Visualizing the Marrow of Science

SCImago Group: Félix de Moya-Anegón
University of Granada, Library and Information Science Faculty, 18071, Granada, Spain. E-mail: felix@ugr.es

Benjamín Vargas-Quesada
University of Granada, Library and Information Science Faculty, 18071, Granada, Spain. E-mail: benjamin@ugr.es

Zaida Chinchilla-Rodríguez, Elena Corera-Álvarez, Francisco J. Munoz-Fernández and Victor Herrero-Solana
University of Granada, Library and Information Science Faculty, 18071, Granada, Spain

This study proposes a new methodology that allows for the generation of scientograms of major scientific domains, constructed on the basis of cocitation of Institute of Scientific Information categories, and pruned using PathfinderNetwork, with a layout determined by algorithms of the spring-embedder type (Kamada–Kawai), then corroborated structurally by factor analysis. We present the complete scientogram of the world for the Year 2002. It integrates the natural sciences, the social sciences, and arts and humanities. Its basic structure and the essential relationships therein are revealed, allowing us to simultaneously analyze the macrostructure, microstructure, and marrow of worldwide scientific output.

Introduction

The construction of a great map of the sciences is a persistent idea of the modern ages. This need arises from the general conviction that an image or graphic representation of a domain favors and facilitates its comprehension and analysis, regardless of who is on the receiving end of the depiction and whether a newcomer or an expert. Science maps can be very useful for navigating around in scientific literature and for the representation of its spatial relations (Garfield, 1986). They are optimal means of representing the spatial distribution of the areas of research while also offering additional information through the possibility of contemplating these relationships (Small & Garfield, 1985). From a general viewpoint, science maps reflect the relationships between and among disciplines; but the positioning of their tags clues us into semantic connections while also serving as an index to comprehend why certain nodes or fields are connected with others. Moreover, these large-scale maps of science show which special fields are most productively involved in research—providing a glimpse of changes in the panorama—and which particular individuals, publications, institutions, regions, or countries are the most prominent ones (Garfield, 1994).

The construction of maps from bibliometric information also is known as scientography. This term was coined by the person in charge of basic research at the Institute of Scientific Information (ISI), George Vladutz, to denominate the graphs or maps obtained as a consequence of combining scientometrics with geography (Garfield, 1986). Although *scientography* is not a widely familiar term, possibly due to the proliferation of terms such as "domain visualization" or "information/knowledge visualization" that make reference to similar notions, in our opinion it is the most adequate term for describing the action and effect of drawing charts of scientific output.

And so scientography, by means of its product known as scientograms, has become a tool and method for the analysis of domains in the sense used by Hjørland and Albrechtsen (1995), consolidating the holistic and realistic focuses of this type of analysis. It is a tool in that it allows the generation of maps, and a method in that it facilitates the analysis of domains, by showing the structure and relations of the inherent elements represented. In a nutshell, scientography is a holistic tool for expressing the discourse of the scientific community it aspires to represent, reflecting the intellectual consensus of researchers on the basis of their own citations of scientific literature.

The present article is the first of three centered on the visualization, analysis, comparison, and evolution of vast scientific domains. Here, we put forth a new methodology

Received January 20, 2006; revised August 14, 2006, June 6, 2007; accepted June 6, 2007

 ● Published online 4 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20683

for visualizing the greatest scientific domain imaginable: the world. Further work will expound the methodology for a comparative analysis of the major geographical and scientific domains of the United States and the European Union. In a third contribution (still under construction), we shall propose a methodology for the dynamic analysis of these same domains.

First, we offer a brief overview of all work to date related with our proposal. We proceed then to outline the methodological development and its validation. After showing the results obtained, an analysis is offered on three levels: the macrostucture, the microstructure, and the marrow of recent scientific output. The ensuing discussion leads us to some brief final conclusions.

Related Works

In Moya-Anegón et al. (2004), we ventured forth with a historic evolution of scientific maps from their origin to the present, and proposed ISI-JCR category cocitation for the representation of major scientific domains. Its utility was demonstrated by a visualization of the scientific domain of geographical Spain for the Year 2000. Since then, other works related with the visualization of great scientific domains have appeared; however, all use journals as the unit of analysis, with the exception of a study based on the cocitation of categories (Moya-Anegón et al., 2005), comparatively focusing on three geographic domains (England, France, and Spain). In contrast, Leydesdorff (2004a, 2004b) classified world science using the graph-analytical algorithm of biconnected components in combination with JCR 2001. Boyack, Klavans, and Börner (2005) applied eight alternative measures of journal similarity to a dataset of 7,121 journals covering over 1 million documents in the combined Science Citation and Social Science Citation Indexes, to show the first global map of science using the force-directed graph layout tool VxOrd. Samoylenko Chao, Liu, and Chen (2006) proposed an approach through the construction of minimum spanning trees of scientific journals, using the Science Citation Index from 1994 to 2001.

In our particular attempt to visualize major scientific domains, we propose the generation of scientograms through cocitation of ISI-JCR categories, pruned by means of PathfinderNetworks (PFNET) layout (Schvaneveldt, 1990), using the algorithm of Kamada and Kawai (1989). This is applied to world scientific output as computed by databases Science Citation Index-Expanded (SCI-EXPANDED), Social Science Citation Index (SSCI), and the Arts and Humanities Citation Index (A&HCI).

Methodology

In processing and depicting the scientific structure of great domains, we further developed a methodology that follows the flow of knowledge domains and their mapping as proposed by Börner, Chen, and Boyack (2003).

Data Source and Processing

Although there may be a number of alternative starting points for such an objective (e.g., *Scopus*: Elsevier, 2005) and points such as the bias in territorial coverage, idiomatic restrictions, and documental typology must be acknowledged, we believe that at present the ISI databases reliably account for world research having international visibility. Proof of this lies in the fact that they are used the world over for formally evaluating research activity. Moreover, the possibility of categorizing references was key for our methodological approach, leading us to discard alternative sources. With this understanding, for strictly investigative purposes, on August 2, 2004, we finished downloading from the *Web of Science* (Thomson Corporation, 2005b)—more specifically from the SCI-EXPANDED, SSCI, and A&HCI—all records of world scientific production published in the Year 2002 (those that in the field "Year" contained the string of characters corresponding to 2002). Because ISI assigns each journal to one or more subject categories, to designate a subject matter (i.e., ISI category) for each document, we also downloaded the *Journal Citation Report* (*JCR*; Thomson Corporation, 2005a), in both its Science and Social Sciences editions, for 2002. The downloaded records were exported to a relational database that reflects the structured information of the documents. This new repository contained nearly 1 million (N = 901,493) source documents: articles, biographical items, book reviews, corrections, editorial materials, letters, meeting abstracts, news items, and reviews that had been published in 7,585 ISI journals (N = 5,876 + 1,709). These were classified in a total of 219 categories, altogether citing 25,682,754 published documents. The information was processed with a PC with a speed of 3 MHz, 512 MB RAM, and 120 MB of hard disk.

Units of Measure

The items of measure used most commonly for the representation of scientific domains are journals, documents, authors, terms, and words. Yet recently, an addition was made to this list, with some broader units such as countries, subject spheres of different levels, institutions, and ISI categories.

One potential complication with units of analysis is the grand total of information if the entire domain is to be represented. If the number of variables or items to be handled is very reduced, we can build visualizations of domains with very small units such as words or descriptors. If this is not the case, we must use broader units of analysis (e.g., documents or authors). Yet, if the amount of information processed is truly very high, it is necessary to resort to units of analysis capable of containing smaller units, as is the case of journals that group documents, authors and terms, or of categories which embrace all the above. This consideration is not new in the field of information visualization, and stems from the physical limitations implicit in representing vast quantities of information in a reduced space. Some authors (e.g., Tufte, 1994, 2001) have analyzed different

DOI: 10.1002/asi

approaches adopted in the face of this and similar problems encountered in the graphic representation of information.

In our view, despite the specified drawbacks, the ISI categories effectively classify documental contents in their databases. As informational units, they are, in themselves, sufficiently explicit to be used in the representation of all disciplines that make up science in general. These categories, in combination with the adequate techniques for the reduction of space and the representation of the information to construct scientograms of science or of major scientific domains, prove much more informative and user friendly for quick comprehension and handling by nonexpert users than those obtained by the cocitation of smaller units of cocitation. The latter would require tagging—usually involving human intervention—of the clusters generated to make for a comprehensible representation.

For these reasons, we used the 219 categories of the JCR 2002 as units of measure, with the exception of "Multidisciplinary Sciences." JCR assigns this category to a specific group of journals of a multidisciplinary nature, such as *Science*, *Nature*, *Endeavor*, or *Proceedings of the National Academy of Sciences*, among others. While this may seem logical and accurate at first, closer consideration shows that works dealing with a given discipline such as Genetics consequently appear cut off from akin categories simply by virtue of having been published in a multidisciplinary journal, and tagged only as such. This problem is not easy to solve without the human touch. The problem of recategorization of documents published in multidisciplinary journals has been dealt with in depth by Glänzel and colleagues (1999a,b). The solution they proposed is to recategorize each one of those documents in view of the most referenced category. We adopted this procedure, with very satisfactory results: Only a few documents had to be recategorized manually because there was a lack of references; however, recategorization of multidisciplinary documents on the basis of the predominant category of the citing documents is an alternative with which we are now experimenting and may possibly incorporate in the near future. The maximum number of categories with which we worked, then, was 218.

Similarity Measures

In light of our previous experience (Moya-Anegón et al., 2004, 2005), we use cocitation as the similarity measure to quantify the relationship existing between each one of the JCR categories.

We have seen in the past that the introduction of measures of standardization in the values of cocitation matrixes, whether using that of Pearson, the cosine function (Salton, Allan, & Buckley, 1994), or Salton and Bergmark's (1979) measure of cocitation normalization, all cause distortions in the visualization of information, as recently described by Leydesdorff and Vaughan (2006). Therefore, after a number of trials, we arrived at the conclusion that using tools of Network Analysis, the best visualizations are those obtained through raw data cocitation as the unit of measure. Yet, it also was necessary to reduce the number of coincident cocitations to enhance pruning algorithm yield. Therefore, to those raw data values we added the standardized cocitation value. In this way, we could work with raw data cocitation while also differentiating the similarity values between categories with equal cocitation frequencies. The key was a simple modification of the equation for the standardization of the degree of citation proposed by Salton and Bergmark:

$$CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}} \quad (1)$$

where *CM* is cocitation measure, *Cc* is cocitation frequency, *c* is citation, and *i* and *j* are categories.

The result is a symmetric matrix of $N \times N$ categories, where N is the number of categories existing in the output of a domain to be visualized. These cocitation matrixes are the base and the origin of our scientograms, which show the structure of the domain represented as well as the relationships and flows of information (i.e., knowledge) within it, between and among disciplines.

Layout

Dimensionality reduction. Over the history of the visualization of scientific information, very different techniques have been used to reduce *n*-dimensional space. Either alone or in conjunction with others, the most common are multidimensional scaling, clustering, factor analysis, self-organizing maps, and PathfinderNetworks (PFNET).

Representing the structure of the scientific output of large domains on a plane is no easy task, whether the domain to be visualized is a region, a state, a country small or large, a continent, or even the world. The adoption of the ISI categories as units of measure implies that the resulting scientograms normally contain over 200 categories in the case of generalistic domains. To display a domain involving such a high number of units that can be easily identified by tags, that show its interactions by means of links, and all this in an intelligible and aesthetically pleasing form for the human eye is a most formidable challenge. Bearing in mind the precautionary message of Hjørland and Albrechtsen (1995): "If users are provided with a system of too many possibilities, without giving priority to the essential connections, the user is overloaded, and the system is ineffective" (p. 416). Then there is the advice of Small (2000): "Despite the loss of structural information . . . the gain in simplicity may for some purposes be worth the sacrifice" (p. 464). And we fully agree with White (2003) in that: "Among techniques, two dimensional PFNET made with raw cocitation counts, and visualized through spring embedders, appears to have considerable advantages" (p. 423).

This process of schematizing information is not new to the current decade. It goes way back to the Middle Ages, known as the principle of simplicity, or Ockham's razor. In its original 15th-century formulation, this principle was

DOI: 10.1002/asi

expressed in Latin: "pluralitas non est ponenda sine necesítate" ("Plurality should not be posited without necessity."), which in common language could be stated as: "Adopt the simplest hypothesis that may explain observations." This law of parsimony has often been taken further, especially since laws of physics began to be expressed in the language of mathematics: The simplest hypothesis is, a priori, that which has the simplest mathematical formulation. In the case of visualization and analysis of a scientific structure, the same minimalistic principle can be applied. Why visualize and analyze a dense and complex structure if we can obtain for study a simpler one, containing the most significant or essential relationships? The methodological challenge presented in these terms is, at any rate, that of the proper choice of significant links. To complete this task with rigor, the option of choice was the PFNET algorithm.

As a consequence of the interdisciplinarity of science, the matrixes proceeding from ISI subject-category cocitation analysis tend to have highly interrelated elements, to such a point that the graphic representation shows a bramble of connections that cannot be studied. In our opinion, PFNET with pruning parameters $r = \infty$, and $q = n - 1$ is the prime option for eliminating less significant relationships while preserving and highlighting the most essential ones, and capturing the underlying intellectual structure in a economical way. Although PFNET has been used in the fields of Bibliometrics, Informetrics, and Scientometrics since 1990 (Fowler & Dearhold, 1990), its introduction in citation was due to the hand of Chen (1998, 1999), who introduced a new form of organizing, visualizing, and accessing information. The end effect is the pruning of all paths except those with the single highest (or tied highest) cocitation counts between categories (White, 2001).

Scalar. There are many different methods for the automatic generation of graphs. The spring embedder type is most widely used in the area of documentation, and specifically in domain visualization. Spring embedders begin by assigning coordinates to the nodes in such a way that the final graph will be pleasing to the eye (Eades, 1984). Two major extensions to the algorithm proposed by Eades (1984) have been developed by Kamada and Kawai (1989) and Fruchterman and Reingold (1991). The criteria for evaluating this type of algorithm are, basically, of an aesthetic nature: the uniform distribution of nodes, the uniform length of the links, the avoidance of crossed links, and so on, all play a fundamental role in the choice of one algorithm or another. While Brandenburg, Himsolt, and Rohrer (1995) did not detect any single predominating algorithm, most of the scientific community goes with the Kamada–Kawai algorithm. The reasons upheld are its behavior in the case of local minima, its capacity to minimize differences with respect to theoretical distances in the entire graph, good computation times, and the fact that it subsumes multidimensional scaling when the technique of Kruskal and Wish (1978) is applied. As Cohen (1997) and Krempel (1999) indicated, the Kamada–Kawai algorithm uses an energy similar to the *stress* of multidimensional scaling as the measure for adaptation to theoretical distances.

We tried out hundreds of representations using the Kamada–Kawai algorithm, and compared results to those obtained with the Fruchterman and Reingold algorithm. The speed in the generation of graphics, capacity for occupying the maximum available space, and the reduced degree of overlapping of links and nodes were the main criteria that led us to opt for the former algorithm. The images shown in Figure 1 make quite clear why we preferred the Kamada–Kawai option.

The result obtained by combining PFNET with the Kamada–Kawai algorithm is as spectacular and visually informative as the map of an underground metro or railroad system:

- At one glance, the center and the outer limits of the system (i.e., domain) can be seen.
- It is easy to get from one station (i.e., category) to another, following the trails or links.
- We can effortlessly see which are the most important nodes in terms of the number of their connections and, in turn, which points act as intermediaries with other lines, as hubs or forking points.

Display

There is no clear expert consensus as to the best format for domain visualization; rather, a wide variety may be used,

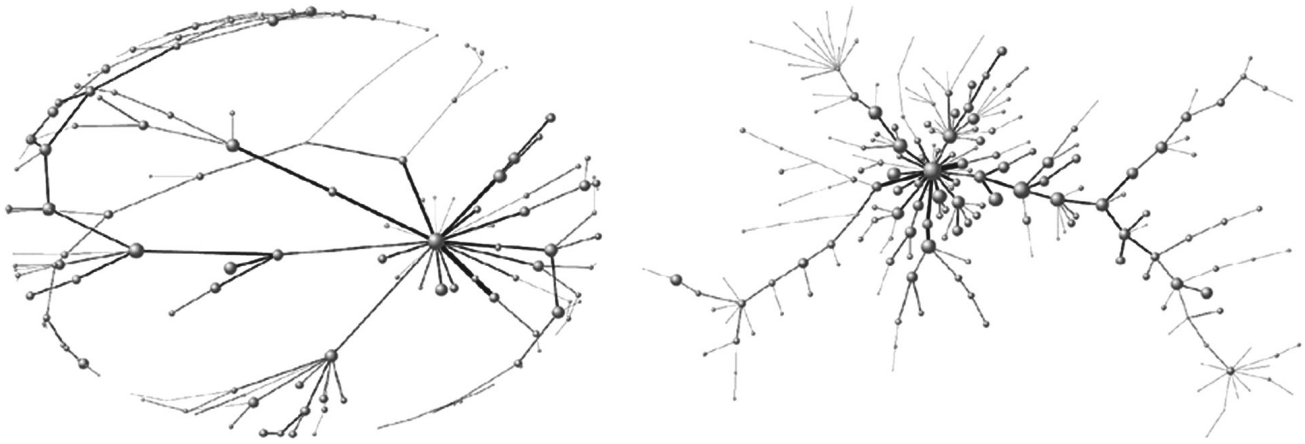

FIG. 1. Scientograms obtained using the algorithms of Fruchterman and Reingold (1991), and Kamada and Kawai (1989), respectively.

DOI: 10.1002/asi

including GIF, JPG, Postscript, Encapsulated Postscript, Virtual Reality Modelling Language, or Scalable Vector Graphics (SVG), among others. In most cases, selection is conditioned by the output format of the programs used by researchers themselves. Still, it is important to obtain quality images with a low weight in bits, so that they can be easily transported over the Web, in light of the growing interest surrounding vectorial graphics and animation.

We find SVG great to work with. Its format is light, quick, ingenious, and free! Being vectorial, it allows one to zoom in and out and wander up, down, or sideways without diminishing the quality of the graphics while also allowing searches for textual information from within points in the image. Moreover, it is backed by firm technical assistance and important underpinnings in the sector as a whole, as well as by individual experts. And as part of the XML standard, it can be used as an interface, facilitating the integration of codes to control the interaction of the graphs and the user. For all these reasons, it was our overriding choice as the visualization format for scientograms of great scientific domains.

So that the scientograms could be displayed in the vectorial format, we exported them to an SVG format using ad hoc software, not commercialized, developed by the SCImago group. This software makes aesthetic and informational touch ups in the scientograms. The tasks it performs are:

- Detection of the superposition of nodes or links, so that they can be repositioned manually.
- Painting the nodes with previously defined colors.
- Tagging each node the tag with the corresponding ISI category.
- Insertion of hyperlinks in nodes and links, to permit the retrieval of related bibliographic information from the relational database (see Moya-Anegón et al., 2005).

Materialization and Validation of the Scientograms

The scientogram of Figure 2[1] shows the structural image of world scientific output in terms of ISI categories for the Year 2002. It resembles a human neuron with a huge axon or central neurite. This scientogram is the visualization obtained as a consequence of applying our methodology to nearly a million scientific documents gathered from ISI databases, then grouped into categories of production for that year.

The links show the most relevant interactions produced between or among categories, and reflect the majority viewpoint of some 2 million scientific authors in light of their nearly 26 million references to other works.

To enhance user comprehension of the scientogram, each sphere is tagged with the name of the JCR category that represents it, and is given a size proportional to the number of documents constituting it in the Year 2002. To help visually establish the relationship between the size of each category and its true output, in the lower left part of the scientogram there is a sphere of reference—a figure scale—with a size equivalent to 1,000 documents. The lines that connect the different spheres are the most significant relationships of cocitation among the categories, the least essential ones having been eliminated with PFNET. As the physical distance between each pair of adjacent categories on the map tends to be constant, the ties are thicker or thinner depending on the intensity of cocitation (i.e., the higher the cocitation, the greater the thickness).

The spatial distribution of the categories in the scientogram is determined by the tandem of raw data cocitation and PFNET. Those categories with a greater number of links (i.e., a higher degree of cocitation) appear in the center. As this number diminishes, the nodes approach the periphery. Just as White (2003) did, though with a greater number of units of measure in our case, we observed that around the most prominent categories, reminiscent of bunches of grapes, one can see the great thematic areas that make up the domain, chained together in explicit sequences. The order that the categories occupy in a chain is by no means arbitrary, reflecting how the subject areas are connected among themselves. In this way, the substructures generated from the prominent subject categories reveal the major thematic areas while the connections among prominent categories reveal how these are interrelated. For example, a second look at Figure 2 allows us to distinguish a huge central cluster surrounded by other smaller ones, distributed all over the surface of the scientogram. If we look even more closely at this central bunch and then at another lower one, we discover the following chain: *Biochemistry & Molecular Biology Neurosciences* ↔ *Clinical Neurology* ↔ *Psychiatry* ↔ *Psychology*. This path indicates that in the scientogram, there are two major subject areas that we could denominate Biomedicine and Psychology, whose most prominent categories, respectively, are *Biochemistry & Molecular Biology* and *Psychology*; which in turn are connected by intermediary categories such as *Neurosciences*, *Clinical Neurology*, and *Psychiatry*. The same can be said, for example, of the chain in the left midsection: *Mathematics Miscellaneous* ↔ *Social Sciences Mathematical Methods* ↔ *Economics* ↔ *History of Social Sciences* ↔ *History*, which shows how Mathematics is indirectly connected with Humanities. These paths are very important, as they are perceived as the thread uniting the overall scientific structure of a domain.

Scientogram Validation

The means in which these scientograms capture the essential structure of the domain and distribute the information over it must be contrasted; that is, just as Boyack et al. (2005) did in their day, we need to validate maps embracing science on such a massive level. And we believe, as they did, that the validation of scientograms of vast domains is an impossible task if one uses traditional methods, based on

[1]Visualization via the Web of the scientograms allows one to carry out a detailed analysis of the scientific structure of a domain without losing any graphic quality at all, thanks to the SVG and the zoom in/out functions. Notwithstanding, it is very difficult to represent over 200 nodes with their respective tags, making them visible and legible in the size and format of a printed scientific journal. To compensate for this problem, together with each figure, we supply an electronic address where readers may view the same figure in real size and high quality.

DOI: 10.1002/asi

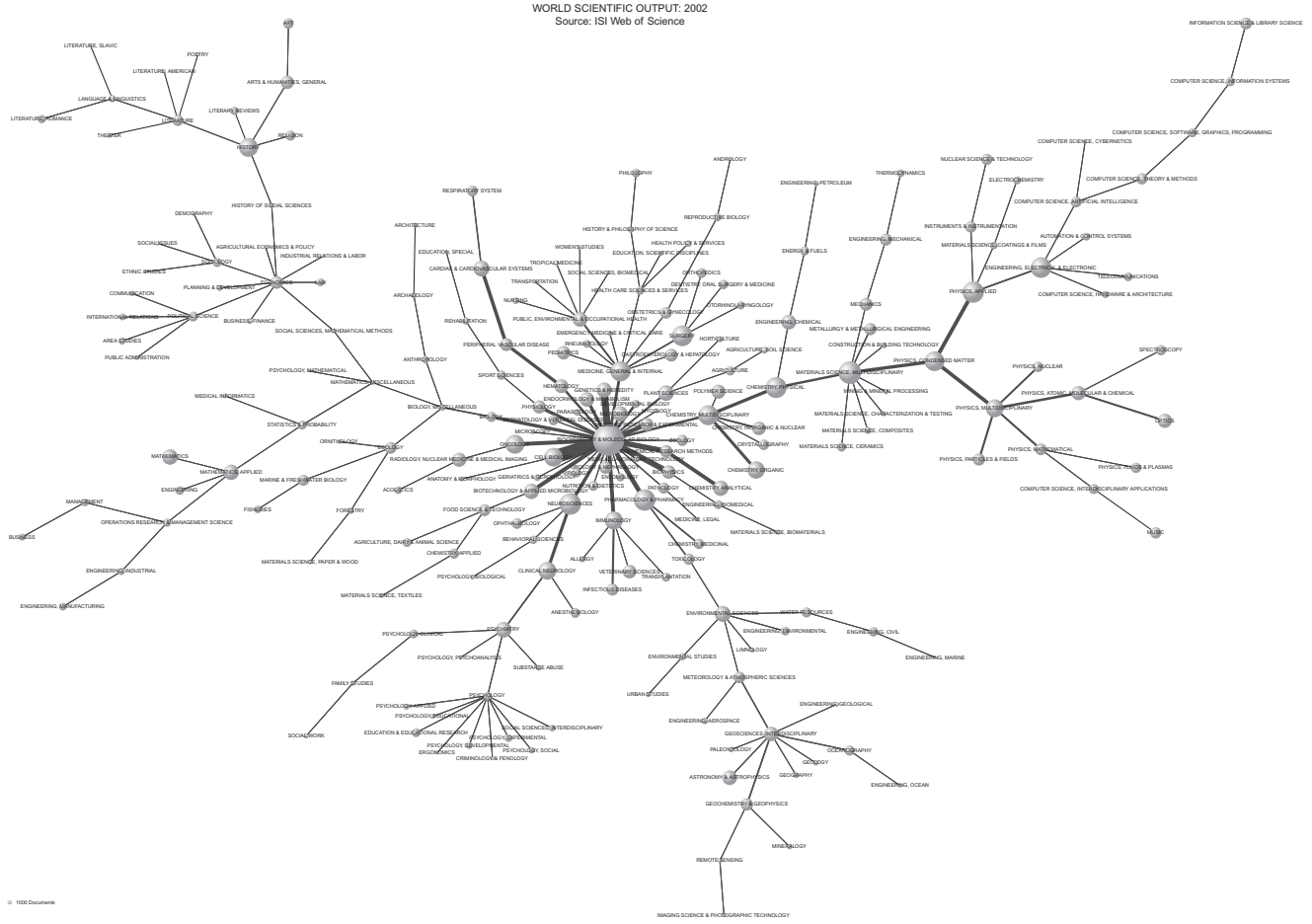

FIG. 2. World scientogram, 2002. Available in real size at: http://www. scimago.es/benjamin/World-2002-2.jpg

qualitative judgment by experts. Far behind us are the days in which a researcher was knowledgeable in all fields of science. Nowadays, the ultimate aim would be to gather, homogenize, and contrast the subjective opinion of 218 experts worldwide, one for each ISI category. Yet, the economic and intellectual investment required for such a feat, not to mention the time factor, made us promptly discard this notion. We finally resorted to a method based on a statistical process—factor analysis—for our validation of findings. Its main features are:

- Factor analysis is conducted on raw data cocitation.
- The number of factors identified is extracted.
- Each factor is tagged.
- The factors identified are transferred to the scientogram.

We stopped extracting factors upon arriving at an eigenvalue ≥ 1,[2] which was done with the scree test.[3] To capture the nature of each factor so as to tag it, we followed the methodology proposed by Moya-Anegón, Jiménez Contreras, and Moneda Carrochano (1998). The factors were first ordered according to their index of weight (i.e., factor loading) in decreasing order, and a cutoff of 0.5 was established for membership, although for denomination, we took into account only those categories of each factor that had a value of 0.7 or more.

So that each one of the subject areas, along with the categories integrating it, can be easily identified, all the categories comprising a common factor were given the same color. For instance, the categories identified in Factor 1 (Biomedicine) appear in light purple, those of Factor 2 (Psychology) are colored emerald green, and so on. Those that belong to more than one subject area are red, the "hot" points of interaction among the subject areas. Finally, dark gray shows the "cold" ones that were not identified by factor analysis and therefore belong to no subject area. Our findings coincide with those of Boyack et al. (2005), in that certain categories are not adequately represented by the documents that make them up (e.g., in Mathematics). This is due to the fact that some journals publish articles whose contents have very little to do with the ISI tag, per se, which is not a problem exclusive to categories but also may stem from the journals themselves.

[2]This simple criterion works quite well, giving results much in accord with the expectations of researchers (Ding, Chowdhury, & Foo, 1999).

[3]The scree test consists of the examination of the line obtained in the graphic representation of the eigenvalues of the identified factors. The extraction of factors comes to a halt when the line of eigenvalues begins to level out, practically forming a line parallel to the axis, with hardly any slant (Lewis-Beck, 1994).

DOI: 10.1002/asi

Whereas factor analysis is a clustering-oriented procedure, PFNET is topology oriented. Yet, they are extremely valuable as complements in the detection of the structure of a scientific domain. Thus, factor analysis is responsible for identifying, delimiting, and denominating the great thematic areas reflected in the scientogram. Meanwhile, PFNET is in charge of making the subject areas more visible, grouping their categories into bunches, and showing the paths that connect the different prominent categories, and finally, the overall topology of the domain.

In summary, this methodology allows for statistical validation of the structural coherence in scientograms of a vast scientific domain. Moreover, it brings into view the large subject areas that make up the domain, providing an image of an intellectual superstructure reminiscent of neural circuitry, which we could call a *factor scientogram* (FS).

Results

The Structure of World Science

Factor analysis identifies 35 factors in the cocitation matrix of 218 × 218 categories of world science 2002. Through the scree test we extracted 16, which we tagged using the previously explained method; these accumulate 70.2% of the variance (Table 1).

The number of categories included in at least one factor is 195. Twenty-three were not included in any factor (Table 2), and 25 belonged to two factors simultaneously (Table 5).

The superposition or overlapping of these results on the scientogram give rise to the FS shown in Figure 3. Next, we invite the observer on an excursion over its surface. Occupying all the central area and most of the upper area is Biomedicine (in purple). Just above it, in deep-sea green, is Health Care & Services, and in a salmon shade, Orthopedics. Agriculture and Soil Sciences takes on a grassy green color.

TABLE 1. Factors of the world science domain, 2002.

Factor	Label	Eigenvalue	% variance	% cumulative
1	Biomedicine	42.255	19.4	19.4
2	Psychology	24.14	11.1	30.5
3	Material Science & Physics Applied	15.472	7.1	37.6
4	Earth & Spaces Sciences	12.655	5.8	43.4
5	Business, Law, & Economy	10.069	4.6	48
6	Computer Science & Telecommunications	8.272	3.8	51.8
7	Agriculture & Soil Sciences	6.815	3.1	54.9
8	Human Studies	6.298	2.9	57.8
9	Chemistry	4.668	2.1	59.9
10	Etiology	4.517	2.1	62
11	Engineering	4.195	1.9	63.9
12	Health Care & Service	3.601	1.7	65.6
13	Applied Mathematics	3.029	1.4	67
14	Nuclear Physics, Particles, & Fields	2.567	1.2	68.1
15	Animal Biology & Ecology	2.321	1.1	69.2
16	Orthopedics	2.16	1	70.2

Going clockwise, we see the rest of the thematic areas. To the center right appears Materials Science and Physics, Applied, in peach. Connected to it by its top section is Engineering, in light yellow, and Computer Science & Telecommunications in hot pink. And connected to its lower part, we can see Nuclear Physics & Particles & Fields in mauve, and Chemistry in brown. The lower central zone holds Earth and Space Sciences, standing out in gray-green, and Psychology in emerald green; just above which we find Etiology, in very light green. In the left center of the display we see, in yellow, Animal Biology & Ecology. Connected to it is Applied Mathematics in dark gray, in the lower part; and in the upper part Business, Law, & Economy, in light purple, and Humanities in sky blue. To establish a quick correspondence between color and the name of each thematic area, see the color code legend in the lower left section of each FS.

Bearing in mind that our scientograms are extremely schematic depictions of the scientific output of a domain, their analysis and interpretation will be based on inferences from the resulting PFNET structure. That is, a category or thematic area occupying a central position in the scientogram will have a more general or universal nature in the domain as a consequence of the number of sources it shares with the rest, contributing more to scientific development than those with a less central position. The more peripheral the situation of a category or subject area, the more exclusive its nature, and the fewer the sources it will appear to share with other categories; accordingly, the lesser its contribution to the development of knowledge through scientific publications. An intermediary position favors the interconnection of other categories or thematic areas. For instance, if the thematic area of Biomedicine disappeared from the scientogram of Figure 3, the rest of the areas would be left disconnected; a similar situation would occur if Biochemistry & Molecular Biology were eliminated. We could say the same of other areas and categories, though the loss of interconnection would be less severe as the positions involved are less central. This broad interpretation of our scientograms not only explains the patterns of cocitation that characterize a domain but also foments an intuitive way for specialists and nonexperts to arrive at a practical explanation of the workings of PFNET (Chen & Carr, 1999).

Macrostructure

When looking at the FS, one can see the combination of just a few thematic areas that are very large in size, connected with many other small ones. This reflects the hyperbolic nature of bibliometric distributions (Small & Garfield, 1985). Another noteworthy aspect is the central–peripheral pattern that the thematic areas adopt in their manner of connection, where a large central thematic area serves as the node of connection to smaller surrounding ones. The conception of a structure formed by a center and a periphery stands as a classical paradigm and appears in many fields of science (Everett & Borgatti, 1999). The existence of a

DOI: 10.1002/asi

TABLE 2. Categories not included in any factor.

ISI Categories	
ARCHAEOLOGY	IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY
ARCHITECTURE	LITERATURE, SLAVIC
BUSINESS	MANAGEMENT
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	MATERIALS SCIENCE, BIOMATERIALS
ENERGY & FUELS	MATERIALS SCIENCE, PAPER & WOOD
ENGINEERING, CIVIL	MATHEMATICS
ENGINEERING, ENVIRONMENTAL	MUSIC
ENGINEERING, INDUSTRIAL	OPERATIONS RESEARCH & MANAGEMENT SCIENCE
ENGINEERING, MANUFACTURING	REHABILITATION
ENGINEERING, MARINE	TRANSPORTATION
ENVIRONMENTAL SCIENCES	WATER RESOURCES
HISTORY & PHILOSOPHY OF SCIENCE	

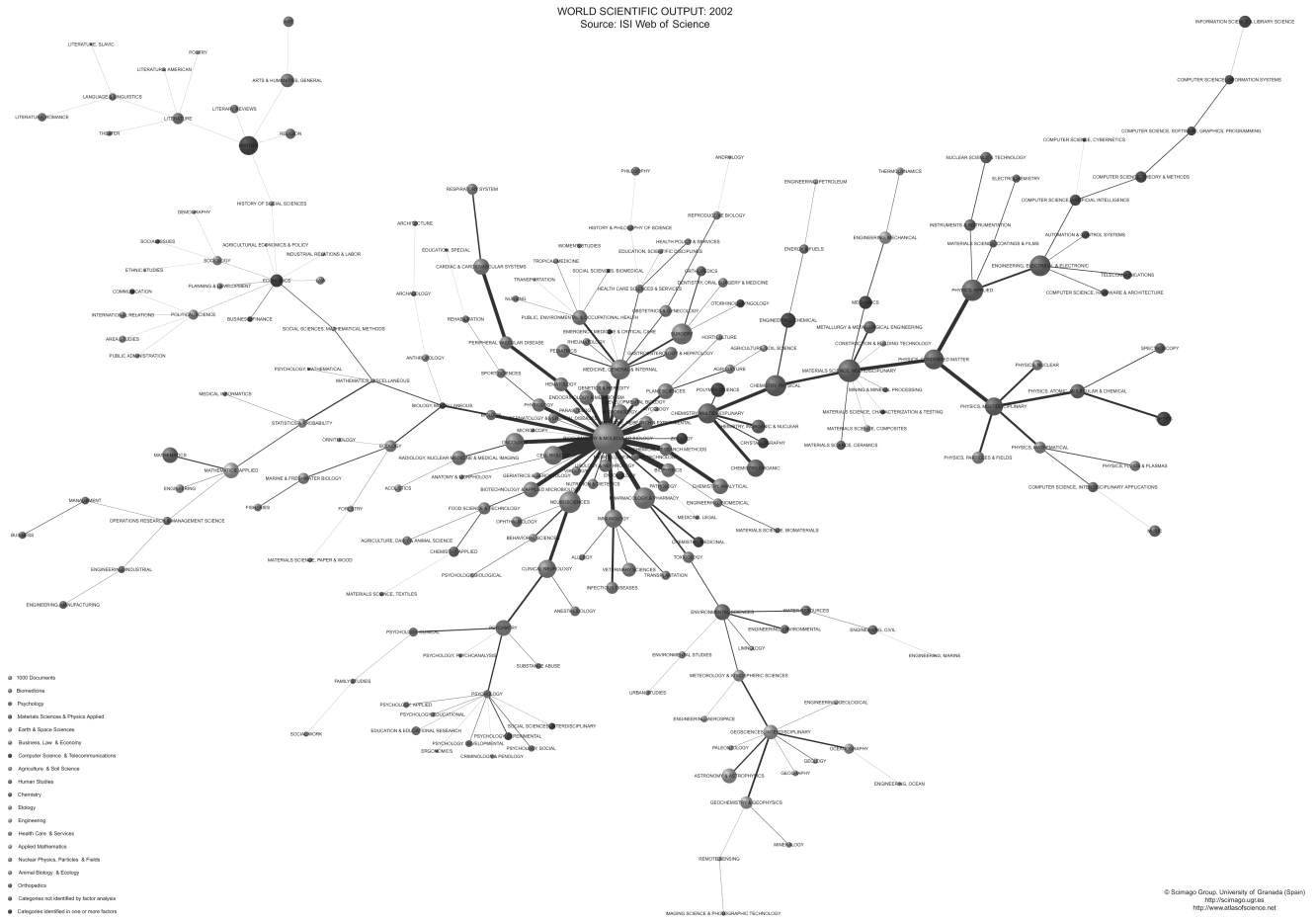

FIG. 3. Factor scientogram of world science, 2002. Available in real size at: http://www.scimago.es/benjamin/World-2002.jpg

structure made up of an active nucleus, formed by a dense and compact grid of categories, creates a striking contrast with a disperse conglomerate of weak interrelations.

From a macrostructural point of view, we can distinguish three major zones. In the center is what we could call *Medical and Earth Sciences*, consisting of *Biomedicine*, *Psychology*, *Etiology*, *Animal Biology & Ecology*, *Health Care & Service*, *Orthopedics*, *Earth & Space Science*, and *Agriculture & Soil Sciences*. To the right, we can see some other basic and experimental sciences: *Materials Sciences & Physics*, *Applied*; *Engineering*; *Computer Science & Telecommunications*; *Nuclear Physics & Particles & Fields*; and *Chemistry*. To the left is the neighborhood of the social sciences, with *Applied Mathematics*, *Business*, *Law*, and *Economy*, and *Humanities*. This scheme of macrostructural vertebration of the sciences can be seen as a typical distribution in the FS displays of developed countries, but evidently differs from the scientific underpinnings of less developed nations, as we

DOI: 10.1002/asi

TABLE 3. Centrality of degree of the thematic areas of world science, 2002.

Thematic Area	Grade
Biomedicine	8
Material Science & Physics Applied	4
Animal Biology & Ecology	2
Business, Law, & Economy	2
Applied Mathematics	2
Chemistry	2
Agriculture & Soil Sciences	1
Earth & Spaces Sciences	1
Etiology	1
Nuclear Physics, Particles, & Fields	1
Human Studies	1
Computer Science & Telecommunications	1
Engineering	1
Orthopedics	1
Health Care & Service	1
Psychology	1

have confirmed with other FSs (Vargas-Quesada & Moya-Anegón, 2007).

At a glance, the most central thematic area is Biomedicine, but to corroborate this, we resort to Social Network Analysis (Wasserman & Faust, 1998) to focus in on the degree of interconnection of the diverse thematic areas involved. This perspective leads us to reconfirm that the most central area is Biomedicine (Table 3).

The centrality of Biomedicine signals it as the area sharing more sources and contributing most knowledge to the rest, lending cohesion to the domain. The identification of Biomedicine as one of the centers of science, its relative position, and its interconnections are nearly identical to the pattern revealed by Boyack et al. (2005) in their map of the backbone of science.

Table 4. Distances with respect to Biomedicine.

Thematic area	Distance
Psychology	1
Agriculture & Soil Sciences	1
Chemistry	1
Etiology	1
Health Care & Service	1
Applied Mathematics	1
Animal Biology & Ecology	1
Orthopedics	1
Material Science & Physics Applied	2
Business, Law, & Economy	2
Earth & Spaces Sciences	3
Computer Science & Telecommunications	3
Human Studies	3
Engineering	3
Nuclear Physics, Particles, & Fields	3

The degree of universality of the rest of the thematic areas will depend on their distance from the center. The shorter this distance is, the greater the involvement in domain evolution. Now, using the paths between thematic areas as the units of measure, we obtain the following ranking of universality, with Biomedicine as the point of reference (Table 4).

Finally, in the FS, we find a series of categories in red (suggestive of friction or points of interaction of different thematic areas). The interdisciplinary categories of the world science domain for 2002 are listed alongside the areas to which they belong in Table 5.

Microstructure

The FS of Figure 3 consists of 218 categories and 217 links that interconnect them. None appears alone or disconnected. As with the thematic areas, note the existence of just

TABLE 5. Categories with double thematic adscription in the World Science Domain, 2002.

ISI Categories	Thematic areas	
BIOLOGY, MISCELLANEOUS	Biomedicine	Animal Biology & Ecology
ENTOMOLOGY	Biomedicine	Animal Biology & Ecology
ZOOLOGY	Biomedicine	Animal Biology & Ecology
CHEMISTRY, MEDICINAL	Biomedicine	Chemistry
EMERGENCY MEDICINE & CRITICAL CARE	Biomedicine	Orthopedics
OTORHINOLARYNGOLOGY	Biomedicine	Orthopedics
BUSINESS, FINANCE	Business, Law, & Economy	Applied Mathematics
ECONOMICS	Business, Law, & Economy	Applied Mathematics
HISTORY	Business, Law, & Economy	Humanities
CRYSTALLOGRAPHY	Materials Sciences & Physics Applied	Chemistry
ENGINEERING, CHEMICAL	Materials Sciences & Physics Applied	Chemistry
POLYMER SCIENCE	Materials Sciences & Physics Applied	Chemistry
MATERIALS SCIENCE, CHARACTERIZATION & TESTING	Materials Sciences & Physics Applied	Engineering
MECHANICS	Materials Sciences & Physics Applied	Engineering
OPTICS	Materials Sciences & Physics Applied	Nuclear Physics, Particles, & Fields
PSYCHOLOGY, MATHEMATICAL	Psychology	Applied Mathematics
COMMUNICATION	Psychology	Business, Law, & Economy
SOCIAL ISSUES	Psychology	Business, Law, & Economy
SOCIAL SCIENCES, INTERDISCIPLINARY	Psychology	Business, Law, & Economy
PSYCHOLOGY, EXPERIMENTAL	Psychology	Etology
SOCIAL SCIENCES, BIOMEDICAL	Psychology	Health Care & Service

DOI: 10.1002/asi

a few large categories and a great number of small ones. The larger ones are seen above all in the center and the right center of the FS, and less so to the left, meaning more production on the part of the categories under medical sciences and "hard sciences" than among the "softer sciences." The pattern of connection that the categories adopt also is of the central–peripheral type: A large central category functions as the central hub of the surrounding categories while maintaining structural cohesion.

There is no doubt that the most central category is Biochemistry & Molecular Biology. This also is demonstrated by its high centrality degree (Table 6).

Again, Biochemistry & Molecular Biology is the category with most shared sources and the greatest share of contributions, demonstrating connectivity and intellectual interchange, emerging as a central axis of the vertebration of science in the Year 2002. If we eliminated Biochemistry & Molecular Biology, the categories around it would be left disconnected, and the semantic structure of the scientogram would be dismantled. Translating the distances to the scientogram and giving each a distinctive color, we can build a new distance scientogram that visually informs, in a quick and easy manner, of the distance of each with respect to the central category (Figure 4). A picture is worth a thousand words.

TABLE 6. Top 16 categories of highest grade, world science 2002.

ISI Category	Degree
BIOCHEMISTRY & MOLECULAR BIOLOGY	31
PSYCHOLOGY	10
MEDICINE, GENERAL & INTERNAL	9
MATERIALS SCIENCE, MULTIDISCIPLINARY	9
ECONOMICS	9
GEOSCIENCES, INTERDISCIPLINARY	8
CHEMISTRY, MULTIDISCIPLINARY	6
ENVIRONMENTAL SCIENCES	6
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	6
PHYSICS, MULTIDISCIPLINARY	5
PSYCHIATRY	5
IMMUNOLOGY	5
ENGINEERING, ELECTRICAL & ELECTRONIC	5
POLITICAL SCIENCE	5
HISTORY	5
LITERATURE	5

The Marrow of Science

The Factor Scientogram is able to reveal the marrow or essence of worldwide scientific divulgation. This is achieved thanks to PFNET's capacity for selecting the most significant

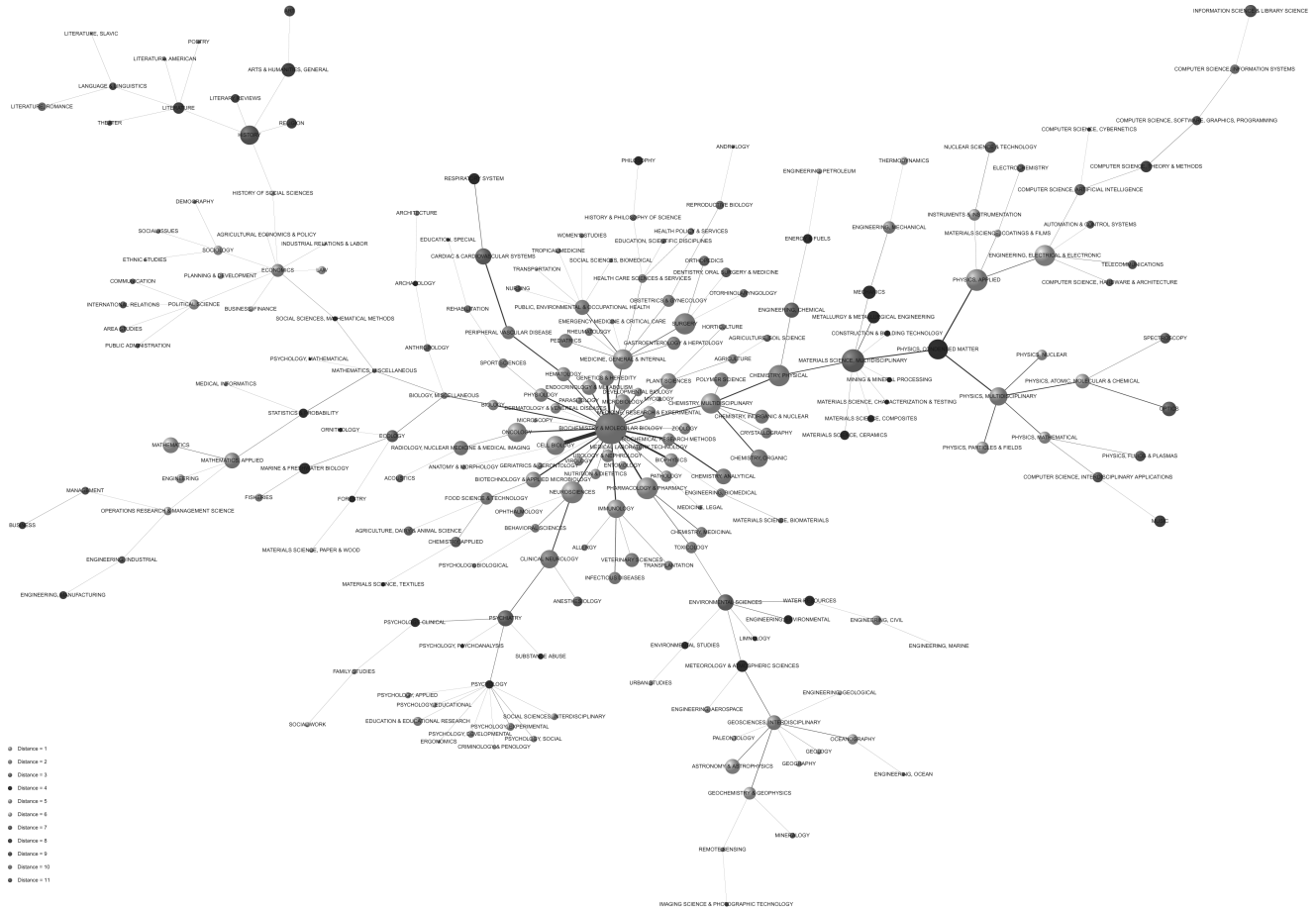

FIG. 4. Scientogram of world scientific distances, 2002, with respect to the central category. Available in real size at: http://www.scimago.es/benjamin/World-2002-dist.jpg

DOI: 10.1002/asi

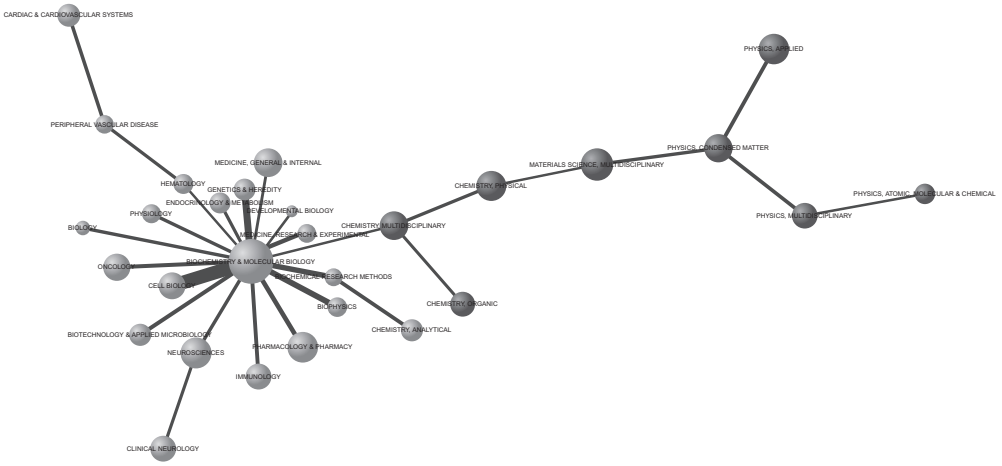

FIG. 5. The marrow of world scientific output, 2002. Available in real size at: http://www.scimago.es/benjamin/World-2002-marrow.jpg

links among categories, together with the graphic possibilities of showing intensity of cocitation by the thickness of the links. Going back to the FS of Figure 3, we see that there are thicker links uniting certain sequences of categories, highlighting the medulla of this domain. To determine which links and categories constitute the spinal column of a domain, we adopt as reference the highest value of the link uniting two thematic areas, and eliminate all those that remain below this cutoff value. The result is the marrow; that is, the part of the structure of knowledge that nourishes and stimulates the whole. There are three thematic areas that stand out in the marrow scientogram of World Scientific Output 2002 (see Figure 5).

Noteworthy is the fact that despite the extreme simplification, there are still some long distance paths such as *Cardiac & Cardiovascular Systems* ↔ *Peripheral Vascular Disease* ↔ *Hematology* ↔ *Biochemistry & Molecular Biology*, indicating the high degree of interdisciplinarity of these categories. The thematic area *Materials Sciences & Physics*, *Applied* appears as a reduced version of its very same structure in the FS display. Despite the schematization, the medulla serves to demonstrate the sequence of a basic structure running from Chemistry Multidisciplinary to Physics Condensed Matter, this in turn serving as a bridge over to Physics Applied and to Physics Atomic Molecular & Chemical.

Discussion

The scientography of vast scientific domains wields the possibility of exploring the state of research from an array of perspectives. On one hand, it offers domain analysts the possibility of seeing the most essential connections between categories of given domain. On the other hand, it allows us to see how these categories are grouped in major thematic areas, and how they are interrelated in a logical order of explicit sequences. Such depictions even can be used by policy makers interested in detecting the strengths and weaknesses of a specific scientific domain by comparing it with others.

Scientograms are a well-designed means of domain visualization, in that they can depict small or large amounts of information: The adoption of the SVG format facilitates the implementation of certain complementary tasks of visualization, such as zooming in and amplifying or reducing any particular area of the scientogram to focus on zones of interest without losing a particle of quality from the original graphics, while traveling within the graph in any direction. Scientograms help to reduce the time visual search for information: The spatial distribution of the information, -occupying the maximum space available, makes the visual search of information very rapid—even in real time, with no need to resort to the zoom. Moreover, and again as a consequence of using SVG, we can quickly locate any chain of text we wish to by means of a search tool incorporated in this format. Scientograms are a good understanding of complex data structure: (a) They make the visualizations self-sufficient in relaying information—little interpretive effort is needed. (b) Scientograms make manifest relationships of which we would otherwise be unaware: PFNET simplifies the relations, showing only the strongest and most relevant ones, so that the structure of the domain is less complicated. (c) Scientograms favor the formulation of hypotheses: The visualizations proposed, which show the semantic/intellectual domain structure in an attractive and comprehensible light, encourage even the nonexpert to theorize about the area depicted, stirring up inferences about interactions that may come about in a certain context. (d) Finally, scientograms would be objects of analysis, debate, and discussion: They can be used by specialists as tools for analysis and debate about the current or past state of a domain.

As the visual result of the consensual opinion of a domain's authors (Vargas-Quesada & Moya-Anegón, 2007), scientograms also are evidence of the evolution of science. True, they cannot predict the future horizons of research, though they may give some clues. Changes over a period of time can reveal tendencies that can be extrapolated to put forth a prognosis of the domain. Their topology, representing

DOI: 10.1002/asi

the structure of the scientific achievements of a specific time period, can be viewed sequentially or dynamically to explore the evolution of a domain (http://www.atlasofscience.net/spanish-evolution.svg).

Scientograms offer new investigators a lasting image of the essential structures of a domain, to complete a mental image already harbored, or become the new point of reference from which an individual perception of the scientific domain can be constructed.

Conclusions

This new methodology for the visualization and analysis of large scientific domains stands as a practical connection of several fields of research, including information visualization, citation analysis, social network analysis, and domain analysis. With very basic means and minimal informational costs, the methodology has allowed us to convey the schematic relationships existing among millions of documents and to generate the complete visualization of the greatest scientific domain feasible: the world. We consider this a very powerful tool not only in view of its capacity to schematize but also because of its facility for representing relational information chained in a series of intelligible sequences, which facilitate and favor comprehension, analysis, and interpretation of the structure of a domain, both for neophytes and for experts. The advantages of scientograms of vast domains are many from the viewpoint of information visualization and analysis. Yet, they also entail two aspects that call for improvement. The first aspect is related to the information used to build the scientograms. Although ISI databases are a most prestigious and adequate means for representing the scientific structure of any domain, the exhaustivity of scientograms would benefit from the incorporation of information proceeding from other sources such as specialized databases and conference reports. The second area for improvement surrounds interpretation: Scientograms are the social and holistic reflection of a domain, yet the final interpretation depends on an individual who is not exempt from some degree of subjectivity. For this reason, we consider it important to continue work on the design of techniques that would enhance the objective components of information representation and limit the more subjective elements.

In practice, Scientograms are being used in the Atlas of Science project (SCImago Group, 2004) as interfaces for access and retrieval of bibliographic information, as tools for analysis and evaluation using a wide variety of bibliometric indicators, and as the starting point for an analytical descent through the structure of science, using online maps of categories and journals. Interested readers may take advantage of free access to all.

Acknowledgments

This work was made possible by public financing through project PNI+D+I: SEJ2004–08358-CO2–01/SOC.

References

Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64, 351–374.

Brandenburg, F.J., Himsolt, M., & Rohrer, C. (1995). An experimental comparison of force-directed and randomized graph drawing algorithms. Lecture Notes in Computer Science, 1027, 87.

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science & Technology, 37, 179–255.

Chen, C. (1998). Bridging the gap: The use of pathfinder networks in visual navigation. Journal of Visual Languages and Computing, 9, 267–286.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing & Management, 35, 401–420.

Chen, C., & Carr, L. (1999). Trailblazing the literature of hypertext: Author co-citation analysis (1989–1998). Proceedings of the 10th ACM Conference on Hypertext (Hypertext'99), (Darmstadt, Germany, 1999), ACM Press, 51–60.

Cohen, J. (1997). Drawing graphs to convey proximity: An incremental arrangement method. ACM Transactions on Computer–Human Interaction, 4, 197–229.

Ding, Y., Chowdhury, G.G., & Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: An author co-citation analysis, 1987–1997. Journal of Information Science, 25, 67–78.

Eades, P. (1984). A heuristic for graph drawing. Congressus Numerantium, 42, 149–160.

Elsevier, B.V. (2005). Scopus. Retrieved August 17, 2005, from http://www.scopus.com

Everett, M.G., & Borgatti, S.P. (1999). Peripheries of cohesive subsets. Social Networks, 21(4), 397–407.

Fowler, R.H., & Dearhold, D.W. (1990). Information retrieval using path finder networks. In R.W. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed placement. Software Practice and Experience, 21, 1129–1164.

Garfield, E. (1986). Towards scientography. Essays of an Information Scientist, 9, 324.

Garfield, E. (1994). Scientography: Mapping the tracks of science. Current Contents: Social & Behavioral Sciences, 7, 5–10.

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999a). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. Scientometrics, 44, 427–439.

Glänzel, W., Schubert, A., Czerwon, H.-J., & Shoepflin, U. (1999b). An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. Scientometrics, 46, 431–441.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain analysis. Journal of the American Society for Information Science, 46, 400–425.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31, 7–15.

Krempel, L. (1999). Visualizing networks with spring embedders: Two-mode and valued data. American Statistical Association, Proceedings of the Section of Statistical Graphics, Alexandria, VA, 36–45.

Kruskal, J.B., & Wish, M. (1978). Multidimensional scaling. Beverly Hills, CA: Sage.

Lewis-Beck, M.S. (1994). Factor analysis and related techniques. London: Sage.

Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. Journal of Documentation, 60, 371–427.

Leydesdorff, L. (2004b). Top-down decomposition of the journal citation report of the social science citation index: Graph and factor-analytical approaches. Scientometrics, 60, 159–180.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. Journal of the American Society for Information Science and Technology, 57(12), 1616–1628.

DOI: 10.1002/asi

Moya-Anegón, F. de, Jiménez Contreras, E., & Moneda Carrochano, M.d.l. (1998). Research fronts in library and information science in Spain (1985–1994). Scientometrics, 42, 229–246.

Moya-Anegón, F. de, Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Herrero-Solana, V., Corera-Álvarez, E., & Munoz-Fernández, F.J. (2005). Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. Information Processing & Management, 41, 1520–1533.

Moya-Anegón, F. de, Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Herrero-Solana, V., Corera-Álvarez, E., & Munoz-Fernández, F.J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics, 61(1), 129–145.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text file. Communications of the ACM, 37, 97–108.

Salton, G., & Bergmark, D. (1979). A citation study of computer science literature. Professional Communication, IEEE Transaction PC-22, 146–158.

Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. Journal of the American Society for Information Science and Technology, 57(11), 1461–1469.

Schvaneveldt, R.W. (1990). Pathfinder associative networks. Norwood, NJ: Ablex.

SCImago Group. (2004). Atlas of science. Retrieved July 28, 2006, from http://www.atlasofcience.net

Small, H. (2000). Charting pathways through science: Exploring Garfield's vision of a unified index to science. In B. Cronin & H.B. Atkins (Eds.), The web of knowledge: A Festschrift in honor of Eugene Garfield (pp. 449–473). Medford, NJ: Information Today.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. Journal of Information Science, 11, 147–159.

Thomson Corporation. (2005a). ISI Journal Citation Reports. Retrieved March 9, 2005, from http://www.isiwebofknowledge.com

Thomson Corporation. (2005b). ISI Web of Science. Retrieved March 9, 2005, from http://www.isiwebofknowledge.com

Tufte, E.R. (1994). Envisioning information. Cheshire, United Kingdom: Graphics Press.

Tufte, E.R. (2001). The visual display of quantitative information. Cheshire, United Kingdom: Graphics Press.

Vargas-Quesada, B., & Moya-Anegón, F. de. (2007). Visualizing the structure of science. New York: Springer.

Wasserman, S., & Faust, K. (1998). Social network analysis: Methods and applications. Cambridge, England: Cambridge University Press.

White, H.D. (2001). Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online. Scientometrics, 51, 607–637.

White, H.D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 54(5), 423–434.

DOI: 10.1002/asi