KEYNOTE ADDRESS

On Extending Informetrics: An Opinion Paper

Howard D. White

whitehd@drexel.edu
College of Information Science and Technology, Drexel University,
Philadelphia, PA 19104 USA

Abstract
This paper suggests the name *bibliograms* for the distinctive core-and-scatter distributions of terms studied in informetrics, scientometrics, and bibliometrics. It argues that naming these distributions as linguistic objects with specific features and demonstrable practical uses will assist in extending informetrics to wider audiences, including librarians and students, for whom existing treatments of the field may be too abstract. It further proposes that bibliograms are closely related to the word association lists, here called *associagrams*, long studied in psycholinguistics and now being imported into research on thesaurus design in document retrieval. It advocates studies that combine bibliograms and associagrams as a way of integrating information science and enriching it with concepts drawn from linguistics and psychology.

A Practical Extension

Although its features are well known to specialists, the central object of study in informetrics has not been named as a linguistic construct. This paper proposes to name it, as a way into discussing how informetrics might be brought to wider audiences, both popular and learned.

The construct is made when substantive terms are ranked by their frequency of co-occurrence with one or more user-supplied seed terms over extended stretches of text (such as a book, a set of related articles, or a large subject bibliography). In recent decades, the ranking has usually been done by computer. Since the ranked co-occurrence counts form a statistical distribution, the constructs are commonly referred to by naming the distribution in statistical or mathematical terms—for example, "empirical hyperbolic," "power-law," "scale-free," "size-frequency," "generalized inverse Gaussian-Poisson," "reverse-J," "core-and-scatter," and so on. The last name is convenient for recapitulating the construct's features:

The "core" consists of relatively few top-ranked terms that account for a disproportionately large share of term co-occurrences overall. The "scatter" consists of relatively many lower-ranked terms that account for the remainder. Typically the top-ranked terms are not tied in frequency, but tied ranks become increasingly common as the frequencies decrease, and a long tail of terms appears in the bottom rank because each co-occurs with the seed term only once.

Early bibliometricians such as A. J. Lotka and S. C. Bradford were interested in analyzing the *form* of distributions like these rather than their detailed verbal *content*, and their descendents have tended to take the same path. Many of these descendents are interested in a science of literatures. They thus look for regularities that generalize across cases and ignore idiographic particulars. The cost of such abstraction, however, is that the verbal content of the informetric distributions is not usually treated as a construct worth studying in its own right. It is treated as a source of frequencies that invite curve-fitting rather than as a communiqué with durable linguistic or psycholinguistic features. Yet for most people it is the *terms* in a core-and-scatter distribution that are interesting, not the content-neutral shapes or zones; they want the meat, not the shell. For them, the main point of the counts in the distribution is that they prioritize some terms over others in a meaningful way.

Without in the least detracting from the achievements of statistical and mathematical informetricians, one can argue that what is needed is a *name* for these ranked lists they deal with — one that conveys the lists as *answers to questions* as well as bundles of items with a particular frequency distribution. The name should be easy to recall and not already widely used to mean something else. It should also be broad enough to cover the different kinds of entities designated in informetric lists—for example, journals, books, descriptors, and authors. I propose that all such

Proceedings of ISSI 2005, pages 442–449.
Edited by P. Ingwersen and B. Larsen.

literature-based lists be called *bibliograms*. Bibliograms consist of: (1) at least one seed term that sets a context, (2) terms that co-occur with the seed across some set of records, and (3) counts of how frequently terms co-occur with the seed by which they can be ordered high to low.

While this addition to the vocabulary may seem superfluous, it has a strategic advantage for the field: it can be used by ordinary people to request what they want in the way of an informetric product. They can say, "I want a bibliogram on..." and then supply the seed term that will generate the kind of associated terms they want. (Most will probably want to look at only a few top terms from the core, but there could be exceptions.) People have no such word now; even the concept behind the word is not widely held. It is, however, a concept that can be lodged in nontechnical minds with many concrete examples. It is teachable and learnable in a way that *might* lead to more rigorous treatments of the field (including the work of Lotka, Bradford, Zipf, Price, Brookes, and so on) but that need not in practical settings.

The question will arise, "Why *biblio*gram, which connotes old-fashioned books?" A possible alternative is "infogram," but that already has other meanings in other contexts. So does the awkward "inforgram." "Webogram" could be used for Web-specific equivalents. However, "bibliogram" is no less suited than "bibliography" to cover electronic writings, and it can likewise cover the innumerable items in print that are still very much with us. It is produced from data properly called bibliographic. It is, moreover, a coinage that is virtually unused.

The test audience for the bibliogram concept is librarians and students intending to enter the information professions. More than one observer has noted that bibliometrics/informetrics is not much used by librarians or taught in library schools (e.g., Wallace 1987, Warning & Emerson 1995, Ungern-Sternberg 1995). The reason may well be that, if librarians or students hear about informetrics at all, the content of core-and-scatter distributions is seldom filled in. (Bradford himself innocently set this trend by not revealing the core journals in lubrication and applied geography.) Nor is that content discussed in a way that would make its value obvious. The way to make it obvious nowadays is to treat computer-generated bibliograms as *recommender systems*—as systems for converting a known seed term into the unknown terms associated with it and (so to speak) recommended by it, which may indeed be informative as answers to questions. ("What journals in lubrication *should* we collect, Mr. Bradford?")

Treating informetrics as closely related to information retrieval accords with the current electronic environment, in which librarians, students, and even the lay public can generate bibliograms without realizing it, as they do when they use Google's new Scholar module. (Similarly, they skim the top of a bibliogram when they look at titles co-purchased with a seed title at Amazon.com.) As Ball & Tunger (2005) point out, an interesting possibility for venturesome librarians is to develop and publicize their ability to deliver specialized informetric products. The bibliogram concept fits nicely with that option.

Besides being recommender systems, computer-produced bibliograms are fast converter systems. They can instantly show the terms associated with seed terms in not one but multiple indexes. For example, the bibliograms of seed authors might rank by frequency their co-authors, the descriptors assigned to their works, or the journals in which they have published. Bibliograms with Library of Congress subject headings as seeds can rank the LC classification codes that are co-assigned with them, or the book titles posted to them that are most widely held by libraries. Bibliograms with descriptors as seeds can rank co-occurring descriptors or associated authors. Fast conversions like these have practical applications in libraries and other information agencies, and they could be used in new popularizations of informetrics to librarians, students, and other nonspecialists.

Courses and workshops could be organized around the bibliogram concept. Table 1 suggests some possibilities for popularizing it as a unifying theme. Many of the applications seen there lead naturally to discussion of more sophisticated products, such as co-citation and co-word maps and various kinds of data mining. The bibliograms in Table 1 can be created with the RANK command in Dialog (or equivalents in other services), the ranking option in OCLC's WorldCat, or inexpensive content analysis software. They lend themselves to both qualitative and quantitative interpretations. Granted, none of this is *research* informetrics, but greater popular awareness of informetric applications could eventually pay off in terms of funding available to serious researchers in the field.

In White (2001) and elsewhere I have discussed particular kinds of bibliograms that are likely to be of special interest in academic settings and that can be produced by academic librarians (through Dialog or other means). These are the personalized products that I call CAMEOs—short for "characterizations automatically made and edited online." They are profiles (hence "cameos") of individual authors in terms of different kinds of noun phrases associated with their names in bibliographic databases—the authors they cite, the authors co-cited with them, the descriptors assigned to the works, and so on.

Table 1. Promotional sketch

Bibliograms can be used to

- suggest additional terms for search strategies
- characterize the work of scholars, scientists, or institutions
- show the subjects associated with a journal or an author
- show the authors, organizations, or journals associated with a subject
- show library classification codes associated with subject headings and vice versa
- show the popularity of items in the collections of libraries
- model the structure of literatures with
  - title terms, subject headings, descriptors
  - author names
  - journal names

and more...

Table 2 shows the top part of a descriptor CAMEO for one of my colleagues at Drexel University. Like most researchers, she was unaware that bibliograms like this can easily be created; even professors who know something about bibliometrics/informetrics may think its uses are limited to impersonal areas like national science policy. But aside from their human-interest value, CAMEO bibliograms of this sort can give researchers (and perhaps their students) the controlled vocabulary needed to conduct literature searches in a particular database—in effect, a personalized thesaurus. For well-known scholars and scientists, they provide clues to intellectual history and, in citation-based CAMEOs, to possible social networks (Otte & Rousseau 2002). They can also be used as probes in interviews with such persons. If made for the faculty of a school or academic department, they can attest to particular subject strengths (and perhaps give evidence of gaps) in analyses of the curriculum (cf. the method in White 2000). However, the point here is that such possibilities rarely occur to academics because they have no concept of what they could ask for, no examples in their heads, no name by which they could make their wishes known.

Table 2. Bibliogram converting an author name into ERIC descriptors

6	Creativity	2	Time
4	Creativity Tests	1	Acceleration
3	Divergent Thinking	1	Anxiety
2	Elementary School Mathematics	1	Beginning Teachers
2	Instruction	1	Behavioral Objectives
2	Mathematics Education	1	Child Development
2	Problem Solving	1	Classroom Techniques
2	Research	1	Cognitive Development
		etc.	

A Theoretical Extension

As stated above, bibliograms are linguistic constructs with distinctive properties. They are not primary products of speaking or writing (such as conversations or letters) but secondary or derivative products that emerge only through analysis. Drawn from written utterances, they partake of relatively fast-moving forces of history and culture. In Saussurean terms, they exemplify *parole* rather than *langue*, but they are not consciously intended; they accrete as bundles of countable noun phrases (types and tokens). While the phrases may be from a single work, they are more often from multiple works. As a rule, they reflect the choices over time of many different authors (or of an individual author in different writings)—choices that determine whether a phrase is absent, present, or repeatedly present across texts. In single texts, the repetition of words and phrases is one way of achieving what discourse analysts and text linguists call "cohesion" (Halliday & Hasan 1976). Across texts, repeated phrases build toward what has analogously been called "intercohesion" (White 2002).

It can be said that a study of bibliograms as linguistic constructs begins as a study of intercohesion—the cross-textual co-occurrence of explicit terms with the seed term. But this quickly leads to interpretations of *implicit* connections among the phrases ranked in the bibliogram—connections inferred by the analyst using specialized domain knowledge or common cultural literacy. Looking at Table 2, for example, it is hard not to infer that the professor's work deals somehow with creativity in teaching mathematics to children. In contrast to "cohesion," discourse analysts use the term "coherence" to denote connections read into a single text from background knowledge (Brown & Yule 1983). Analogously, valid thematic and logical ties that are implicit across texts can be taken as examples of "intercoherence" (White 2002). Intercohesion and intercoherence are the forces that bind bibliograms—and the literatures from which they are drawn—together.

Intercoherence is related to relevance, often said to be the central idea in information science. Generally, relevance is invoked in describing the relation between a person's request and the documents that supposedly fulfill it. However, Saracevic (1975) called the bibliometric/informetric distributions themselves "relevance-related," meaning that bibliograms record relevance as it is perceived over time by authors, journal editors, and subject indexers. Broadly speaking, what these persons are doing is *associating terms*, and their habits and reasons in doing so fall within the purview of linguistics and psychology as well as informetrics. Thus, whether one speaks of relevance or intercoherence, bibliograms are objects that can be investigated within the psychology of verbal associations. It seems eminently possible to relate them to research in word association norms (a tradition dating back to the nineteenth century), although this is not usually done.

The most compelling reason for relating bibliograms to word association norms is that the counts of the words or phrases in the two kinds of lists both have core-and-scatter distributions. It is tempting, in fact, to conjecture that bibliograms (from literatures) and lists of word association norms (from people) are internally coherent for broadly similar reasons. Accordingly, I will here call tabulated word association norms "associagrams" to stress the parallel with "bibliograms." The word or phrase that evokes an associagram is called the stimulus, which is equivalent to the seed term in bibliograms.

Consider Table 3, which reproduces the associagram produced in response to the stimulus word "sardine" by 102 students in a study by Marshall & Kofer (1970). The responses seem intercoherent with (or relevant to) "sardine" in identifiable ways. A recent popularization of linguistics gives four kinds of responses that tend to turn up in associagrams: (1) terms in the same semantic hierarchy as the stimulus—superordinates, coordinates, and subordinates; (2) attributive terms modifying the stimulus; (3) part-whole relations; and (4) terms identifying uses of whatever the stimulus word denotes (Miller 1996: 158). Aitchison (1987: 74-75) narrows lists of such responses to coordinates, collocates, subordinates, and synonyms of the stimulus. Some of these categories fit words in Table 3, but neither set quite captures the cluster made up of "subway, crowded, tight, close, cramped, crowd, squashed, squeeze" that is based on metaphorical associations with "sardine." But at least they constitute a working theory about the *content* of associagrams—a theory that attempts to make sense of the term-column. Informetricians, in contrast, might heed only the count-columns in such data, seeking generalizable explanatory power that is independent of content. A more fruitful course might be to focus on *both* columns in trying to answer questions important to information science, such as "What terms rise to the top of the distribution and why?" It is interesting, in other words, to know both the winner of the election and the winner's characteristics.

Table 3. Terms 102 students associate with the stimulus "sardine"

39	fish
24	can
5	oil
5	sandwich
3	oily
3	subway
2	crowded
2	good
2	smell
2	tight
1	anchovy, bones, close, cramped, crowd, horrible, onions, salmon, salty, slimy, squashed, squeeze, stuffed, tuna

Source: Marshall & Cofer 1970: 335

In a manner of speaking, associagrams suggest how respondents spontaneously *index the stimulus,* with the counts weighting some kinds of indexing over others. In the many associagrams found in Marshall & Cofer (1970), the top responses are very commonly words in the same semantic hierarchy as the stimulus. For example, names of specific birds most frequently elicit the association "bird" (exceptions: "canary"—"yellow"; "chicken"—"food"; "ostrich"—"feathers"). Names of specific fish most frequently elicit the association "fish." In effect, pluralities or majorities of respondents demonstrate that they know what the stimulus term *means*; the superordinate term defines it, and the semantic relation is conventional. It is only in the lower ranks of the associagram that we see relations *not* part of the conventional semantic hierarchy of "sardine"—relations that might be called more creative, such as "subway."

Table 4. Terms medical personnel associate with the stimulus "Serious adverse drug reactions"

17	Side effect(s)
3	Serious adverse event(s)
2	*Bivirkning(er) (Danish)*
2	Adverse effects
2	QT prolongation
2	Toxic
2	Toxicological effect(s)
2	Toxicology
1	Adverse event, Allergic reactions, Cardiac arrest, Cardiovascular side effects, Case report, Death, Disquality, Drug interaction, EPRS, Fatal Events, Fatality, FSPV, Genetic polymorphism, Hospitalisation, Life threatening, Malreaction,
.	Mechanism based inhibition, Metabolite, Mutagenic effects, Negative reaction,
.	OT incapacitation, Periodic safety update report, Pharmacovigilance, Polypharmacy, Product safety, PSUR, Relationship, Safety reactions, Serious events, Serious side effects, SPC, Sudden death, Torsade de Points, Toxic effects

Source: Nielsen 2002 and personal communication

Within information science, Nielsen has recently elicited word associations from respondents as an alternative way of creating thesauri for document retrieval. In Table 4, taken from her dissertation (Nielsen 2002, with counts supplied at my request), one sees again that the top responses from Danish

medical personnel simply define or paraphrase the stimulus. The less conventional, more creative responses in the lower ranks again suggest how the stimulus phrase might connect to terms in other semantic hierarchies.

It would appear that analysis of this sort could be applied to bibliograms of the descriptors that co-occur with a seed descriptor. To be sure, indexers work under different instructions than the people who participate in word-association trials: the writings to which indexing terms are assigned are much more heterogeneous as stimuli than single words or phrases. But the resulting word-and-count lists are comparable, and bibliograms of indexers' choices might yield empirical data toward a clearer understanding of their behavior than we now have.

Table 5 contains bibliograms for the seeds "scientometrics" and "informetrics" in the INSPEC database as vended by Dialog. Again, the top four terms in each list appear to be members of the same semantic hierarchies as the seed terms, and only the lower terms suggest different hierarchies. In discussing this phenomenon in White & McCain (1989: 125), I wrote "Within a given distribution, core maintains identity; scatter individualizes. Core contains redundancy; scatter variety." Indexers seem to associate seeds most frequently with terms that overlap their meaning—terms that could be used in defining the seeds, just as we saw with the associagrams earlier. Moreover, the top-ranked terms in these bibliograms seem relatively easy to apply in the context of the seed term; they lend themselves to conjectures about "least effort" responses and "machinelike indexing by people" that are already familiar in information science (cf. Poole 1985, Montgomery & Swanson 1962). They also look like examples of what psychologists call convergent thinking, in that they can readily be justified as "correct" responses to the seed, whereas lower-ranked terms from other semantic hierarchies depend more on indexers' divergent judgments in capturing the topics of writings.

Table 5. Partial bibliograms of two fields as portrayed with INSPEC descriptors on Dialog

Scientometrics		*Informetrics*	
23	Information Analysis	9	Information Retrieval
13	Citation Analysis	8	Citation Analysis
5	Bibliographic Systems	8	Information Analysis
5	Information Science	8	Information Science
5	Social And Behavioural Sciences	5	Internet
5	Statistical Analysis	4	Information Resources
4	History	3	Data Mining
3	Bibliographies	3	Information Services
3	Classification	3	Statistical Analysis
3	Literature	2	Bibliographic Systems
3	Physics	2	Data Analysis
3	Publishing	2	Graph Theory
3	Research & Development Management	2	Information Retrieval Systems
		2	Statistical Distributions

The contrast being drawn here may be summed up as one between terms that are *semantically* related to the seed (or stimulus) term and terms that are *associatively* related to it. Broadly speaking, we know semantic relationships by knowing a language, including both the mother tongue and specialized jargons it contains, such as the jargon of football or cookery. We know associative relationships by knowing that quite different things have often been discussed in the same context, such as anthrax and U.S. postal workers, or unicorns and maidens. The potential for surprise—for becoming newly informed—is much greater with associative relationships. Vast numbers of these we know insufficiently, which is one reason we do literature searches—or perhaps consult bibliograms.

Semantic relationships are commonly revealed in thesauri, such as WordNet for English as a whole or the MeSH thesaurus for the field of medicine. Associative relationships are infinitely more numerous and cannot be codified except on an *ad hoc* basis. The big conventional thesauri exist to show semantically related terms in the same or neighboring hierarchies. Within hierarchies, broader,

narrower, and related terms are potentially substitutable with OR logic. (Loosely, these are "paradigmatic" relations.) After a thesaurus is compiled, terms from different hierarchies can be associatively related by being combined with AND logic. (Again loosely, these are "syntagmatic" relations.). A paper from psycholinguistics that explores differences between semantic and associative relationships of terms as sketched here is Maki, McKinley, & Thompson (2004), which also has useful pointers to earlier work. These differences have also been explored for information retrieval in an innovative paper by Nielsen & Ingwersen (1999).

Both semantic and associative relationships manifest themselves in associagrams and bibliograms. Either can furnish material for small thesauri tailored to specific work sites or experimental settings, as Nielsen & Ingwersen (1999) report. Perhaps their most interesting data in the present context come from comparing associagrams and bibliograms made from the same stimulus/seed terms in research conducted to develop a thesaurus for an international food company. They found "a surprisingly low degree of overlap; on average 31%" when terms from authors and indexers were compared with terms from specialist employees of the food company. They found as well that, compared to terms from the *Food Science and Technology Thesaurus,* the associagram terms supplied by employees were more numerous, more specific, and from a greater number of hierarchies. Similar results appear in Nielsen (2004).

This is exactly the direction in which research should go if informetrics is to develop a psycholinguistic side and move closer to the other major area of information science, document retrieval (cf. Schneider & Borland 2004a, b). It also suggests an interesting way to introduce informetrics in library schools: have the students jointly create an associagram on some stimulus term and then use the same term as the seed for obtaining a bibliogram; let them compare the results and discuss their implications for online searching.

The present paper will further thinking along these lines if it succeeds in giving two important constructs, the bibliogram and the associagram, names.

References

Aitchison, J. (1987). Words in the Mind; An Introduction to the Mental Lexicon. Oxford, England: Blackwell.

Ball, R., & Tunger, D. (2005). Bibliometric analysis—A new business area for information professionals in libraries? Paper in review for publication.

Brown, G., & Yule, G. (1983). Discourse Analysis. Cambridge, England: Cambridge University Press.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longman.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). Behavior Research Methods, Instruments, & Computers, 36, 421-431.

Marshall, G. R., & Cofer, C. N. (1970). Single-word free-association norms for 328 responses from the Connecticut cultural norms for verbal items in categories. In Norms of Word Association, L. Postman & G. Keppel (Eds.). (pp. 321-360). New York: Academic Press.

Miller, G. A. (1996). The Science of Words. New York: Scientific American Library.

Montgomery, C., & Swanson, D. R. (1962). Machinelike indexing by people. American Documentation, 13, 359-366.

Nielsen, M. L. (2002). The Word Association Method: A Gateway to Work-Task Based Retrieval. Doctoral dissertation. Åbo : Åbo Akademi University Press.

Nielsen, M. L. (2004). Task-based evaluation of associative thesaurus in real-life environment. Proceedings of the 67th ASIS&T Annual Meeting. (pp. 437-447). Medford, NJ: Information Today.

Nielsen, M. L. & Ingwersen, P. (1999). The word association methodology—A gateway to work-task based retrieval. Final Mira Conference on Information Retrieval Evaluation. Retrieved April 17, 2005 from http://ewic.bcs.org/conferences/1999/mira99/papers/paper6.pdf

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science, 28, 441-453.

Poole, H. L. (1985). Theories of the Middle Range. Norwood, NJ: Ablex.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321-343.

Schneider, J. W., & Borland, P. (2004a). Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations. Journal of Documentation, 60, 524-549.

Schneider, J. W., & Borland, P. (2004b). Identification and visualization of 'concept symbols' and their citation context relations: A semi-automatic bibliometric approach for thesaurus construction and maintenance. In Nord I & D, Knowledge and Change—Proceedings of the 12th Nordic Conference for Information and Documentation, M. Hummelshøj (Ed.). (pp. 44-56). Aalborg, Denmark: Biblioteksarbejdes Skriftsserie.

Ungern-Sternberg, S. von. (1995). Applications in teaching bibliometrics. Proceedings of the 61st IFLA General Conference. Retrieved April 17, 2005 from http://www.ifla.org/IV/ifla61/61-ungs.htm

Wallace, D. P. (1987). A solution in search of a problem: Bibliometrics & libraries. Library Journal, 112 (May 1), 43-47.

Warning, P., & Emerson, P. (1995). Cocitation analysis: Using bibliometrics to bring academics and information professionals together. LASIE, 25(4-5), 84-89.

White, H. D. (2000). Computing a curriculum: Literature-based domain analysis for educators. Information Processing & Management, 37, 91-117.

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

White, H. D. (2002). Cross-textual cohesion and coherence. The CHI 2002 Discourse Architectures Workshop. Retrieved April 17, 2005 from http://pliant.org/personal/Tom_Erickson/DA_White.pdf

White, H. D., & McCain, K. W. (1989). Bibliometrics. Annual Review of Information Science and Technology, 24, 119-186.