January 2012

Bibliometrics as a Tool for Research Evaluation

Bibliometrics as a Tool for Research Evaluation

Chapter in forthcoming book: Al Link & Nick Vornatas, *Handbook on the Theory and Practice of Program Evaluation*, Edward Elgar, 2012.

Diana Hicks and Julia Melkers

School of Public Policy

Georgia Institute of Technology

Atlanta GA, 30332-0345

March 2012

Introduction

Bibliometrics, or the study of publication-based output, is a method widely applied in evaluation. As pressure for quantitative evidence of the success and impact of research efforts has grown, the use of bibliometrics has spread. The most common and accepted use is for the analysis of the output of basic research. However, bibliometrics is also useful as a partial indicator of overall R&D output, and of the productivity and impact of funded research teams and centers. The bibliometric evaluation of research and development activities remains one of the most challenging issues in program evaluation despite the effort devoted over the last few decades to develop and test reliable and accurate measures of research output.

Fundamental issues arise in the use of bibliometric measurements of research and development activities. For example, a potentially long time frame – easily a decade or more - is needed to fairly judge the results of research programs. The nature of research, and particularly basic research, does not always permit immediately perceivable results but funding bodies require evidence of success after perhaps two or three years. In addition, R&D is a cumulative, networked process. Results most often build upon other results produced by other groups and programs. Success is often difficult to credit in its entirety to a single source, which is a problem if agencies need to claim ownership. Additionally, the results of any single program will have a wide variety of impacts on future work. Multiplicity of impacts presents difficulties in deciding which outcome to evaluate, and makes tracing particular research findings almost impossible.

As a result of the difficulties associated with accurate measurement of the outputs of the R&D process, most measures are used as proxies for a particular aspect of the output. This in itself presents problems as the accuracy and reliability of different measures varies. The purpose of this chapter is to provide a background and survey of bibliometric approaches to the analysis of R&D. This chapter is by no means the first to treat this topic, nor is it the most in depth treatment. Readers interested in a full length discussion of bibliometrics in evaluation might consult Moed (2005) *Citation Analysis in Research Evaluation* or for a broader treatment, De Bellis (2009) *Bibliometrics and Citation Analysis*, see also REPP, 2005 and Bornmann et al. 2008.

This chapter begins with a brief history of the development of bibliometric analysis. Bibliometrics encompasses several types of analysis including: publication counts, citation counts, co-citation analysis and scientific mapping. These will be combined in the best evaluations, and in addition qualitative methods will be used. Each of these types of analysis will be discussed in turn. Data sources and their use will also be examined. Finally, examples are given in which program goals are translated into bibliometric terms for measurement in real world evaluations. We conclude that creative application of bibliometric analysis can provide insight into the achievement of many different types of R&D program goals.

Scope and History

Bibliometric approaches to assessing scientific output may be traced back to the turn of the century. The first use of bibliometrics is commonly attributed to Cole and Eales in their 1917 paper entitled: *The history of comparative anatomy: Part 1.-a statistical analysis of the literature*. However, until the launch of Eugene Garfield's *Science Citation Index* (SCI) in 1961, bibliometric analysis remained extremely laborious and so limited in scope and scale (Garfield, 1964). The pioneering work of Derek de Solla Price using the SCI did a great deal to establish the potential of the techniques (Price, 1963). Other pioneers, such as Henry Small, helped to develop techniques, such as co-citation analysis, that increased the sophistication and range of questions that could be studied using SCI data (Small, 1980.) Francis Narin laid the groundwork for using publication and citation counts as indicators of scientific output and prestige, establishing that the measures correlated with peer review measures of impact (Narin, 1976). Scholars gradually recognized the potential inherent in a database of scientific papers and their references to provide an overview of the scientific landscape, examine the characteristics of the scientific community and chart the development of knowledge. Bibliometric research groups engaged in evaluation were established at CHI Research in the U.S. (Narin); at SPRU, University of Sussex UK (Martin & Irvine); University of Leiden in the Netherlands (Van Raan, Moed etc.); Australian National University in Australia (Bourke & Butler); CINDOC in Spain (Gomez), and ISSRU at the library of the Hungarian Academy of Sciences (Braun). By the mid-1980s, the Office of Technology Assessment argued that bibliometric indicators could be of assistance in research evaluation, when combined with peer review, and advocated their wider application (OTA, 1986). This was a departure from the subjective methods that had dominated research evaluation. Bibliometrics provided a body of information about the scientific community that was not uniformly available through methods such as peer review and narrative accounts of scientific developments. The SCI and Garfield's company, ISI, was acquired by Thomson-Reuters, and today is known as the Web of Science (WoS). In 2004 Elsevier launched a second citation database, Scopus, bringing competition to what had been a monopoly market. Today bibliometrics are routinely combined with qualitative methods in evaluations of research of all types (ESF, 2009).

The research tradition in bibliometrics is well developed and remains vibrant. Central journals include *Scientometrics*, *JASIST*, and *Journal of Informetrics*. Biennial conferences include: International Conference for Informetrics and Scientometrics and the International Conference for Science and Technology Indicators. Contributions to the field come not only from bibliometric specialists but from information scientists, research evaluators and natural scientists. For example, the recent introduction of the h-index metric came from a physicist. Frontier work involves examining the possibility for new metrics such as download counts or social media mentions as well as employing data visualization.

Bibliometric techniques have been applied to evaluations for several decades in many OECD countries including the U.S.[1] Bibliometrics is simply a type of social science archival research method and has been endorsed through methodological investigation (Narin, 1976), widespread use by governments (ESF, 2009; Hicks et al., 2004; Hicks, 2011) and in sober assessment (OTA 1986). Proponents have long argued for the utility of the techniques in evaluation when combined with peer review processes (Martin & Irvine, 1983; Moed, 2005). Nevertheless, because the techniques are used to evaluate academics, who like anyone else would prefer not to be evaluated, debate over the strengths and weaknesses of the measures does not settle (Adler et al., 2008; MacRoberts & MacRoberts, 1996). All would agree that

[1] The passage of the Government Performance and Results Act (GPRA) in the early 1990s seemed to necessitate widespread, systematic application of bibliometrics to measure Federal research program performance (Narin and Hamilton, 1996). This did not happen as the GPRA activities of NSF and NIH ended up focusing on articulating strategic goals and benchmarking process – such as peer review, managing capital expenditure etc. (Cozzens, 2007).

bibliometric analysis is most appropriately applied to research whose primary output takes the form of publications, and that peer review should also be a component of any evaluation using bibliometrics at the individual, group, department or field level. Disagreement centers on what citations measure (quality would be ideal, but impact is more realistic), distortions of citation counts (high rate achieved through negative citation, citation circles, self-citation etc.) or the value of contributions that do not appear in papers (database curation, creation of new materials or organisms, or increased human capital for example).

Other challenges are common across evaluation techniques, including the uncertainties associated with R&D, its multiple consequences, cumulative nature, interdisciplinarity and transferability. A critical problem in evaluating R&D activities is the long and often uncertain time frame in which "results" may be observed compounded by the fact that uncertainty is a requirement in the basic research process. Basic research is exploratory, with little idea of what form the final applications might be, which in turn makes it difficult or impossible to assign quantitative values to the uncertain outcomes. Similarly, the success or the precise nature of these "results" is not only difficult to identify, but it is difficult to know when a particular research project has been completed and what the results are. This is especially true in basic research. To add to these hurdles, research often results in unexpected or not obvious results, such as enhanced communication between researchers. R&D is also a cumulative process, with the results of research growing upon each other. While the worth of a single project in isolation may be limited, its results may be critical to the development of another unrelated project. Finally, it is difficult to trace the transfer of this knowledge.

Against these difficulties, stands the public record of research progress established when research results are published in scholarly journals. And the incentives to do this are quite explicit as promotion and hiring are based on the strength of the published record as is, in some countries, increasing amounts of university core funding. Publications serve to record in measurable terms the results of publicly funded research. In many settings, it may be the only quantifiable measure of productivity available. Bibliometric analysis makes use of this tangible and quantifiable output of the scientific enterprise in order to answer fundamental questions about the output and quality of scientific research.

USEFUL MEASURES IN RESEARCH EVALUATION

Publication Counts and Mapping

Publication of journal articles endures as a central feature of scientific work, despite the importance of tacit knowledge in laboratory work and the rise of blogging. We can understand this by remembering an anthropologist's first observations of laboratory life, which established the centrality of documents in scientific life. The anthropologist was able to portray laboratory activity as "the organization of persuasion through literary inscription" (Latour & Woolgar, 1979, p. 88) - and to understand scientists as readers and writers of documents. 'Documents' here has a wide interpretation that includes inscriptions written on tags on bottles; the output of instruments, whether printed or electronic; charts and graphs derived from the instruments' output; and reports or published papers.

> *Almost without exception, every discussion and brief exchange observed in the laboratory centered around one or more items in the published literature. In other words, informal exchanges invariably focused on the substance of formal communication. . . much informal communication in fact establishes its legitimacy by referring or pointing to published literature.*
>
> *Every presentation and discussion of results entailed the manipulation either of slides, protocol sheets, papers, pre-prints, labels, or articles. Even the most informal exchanges constantly focused either directly or indirectly on documents. Participants also indicated that their telephone conversations nearly always focused on the discussion of documents; either on a possible collaboration in the writing of a paper, or on a paper which had been sent but which contained some ambiguity, or on some technique presented at a recent meeting. When there was no direct reference to a paper, the purpose of the call was often to announce or push a result due to be included in a paper currently being prepared. (Latour & Woolgar, pp. 52-3)*

Scientific work connects heterogeneous elements such as formal and informal communication. Papers are useful precisely because in scientific work they are part of such heterogeneous assemblages. The heterogeneous assemblages found in research are the focus of Hilgartner and Brandt-Rauf's analysis of data streams. "Data" in their scheme, are "not the end-products of research or even . . . isolated objects, but . . . part of an evolving data stream". They:

> *use the word data as a technical term, encompassing both inputs and outputs of research, as a shorthand for 'information and other resources produced by or needed for scientific work.' This definition is meant to include not only experimental results, but also instrumentation, biological materials and other samples, laboratory techniques, craft skills and knowledge, and a wide range of other information and know-how. (Hilgartner & Brandt-Rauf, p. 7)*

Data streams are thus seen as "chains of products", and papers are part of the chain. A data stream comprises objects and information, and the information can be tacit, unpublished or published. Neither the objects nor the tacit knowledge can be communicated in a publication. However, a paper describing research points to other elements of the data stream and thus indicates that the authors possess certain tacit knowledge and materials. Readers learn the area in which the researchers work, their collaborators, their funding sources, the names of the materials used, the techniques used to manipulate them, and the astute reader assesses the technical quality of the work. Readers are alerted to the existence of underlying tacit knowledge, skills, substances and so on possessed by the authors. Published papers thus are indicators of the presence of productive, unpublishable resources.

Papers help move knowledge in two ways, the first being the obvious one of conveying useful information and the second being signaling. Papers signal the presence of other elements in the data stream. So for areas in which papers are produced, they are extremely useful in research evaluation as indicators of the presence of everything else that constitutes the substance of scientific and technical advance. For this reason, the most basic of bibliometric techniques involves counting scientific publications published by a researcher, research group, program, university or country.

Many studies have established publication counts as a reasonable proxy of scientific output, finding a strong correlation between publication counts and peer review results (Narin, 1976; Stephan, and Levin, 1988; King, 1987). The acceptance of this measure as indicative of scientific productivity is illustrated by its wide use. However, publication counts are never used alone in any serious evaluation setting. Other bibliometric metrics as well as qualitative information are also brought to bear.

Publication counts are most useful for providing an indicator of the volume of research output. Their primary drawback is the limited information provided about the quality of the output. Analysts often use the perceived quality of the journal in which a paper is published as an indicator of a paper's quality. Web of Science provides for every journal a calculation of its average citation rate called the "impact factor". This measure has been widely used in evaluations. And while this has some merit, it is also controversial. Citation counts of papers in even highly regarded journals vary over orders of magnitude; impact factors are determined by technicalities unrelated to article quality; and impact factors depend on field – the highest impact factors are found in rapidly expanding areas of basic research with a short lived literature with many references in each article (Seglen, 1997).

Mapping of research activities and related outputs has grown in popularity in recent years with the growth in computing power and accessible software (see Börner et al., 2003 for a comprehensive introduction to the techniques). These advances have significantly added to the bibliometric toolkit by allowing for additional visualization, as well as statistical representation of the character of publication output, as well as the cited works. A number of maps of the entire scientific enterprise now exist that depict the landscape of scientific disciplines and outputs. Rafols, Porter and Leydesdorff (2010) have created an open source resource that allows an analyst to take a set of publications, create a map of the papers by scientific discipline and then overlay that map on their map of science as a whole. For research evaluation, this tool can be especially powerful in visually representing the placement and impacts of a research group's work within the larger disciplinary landscape. On a larger scale, Rafols et al. have used this approach to produce empirical evidence that interdisciplinary research is at a disadvantage in national evaluations of university research output (Rafols et al., 2011). On a smaller scale, this approach may be used to track the productivity and interdisciplinarity of the research outputs of a team or center over the life of a grant.

Akin to publication counts is analysis of co-authored papers. Co-authorship patterns can be used to investigate and characterize the networks of the scientific community. In evaluation one might be interested in whether scientists co-authored more frequently after a program or center has been funded if the aim of the program were to encourage collaboration. For many large-scale collaborative funded efforts, evaluation may focus on the patterns of co-authorship across institutions or institutional units, regions, or other specific groups. For example, the European Union might be interested in whether the rate of collaboration between countries in Northern and Southern Europe increased over time and whether any such effect could be traced to EU programs that sought to encourage this activity.

Citation Counts

While publication counts measure output, citation counts go one step further and address scientific impact based on the number of times that subsequent papers reference a particular earlier paper. One scholar refers to citations as "frozen footprints in the landscape of scholarly achievement; footprints which bear witness to the passage of ideas" (Cronin, 1984). Proponents of citation counts as an indicator argue that the most important works will have the highest number of citations. Critics counter that authors cite earlier work for any number of dubious reasons, making citation counts meaningless as indicators of scientific quality. Over the past few decades, many leading scholars have wrestled with the theoretical question: what do citation counts measure? Table 1 offers a compressed summary of the classic contributions.

Crucial to a sophisticated understanding of citation counts is that most agree that citation counts do not signify scientific quality in any simple way. Since natural scientists "tend to think of scientific contribution as somehow inherent" in a paper (Aksnes and Rip, 2009, p. 898) and thus what citation counts **should** signify, they dismiss citation counts as useless. Early sociological contributions argued that scientists were rational in their behavior and referenced prior work in a way closely connected to its inherent quality. This view fell out of favor, and sociologists subsequently argued that scientific knowledge is made through a social process and many political and other factors influence scientists' behavior, including referencing behavior. Some were accused of arguing that was all there was. Others maintained that referencing combines a myriad of motivations, and so citation counts measure an inextricable mixture of social and cognitive influence. Martin and Irvine (1983) offered the most developed explanation of this mixed point of view in support of evaluative use of citation counts and proposed that "impact" be the term used to interpret citation counts, and that impact be viewed as both a social and cognitive influence on subsequent research. Their idea did prevail to some extent, and impact remains the soundest way of interpreting citation data.

Table 1 – Theoretical perspectives on the meaning of citation counts[2]

Author	**References are**	**Cites measure**
Garfield (1979)	Supportive, illustrative or elaborative of points in a document	Importance
Small (1978)	Elements in a symbol making process	Highly cited papers are concept symbols
Merton (1996)	Registration of intellectual property & peer recognition	Intellectual influence
Cole & Cole (1967)		Socially defined quality
Gilbert (1977)	Tools of persuasion	Authoritativeness
Cronin (1984)	A reflection of authors' personalities and professional milieu	Unclear, complex interplay of norms and personal factors
Martin & Irvine (1983)	Influence, social and political pressure, awareness	With matched groups, differences indicate differences in influence
Zuckerman (1987)	Response to Gilbert - motives and consequences analytically distinct	Proxies of more direct measures of influence
Latour (1987)/ Luukkonen (1997)	Resources authors wield to support their knowledge claims in a dynamic and hostile environment	Usefulness to subsequent authors in both social and cognitive dimensions
Cozzens (1989)	Reward, rhetoric, communication intersect in refs – rhetoric first	Recognition, persuasiveness, awareness
White (1990)	Acknowledgements of related documents	Co-cites = historical consensus of important authors and works
Van Raan (1998)	Partly particular, but in large ensembles biases cancel out	highly cited = top research
Wouters (1999)	Product of scientist	Product of indexer

There has also been a great deal of empirical work examining the meaning of citation counts. One stream examines the context in which references are found in citing documents. Bornmann and Daniel (2008) review this literature in depth finding great variability in the citation motives and types uncovered. They conclude that the studies reveal no consistent findings, in part because their methods are variable and not always of high quality. In the end they agree with Van Raan that citations are not so random as to be meaningless, and citation counts are therefore useful in evaluation. Another stream points to the correlations between citation counts and peer review based measures of esteem to support the use of citation counts as proxies for these more labor intensive methods (Narin, 1976). Still another stream of work asks scientists about their own highly cited work. In a recent contribution, Aksnes and Rip (2009) surveyed 166 scientists about their highly cited papers, finding that scientists were well aware of a standard repertoire of criticisms of citation counts which they deployed to explain low counts of their papers more often than they explained low counts by the paper lacking importance. However, high citation counts to their papers were explained using the importance of the paper. Scientists seem to recognize that quality as well as communication dynamics both play a role in contributing to scientific progress. In the case of their own citation counts, the communication dynamics are seen as illegitimate when a paper is not well cited, but are seen as part of the process of influencing scientific progress when a paper is highly cited.

[2] Adapted from Moed (2005) page 194.

The complexities of citation dynamics mean that evaluators are more comfortable using citation data on groups or programs than they are on individuals. Nevertheless, scientists themselves have become quite enamored with the individual-level h-index measure in recent years. The h-index reports for an individual, the number of papers that have earned more than that number of citations. Thus an h-index of 10 would indicate that a scientist has published 10 papers with more than 10 citations. One strength of the index is that it deals well with the nature of the citation distribution, which is a power law. However, it also favors older scientists who have had a long enough career to amass a large number of papers and citations. The number is also unstable against seemingly trivial changes in the paper-citation count, and will go on increasing even after a scientist dies. The h-index has been much debated in the pages of *Scientometrics* and *JASIST* and many alternative formulations have been proposed.

An important issue in using citation counts is proper normalization. Citation rates vary by field and by year of publication, with older papers having had more time to accumulate citations. Therefore, analysts compare the actual citation counts to the rate to be expected. To normalize for year a fixed citation window is used. In other words, citations to a paper are summed over the year of publication and three or four years after.

Normalization for field of publication requires a choice: do you compare a paper's citation count to others of the same age in the same journal, the same specialty, the same field, or the same discipline? A careful analysis by Zitt et al. (2005) showed that the level chosen makes a difference. For example, a paper can be in the top 10% most cited in its journal, but if the journal is low impact, it might not make it into the top 10% in its specialty. Or a paper could not be in the top 10% in its journal but be in the top 10% at the field level if it is in a high impact journal. In addition, fields can be defined differently. One could use fields as defined in WoS, or a custom group of papers built for a particular evaluation. One problem with WoS fields is that they exclude papers in high impact multi-disciplinary journals such as *Science*, *Nature* or PNAS.

Even the arithmetic used to compute the aggregate, normalized citations per paper metric has been disputed. A scientist who was subject to an evaluation and who was unhappy with the figures produced by a leading bibliometric group turned to another bibliometric analyst to explore possible problems with the method. The result was a high voltage dispute over the best way to calculate an average, with the method used by the leading evaluation group generally agreed to be wrong (Bornmann, 2010; Opthof & Leydesdorff, 2010; Van Raan et al., 2010). The existence of this dispute points to the dangers of engaging in quantitative evaluation of subjects who themselves are quantitatively adept and highly critical. One's method must be sound and the numbers correct because they will be challenged.

As part of this dispute, the problems of using any sort of average on data distributed in a power law has been raised and suggestions put forward that one should instead rely on percentiles, for example, the share of papers in the top 10% most cited. If the share exceeds 10%, that is favorable performance. This discussion echoes the early method of partial indicators proposed by Martin and Irvine (1983) in which publication counts, total citations, citations per paper, and share of papers in the top 1% and top 10% were all interpreted to arrive at a judgment of bibliometric performance and impact.

Co-citation analysis and mapping

Co-citation analysis, developed in the early 1970's, is a method that identifies pairs or groups of articles that are often found in the same reference lists of subsequent papers. From these pairs or groups of articles a "cognitive structure" is derived (Small, 1980). This cognitive structure is believed to be linked to the theoretical concept of "invisible colleges" said to describe the informal communication links that develop among the most productive and influential scientists (Price, 1965; Crane, 1972.) The individuals in these invisible colleges tend to be highly influential and research leaders and their informal communication forms links between research groups. Understanding these invisible colleges is important because they are critical in the transfer of knowledge within the scientific community (Crane, 1972). Surveys of scientists suggest that the co-citation structure of a research field is a fair representation of how it is perceived by its members (Mullins et al, 1977; McCain, 1986), though this affirmation may arise from the enhanced persuasiveness of the graphical representation and the perceived authority of the method of construction (Hicks, 1987 & 1988).

Increasingly, attention is focusing on networks in science to help explain scientific developments. Co-citation analysis is most useful in monitoring the cognitive distance between subfields of science because it helps define research problem areas that appear to be the basic units of scientific activity as well as providing a measure of the level of interaction between these areas. This results in the collective cognitive representation of the organization of science. This can be done on a large scale (science as a whole) or a small scale (scientific specialties or subfields, research centers, or teams). For example, in the context of the evaluation of centers or research teams, co-citation analysis can also reflect knowledge exchange and transfer within that particular group. This notion of “knowledge coalescence” or “knowledge coupling” has been identified as a useful indicator of the maturation of a research center (Youtie, Kay, and Melkers, 2011.) “Knowledge coupling” occurs when a set of common citations in center publications increase over time, reflecting the sharing of information among team members.

The results of co-citation analyses are displayed as metaphorical maps of the scientific landscape. The "map" presents a projection into two dimensions of the position of scientific areas in relation to each other in n-dimensions. On the map are circles representing clusters of papers and lines between clusters that are strongly linked by citations or words in common. At times the metaphor has been extended into a third dimension with height representing another metric, such as number of citations to the papers in a cluster. The maps are interpreted as the location of specific fields within the realm of science as well as the "distance" between specialties (Rip, 1988.) A variety of complex clustering methods are used to construct maps, and recent work has sought to determine the best methods (Gmür, 2003; Klavans & Boyack, 2005).

Data Sources For Bibliometric Analysis

Research evaluation focuses on institutionalized research groups - funded through a specific funding program or source, affiliated with a research center or grant or a department. The first step in the evaluation is compiling the bibliography of the research group. The Web of Science or Scopus can be searched for group papers, though the difficulty of such a process should not be underestimated. Careful name cleaning and matching must be undertaken to create a reliable bibliometric data set. In house publication lists - annual reports, curriculum vitae, etc. - provide an alternative. Though here one must be cautious in assuming the list is complete.

Some groups may be more diligent in list construction than others; this administrative difference between groups could appear as a research performance difference if the analyst using in-house lists is not careful. In most cases additional publications will be found in the databases.

Compiling a sound bibliography for evaluation purposes requires awareness of several methodological caveats. First, scholarly journals publish material that is not considered original contributions to knowledge such as editorials, book reviews, corrections, bibliographies, meeting abstracts etc. This material is indexed in databases. In bibliometric work one wants to restrict the count to original contributions to knowledge such as articles, notes and letters. Analysts differ as to whether reviews and conference proceedings articles should also be counted as original contributions to knowledge, so sometimes they are included and sometimes not. In some fields, these contributions may be more highly valued than in others and insensitivity to this point when designing the bibliometric approach can cause problems later.

Second, crucial for the integrity of a publication count and related bibliometric analysis is restricting it to peer reviewed material. This is not a problem when WoS is the basis for the data because journals must meet stringent requirements to be indexed in WoS, including using a peer review process to screen submissions. But publication counts also can be made using in-house publication lists, institutional repositories, e-print archives such as arXiv etc. Since peer review is not a criterion for accession in these resources, counts could be inflated by a determined researcher bent on extensive vanity publishing or self-plagiarism.

Third, the limits of the source used to compile the publication list must be kept in mind. WoS in particular, focusing as it does on the most important journals, often has been criticized for insufficient coverage of non-English language material. This should not be of great concern in U.S evaluation studies. The databases offer a classification of journals into fields, but this may be found wanting if examined too closely. The strengths of the databases lie in their easy to use format (relatively speaking), their coverage across all fields of science and back in time as well as the very accession criteria that keep some material out, that is they cover the best quality, peer reviewed journals. In contrast, in-house lists may not be in database format, may be partial, may be dated or cover only recent years and may include non-scholarly material. Against this, in-house resources are helpful in differentiating between grant or center-affiliated work relevant to the evaluation and other publications from group researchers irrelevant to the evaluation.

Online sources are clearly attractive for gathering large sets of publication data. Yet, caution must be observed in the selection of sources. For example, Google Scholar, because it offers such a superior method of finding articles and is so easily available, can seem like an easy tool to conduct an evaluation. However, Google Scholar is in fact not a good tool in which to build an evaluation. This is because databases such as WoS and Scopus are highly structured, with author information differentiated from institutional affiliation and from the body of the article itself. Such structure, or metadata, is not needed for simply finding articles on a topic of interest to the user. Therefore, Google Scholar has never focused on constructing accurate metadata. Google Scholar coverage is also problematic. Google Scholar coverage is never explicitly stated whereas WoS and Scopus have clear and consistently applied criteria for accession to the database, placing a quality filter on the work that is indexed. Meho and Yang (2007) undertook a bibliometric study using WoS, Scopus and Google Scholar and counted the hours needed to collect, clean and standardize the data. WoS was the easiest to use at 100 hours, Scopus required 200 hours and Google Scholar 3,000 hours for the same job. They also determined the citations missed by each database due to database error. WoS missed 0.2%, Scopus 2.4% and Google Scholar 12%. WoS & Scopus failures were traced to incomplete cataloguing of reference lists. Google Scholar failures were traced to inability to match searched words and ignoring reference lists in documents if the keywords: "Bibliography" or "References" were absent. However, Google is always improving and launched personal citation profiles in 2011. Strong individual disambiguation technology and accurate author identification at the individual level seems to underlie this tool. If structured metadata could be downloaded from Google Scholar, this would change the evaluation landscape going forward.

An Illustration: Applying Bibliometric Analysis in Practice

To illustrate the ways in which bibliometric analysis can be used to evaluate a research group, we present several examples using data drawn from research evaluations in multi-year programmatic and center-based research efforts. Some examples are drawn from a ten-year retrospective evaluation of a major program effort, while others are drawn from on-going evaluations that occur throughout the funded research period. All involve multi-disciplinary and multi-institutional teams.

The overall evaluation designs from which these examples derive take a multi-methodological approach in assessing the development and outcomes of the research program/center, placing the bibliometric analysis within a larger analysis of existing data, case studies, interviews, and surveys. This is important to note because the bibliometric analysis in all cases is intended to complement the interviews and focus groups, regular surveys of researchers etc. For the ten year retrospective, bibliometric data provided important evidence of the outcomes and impacts of a program that had supported fifteen centers. The evaluation did not focus on individual centers, but instead took the programmatic view of the contributions of the centers overall. In other examples, we draw from on-going work that provides formative evaluation of multi-year research centers. In these evaluations, the placement of bibliometric analysis within the evaluation design accounts for the development and maturation of each center by scheduling it a midpoint and then again near the end of the funding period. Because bibliometric data have high salience for the funder and the center director, it is perceived as important summative evidence of center productivity and impacts. Yet, some formative value is also recognized, where early and mid-term assessment of the number and characteristics of the publication outputs of a center or multi-year large scale research project (almost a virtual center) may be useful in research management.

Developing the Bibliometric Data Set

The development of a bibliometric data set must be driven by the scope of the evaluation, as well as the characteristics of the research endeavor. In our examples, the bibliometric dataset for the ten year retrospective analysis was developed by extracting Web of Science bibliometric records for articles identified in a large set of center-level reports. The bibliometric dataset for the on-going evaluation example was developed using a combination of compiled publications from the Web of Science, center-affiliated researcher curriculum vitae, and center annual reports. First, for the initial Web of Science search, the research team used researcher curriculum vitae to correctly identify the publications of the authors(s). Searches combined name, researcher affiliation, typical publication outlets and discipline to correctly identify publications. The evaluation team gathered publication records for the years in which the Center had been funded, but also the five years preceding that funding in order to compare pre-center publications to center publications. Finally, the publication data were coded as "center affiliated" or not, based on their inclusion in the center annual reports. This coding relied on the accurate reporting of Center personnel, and was confirmed with researchers.

For all bibliometric examples, once records were extracted from WoS, the data were imported into Vantage Point software (http://www.thevantagepoint.com/) enabling the subsequent cleaning and organizing of the data for analysis. At a later point in the evaluation, the data were also input to the open-source data platform that allows for the analysis of publications and citing articles within the Map of Science, as well as calculating aspects of the disciplinary characteristics of these works (Rafols, Porter and Leydesdorff, 2010).

Publication Counts and Co-Authorship Patterns

The first question asked in a bibliometric evaluation is: "Are the funded researchers producing?" Basic time series publication counts capture productivity and are critical to understanding trends in outputs. In an external evaluation of a multi-year research project or center, publication counts can be most useful at the funding midpoint (as baseline) and continue through funding completion (as summative evidence). Expectations are that publication counts will show some activity early on, but increase over time. While other evaluative approaches can provide useful ongoing formative and summative input, bibliometric analysis is constrained by lag factors in publication and subsequent visibility and impacts captured through citation analysis. A retrospective analysis can however benefit from the longitudinal view of center or project evaluation by capturing publication activity over time. Figure 1 illustrates the ideal result in which output from a group of centers funded through a program increased steadily over time. Depending on the timing of the evaluation, the data for the final year (as is the case here) may drop off due to incomplete data for that time period. From these data, the institutionalization of the centers, and increases in productivity demonstrated a "healthy system of scientific activity" where the "research programs hosted by centers were able to gain momentum." From this, the evaluator may conclude that the centers increased in productivity. However, additional conclusions are limited without additional data breakouts, for example, by discipline, center, faculty rank, or other characteristics that may be important in the overall center evaluation design.

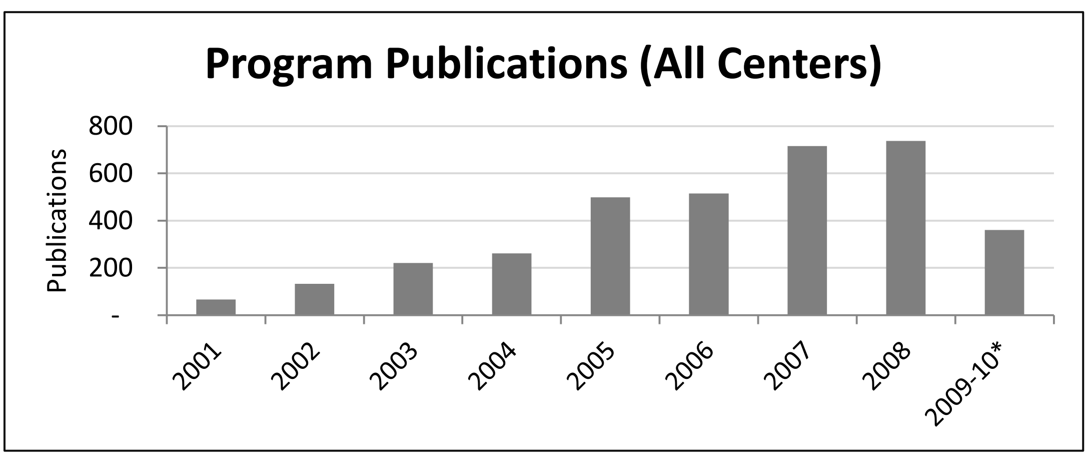
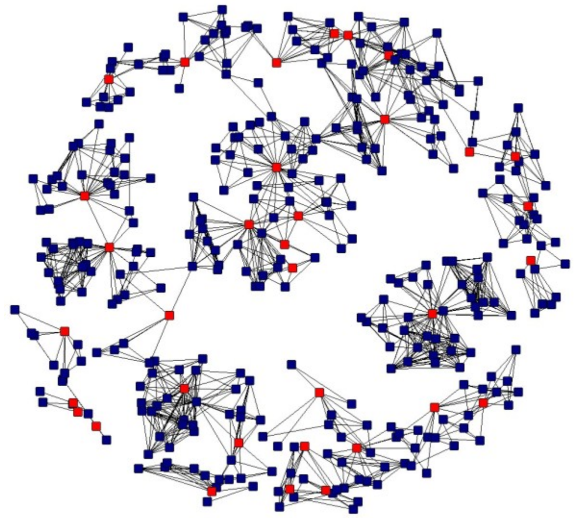
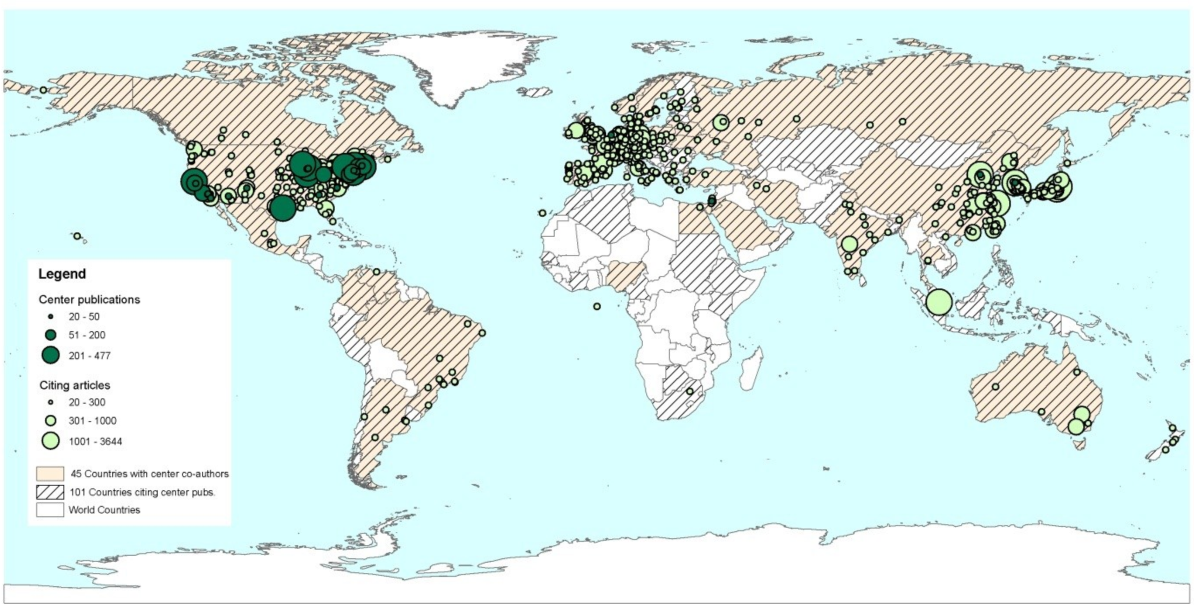


Figure 1: Retrospective Analysis -- Publication Counts

Simple publication counts show these increases in activity and output. Yet, for many funded research programs, the composition of the co-authoring teams is also critical in addressing the goals and objectives of the program. For example, many large research grants include multiple institutions, where expectations are that cross-institutional (or even cross regional or cross national) knowledge exchange and generation occur. Other programs may prioritize cross-disciplinary research and related academic production. Bibliometric data can be compiled in numeric form, but also graphical form (Figure 2), to demonstrate these linkages. In the top graphic in this figure, the darker nodes represent program-supported/affiliated researchers, while the lighter nodes show co-authors outside of the program. Overall, the data show little co-authorship among program affiliated researchers, but extensive co-authorship ties outside of this group. The evaluator may come to two competing conclusions, depending upon center strategy. On one hand, these data may show that the center has not developed very successful co-authorship ties within the program group. Conversely, depending on program strategy, these same data may be used to demonstrate the affiliation of program researchers with other institutions. Further, if multiple graphs and related statistics are shown over time, additional insight into the evolution of collaborative teams may be quantitatively assessed. A similar graphical depiction could be done by coding the co-authorship nodes by institution, sector, industrial partners, discipline, or other category. Again, this data breakout should be linked to the evaluation questions and strategy of the research program.



Program vs non-program affiliated co-authorship



Geographic co-authorship ties and citing authors

Figure 2: Examples of Graphical Displays of Co-Authorship Bibliometric Data

Citation & Co-Citation Analysis

In research evaluation, capturing the impacts of the funded research are critical in assessing the value of the research overall. As research results are published, their visibility and use within the scholarly community is indicative of their value and impact. Citation counts are generally used to address this scientific impact based on the number of times that subsequent papers reference a particular earlier paper.

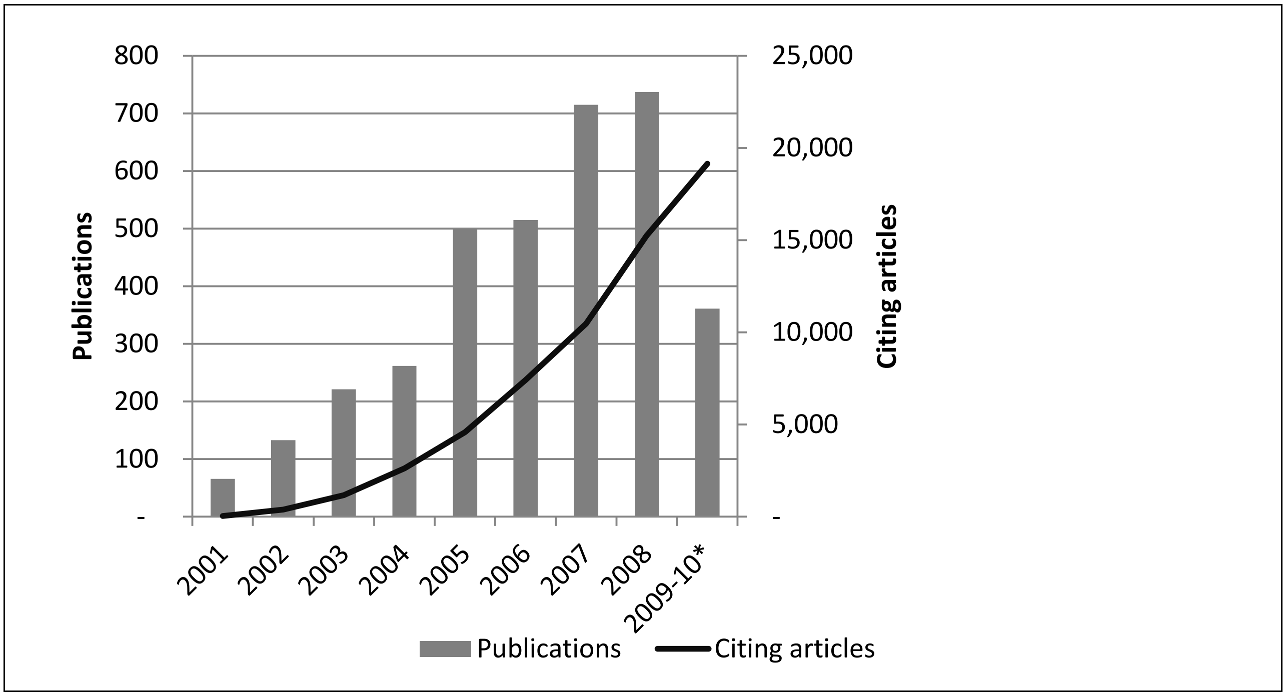


Figure 3: Example: Citation Counts Over Time (Program Level)

In Figure 3, citation counts (line graph) are shown to build over time, which is an expected function of not only the growth of program publications (bars), but also increases in visibility and knowledge spread. The evaluator may conclude that centers that have produced the publications shown in the figure have also demonstrated an increasing impact of this work as the centers developed and matured. In a research program evaluation, tracking these impacts are central to assessing the core impacts of the program's research activities and outputs. It communicates changes in magnitude and reach of program research.

Citation counts are much more powerful when placed in context. To do this one needs a sense of what is the expected rate of citation for a paper in the same field published in the same year. Ideally the group being evaluated will achieve a higher than expected rate of citation. In the example here, the centers being evaluated published 128 papers in 2002. There were 35,000 papers published in the field as a whole in 2002. By 2009, the 128 center papers published in 2002 had achieved an average citation per paper of 102, which considerably exceeded the field average of 22 citations per 2002 paper. In addition, 15 center papers are found among the top 0.1% most cited papers, representing 4% of the 128 papers the centers published in 2002. Thus the center is having a much higher than expected impact on the field. We would expect 1 center paper to be found among the 0.1% most highly cited. In fact, 15 are cited at the 0.1% level - an outstanding result.

Citation data can also be analyzed along other dimensions. For example, understanding the characteristics of the citing authors, and their affiliations (discipline, geographic location, and institution) can add important depth to the evaluative analysis. Disciplinary analysis can be especially important when disciplinary and cross-disciplinary impacts are part of the programmatic or center goals. For example, research centers or large grants are typically expected to make significant impacts on the field or fields of inquiry, and to have a broad impact geographically. Similar to publication counts, there is typically some interest in the development and growth of citation rates over time. As noted earlier in this chapter, these counts may be most useful and have the greatest face validity when citations are considered to represent "impact," both as social and cognitive influences on subsequent research (Martin and Irvine, 1983).

In research evaluation of programs and centers, the question of "added value" often arises. Is the work of the researchers funded through the specific initiative greater than the sum of its parts? Is the research that emerges from constructed and publicly supported research groups different and higher impact than the work the individual scientists were doing otherwise? Bibliometric analysis can also be useful in addressing this question using various comparisons in the analysis. Publication counts and related characteristics can be examined alongside a comparison group – a matched sample reflecting the research focus or expertise, or other values. Citation patterns can be also be explored in various way – compared to matched comparison groups, or even to the non-project publications of cited authors.

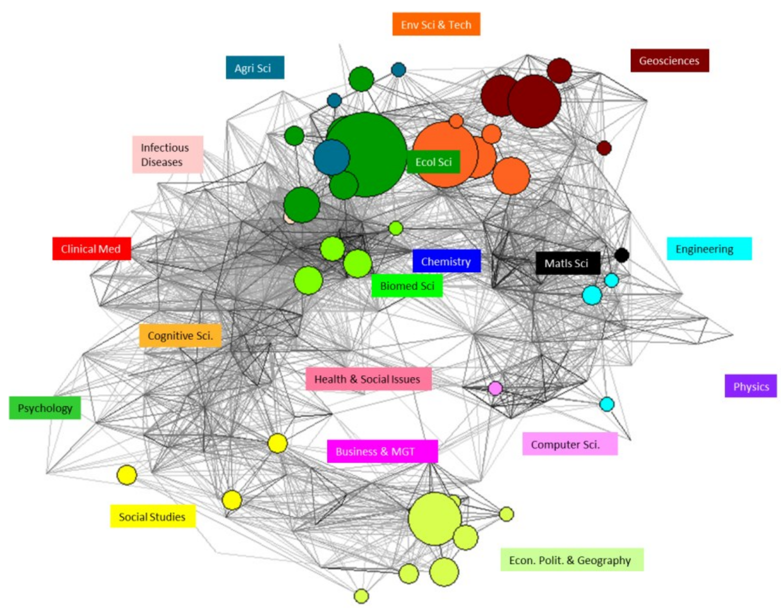
Related is the issue of intellectual leadership and visibility in any research program. To what extent are "star" scientists, or more established and visible researchers driving the production, but also the impact (citations) of the research group? Research evaluations may be designed to differentiate researchers with exemplary production and visibility from others in order to understand individual contributions to overall impacts of the research group. For example, in some evaluation scenarios, the question arises of whether a "star scientist's" work explained the growth of citations of center publications. If publications and related citations are broken out to separate this individual's work from others, "center-wide" impacts may be more accurately assessed. This example highlights the importance in having an appropriate evaluation research design in using bibliometric analysis, but also the range of ways that it may be used to answer large and small scale evaluative questions.

Finally, as noted earlier in this chapter, there is increasing interest and capacity for mapping of bibliometric data. This mapping can be organized around a range of variables, including geographic location (as shown in Figure 2 above). For example, geographic reach can be important in demonstrating regional, national, and even global knowledge flows and impacts as well as research visibility. Citation counts by state, country, or continents (and changes over time) can provide additional depth to the citation count data. Mapping software can be used to effectively communicate these results. It is particularly powerful to show citing authors across the globe and in a number of countries.

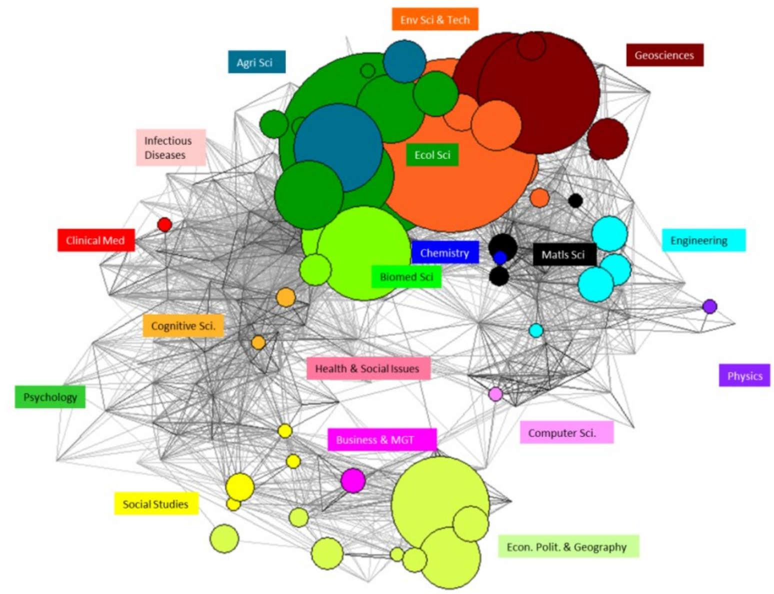
Another approach to mapping capitalizes on the development of the spatial depiction and relationship of scientific disciplines. In many large-scale funded research projects and centers, some aspect of cross-disciplinary interaction and production are of interest because they represent a specific type of knowledge integration and development. The expectation is typically that the linking of varied disciplinary perspectives on common sets of issues will yield frontier research and innovation. Research evaluation professionals have struggled with mechanisms to capture and represent interdisciplinarity (for extensive discussion on these challenges, see Wagner et. al., 2011). Recent advances involving the disciplinary coding of cited references, allow for the statistical representation of interdisciplinarity of published works (Rafols et al., 2011). Within a research evaluation context, these mechanisms may be used to illustrate the disciplinary relationships of key researchers.

For example, the upper map in Figure 4 shows the publications of the first four years of a multi-institutional and multi-disciplinary research program within the map of science. This program is multi-disciplinary, including physical and biological sciences, as well as social sciences. The size of the nodes reflect the number of publications (publication count) in a given disciplinary area, whereas their placement on the map reflects their relationship to the disciplines. As shown in the lower portion of Figure 4, examining publication patterns by discipline together with similar analysis of citing articles, can show disciplinary reach and expansion.

From the data shown in Figure 4, the evaluator may conclude that the research produced through this program is having an impact on broader disciplinary communities not only in ecological and environmental science and geosciences but also in economics, politics and geography. The pattern of citing across disciplines indicates how program impact expands in the next generation of publications. The evaluator may conclude that the impacts of the research are reaching beyond the disciplinary focus of the researchers and showing some indication of cross disciplinary impact. Finally, the data shown in Figure 4 reflect a summative "snapshot." However, tracking these relationships over time can provide additional insight into program performance and impact. While not shown here, retrospective analysis capturing publication and citation activity can use similar methodologies for more years and even more institutional groups. While these graphical depictions can be powerful in communicating results, additional detail is also possible using statistical summaries of the publications shown. Based on recent work of Rafols, Porter and Leydesdorff (2010), the statistical summary of the disciplinary compositions can also be added to the evaluative design and analysis.



Disciplinary Placement of *Program/Center* Publications



Disciplinary Placement of *Citing* Publications

Figure 5: Example: Program Publications and Citing Publications in the Map of Science

Cautionary Guidance in the Application of Bibliometrics in Research Evaluation

As with any sophisticated methodology, bibliometric research evaluation can be done well or it can be done badly. We offer the following guidelines so that evaluators are aware of the issues that might arise in conducting a bibliometric evaluation.

First, bibliometric tools are not appropriate for all levels of analysis, nor in all evaluative settings. Although becoming more popular with individual scientists who use the h-index and Google Scholar Citation Profile, the individual level of analysis requires great caution (Cole, 1989; Martin and Irvine, 1983.) Cole presents several reasons why the performance of individual scientists, and particularly the quality of the performance of individual scientists, should not be judged primarily through citation analysis. The most important reason for this is the roughness of citations as a quality indicator. Cole argues that the potential for a large margin of error is great. Some scientists may be considered by other scientists to do high quality work but may not be highly cited. The remedy for this problem is to 1) analyze research groups, not individuals and 2) combine citation analysis with other methodologies (Cole, 1989; Martin and Irvine,1983.) This also underscores the use of bibliometrics to evaluate teams or centers, rather than to identify the most impactful team members.

Second, care should be taken in the structuring of bibliometric analysis for evaluating interdisciplinary research teams, and for some disciplines. Funded research teams are increasingly multi-disciplinary and recent advances in bibliometrics allow for additional layers of analysis that captures the richness and character of these academic outputs. Yet, interdisciplinary publication patterns differ from those of individual disciplines, and citation rates differ across disciplines Further, care should be taken to draw from the appropriate publication data-sources for the analysis.

Third, a related point is that publication norms differ by field – some disciplines value conference proceedings and others books and book chapters. Therefore, a database of journal articles alone cannot accurately represent output in these areas (Hicks, 2004). The ability to conduct meaningful bibliometric analysis varies across fields, partly due to disciplinary traditions, but also in the ability to access accurate publication data.

Fourth, in the evaluation of research groups, it is always a challenge to differentiate grant or center-related publications from other work and additional coding of publications may be necessary to identify the correct units of analysis. This adds an additional layer of data identification in the evaluation process. Related, in citation analysis, the status of self-citations should be carefully considered. It may be important to remove them from citation counts to demonstrate outside impact. Alternatively, self-citations might be valued evidence of knowledge sharing within the research group.

Fifth, currency is always a challenge in bibliometric evaluation. After funding has been obtained, it takes time to perform the research, write it up, submit, revise, and for it to appear in a journal. After this, more time is needed for papers to become well known and incorporated into the ongoing research of the field and so to gather citations. In fast moving fields citation counts typically peak in the second or third year after publication, though citations build over a decade or more in the social sciences. Even in fast moving fields there will be "late bloomers", that is papers that become highly cited after a decade or more (Rogers, 2010). In time-bound grant or short-term research centers (ten years or less), bibliometric analysis will not be meaningful in the early stages of team institutionalization. Funders however, need to establish the impact of their programs in the short term. This tension always marks bibliometric evaluation and is one reason why the method benefits from being combined with other ways of collecting more immediate feedback on program performance. Longitudinal bibliometric analysis that focuses on changes over time in the center or team can enhance the picture obtained from such short-term measures.

Finally, we have not addressed all of the different ways that bibliometric data and analysis are used in research evaluations. We have purposely focused on core metrics and approaches that are well accepted and meaningful. For example, although not addressed here, an additional dimension can be added to bibliometric evaluation by assessing the extent to which scientific papers are cited in patents. Narin and CHI Research demonstrated in the late 1990s that US patents cite increasing amounts of scientific literature (Narin et al, 1997). The cited literature was from top universities, published in top journals and often was very well cited in the scientific literature as well. They interpreted this as a sign of the increasing linkage between science and technology. Patents in highly scientific areas of technology contain as many citations to the basic research as the scholarly literature. Variation in the rate of citing to scientific literature between fields of technology or countries or over time can indicate difference in how "high-tech" is a field or country.

Conclusion

Creative use of bibliometric analysis in evaluation offers an unparalleled opportunity to take advantage of the rich information embedded in the written products of scientific work to track the output and influence of funded scholars. Many metrics and techniques have been developed: from publication and citation counts to percentile rankings, h-index, impact factor, maps of the knowledge landscape, maps of geographical distribution, and metrics of interdisciplinarity and specialization. Analysis can demonstrate evolution over long periods of time, and can draw quantitative comparisons among subgroups or with others anywhere in the world. It would be dangerous to consider such data and analysis as "easy" however. Careful attention to detail and method are required to produce robust results.

Acknowledgements

Many thanks to the research organizations that supported the evaluative data presented in this chapter. Special thanks go to Juan Rogers, Eric Welch, and Jan Youtie for examples presented in this chapter. Thanks also to Luciano Kay, Agrita Kiopa and Kaspars Berzins, graduate assistants, for their contribution to this work.

References

Adler, R., Ewing J., and P. Taylor. 2008. *Citation Statistics*. International Mathematical Union http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf.

Aksnes, Dag W. and Arie Rip. 2009. "Researchers' perceptions of citations." *Research Policy* 38:895-905.

Bornmann, L., Mutz, R., Neuhaus, C. and Daniel, H.-D. 2008. "Use of citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results." *Ethics in Science and Environmental Politics* 8:93–102. doi:10.3354/esep00084

Bornmann, Lutz and Hans-Dieter Daniel. 2008. "What do citation counts measure? A review of studies on citing behavior." *Journal of Documentation* 64:45-80.

Bornmann, Lutz. 2010. "Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff paper." *Journal of Informetrics* 4:441-443.

Börner, Katy, Chaomei Chen and Kevin Boyack. 2003. Visualizing Knowledge Domains. In *Annual Review of Information Science & Technology*, volume 37, edited by Blaise Cronin, 179-255. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Cole, F. J., and N.B. Eales. 1917. "The history of comparative anatomy: Part 1.-a statistical analysis of the literature." *Science Progress in the Twentieth Century: A Quarterly Journal of Scientific Work & Thought* 6:578-597.

Cole, S. and Cole, J.R. 1967. "Scientific output and recognition: A study in the operation of the reward system in science." *American Sociological Review* 32:377-390.

Cole, S. 1989. "Citations and the Evaluation of Individual Scientists." In *Trends in Biochemical Sciences* 14:9-13.

Cozzens, Susan E. 1989. "What do citations count? The rhetoric first model." *Scientometrics* 15:437-447.

Cozzens, Susan E. 2007 "Death by Peer Review? The Impact of Results-Oriented Management in U.S. Research." In *The Changing Governance of the Sciences* edited by Richard Whitley and Jochen Gläser, 225-244. Dordrecht: Springer.

Crane, D. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communication*. Chicago: University of Chicago Press.

Cronin, Blaise. 1984. *The Citation Process: The role and significance of citations in scientific communication*. London: Taylor Francis.

De Bellis, Nicola. 2009. *Bibliometrics and Citation Analysis.* Lantham MD: Scarecrow Press.

ESF (European Science Foundation). 2009. *Evaluation in National Research Funding Agencies: Approaches, Experiences and Case Studies*. Strasbourg: Member Organisation Forum on Ex-Post Evaluation of Funding Schemes and Research Programmes.

Garfield, Eugene. 1964. '"Science Citation Index" – A New Dimension in Indexing.' *Science*, New Series, 144, 3619: 649-654.

Garfield, Eugene. 1979. *Citation Indexing*. New York: Wiley.

Gilbert, Nigel. 1977. "Referencing as persuasion." *Social Studies of Science* 7:113-122.

Gmür, Markus. 2003. "Co-citation analysis and the search for invisible colleges: A methodological evaluation" *Scientometrics* 57:27–57

Hicks, Diana M. 1987."Limitations of Co-citation Analysis as a Tool for Science Policy." *Social Studies of Science* 17:295-316.

Hicks, Diana M. 1988. "Limitations and more Limitations of Co-citation Analysis/Bibliometric Modelling: A Reply to Franklin." *Social Studies of Science* 18:375-384.

Hicks, Diana. 2004. "The Four Literatures of Social Science." In *Handbook of Quantitative Science and Technology Studies*, edited by Henk Moed, W. Glänzel and U. Schmoch, 473-496. Dordrecht: Kluwer Academic.

Hicks, Diana. 2012. "Performance-based University research funding systems." *Research Policy* 41:251-261.

Hicks, Diana, H. Tomizawa, Y. Saitoh, S. Kobayashi. 2004. "Bibliometric techniques in the evaluation of federally funded research in the United States." *Research Evaluation* 13:76-86.

Hilgartner, S. and S. I. Brandt-Rauf. 1994. "Data Access, Ownership, and Control Toward Empirical Studies of Access Practices." *Science Communication* 15:355-372.

King, J. 1987. "A Review of Bibliometric and Other Science Indicators and Their Role in Research Evaluation." *Journal of Information Science* 13: 261-276. Available: <Go to ISI>://A1987K402700001

Klavans, Richard, Boyack, Kevin W. 2005. "Identifying a better measure of relatedness for mapping science." *JASIST* 57:251-263.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Open University Press: Milton Keynes.

Latour, Bruno and S. Woolgar. 1979. *Laboratory Life*. Beverly Hills: Sage.

Luukkonen, Tertu. 1997. "Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis." *Scientometrics* 38:27-37.

MacRoberts, M.H. and B.R. MacRoberts. 1996. "Problems of Citation Analysis." *Scientometrics* 36:435-444.

Martin, Ben R. and J. Irvine. 1983. "Assessing Basic Research." *Research Policy* 12:61-90.

McCain, K. W. 1986. "Cocited author mapping as a valid representation of intellectual structure." *Journal of the American Society for Information Science* 37:111-122.

Meho, L.I. and K. Yang. 2007. "Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus Scopus and Google scholar." *Journal of the American Society for Information Science and Technology* 58:2105-2125.

Merton, Robert K. 1988. "The Matthew Effect in Science II: Cumulative Advantage and the Symbolism of Intellectual Property." *ISIS* 79:607-623.

Moed, Henk. 2005. *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

Mullins, N. C., Hargens, L. L., Hecht, P. K., and E.L. Kick. 1977. "The group structure of cocitation clusters: A comparative study." *American Sociological Review* 42:552-562.

Narin, Francis. 1976. *Evaluative Bibliometrics*. Cherry Hill, NJ: Computer Horizons Inc.

Narin, Francis, and Kimberly S. Hamilton. 1996. "Bibliometric Performance Measures." *Scientometrics* 36:293-310.

Narin F, Hamilton KS, Olivastro D. 1997. "The increasing linkage between US technology and public science." *Research Policy* 26:317-330.

Opthof, T., and Leydesdorff, L. 2010. "Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance." *Journal of Informetrics* 4:423–430.

OTA. 1986. *Research Funding as an Investment: Can We Measure the Returns?*-A Technical Memorandum. Washington, DC: U.S. Congress, Office of Technology Assessment, OTA-TMSET-36.

Price, Derek de Solla. 1963. *Big Science, Little Science*. New York: Columbia University Press.

Price, Derek de Solla. 1965. "Networks of scientific papers." *Science* 149:510-515.

Rafols, I, Porter, A. L. and Loet Leydesdorff. 2010. "Science Overlay Maps: A New Tool for Research Policy and Library Management." *Journal of the American Society for Information Science and Technology* 61:871–1887.

Rafols, Ismael, Loet Leydesdorff, Alice O'Hare, Paul Nightingale, and Andy Stirling. 2011. How journal rankings can suppress interdisciplinary research. A comparison between Innovation Studies and Business & Management. arXiv:1105.1227v3 [cs.DL]. Accessed February 3, 2012.

REPP. 2005. *Quantitative indicators for research assessment, a literature review*. Canberra: Australian National University.

Rip, A. 1988. “Mapping of science: possibilities and limitations.” In *Handbook of Quantitative Studies of Science and Technology* edited by A.F.J. Van Raan. North-Holland: Elsevier.

Rogers, Juan D. 2010. “Citation Analysis of Nanotechnology at the Field Level: Implications for R&D Evaluation.” *Research Evaluation* 19:281-290.

Seglen, Per O. 1997. “Why the impact factor of journals should not be used for evaluating research.” *British Medical Journal* 314:497.

Small, H. 1978. “Cited Documents as Concept Symbols” *Social Studies of Science* 8:327-340.

Small, H. G. 1980. “Co-citation context analysis and the structure of paradigms.” *Journal of Documentation* 36:183-196.

Stephan, P.E. and Levin, S.G. 1988. “Measures of scientific output and the age-productivity relationshsip.” In *Handbook of Quantitative Studies of Science and Technology*, edited by A.F.J. Van Raan. North-Holland: Elsevier.

Van Raan, A.F.J. 1998. “In matters of quantitative studies of science the fault of theorists is offering too little and asking too much.” *Scientometrics* 43:129-139.

Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J., and Waltman, L. 2010. “Rivals for the crown: reply to Opthof and Leydesdorff.” *Journal of Informetrics* 4:431–435.

Wagner C.S., Roessner J.D., Bobb K., Klein J.T., Boyack K.W., Keyton J., Rafols I., Borner K. 2011. “Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature” *Journal of Informetrics* 5: 14-26.

White, H.D. 1990. “Author co-citation analysis: overview and defense.” In *Scholarly Communication and Bibliometrics*, edited by C.L. Borman, 84-106. Newbury Park: Sage.

Wouters Paul. 1999. *The Citation Culture*. Thesis, University of Amsterdam.

Youtie, Jan, Luciano Kay and Julia Melkers. 2011. Using bibliographic coupling as a measure of centerness in R&D assessment. Working paper. Georgia Institute of Technology.

Zitt, Michel, et al. 2005. “Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation.” *Scientometrics* 63:373-401.

Zuckerman, H. 1987. “Citation analysis and the complex problem of intellectual influence.” *Scientometrics* 12:329-338.