noted n(k, i). Let Min(k, i, j) = Min(n(k, i), n(k, j)). For example, if i contains 5 words on the "theoretical context and hypothesis" and j contains 12 words on this dimension, Min(k, i, j) = 5 and only those words will be available for matching *i* and *j* on this dimension, whatever the particular coupling formula. The active part of the total list L on which matching i-j in usual coupling formula can take place is $\sum_{k} Min(k, i, j) \leq L$. The fraction $L-\sum_{k} Min(k, i, j)$ is not available for matching. This expresses a quite simple notion that under the hypothesis of separation of dimensions, the global term sharing measure, dimension by dimension, is bounded by the alignment of dimensions in terms of word abundance. The precise effect depends of course on the weighting system used in particular coupling formula. In summary, a high proximity between articles requires both a similar balance of semantic dimensions and an alignment within the (main) dimensions.

Although not as clear-cut—any cited article already comprising a mix of dimensions—, most cited references can chiefly be classified using similar categories: theoretical background, methodology/instruments, field of application, usually referred to in specialized sections of article.

Then, proximity calculations, in both universes, depend on the balances of dimensions in the "scientific mix," with no guarantee that word and citations will establish, in a given article, the same balance. In a dynamic perspective, the variations of the respective importance of the dimensions could be witnessed in the evolving structure of terms and cited references. In growing areas, both terms and references may exhibit a similar tendency to reflect the differentiation of topics. Conversely, if one dimension of the scientific mix is getting stabilized-say, methodology, which is getting standardized in a particular area-, then one expects that number of both the terms and the references relative to methodology will decline. The same is true for merges of topics, well captured by cocitations and possibly echoed by fusion of vocabulary. However, the possibly different rhythm of compliance of terms and citations, with the changing balance of interests in life cycles of themes, is not easy to capture and a convergence in this respect cannot be taken for granted.

Therefore, in dynamic as well as static terms, the analogy of approaches, holistic in both cases, may obscure systematic differences as to how semantic dimensions are covered, according to balances that moreover can differ across scientific areas. One might risk a few hypotheses:

- The semantic balance of references probably exhibits more degrees of freedom than its linguistic counterpart. If so, then the references tend to inflate the main dimensions of the articles' argument, more than the word abundance regulated by the rhetoric structure of the papers sections. In such cases, this main dimension carries the classifying power of citations, for either unification or discrimination.
- According to a widely accepted idea, popularized by Narin, Pinski, and Gee (1976) in their studies of scientific journals, citations, all things equal, privilege up-stream idea, for example, cognitive aspects or key methodology rather than applications. We could expect, for example, that applied works will pay tribute to theoretical papers more than the reverse.
- When all dimensions are strongly dependent of each other, for example, in emerging areas where theories, methods, and objects are specific, citations and words can converge, with respects to statistical constraints on distributions.

In absence of general rules, tools for comparison and combination of the two approaches are appealing. An extreme convergence would suggest that both ways are substitutable and robust, a moderate convergence that they are complementary, and a low convergence could inspire suspicion about the reliability of mapping based on bibliometrics. Depending on areas, local convergence suggests to build mixed superclusters where "strong forms," apparent from both methods, are found and to be more cautious in the interpretation of diverging schemes.

Methods

Sources and Data

The original data come from the *Web of Science (WoS)* offline database at Observatoire des Sciences et Techniques

(OST, Paris), implemented on contract with Thomson Reuters. The dataset was collected by a hybrid method that involves a lexical query, complemented by a citationbased extension process, detailed in the above-quoted articles (Zitt & Bassecoulard, 2006; Laurens et al., 2010). The set used here represents approximately 168,000 articles for the years 1999–2003. This relatively ancient set, especially in the context of a rapidly evolving field, is kept, for the sake of continuity with previous experiments, because the present article is focused on methodology and not on the current development of this field.

Clustering Stage

Before addressing the core of the comparison, let us summarize the methods and techniques used in this particular experiment. It should be stressed that the further comparison stages could be carried out on the outcomes of any clustering technique with good properties as well, but the delineation of clusters is, of course, sensitive, to a certain extent, to the particular technique picked.

Citation-based clustering. The general principle consists in connecting articles through their proximity, in a "bibliographic coupling" rationale rather than through an intermediate stage of cocitation. Background of citation-based clustering, by either cocitation or coupling, has been recalled above.

A variety of data analysis methods are available for clustering. The method adopted here is an axial k-means technique (AKM or KMA), developed by one of us and implemented in the software Neuronav (Diatopie). It is a variant of k-means partitioning methods (Lloyd, 1982, building on an Bells's lab unpublished paper by the same author, 1957; McQueen, 1967), which brings significant improvements (Lelu, 1994, 2008), giving rise to concept-vectors and document orthogonal and oblique projections on each cluster axoïd. It is related to the spherical k-means family (Domengès & Volle, 1979), with a robust theoretical basis. AKM metrics is based on the Hellinger distance: Any raw data-vector **x** with components x_i , i = 1, ... I, is turned into a normalized **z** one, the components of which write:

$$z_i = (x_i/x_i)^{1/2}$$
 where $x_i = \sum_i x_i$

In this spherical dataspace, comparing distances amounts to compare vector angles:

 $d(\mathbf{z}_1, \mathbf{z}_2)^2 = 2(1 - \cos(\mathbf{z}_1, \mathbf{z}_2))$, where $\cos(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^t \mathbf{z}_2$ and \mathbf{z}^t stands for the transpose of column vector \mathbf{z} .

This distance satisfies the property of distributional equivalence. Like other k-means methods, the variant AKM is quite powerful and can process relatively large sets of data, but it is sensitive to the initial draw of centroids. However, this weakness is alleviated here, because of the low number of clusters targeted (50), compared with the ~100 000 documents clustered. Bassecoulard, Lelu, and Zitt (2007) have reported the methodology, some features of clusters and the thematic map of the dataset under study. We prolonged this analysis by a clustering process with the same grain (M = 50), built on word-base similarity. Note that other methods might have been used instead for clustering or partition, involved in the present comparison, provided the results are of good quality.

Word-base clustering. The term extraction was carried out under the software Neuronav by Diatopie (S. Aubin) and some complements under SAS software by Lereco. Morphosynctactic labeling and stemming are conducted on a dictionary of forms, and two heuristics: In case of ambiguity, the most frequent label is used; then, the forms that are absent in the dictionary are considered as nouns. Multiterms are obtained from syntactic motifs, under constraint of noninclusion in specific lists (e.g., to avoid *"remarkable results"* or *"improved method"*). The process privileges high precision, around 95% versus a careful manual indexing. The recall is lower (ca. 50%), not a serious shortcoming given the high degree of redundancy of terms within an abstract.

For the clustering stage, the same principle as for citationbased clustering was adopted, namely, a direct classification of articles by lexical coupling rather than through an intermediary coword or cocitation stage. The same options used in the citation round were also applied to the metrics: the classification method, the number of clusters set to the same value, M = 50, to facilitate comparison. In the following, ccluster designates clusters of articles from citation analysis (bibliographic coupling) and w-cluster designates a cluster of articles from word analysis (lexical coupling). The original maps of c-clusters and w-clusters are shown in Figure 2A and Figure 2B in the Results section. For simplicity reasons, c-clusters as well as w-clusters result from a strict partition (no overlaps), but this constraint can be lifted in many methods, including AKM. The overlaps we are interested in are the intersection between c-clusters and w-clusters.

It should be recalled that the retrieval of literature by a hybrid method—as in the data collection for the present experiment—alleviates a possible bias of classical queries methods. A query system on large fields is often a union of elementary queries that are expected to cover particular subfields/areas, and then to predetermine, to a certain extent, the structure that will emerge from clustering. If our initial query system was indeed based on unions, then the sequence lexical query-citation expansion, mentioned above, smoothes the boundaries of the field and then reduces the structuring bias of initial queries.

The Table of Clusters Overlaps as a Tool for Comparison

Our aim is to compare the outputs of clustering processes, conducted on citation and word proximity, with the same set of scientific publications. The main difference here does not lie in methods/options of metrics used—identical by construction as seen above—but in the nature of the original linkage used for structuring, bibliographic coupling, and lexical coupling.

For memory's sake, many kinds of data analysis can explore the convergence of partitions or classification trees, including full-range analyses that consider the original similarities among the N documents. Examples of complete analysis involve a pair comparison of distances between articles in either approach or a comparison of classification trees (Guénoche & Garetta, 2002), yielded by lexical coupling and bibliographic coupling on this full set of documents. One can, for example, compare the fate of any couple of documents (k, l) in two classifications. Such analyses aim solely at a comparison of approaches, not a combination. The present work is rather a cross map of literature overlaps that are generated by combining lexical and citation clustering. Various applications of the crossmaps to scientific literature were sketched by Morris and Yen (2004).

Here, we start from the cluster level (M = 50 clusters in each approach). We assume c-clusters in rows and w-clusters in columns. In the contingency table of dimension $M \times M$, each cell *ij* contains the intersection of c-cluster *i* and wcluster *j* (number of articles); derived tables may be with normalized intersections, either similarity (Jaccard, Ochiai) or probability measures. For any type of content, direct or normalized measures, a color scale may help to visualize the strength of intersections; here, we will use shades of grey.

The table of intersections proves rich in information. It enables us to study the convergence of the citation and word approaches in a simple way, at first, by independence measures like chi-square and derived indexes (Phi). For example, if all cells are zero except M heavy cells, none of these with the same i or the same j, then the breakdowns are identical. Although this cell-level measure provides a first piece of information, rearranging rows and columns, by proximity, could give us a structured landscape to watch at several levels and give us a more complete insight on the overall convergence of c-classifications and w-classifications.

Hybrid profile distances on cc and ww. There are two basic options for computing proximity between rows, and between columns. We use either the sole information contained in the intersection table or additional information, namely, the original distances between c-clusters, on the one hand, and w-clusters, on the other hand, calculated at the initial clustering stage in either universes. We will go back hereafter to the latter process. Let us now focus on the former, where proximity between w-clusters (here rows) is assessed by their profile towards c-clusters: so to say, citation structure is used to group w-clusters, and conversely. For example, two c-clusters i = 1, 5 that both overlap, on the same profile, with four w-clusters j = 4, 7, 24, 39 will be deemed very close. Two c-clusters that overlap with completely different sets of w-clusters will exhibit a zero similarity. As the table of intersections is a contingency table from two partitions of the same set, correspondence analysis (CA; Benzecri, 1973), which relies on the good properties of the chi-square distance, may be used. The alignment of axes in both spaces is a central feature of the methods, allowing a projection of the two types of items, but the proximity between projections of individuals (say, w-clusters in rows) with variables (c-clusters in

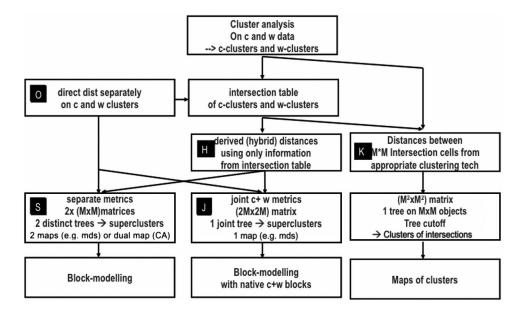


FIG. 1. Five sequences. O–S (original dist+separate metrics); H–J (hybrid dist+joint metrics), O–J (original dist+joint metrics); H–S (hybrid dist+separate metrics); K joint clustering: direct clustering of intersections.

columns) is not directly interpretable. In those representations, the visual perception of the global agreement between c-clusters and w-clusters (see Figure 4 in the results section), after their projected distance, is poor. Some extensions of CA are based on the Hellinger's distance (Rao, 1995).

Matrix Reordering

Matrix reordering is a long-lasting and still very active research field, under varied names: seriation, biclustering, and heat-maps. Bertin was a pioneer of *seriation* studies (Bertin, 1967), who inspired many subsequent works, e.g., Wilkinson (1979) and Marcotorchino (1987). The latter coined the term "block-seriation" for his linear optimization process, while reordering is often based on a hierarchical clustering on rows/columns or a factor analysis. The term biclustering stems from Mirkin (1996), with Hartigan (1972) as a forerunner. Heat maps (Eisen, Spellman, Brown, & Botstein, 1998) appeared as a special application of reordering to gene expression data.

Combined metrics. For the present application, we tried to investigate joint metrics, leading to hybrid super-clusters and unambiguous blocks. In this approach, the distance between a c-cluster and a w-cluster (noted c-w) can be expressed as a function of c-c and/or w-w distances. We reached acceptable results in practice, as shown in the following. The detail of calculation is given in Appendix A. An alternative technique that takes advantage of AKM properties, and is likely to be more rigorous and powerful, would present some advantages (see below).

Figure 2 summarizes some possible protocols, according to the choice (a) of metrics: original distance in the two networks ["O"] versus hybrid distance from the intersection table ["H"] and (b) of treatment (separate treatment of cclusters and w-clusters ["S"] versus joint treatment using a c-w distance ["J"]. This yields four typical sequences: O–S (original distance + separate treatment); H–J (hybrid distance + joint treatment); O–J (original distance + joint treatment); and H–S (hybrid distance + separate treatment). Another pathway of joint clustering, K, based on direct classification of intersections, is mentioned but not applied in the present work.

Assuming that both word-based clustering and citationbased clustering offer consistent views of the same universe, then we wish to compare and possibly combine the words and citation perspectives. Among the four ways depicted in Figure 2, "O–S" is classical in block-modeling. Here, we investigated the (O–J) and (H–J) ways with particular attention on to the latter. The fact that many ways of reordering give usable but perfectible results (the problem has an exponential complexity) lead many software packages to allow manual optimization, in the spirit of Bertin's recommendations (Bertin, 1977). Intentionally, for comparison purposes, we made no attempt here to optimize visualization manually.

Reordering and Super-Clusters

Basing ourselves on the joint metrics described in Appendix A, we reordered rows and columns in two stages:

 Hierarchical clustering of c-clusters (here in columns) together with w-clusters (rows) into super-clusters. We chose the group average distance, generally considered robust, to generate the super-clusters at all levels. A supercluster typically entails c-clusters and w-clusters. Then, on the reordered matrix, a super-cluster typically embodies cells within two orthogonal stripes, one of contiguous rows

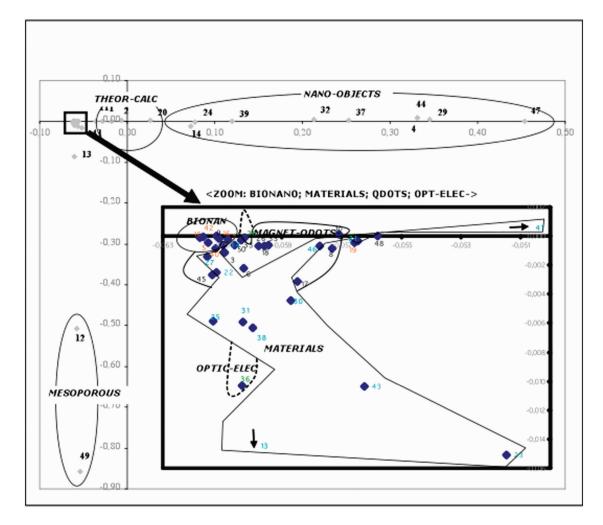


FIG. 2A. Original correspondence analysis (CA) map of citation-based clusters (1st plane of CA). CA is typically based on a table of contingency categories-criteria. Here, the intersection table of w-clusters and c-clusters is used as the contingency table fed to CA. Despite the dual nature of CA, the proximity between projection of individuals (w-clusters in rows) and projections of variables (c-clusters in columns) is not directly interpretable. Grouping in macrothemes is apparent from the display.

and one of contiguous columns, where the crossing area represents the block. Two super-clusters at a given cut-off level tend to be overlapping in terms of cells.² Super-clusters with complete discrimination are those where crossing stripes are empty except in their common block. There is a perfect correspondence in this case, at the level considered, between w-grouping and c-grouping.

2. To finalize the reordering on c-clusters and w-clusters, a good representation among all those compatible with the hierarchical clustering found, is sought. We used a unidimensional MDS, under constraint of the hierarchical tree: An average of components (leaf) positions on the axis is calculated for clusters at all levels in the tree, and at each dichotomy, starting with the root level, and the two

branches are ranked depending on their average position.³ This process aims only at bringing an acceptable organization of blocks. The super-clusters and related blocks built on the joint metrics are unambiguously determined and do not need an external decision to choose among many combinations of w-clusters and c-clusters, in contrast with sequences O–S and H–S.

The outcome of a matrix reordering, for contingency tables, may be called a "pseudo-map," where both the size of objects, cells or blocks, and their visual distances are interpretable. Each cell visualizes the size (absolute or relative) of the intersection set between a c-cluster and a w-cluster, and the number of cells that separates two such cells, after reordering, row-wise and column-wise (city block distance) can be loosely interpreted as an image of their dissimilarity.

²If necessary a nonoverlapping scheme may be derived from this representation. The outcome is a nonoverlapping and nondichotomic classification of cells/blocks of successive levels. For example, at a given level, a block will aggregate two blocks of the last level, plus the two pseudo-blocks of previously unassigned cells. Note that the process is different from a treatment of cells as individual objects, submitted to a classification process (pathway K on Figure 2).

³In this particular experiment, the constraint of the hierarchical tree was ignored for the 10 first levels (close to the root) the ranking of which is left to the MDS. Groupings at this level involve quite weak links. This option can be cancelled if one wishes a complete hierarchy.

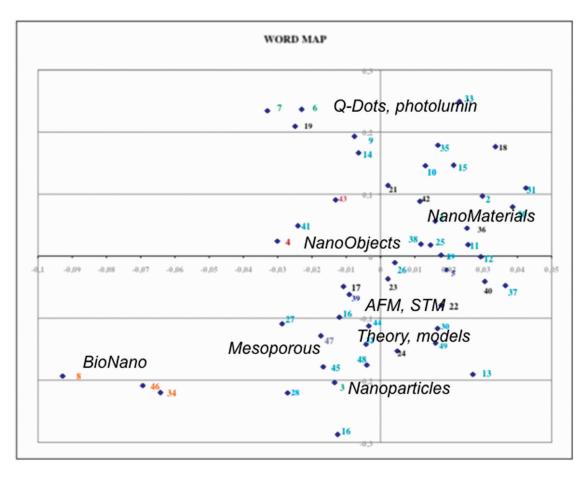


FIG. 2B. Original correspondence analysis (CA) map of word-based clusters (1st plane of CA). Large themes are less clearly delineated than above and borders are not represented.

With the help of this visualization, let us now focus on the blocks defined by those cells common to the c-clusters and the w-clusters that compose the super-cluster. A block appears a "strong form" of hybrid nature, retrieved by both citation and lexical analysis, if it concentrates most of the heavy cells of the corresponding rows and columns. Blocklevel indexes of retrieval can be coined—a simple one being the geometric average of the proportion of recovered articles in both directions, the Ochiai index, already employed for the color coding of individual cells in Figure 5 in the Results section.

Original distances, on which the initial 50 clusters on each side were built, may be used instead of hybrid distances. A distinct advantage when one works on original distances rather than hybrid profiles distances is that the logic of separation of c and w pathways is maintained. In this respect, the path O–S, classical in block modeling, is the most orthodox if we adopt a strict posture of discrimination between word and citation approaches throughout the process. The weak point is that it does not yield a unique hierarchy of super-clusters/blocks but many possible combinations. More suited, perhaps, to comparison purpose, it is not so well adapted to combination of methods.

In the present exploration, we eventually limited ourselves to the sequences H–J and O–J. Hybrid profile distances group c-clusters on their intersection profile with w-clusters, hence, on a lexical logic and, conversely, for w-clusters. Hybrid profiles' distances are likely to be efficient for reordering the intersection matrix and creating explicit super-clusters, but the interpretation of combinations reached is sometimes delicate. Outcomes of the sequence O–J are also mentioned, O–J keeps the separation of techniques a bit longer. A practical shortcoming of the latter sequence is the fairly large distributional differences in original distances of c-clusters, on the one hand, and w-clusters, on the other hand.

A more satisfactory way to deal with original distances is the spherical factor analysis, taking advantage of the AKM clustering. It is possible to represent the cluster axoïds directly in their *common* document space, as well as in their distinct attribute space—thus, doing without the intersection matrix—, a process that will be developed elsewhere.

An alternative exploitation of the table of overlaps could take advantage of this method: It would start from the individual cells—here, 2,500 cells—instead of rows/columns as studied here (pathway K on Figure 2). Two-stage or "hybrid" hierarchical process, where strong forms (stable sets) are submitted to a final clustering stage, is a classical practice in cluster analysis of large datasets. Strong forms (Diday, 1971) may result from a variety of treatments including repeated trials (k-means or moving clouds for example) or crossing of outcomes from different techniques. In our context, intersections could be considered as elementary strong forms, needing the construction of a distance between individual cells or blocks, that could be provided by the spherical factor analysis just mentioned.

Convergence of approaches

The degree of independence of the lexical and citation breakdowns, regardless of the arrangement of rows and columns, can be measured by classical indexes typically of the chi-square family, for example, the Phi-index and the contingency coefficient C. Visually, with an appropriate grey scale, the dependence will be associated with high contrasts between, say, "black cells" (high absolute or relative overlap) and "white cells" (no or negligible overlap). An identity between c-clustering and w-clustering would result in M black cells and the rest white; independence would result in a quasi-uniform grey rendering. Global independence measures are limited to the initial cut-off level, the M-level clusters, and say nothing about the higher level structures, represented by the super-clusters.

The reordering helps to visualize another phenomenon, inherited from the respective properties of c-clusters and wclusters size distribution. Let us assume an efficient reordering. Intuitively, a strong convergence of c-classifications and w-classifications will result in a quasi-diagonal scheme. On the $M \times M$ table built here, an exact diagonal will be reached only if the size distribution of c-clusters and w-clusters is alike. In case of convergence with distributional differences, the "black cells" will follow a bent or "snaky" curve instead of an alignment on the diagonal. If, for example, we imagine a quite skew distribution of c-clusters size (say, a big cluster and many little ones), while the w-clusters size is more evenly distributed, then we will get a long vertical rectangle in some place, compensated by sequences of small horizontal ones. The total number of c-clusters and w-clusters being identical in the present study, if citations prove more discriminating in a particular area (say, nanotubes), then it will tend to be compensated in another area. In the following, the convergence of approaches will be tested after the local density in the neighborhood of cells with strong overlap (dark cells), density that reaches high values for continuous agglomerations, typically along a diagonal scheme.

One clue of approximate convergence is the accumulation of grey-embedded blocks, with mitigation when the scale enlarges, along a dark curve (approximately) diagonal. Scattered structures, in contrast, suggest a divergence of logic, keeping in mind the limitations of the block modeling practiced. Globally, the landscape generated and the system of grids that could be materialized on an appropriate display exhibit some analogy with the medieval sea charts, "portolans," as seen in the Results section.

Results

General Outcomes of Clustering, Citations Versus Lexical

Let us first have a look at the two basic maps of 50 clusters, established by bibliographic coupling and lexical coupling (Figures 2A–B). The display is based on a principal component analysis of the AKM clusters in their descriptor space. These two-dimensional (2-D) projections look quite different. The general organization is more apparent on the citation-based classification: While w-clusters are scattered across the diagram, c-clusters are structured in macrogroups, contributing heavily to one factor, easy to identify—nanotubes, bionano, and theoretical aspects.

The size distribution of clusters is shown Figure 3. In contrast with the native distribution of cited items or words, it reveals more skewed for c-clusters than for w-clusters. The citation-based method produces a few very big clusters and a majority of comparatively small clusters. This is not paradoxical: In many cases, the abundance of citation in the medium-low part or low tail of citation distribution is able to identify small research areas and communities (an example is the discrimination of several fronts within nanotubes). This is compensated (with the rule of 50 clusters on both sides) by a few very big clusters, such as gene delivery, 2-D and 3-D arrangements and self-assembly, quasi-crystals, and magneto-resistance. Citation networks can produce, at a given threshold, quasi-cliques in a quite large range of size, especially for integrated communities basing on a strong common repertoire.

Respective distributional features of citation coupling or word coupling—and the particular classification method used, here AKM—influence the final rendering of the reordered matrix, especially the diagonal area.

A Classical Display: The Field Landscape/CA Map

The map of the first factors plane is shown Figure 4. As previously mentioned, the distance between c-clusters, or between w-clusters, is directly interpretable with respect to the stress of projection, while the distance between a c-cluster and a w-cluster is not. The general structure shows a three-branched star, around a heavy kernel, evoking the original c-map of Figure 2A. Although giving an appealing view of the general landscape, the CA display alone is not really appropriate to compare and combine c-clusters and w-clusters, a goal more efficiently achieved by elaborating on the intersection matrix.

Intersection Table

The general aspect of the full intersection table is shown (after the reordering of rows and columns) in Figure 5A, with

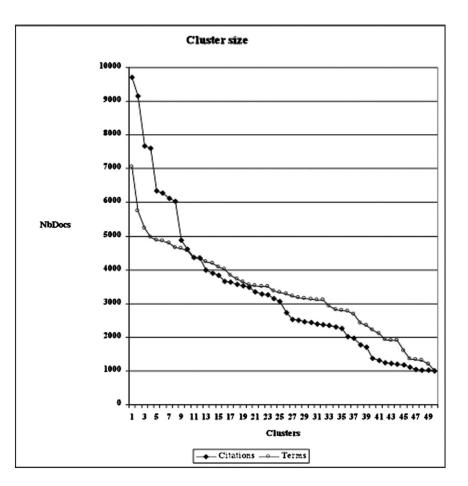


FIG. 3. Cluster size distribution. c-clusters and w-clusters are separately ranked by size.

grey shades based on the intensity mutual inclusion (Ochiai Index). Figures 5B and 5C are zooms on two parts of the figure, top-left and bottom-right of Figure 5A. Independently from the particular reordering:

The independence of breakdown at the cluster level was tested by usual measures on contingency tables $R \times C$, based on chi-square Q.

Phi Coef. = $\sqrt{(Q/n)}$ Phi = 1.96 (here max 7)

Contingency Coef.: $\sqrt{(Q/(Q+n))}$ CC = 0.89 (here max \cong 1)

In both cases, independence is strongly rejected p < 0.0001.

Beyond the global balance that is expected from the cross-tabulation, we observe contrasts in the distribution of concentration along rows and along columns. Scrutiny of Herfindahl and entropy-based coefficients lead to the same conclusion: c-clusters exhibit a more even distribution of concentration, w-clusters a broader range of concentration. Besides, if we sort clusters by diversity of their overlaps, the top c-clusters (Ti02surface/ particles/ films; nanobelts-ribbons; photonic/ colloidal crystals; molecular electronics-principles; fullerenes) reveal less diverse than the top w-clusters (X-ray, Molecular_dynamics/

Simulations, Magnetoresistance, and two clusters on modelling).

Reordering

The maps shown Figures 5 and 6 are intentionally left in the configurations that are produced automatically. Figure 5 details the outcome of the sequence H–J.

The diagonal area appears heavy, especially at the two ends: the blocks in the top-right area (the orientation is arbitrary) represent bionano, mesoporous structures, two blocks devoted to self-assemby, electrochemistry and films, AFM-STM, and a sub-block of nanomaterials, metal particles, and crystals; bottom-right blocks are the home of quantum dots and Kondo affects (bottom), the carbon nanotubes, alloys, adsorption-catalysis, and a sub-block of nanomaterials, surface, and carbon films. The central sub-block of nanomaterials, theoretical studies, is fuzzier. An organization of larger blocks is guessed at, but the concentration of grey cells decreases off the diagonal area. This general view brings evidence that the two c and w representations are fairly similar but not substitutable.

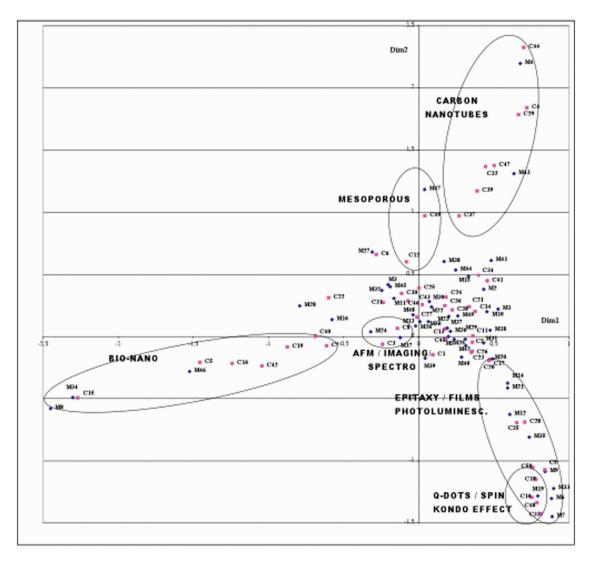


FIG. 4. Correspondence analysis map: c-clusters and w-clusters. Correspondence analysis typically exploits a table of contingency categories-criteria, namely, here the intersection map of w-clusters and c-clusters. Despite the dual nature of CA, the proximity between projection of individuals (w-clusters in rows) and projections of variables (c-clusters in columns) is not directly interpretable.

Figure 5C is a simple zoom on the bottom-right section of Figure 5A. Figure 5B zooms on the top-left, using a different display that conveys some visual analogy with early manuscript sea charts, used during centuries XII-XVI. The "portolans" showed shores and ports in some detail, as well as a web/grid of compass directions, radiating from several points. Here, lexical and citation clusters figure the ports on two mainland shores. Forming a roughly diagonal archipelago, a few directions starting from each island (the major intersections), are of particular interest: A multilevel grid connects ports in the shores to islands in the archipelago; the diagonal directions along the archipelago invite a visit along the universe's main dimension. Portolans had many limitations, due to data and the use of approximate planar representation before Mercator projection. Here, too, is the reduction of information drastic, in comparison, say, to CA rendering; but, the clear information on intersections made it useful for understanding a combined bibliometric vision of a field.

Using a large-scale visualization Figure 6 compares the outcome of H–J and O–J sequences obtained by the process (respectively top-left and bottom-left). It also displays a basic map obtained by the first factor of Correspondence Analysis (top-right), and a variant of the latter map, starting with the CA order and complemented by an automatic research of a maximum load for the diagonal (bottom-right). The top-left map presents a good concentration, avoiding most outsiders.

Structural convergence and goodness of rearrangement. We measured the level of grouping of high-density cells, regardless of the sinuous or curved shape of the diagonal path. Using a rule of thumb, we selected top-quantiles cells (respectively 5% and 8%) on the criterion of overlap and tested the number of other top-quantile cells and the average

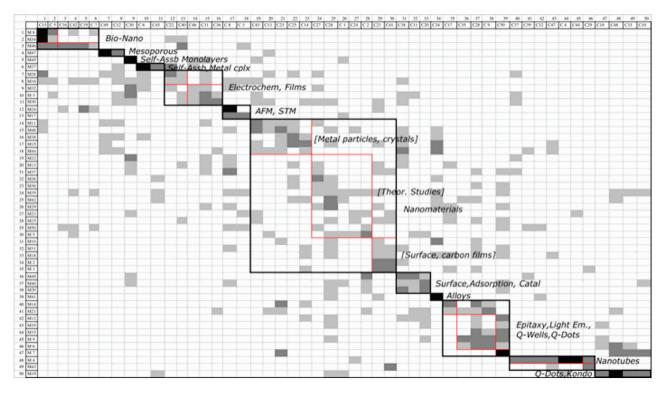


FIG. 5A. Reordered matrix: general: general appearance. Two large parts are detailed in Figures 5B and 5C.

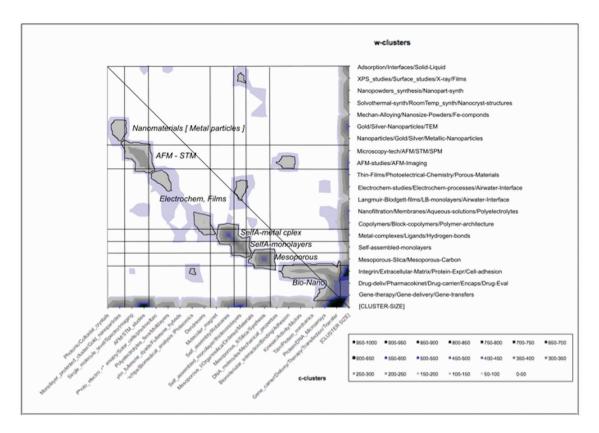


FIG. 5B. Reordered matrix—zoom on top-left part of 5A. The display is a surface graph, with the overlap index (Ochiai) as z-axis values, displaying the archipelago structure on a sort of Portulan chart. Shores are given an "altitude" proportional to the size of c-clusters or w-clusters (values not comparable to overlap value). On a table based on gross intersections rather than relative overlap measures (Ochiai), altitude of shores and islands may use the same scale. For convenience, the y-axis of 5A are reversed.

			TiO2_surface/Particles/Films/Catalysis	Density_functional_theory_studies/Self_diffusion/Modelling	Electronic_structures/Calculations	z Quasicrystals/Amorphous alloys	Poreus_săcon_properties	x Semiconductors/Ferromagnetism	Diodes/Optical_properties/Applications	# Band_Structure/Parameters/Gap	2D-3D_ordering'Self_assembledqQuantum_dots	5 Carbon_nanotubes_functionalization	Earbon_nanotubes/Hydrogen_storage	Carbon_nanotubes/Mechanical_properties	carbon_nanotubes/Production/Properties	E Single_walled_carbon_nanotubes/Electronic_properties	5 Single_walled_carbon_nanotubes/Synthesis/Raman/Arc_method	5 Carbon_nanotubes_growth	Spin_injection/Filter/Polarization/Transport	a Quantum_dost/Magnetic_fields/Coulom/Interactions	a Quantum_dots/Bits/Josephson	g Kondo_effect
Adsorption_analysis/Adsorption_desorption/Catalysis/Adsorption_methods			C38	CII	C39	C34	C17	C35	C28	C 9	C50	C37	C29	C32	C47	C4	C44	C29	C18	C48	C33	C10
STM_studies/		M49		-	-	Su	irfac	.e, A	dso	orpti	on.	Cat	alvs	is							-	
Surface_structures/Surface_grown_films/Surface_adsorption/LEED		M40	-			-	-	-	-			-	-	-							-	
Alloys/Grain_boundary/Microstructures/Metallic_alloys/Magnetic_properties	-	M20		_			Al	loys					-									
Photoluminescence/SiO2_substrates/Light_emission/Films	-	M41											-	-								_
Band gap/Electronic structures/Conduction band/Photoemission		M14			-				-													
Growth_behaviour/Epitaxial_growth/Films		M21				-																
Epitaxial_growth/Thin_films/RHEED_studies		M12													ight		issi	on,				
Molecular_beam_epitaxy_growth/Qdots_formation		M35 M33				-						Q-	wei	IS, (2-D	ots						
Molecular_beam_epitaxylEpitaxial_growth/GaAs		M9				-												-				
Photoluminescence/Optical_properties/Quantum_wells/Quantum_wires		Mo																				
Quantum_dots/Structural_properties/Optical_properties		M7				-		_														
Carbon_nanotubes/Single_wall_nanotubes /Synthesis/Growth/Characterization	-	M4																	Ca	rboi	0	
Carbon_nanotube_field_emission		M43																			ubes	s
Quantum_dots/Kondo_effect/Electron_transport/Quantum_wires	-	M19												C	-do	ts, k	Kon	do				

FIG. 5C. Reordered matrix-zoom on the bottom-right part of 5A. This zooming uses the same display as Figure 5A.

density in their neighborhood (contiguous or next contiguous cells). The performance is about three times the expectation (randomly arranged matrix). The density measure embodies the latent organization of data and the capability of the reordering to reveal this organization. There may be some systematic underestimation of the degree of organization of data if the reordering process falls far from the optimum: A better heuristic could probably enhance the visual perception of the diagonal/block structure.

Let us browse the variety of macrostructures uncovered.⁴ We focus on the outcome of hybrid profile distances (sequence H–J) and on occasion briefly mention the outcomes of the sequence O–J.

Converging areas. Strong cells are found in the intersection table, but no real one-to-one cluster correspondence. In many cases, a good agreement is reached at a more aggregate level, on either square or rectangular blocks. Let us first envision the blocks exhibiting square or quasi-square shapes. In the best cases, few external connections exist outside the block, row-wise and column-wise. A block that is not decomposable at a finer-grain level, displaying a quasi-uniform grey area, shows that terminology and citations propose different breakdowns of this block, as if the balances of the "scientific mix" discussed above were treated in a different way. At a

larger scale, the level of the block, both methods retrieve the theme.

Examples of quasi-square blocks are as follows:

Scanning tunneling microscopy (AFM; 1981) and atomic force microscopy (STM; 1986) are historical pillars of the domain, making it possible to image and manipulate atoms and nanoscale structures. Although transversal to many applications, the theme is the support of two self-standing clusters, in both c and w approaches. The c-clusters are AFM/STM studies and Single_molecule_level/Spectroscopy/Imaging. The wclusters are AFM studies/AFM imaging and microscopy techniques/AFM/ STM/ SPM. SPM denotes scanning probe microscopy, the branch of microscopy initiated by SPM and now encompassing many techniques. These structures are strongly overlapping, but the generic character of the technique is apparent from quite a few intersections off the diagonal. This is especially the case for the c-cluster AFM/STM studies, bigger and with a more scattered profile, especially intersections with the w-clusters of the electrochemistry and nanomaterials block. This is perhaps an illustration of a hypothesis stated in the context section, the strength of a pathbreaking methodology in constant development and applied in many areas.

In the sequence O–J, the block rightly captures another dedicated w-cluster, AFM_studies/Thin_Films/ Surface_studies (quite scatted in its c-profile) but loses homogeneity on the c-cluster side by accretion of two bio-oriented clusters.

⁴Warning: As usual for research themes or front naming, titles are given from central or specific words in the cluster, which does not imply that the title is representative of all documents in the cluster.

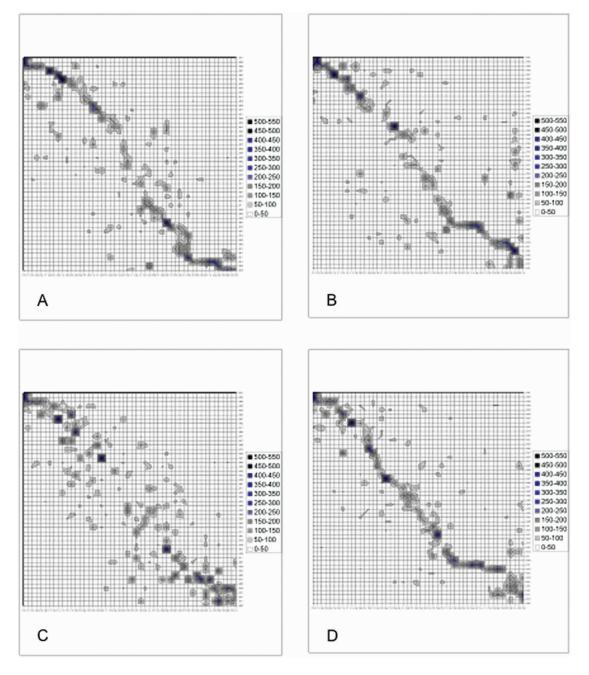


FIG. 6A–D. Visual comparison of reorderings. Scale: Ochiai \times 1000. Panel B makes use of the original distances, others are based on information from the intersection table only (sequence O–J).

A (top-left): from sequence H–J (a schematic display of Figure 5A).

B (top-right): from sequence O-J, based on original distances.

C (bottom-left): CA 1st factor ordering.

D (Bottom-right): CA 1st factor ordering, with local automatic optimization.

C and D do not comply with the tree structure of H–J or O–J. No display presents both properties of a quasi-continuous diagonal region and a visual regularity of the gradient off-diagonal. A and C have the second property but not the first one, D the first one but not the second one, and B none of them. Overall, A presents the soundest repartition.

Surface studies, adsorption, catalysis. Sub-block within *nanomaterials*, with grey cells: three w-clusters discriminate on type of phenomenon and instruments (Adsorption_analysis/Desorption/Catalysis/Adsorption_methods; STM_studies; Surface structure/grown_films/Adsorption/LEED). Among the three c-clusters, one addresses the type of materials, TiO2_surface/Particles/Films, the two others

discriminate on theoretical aspects, Density_Functional_ Theory/Self-diffusion/Modelling; Electronic structures/ calculation, a capability is expected from citation techniques. A nearby c-cluster devoted to theoretical studies is left outside the block.

In the sequence O–J, the same clusters are contiguous but not all retrieved as a super-cluster.

Metal nanoparticles, crystals, nanoparticles. Another sub-block of nanomaterials: five w-clusters discriminate (loosely) on materials: Nanoparticles/gold/silver/TEM (transmission electron microscopy); Gold/silver/metallic nanoparticles; mechanical alloying/nanosize powders/ Fe_compounds; Solvothermal synthesis/room temp; synthesis/ nanocrystalline structures; Nanopowder synthesis/ nanoparticles synthesis. Five c-clusters on materials and objects: Monolayer/Gold particles; Photonic/colloidal_ crystals; Nanocrystals; Nanocomposites/Clay_polymers/ Hybrids; Nanobelts/ Nanoribbons/ Nanowires. This suggests that communities are based on investigation by type of object (see also nanotubes below). The block is rather homogeneous, with a particular proximity to some outsiders with similar profile such as c-clusters TiO2_surface and Quasi_crystals/ Amorphous_alloys.

The sequence O–J shows a rearrangement within a larger block containing mesoporous structures.

Bionano, a particularly promising area within nano sciences. Peripheral to the central themes, bionano is clearly delineated as a block, but the internal structure again indicates different logics. An efficient sorting in terms of usage is offered by two of three large-size w-clusters: Gene therapy/Gene delivery/Gene transfer; Drug delivery/ Pharmakinetics/Drug_carrier/Encapsulation. The third one is Integrin/Extracellular matrix/ Protein_expression/ Cell_ adhesion. The five c-clusters discriminate medical applications of gene therapy and drug delivery (grouped into a very big cluster), on the one hand, and underlying protein properties and specific tools, on the other hand: Protein/DNA_microarrays; Titin/Protein_mechanics; Kinesin/Activity/Motors; Biomolecular_interaction/Binding/ Adhesion. This result suggests the capability of citations to discriminate research communities, whatever the size, contrasting with the ability of terminology to point out typical domains of application. The c-clusters, especially the first two ones, have some connections with off-diagonal w-clusters (such as Copolymers/Polymer_architecture, AFM_studies). The c-cluster microchips/ Biomedical_analysis/Proteomics is classified off-diagonal, within the block Electrochemistry/Films, favoring technical proximity.

Bionano appears as a strong form, whose clusters are best grouped on hybrid distances (sequence H–J) than original ones (sequence O–J).

Also peripheral with respect to the core of nanoscience, the block "*mesoporous structures*" (top-left in Figure 5A) is fairly well-delineated by the sequence H–J, with a single w-cluster (Mesoporous Silica/ Mesoporous carbons) and two c-clusters (Mesoporous/ Silica/ Synthesis; Mesoporous/ Organo-silica/ Ordered_materials) with external traces.

In the sequence O–J, those clusters are merged with other nanomaterials.

Electrochemistry as a block is moderately homogeneous (a few white cells) but two sub-blocks, Dendrimers/ Microchips and Photoelectrochemistry/Polyelectrolytes, contain mostly grey cells. In the first sub-block, the two c-clusters are Bio-medical_biochips/ Proteomics and Dendrimers (large

branched molecules), both with extensive biomedical applications. The two w-clusters distinguish between Copolymers and Nanofiltration/Membranes. In the second sub-block, one comes across three clusters on both sides, c-clusters highlight Porphyrin_Fullerene_dyads/Fullerene_hybrids, Polyelectrolyte_Films/Multilayer, Photoelectrochemistry/ Solar_ Cells; w-clusters are Langmuir-Blodgett films/Air-water interface, Eletrochemical_studies/Electron_transfer, Thin_ Films/Photoelectrochemistry. Giving a more precise account of the differences would need a deeper analysis of the articles contents. The w-clusters of the block show many more intersections off-diagonal than the c-clusters, including the intersections with the block Bionano already noticed.

The sequence O–J retrieves approximately the same cluster, with the addition of self-assembled monolayers, however, and metal complexes rejoin copolymers.

The block *epitaxy-light emission*, *Q-Wells*, *Q-dots* deals with emerging models of semiconductors. Epitaxy refers to various techniques of depositing monocrystalline layers. Molecular beam epitaxy (MBE) was a precursor of the crystal depositing techniques. Quantum dots, quantum-wires, and quantum wells cover electron-confinement in 3-D, 2-D, or one dimension. The convergence at the block level is moderate, but the number of cells involved (five c-clusters, eight w-clusters) can contribute to its heterogeneity. The w-cluster Quantum-Dots/ Structural_properties/ Optical_properties, already mentioned, is particularly marginal, with a unique but strong connection within the block, with the c-cluster 2D-3D_ordering. The sequence O–J cuts off the cluster, with epitaxy and photoluminescence on one side, and Quantum Dots/ Kondo effects on the other side.

Nearby, the apparent one-to-one correspondence between *Quasicrystals_Amorphous_Alloys* (c-cluster) and Alloys/ Grain_bourndary (w-cluster), also found in the O–J sequence, is a bit misleading because the c-cluster, especially, traces a long wake with close w-clusters such as Amorphous_Alloys and more remote ones such as Solvothermal synthesis and Nanopowders_synthesis.

Converging areas with asymmetry. A block is again clearly delineated but with an elongated shape, evoking microcommunities broken down by citations, while the textual approach is less discriminating. Two example of blocks, close to each other, are found in the bottom-right area of Figure 5.

One is located in an emblematic area of nanosciences, the *Carbon nanotubes*, the family of nanoobjects, which emerged next to fullerenes, also covering early research on flat mono-layer lattices (graphenes). It forms a fairly well-delimited block, letting few components in the stripes outside the crossing area. Citations split the theme into seven clusters, each addressing a particular type of object or properties, an indication of distinct research areas and possibly communities: Carbon_nanotubes/Functionalization; CN/Hydrogen_storage; CN/Mechanical_properties; CN/Production/Properties; Single_wall_CN/Electronic_properties; Single_wall_CN/Synthesis/Raman;CN/Growth. The lexical way finds only two clusters, one is CN/field-emission, the second taking on CN/ Single wall CN/ Synthesis/ Growth/ Characterization. The block carbon nanotubes is exactly retrieved by the sequence O–J.

Another example is the block quantum dots, Kondo effect (bottom-right). The Kondo effect refers to the relation of resistivity to temperature in metals, due to magnetic impurities. The block is based on a single w-cluster (Quantum_dots/ Kondo effect/ Electron transport/ Quantum wires) because the reordering groups a second candidate (Quantum_dots/ Structural properties/Optical properties) within the Epitaxy block-because of a heavy intersection with the c-cluster 2D-3D_reordering. Corresponding c-clusters discriminate between Spin_injection/Filter/Polarization/Transport; Quantum_dots/Bits/Josephson; Kondo_effect; Quantum_dots/ Magnetic_field/Coulomb_interactions. The sequence O-J groups the two c-clusters but with a slight rearrangement of c-clusters, adding the c-cluster 2D-3D_ordering in the same block, instead of Spin_injection/ Filter/ Polarization/ Transport, which remains contiguous.

A c-cluster and a w-cluster are given the same title, "*self-assembled monolayers*," a core topic in the bottom-up approach of nanosciences. However, whereas the w-cluster is pretty much centered on its c-cluster counterpart, the c-clusters is much larger and encompasses many other intersection cells, especially with the Electro-chemistry/Films block. Again, the suggestion here is an intellectual base apparent from citations, with several areas of exploitation. An immediate neighbor block combines the c-clusters Self-assembly/Rotaxane and the smaller Molecular_magnets with the w-cluster Metal_complexes/Ligands/Hydrogen_Bonds. The sequence O–J integrates the two cells, kept separate, into two larger blocks, devoted to electrochemistry and fullerenes, and to Metal_complexes and Polymers.

Diverging areas. In some areas, the logic of c-clusters and w-clusters seems quite different, and the reordering does not achieve a homogeneous block. A good example is entitled "Nanomaterials: Theoretical studies," a consistent theme when retrieved by citation clustering alone. In the hybrid analysis, it forms a sub-block of nanomaterials. The c-clusters Magnetoresistance (giant magnetoresistance is an important topic of "top-down" approaches in nanotechnology) and Theoretical studies exhibit some intersections outside the block, but the phenomenon is striking for the w-clusters of the block, quite scattered in their intellectual base, seemingly attracted by the various fields of experiment or application. This is the case, in particular, for clusters with a focus on modeling and simulation. In this case, citation clusters are likely to be more reliable to discriminate areas of research, while word-clusters connect theoretical aspects with various applications or orientations. Another sign of the loose organization of the block is the fact that the sequence O-J, unusually, does not retrieve this grouping, the components being scattered; however, the c-clusters Fullerene, Molecular_electronics and Glass_transition are kept close to each other.

Conclusion

Our purpose in this work was to investigate novel avenues for hybrid classification and mapping, exploiting lexical and citation relations. For methodological reasons, the study was conducted on data on nanosciences that were previously collected and the landscape depicted is a bit out of date now. Similarly, the data used (SCI-WoS) do not allow a complete coverage of the subject, namely, the environmental, health, and ethical implications of nanotechnology. The focus of our work is on the methodology.

There are several ways to achieve a mix of the two classical, powerful methods of thematic mapping, for example, an early mix of citation and word-based metrics, addressed recently by scholars, or a separate construction of clusters/ groups and a comparison of the outcomes. Again, several variants are possible, keeping the word/cite separation or not at the various stages in the process, depending also on the balance of objectives: comparison or combination. Given the variety of ways to exploit such intersection tables by matrix reordering, we picked pragmatic methods, aiming at an unambiguous construction of hybrid blocks, allowing both a comparison and an easy combination of clusters.

The final display may be interpreted as a pseudo-map of the field, carrying a type of information different from a factor analysis layout. The reduction of dimensions is drastic, and the field appears structured along a main route loosely following the diagonal of the matrix. The loss of information on other dimensions, if compared with CA rendering, for example, is compensated by a more precise account of the convergence/divergence of the two approaches. Exhibiting some similarity with old sea charts (portolans), the resulting pseudo-maps are perfectible in many respects, but offer a stimulating way of bibliometric mapping.

Improvements may be sought in several directions. We limited ourselves to the exploitation of titles and abstracts. More focused lexical units along with "citations in context," based on full text, could be appealing. The informetric properties of the two underlying networks, influencing the distribution of cluster size, also remain an open question. Besides, the comparison/combination involves a sequence of stages with several methodological options, sources of artefacts: metrics, clustering/factor techniques, type of hybridization, visualization. A more fundamental point, the semantic and social balances conveyed in vocabulary versus bibliographic references, need in-depth studies, going beyond the purpose of this work. No strict corroboration can be claimed, given the limits of the exercise, but the findings tend to illustrate a classical idea on citations, the higher precision in the delineation of micro-communities based on a cognitive base or a large research question, while terminology mirrors the various aspects of the scientific mixes, including criteria of usage. These tendencies do not mean a clearcut divergence between lexical clusters driven by products, applications, or stakes, and citation clusters designed by theory or methodology. In this respect, the field studied, where nanosciences and nanotechnology are often held synonymous

with an integration of theory, instruments and products (nanomaterials, nanoobjects), is perhaps not the easiest one to test these discrepancies, and applications to other fields with a different heuristic structure are foreseen.

Extensions to other types of linkage may be added: the natural candidate is the network of authors, with a limit, however, due to the thematic range and mobility of individual scientists. The corresponding overlap table is tridimensional, with an expected concentration, after reordering, on a diagonal of the cube. Without adding the third dimension of authoring proximity, the discriminative content of each partition with respect to the community features could not be addressed at this stage of the work but is quite interesting. However, some spurious agreement might appear between authoring and citation-based approaches, especially if no correction for self-citations is introduced.

As to the visualization, there is little doubt that more elaborate techniques could achieve better results, but the outcome is satisfactory. Nonoptimal reordering may lead to the underestimation of the convergence of the two approaches, word or cite-based.

Despite some limitations, to overcome in further investigations, they are not likely to question the main finding, which is a good overall convergence of citation and word mapping. The diagonal archipelago in our "pseudo-map" portolan suggests that bibliometric mapping is robust, but the convergence remains partial, not only as the result of local differences in scale effects. A complete convergence would mean substitutability, making it pointless to mobilize two approaches. This is obviously not the case—with respect to the possible underestimation of convergence noted above—and we have suggested some interpretations of the local divergences. These conclusions need to be confirmed by applications to other fields with different features—an experience is ongoing in the field of genomics, with first indications of similar conclusions.

It is quite important for science mapping that, together with its sociological role in stimulation of experts and scientists' discussion, it can rely on a firm basis as a descriptive tool with scientific value. If completely different bases (lexical and citation networks) lead to relatively convergent but not substitutable views, then we hold a strong argument in favor of the reliability of bibliometric mapping of science, provided that a careful use is made thereof.

Acknowledgments

We are indebted to Henry Small (Thomson Reuters) for helpful discussions, and to anonymous referees for several suggestions.

References

- Aksnes, D.W., & Rip, A. (2009). Researchers' perceptions of citations. Research Policy, 38(6), 895–905.
- Bassecoulard, E., Lelu, A., & Zitt, M. (2007). Mapping nanosciences by citation flows: A preliminary analysis Scientometrics, 70(3), 859–880.

- Benzecri, J.-P., ed. (1973). L'Analyse des données 1 L'analyse des correspondances (Vol. 1) [Data analysis 1: Correspondence analysis]. Paris: Dunod.
- Bertin, J. (1967). Sémiologie graphique; Les diagrammes, les réseaux, les cartes [Semiology of graphics: Diagrams, networks, maps]. Paris: Mouton Gauthier-Villars.
- Bertin, J. (1977). La graphique et le traitement graphique de l'information. [Graphics and graphical treatment of information]. Paris: Flammarion.
- Börner, K., Chen, C.M., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37(1), 179–255.
- Boyack, K.W. (2004). Mapping knowledge domains: Characterizing PNAS. Proceedings of the National Academy of Sciences of the United States of America, 101, 5192–5199.
- Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351–374.
- Braam, R.R., Moed, H.F., & Van Raan, A.F.J. (1991). Mapping of science by combined cocitation and word analysis. I. Structural aspects. Journal of the American Society of Information Science, 42(4), 233–251.
- Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. Journal of the American Society for Information Science and Technology, 61(6), 1130–1143.
- Callon, M., Courtial, J.P., Turner, W.A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Social Science Information, 22(2), 191–235.
- Callon, M., & Law, J. (1982). On Interests and their transformation: Enrolment and counter-enrolment. Social Studies of Science, 12(4), 615–625.
- Chen, C.M. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing & Management, 35(3), 401–420.
- Chen, C.M. (2003). Mapping scientific frontiers: The quest for knowledge visualization. London: Springer-Verlag.
- Cozzens, S.E. (1989). What do citations count—The rhetoric-1st model. Scientometrics, 15(5–6), 437–447.
- Cronin, B. (1984). The citation process: The role and significance of citations in scientific communication. London: Taylor Graham.
- Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques [A new method in automatic clustering and pattern recognition: The moving clouds method]. Revue de Statistique Appliquée, 19(2), 19–33.
- Domengès, D., & Volle, M. (1979). Analyse factorielle sphérique: Une exploration. [Spherical factor analysis: An exploration]. Annales de l'INSEE, 35, 3–84.
- Egghe, L. (2000). New informetric aspects of the Internet: some reflections many problems. Journal of Information Science, 26(5), 329–335.
- Eisen, M.B., Spellman, P.T., Brown, P.O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95(25), 14863–14868.
- Garfield, E. (1967). Primordial concepts, citation indexing and historiobibliography. Journal Library History, 2, 235–249.
- Garfield, E. (1998). Random thoughts on citationology. Its theory and practice—Comments on theories of citation? Scientometrics, 43(1), 69–76.
- Garfield, E., & Sher, I.H. (1993). Keywords-PlusTm—Algorithmic derivative indexing. Journal of the American Society for Information Science, 44(5), 298–299.
- Glänzel, W., & Czerwon, H.J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. Scientometrics, 37(2), 195–221.
- Granovetter, M. (1983). The strength of weak ties: A network theory revisited. Sociological Theory, 1, 201–233.
- Guénoche, A., & Garreta, H. (2002). Representation and evaluation of partitions. In A.S.K. Jajuga & H.-H. Bock (Eds.), Classification, clustering, and data analysis (IFCS 2002) (pp. 131–138). Cracow, Poland: Springer.

- Hartigan, J.A. (1972). Direct clustering of a data matrix. Journal of the American. Statistical Association, 67(337), 123–129.
- Janssens, F., Glanzel, W., & De Moor, B. (2008). A hybrid mapping of information science. Scientometrics, 75(3), 607–631.
- Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14, 10–25.
- Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge, MA: Harvard University Press.
- Laurens, P., Zitt, M., & Bassecoulard, E. (2010). Delineation of the genomics field by hybrid citation-lexical methods: Interaction with experts and validation process. Scientometrics, 82(3), 647–662.
- Law, J., & Williams, R.J. (1982). Putting facts together: A study of Scientific persuasion. Social Studies of Science, 12(4), 535–558.
- Legendre, P., & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. Oecologia, 129(2), 271–280.
- Lelu, A. (1994). Clusters and factors: Neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday & Y. Lechevallier (Eds.), New approaches in classification and data analysis (pp. 241–248). Berlin, Germany: Springer-Verlag.
- Lelu, A. (2008). La méthode de classification non-supervisée K-means axiales. Document LORIA (Vol. 6, pp. 14–25):INRIA.
- Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. Journal of the American Society for Information Science, 48(5), 418–427.
- Leydesdorff, L. (1998). Theories of citation? Scientometrics, 43(1), 5–25.
- Leydesdorff, L. (2004). The university-industry knowledge relationship: Analyzing patents and the science base of technologies. Journal of the American Society for Information Science and Technology, 55(11), 991–1001.
- Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. Journal of the American Society for Information Science and Technology, 61(6), 1105–1119.
- Lloyd, S.P. (1957). Least squares quantization in PCM (No. Published in S.P.). Bell Telephone Laboratories Internal Technical Report.
- Lloyd, S.P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137.
- Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. Scientometrics, 38(1), 27–37.
- Marcotorchino, J.-F. (1987). Block seriation problems: A unified approach. Reply to the problem of H. Garcia and J.M. Proth from Applied Stochastic Models and Data Analysis, 1(1), 25–34, 1985. Applied Stochastic Models and Data Analysis, 3(2), 73–91.
- Marshakova, I.V. (1973). Document coupling system based on references taken from ScienceCitation Index (in Russian). Nauchno-TeknicheskayaInformatsiya, 2(6.3).
- McCain, K.W. (1983). The author cocitation structure of macroeconomics. Scientometrics, 5(5), 277–289.
- McCain, K.W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. Journal of the American Society for Information science, 42(4), 290–296.
- McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L.M. LeCam & J. Neyman (Eds.), Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). Berkeley: University of California Press.
- Mirkin, B. (1996). Mathematical Classification and Clustering. Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Morris, S.A., & Yen, G.G. (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. Proceedings of the National Academy of Sciences of the United States of America, 101(suppl. 1), 5291–5296.
- Morris, S.A., Yen, G.G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. Journal of the American Society for Information Science and Technology, 54(5), 413–422.
- Mullins, N.C., Hargens, L.L., Hecht, P.K., & Kick, E.L. (1977). The group structure of cocitation clusters: A comparative study. American Sociological Review, 42, 552–562.

- Narin, F., Pinski, G., & Gee, H.H. (1976). Structure of biomedical literature. Journal of the American Society for Information Science, 27(1), 25–45.
- Noyons, E.C., Buter, R.K., Hinze, S., van Raan, A.F.J., Schmoch, U., Heinze, T., et al. (2003). Mapping excellence in science and technology across Europe: Nanoscience and nanotechnology (Draft Report No. EC-PPN CT 2002-0001), European Commission.
- Noyons, E.C., & van Raan, A. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. Journal of the American Society for Information Science and Technology, 49(1), 68–81.
- Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing & Management, 12(5), 297–312.
- Polanco, X., Grivel, L., & Royauté, J. (1995). How to do things with terms in informetrics: Terminological variation and stabilization as science watch indicators. In M. Koenig (Ed.), Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics (River Forest IL, June 7–10) (pp. 435–444). Medford NJ: Learned Information.
- Prime, C., Bassecoulard, E., & Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. Scientometrics, 54(2), 291–308.
- Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Questiio, 19(1–3), 23–63.
- Rip, A. (1988). Mapping of science: Possibilities and limitations. In A.F.J. v. Raam (Ed.), Handbook of quantitative studies of science and technology (pp. 253–273). Amsterdam: Elsevier Science Publishers.
- 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(4), 265–269.
- Small, H. (1986). The synthesis of specialty narratives from co-citationclusters. Journal of the American Society for Information Science, 37(3), 97–110.
- Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster—Recombinant-DNA. Scientometrics, 2(4), 277–301.
- Small, H., & Griffith, B.C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. Science Studies, 4, 17–40.
- van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. Scientometrics, 68(3), 377–393.
- White, H.D. (2001). Authors as citers over time. Journal of the American Society for Information Science and Technology, 52(2), 87–108.
- White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32(3), 163–172.
- Wilkinson, L. (1979). Permuting a matrix to a simple pattern. In Proceedings of the Statistical and Computing Section of the American Statistical Association (pp. 409–412). Alexandria, VA: American Statistical Association.
- Wouters, P. (1997). Citation cycles and peer review cycles. Scientometrics, 38(1), 39–55.
- Zitt, M. (1991). A simple method for dynamic scientometrics using lexical analysis. Scientometrics, 22(1), 229–252.
- Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. Scientometrics, 30(1), 333–351.
- Zitt, M., & Bassecoulard, E. (1996). Reassessment of co-citation methods for science indicators: Effect of methods improving recall rates. Scientometrics, 37(2), 223–244.
- Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. Information Processing & Management, 42(6), 1513–1531.
- Zitt, M., Lelu, A., & Bassecoulard, E. (2008, September). Hybrid maps of scientific fields (terms and citations): An application to nanosciences. Paper presented at the Tenth International Conference on Science & Technology Indicators: Excellence and Emergence—A New Challenge for the Combination of Quantitative and Qualitative Approaches, Vienna, Austria.

Appendix A

We start from this intersection table. Let n(ij) the value of the cell at the intersections of w-cluster i (row) and c-cluster j (column); d(j, j') the distance between c-clusters clc(j) and clc(j'); d(i, i') the distances between w-clusters clw(i) and clw(i').

Distances *d* are calculated using Hellinger's formula (e.g., Legendre & Gallagher, 2001), for example:

 $d(j, j')^2 = \sum_i (\sqrt{p(j, i)} - \sqrt{p(j', i)})^2$ where p(j, i) are the profile values like:

 $p(j,i) = n(j,i) / \sum_{i} n(j,i)$

Note: $\max(d(j, j')) = \sqrt{2}; \min(d(j, j')) = 0$

We try to find surrogate distances between a w-cluster clw(i) and a c-cluster clc(j) that first are commensurable with distances d(i, i') and d(j, j') and second do not depend only on the local overlap of *i* and *j*, which may be sensitive to particular cut-off artefacts at the classification stage, but also on their neighborhood.

i, *j* vary from 1 to M.

Proxy $D_c(ij)$ based on length-2 paths linking cluster clw(i) and cluster clc(j) through all clc(j') for k = 1..M:

paths: $\operatorname{clw}(i) - \operatorname{clc}(j' = 1..M) - \operatorname{clc}(j)$

 $D_c(ij) = \sum_{j'} w(i, j') d(j, j')$

Proxy $D_w(ij)$, analogous to $D_c(ij)$, based on M length-2 paths linking cluster clw(j) and cluster clc(j) through all clw(i') for k = 1..M:

clw(i) - clw(i' = 1..M) - clc(j).

 $D_{w}(ij) = \sum_{i'} w(j, i') d(i, i')$

w(i, j') is a weight depending on the profile of the cluster *i* with the c-clusters *j'* in the matrix of intersections. A satisfactory weight is the product of row and column percentages per cell along the row *i*, taking into account the weight of cell (i, j') in both the profile of intersection of *i* and the profile of intersection of *j'*.

 $e(i, j') = (n_{i,j'} / n_{i.}) \cdot (n_{i,j'} / n_{.j'}) = n_{i,j'}^2 / (n_{i.}n_{.j'})$ $w(i, j') = e(i, j') / \sum_{j'} (e(i, j'))$ and similarly for w(j, i').

In the limit case where a c-cluster and a w-cluster are identical, with the same contents in terms of publications, D_c (*ij*) and $D_w(ij)$ are reduced to the direct zero distance, with w(i, j') = w(j, i') = 1. Built as weighted averages of the profile distances d, D_c and D_w can be used in the same scheme of distances. As expected from the averaging process:

- the mean of D_c (*ij*) over the table is close to the mean of d(j', j), the mean of $D_w(ij)$ close to the mean of d(i, i').
- the standard deviations are lower for D_c and D_w .

The following table represents some aspects of the four distributions on our particular dataset. They are quite similar, however, the lower deviation for c-w distances disfavors these pairs in the early clustering stages in ascending classifications. The distances of each kind (c-c, w-w, c-w) can be standardized or rank-transformed. The latter option was picked. We also chose to aggregate $D_c(ij)$ and $D_w(ij)$ in a single measure $D_{cw}(ij)$, with two options: geometric mean(D_c , D_w) or min(D_c , D_w). The outcome is a 2 × M, matrix containing distances c-c, w-w, c-w, where a given c-cluster can be associated with a c-cluster and with a w-cluster with the same probability.

This process is a provisional solution planned to be replaced by a development of AKM.

Distances Statistics

On our particular dataset, the overall correlation between $D_c(ij)$ and $D_w(ij)$ is >0.86, resulting in correlation of D_{cw} with components >0.97 (geometric mean) and >0.94 for (min). (see Table A1)

TABLE A1.

Dist	Ν	Mean	Std dev	Skewness	Min		
$\overline{d(j, j') c-c}$	1275^{a} (k = j incl.)	0.831	0.283	-0.79	0 (0.54 excl. diag)		
d(i, i') w–w	1275^{a} (i = h incl.)	0.835	0.248	-1.42	0 (0.23 excl. diag.)		
$D_c(ij)$ c-w	2500	0.855	0.174	-0.56	0.03		
$D_w(ij)$ c-w	2500	0.905	0.209	-0.46	0.02		
$D_{cw}(ij)$ geom.mean	2500	0.878	0.186	-0.52	0.046		
$D_{cw}(ij)$ min	2500	0.835	0.186	-0.52	0.02		

^aDiagonal terms k = j or i = h are included because they are used in computation of $D_c(ij)$ and $D_w(ij)$.

Appendix B

List of Clusters

C-CLUSTERS (by order of Figure 5)	#	W-CLUSTERS	
Gene_carrier/Delivey/Therapy/Transfection/Transfer	C15	Gene_therapy/Gene_delivery/Gene_transfer	M8
Protein/DNA_Microarrays	C5	Drug_delivery/Pharmacokinetics/Drug_carrier/Encapsulation/ Drug_evaluation	M34
Titin/Protein_mechanics	C16	Integrin/Extracellular_matrix/Protein_expression/Cell_adhesion	M46
Kinesin/Activity/Motors	C42	Mesoporous_silica/Mesoporous_carbons	M47
Biomolecular_interaction/Binding/Adhesion	C19	Self-assembled_monolayers	M45
DNA_molecules/Mechanical_properties	C 7	Metal_complexes/Ligands/Hydrogen_bonds	M27
Mesoporous_II/Silica/Synthesis	C49	Copolymers/Block_copolymers/Polymer_architecture	M28
Mesoporous_I/Organosilica/Ordered/Materials	C12	Nanofiltration/Membranes/Aqueous_solutions/Polyelectrolytes	M16
Self_assembled_monolayer/thiolmonolayers	C30	Langmuir-Blodgett_films/Langmuir_monolayers/Air/water_interface	M32
Self_assembly/Rotaxanes	C6	Electrochemical_studies/Electrochemical_processes/Electron_transfer	М3
Molecular_magnet	C45	Thin_films/Photoelectrochemistry/Porous_materials	M3(
Dendrimers	C22	AFM_Studies/AFM_imaging	M24
Microchips/Biomedical_analysis /Proteomics	C40	Microscopy_techniques/AFM/STM/SPM	M17
Porphyrin_fullerene_dyads/Fullerene_hybrids	C46	Nanoparticles/Gold/Silver_nanoparticles/TEM	M11
Polyelectrolytes_fims/Multilayers	C31	Gold/Silver_nanoparticles/Metallic_nanoparticles	M48
Photo_electro_chemistry/Solar_cells/photovoltaic	C36	Mechanical_alloying/nanosize_powders/Fe_compounds	M38
AFM/STM_studies	C8	Solvothermal_synthesis/Room_temperature_synthesis/ Nanocrystalline_struct	M15
Single_molecule_level/Spectro/Imaging	C3	Nanopowders_synthesis/Nanoparticles_synthesis	M44
Monolayer_protected_cluster/Gold_nanoparticles	C43	XPS_studies/Surface_studies/X-ray/Films	M22
Photonic/Colloidal_crystals	C13	Adsorption/Interfaces/Solid-liquid	M13
Nanocrystals	C23	AFM_studies/Thin_films/Surface_studies	M37
Nanocomposites/Clay_polymer_hybrids	C25	Glasses/Glass_transition/Phase_transition/themosensitivity	M26
Nanobelts/Nanoribbons/Nanowires_fabrication	C14	X-ray_studies/Grazing_incidence_X-ray	M36
Glass_transition/Glass_forming	C27	Modelling/Simulation/Model_systems/Transport	M39
Magnetoresistance	C26	Room_temperature/Magnetoresistance/Temperature_dependence	M42
Molecular_electronics/First_principles	C1	Thin_films/Ultrathin_fims/Thickness	M29
Fullerene/C-60/Production/Formation/behavior	C24	Raman_scattering/Raman_spectroscopy/SERS	M23
Theoretical_studies/First_principles/Ab_initio_studies	C2	Ion_implantation/Ion_beam/	M25
Carbo_nitride_films/MechanicalTribological_properties	C21	Molecular_dynamics_simulations	M50
Carbon_films/Diamond_like/Amorphous	C41	Ab_initio_studies/Models/Calculations/Density_functional_studies	M5
TiO2_surface/Particles/Films/Catalysis	C38	Amorphous_alloys/Mechanical_alloying/Heat_treatment/ Thin_films/X-Ray	M10
Density_functional_theory_studies/Self_diffusion/Modelling	C11	Surface_morphology/Thin_films/Morphology_formation/roughness	M31
Electronic_structures/Calculations	C20	TEM_studies/Structure_studies/Nanostructures/Dislocations	M18
Quasicrystals/Amorphous alloys	C34	Nitride_films/Sputtering/Annealing/X-Ray	M 2
Porous_silicon_properties	C17	Diamond_films/Silicon films/Nanocrystalline/Vapor_deposition	M1
Semiconductors/Ferromagnetism	C35	Adsorption_analysis/Adsorption_desorption/Catalysis/ Adsorption_methods	M49
Diodes/Optical_properties/Applications	C28	STM_studies/	M40
Band_Structure/Parameters/Gap	C9	Surface_structures/Surface_grown_films/Surface_adsorption/LEED	M20
2D-3D_ordering/Self_assembledqQuantum_dots	C50	Alloys/Grain_boundary/Microstructures/Metallic_alloys/ Magnetic_properties	M41
Carbon_nanotubes_functionalization	C37	Photoluminescence/SiO2_substrates/Light_emission/Films	M14
Carbon_nanotubes/Hydrogen_storage	C39	Band_gap/Electronic_structures/Conduction_band/Photoemission	M21
Carbon_nanotubes/Mechanical_properties	C32	Growth_behaviour/Epitaxial_growth/Films	M12
Carbon_nanotubes/Production/Properties	C47	Epitaxial_growth/Thin_films/RHEED_studies	M35
Single_walled_carbon_nanotubes/Electronic_properties	C4	Molecular_beam_epitaxy_growth/Qdots_formation	M33
Single_walled_carbon_nanotubes/Synthesis/Raman/ Arc_method	C44	Molecular_beam_epitaxy/Epitaxial_growth/GaAs	M9
Carbon_nanotubes_growth	C29	Photoluminescence/Optical_properties/Quantum_wells/Quantum_wires	M6
Spin_injection/Filter/Polarization/Transport	C18	Quantum_dots/Structural_properties/Optical_properties	M7
Quantum_dost/Magnetic_fields/Coulom/Interactions	C48	Carbon_nanotubes/Single_wall_nanotubes /Synthesis/Growth/ Characterization	M4
Quantum_dots/Bits/Josephson	C33	Carbon_nanotube_field_emission	M43
Kondo_effect	C10	Quantum_dots/Kondo_effect/Electron_transport/Quantum_wires	M19

Note. NB = no match c-clusters – w-clusters by rows.