August, 2004

Bibliometric evaluation of federally funded research in the United States

Diana M Hicks

Research Evaluation, volume 13, number 2, August 2004, pages 78–86, Beech Tree Publishing, 10 Watford Close, Guildford, Surrey GU1 2EP, England

Evolving indicators

Bibliometric techniques in the evaluation of federally funded research in the United States

Diana Hicks, Hiroyuki Tomizawa, Yoshiko Saitoh and Shinichi Kobayashi

Research evaluation in the USA historically tended to rely more heavily on peer review than on bibliometric method, but interest in quantitative methods including bibliometrics appears to be growing. In this paper, we discuss the use of bibliometric techniques of research evaluation by the US federal government over the past decade. Within the past decade, commentators have pointed to something of a rebirth of interest in evaluation along with pressure on agencies to develop quantitative indicators. Evaluation of economic and societal outcomes of research has become a priority. Bibliometric method continues to evolve in response to these needs and therefore often finds application in evaluations of federal agency research.

Diana Hicks is at the School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, USA and at CHI Research, Inc., 10 White Horse Pike, Haddon Heights, NJ 08035, USA; email: diana.hicks@pubpolicy.gatech.edu

Hiroyuki Tomizawa, Yoshiko Saitoh and Shinichi Kobayashi are at the National Institute of Science and Technology Policy (NISTEP), Ministry of Education, Culture, Sports, Science and Technology (MEXT), Tokyo, Japan.

THIS PAPER REVIEWS from a methodological perspective a selection of bibliometric evaluations conducted for the US federal government over the past decade. The extent and characteristics of bibliometric research evaluations in recent years reflects the policy context, which will be discussed in the next section.

The paper focuses on recent US studies, but we should place this work in its international context historically. Although large-scale, methodologically pioneering bibliometric evaluations have originated in the USA, commentators historically tended to look to Europe where governments routinely commission bibliometric evaluations from academics. Commentary therefore lamented the lack of use in the USA, pondered the cause and expressed belief that peer review and bibliometric methods should be combined. Nevertheless bibliometric methods were extensively covered in a 1985 report from the Office of Technology Assessment (OTA). The OTA and others at the time reported at length on CHI's work with the National Institute of Health (NIH) (discussed below). Six years after the first OTA report in 1991, the OTA again reported on research evaluation and this time found little advance in methods and focused more on lack of use and how studies should be structured. Both studies concluded that quantitative methods were little used in the USA in comparison with Europe, where the techniques were pervasive and practiced by academics. OTA 1985 included pages of discussion of Martin and Irvine's (University of Sussex) bibliometrics and a bibliography compiled for OTA 1991 contains a large European component (Averch, 1993).

 0958-2029/04/020078-9 US$08.00

Nevertheless, the techniques were used. In the mid-1980s, Logsdon and Rubin (1988) interviewed 44 people responsible for research management and research evaluation in 10 federal agencies. They reported widespread use of peer review, with almost every agency they contacted using the technique. Logsdon and Rubin noted that this situation is in line with the recommendations the National Academies made in a 1982 report examining evaluation methods.[1] Logsdon and Rubin also found extensive use of bibliometrics; half of the agencies used bibliometrics in some form (Logsdon and Rubin, 1988, Table 2).

Bibliometric techniques seem to have been in use but not perceived to be in use. Two factors may underlie this disparity: a fragmented system and publication habits. US research funding agencies and so research evaluation are famously fragmented in comparison with other countries. Also evaluation performers are varied and scattered, including multiple in-house capabilities as well as contracting out to universities, non-profits and companies (Melkers and Roessner, 1997). The more centralized activity found in European countries is more visible.

Traditionally in the USA, companies such as Thomson-ISI, SRI and CHI Research, Inc. innovated in bibliometric methodology. Their counterparts in Europe and Australia were academics.[2] These American performers may have been less visible in the USA because they published less, or published in places less visible to the US academics who were not involved in bibliometrics but who wrote reviews of evaluation. This may have underpinned a certain reluctance by American academics to accept bibliometrics as a methodology let alone as an area in which foreign academics and US firms led. While US academic studies deploying econometric methods became ever more sophisticated, some US bibliometric work has had an improvisational flavor and languished far behind the state-of-the art. With interest increasing in databases as access broadens, US academia is becoming more involved in bibliometrics — using both patents and papers.

This comes at a time in which overall interest in research evaluation appears to be increasing in the USA. Before examining examples of recent bibliometric evaluations, we explore the policy context and the factors behind the changing nature of bibliometric research evaluation.

Context

In recent years, research evaluation at the federal level has increased. Melkers and Roessner (1997) point to a 'rebirth' of evaluative activities both at the federal and state levels. Contributing to renewed interest in evaluation have been new accountability requirements placed on federal agencies under the GPRA and PART programs. CHI's records suggest that bibliometrics has been part of this rebirth; Figure 1 reports the number of studies conducted by CHI Research for agencies of the federal government in three eight-year time periods: 1977–84, 1985–92 and 1993–2000. The timing of the OTA reviews is noted. Several prominent evaluations that included bibliometrics were conducted at the federal level in the late 1990s; they are discussed below.

The increased interest in evaluation has been accompanied by a growing emphasis on evaluating research outcomes as opposed to research outputs. Research outputs are the traditional products of research such as new knowledge and scientific excellence as manifested in publications and scientific reputations. The term 'outcome' is used to mean the effect of research on society or the economic benefits, new technologies, environmental improvements, etc. that the country gains from its investments in research. Today research for the sake of research is of considerably less interest than is research more directly connected with technology or other application. Evaluating outputs has been the strength of peer review; the movement to outcomes sidelines peers in favor of users. Similarly, traditional paper–paper bibliometrics has been seen as somewhat

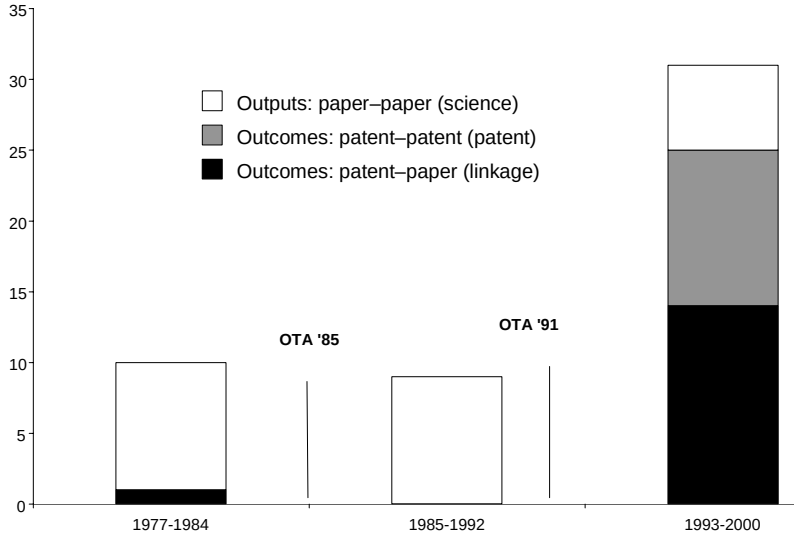

Figure 1. Number of evaluations conducted by CHI for the US federal government

irrelevant to the evaluation of research outcomes (Georghiou and Roessner, 2000; quoting Cozzens *et al*, 1994).

However, bibliometrics has evolved as research has changed. The emphasis on outcomes in research evaluation in part represents an evolution in the nature of research so that today much more research is much more closely linked to technology, in the life sciences or information technology, for example. Narin has used evidence based on patterns of patents citing papers to argue that science and technology are becoming intermingled (Narin and Noma, 1985). Stokes argued that science policy must transcend the basic/applied distinction because so much work today is both basic and relevant to application, as was Pasteur's work (Stokes, 1997). CHI's examination of papers cited in patents supports Stokes argument in that the work cited is in basic journals, and tends to be highly cited in scientific papers as well. *Nature* and *Science* are two of the journals most intensively cited in patents (Hicks *et al*, 2000). The increase in top-quality research with close links to outcomes makes it easier to demonstrate those links because the increasing rate of citation from US patents to papers has made more visible the connections between research and technology. Patent-paper bibliometrics has thus become available as an outcome-related evaluation tool, and patent-paper bibliometrics account for a significant portion of CHI's work for the federal government in the 1990s — see Figure 1.

Another factor may be at work to increase the use of non-peer review evaluation in general and bibliometrics in particular, that is increased complexity in the nature of granting programs. The focus on outcomes has accompanied policy interest in science–industry links. This has led to policy and programs to try to increase linkage — such as Cooperative Research and Development Agreements (CRADAs) that provide a framework for national laboratories to work with industry with the goal of increasing technology transfer.

Evaluations conducted in the late 1990s reflect this. Evaluations conducted by Abt Associates address the new complex programs (small business innovation research; design, manufacture, and industrial innovation program; science and technology centers program; engineering research centers). Mechanisms such as these, with goals that go beyond fostering research excellence in traditional scientific disciplines, may well need a more comprehensive and structured external evaluation than is provided by peer review. Traditional paper–paper bibliometrics alone tends also to be insufficient, which may explain a change in CHI's work. In the 1980s, CHI's paper–paper or science bibliometrics tended to stand alone. In the later half of the 1990s, CHI's paper–paper bibliometrics were combined with case study methodology in evaluations performed collaboratively. This movement to multiple methods suggests that more complex methods are required to evaluate more complex programs.

In order to illustrate some of the themes discussed here we turn to examples of US evaluations that have used bibliometrics. We begin by describing work by CHI Research for the National Institutes of Health in the early 1980s, followed by a study from the National Academies exhibiting tension between peer and bibliometric evaluation. The harmonious combination of qualitative and quantitative methods is illustrated by two studies with contrasting sophistication on the qualitative and quantitative sides. We also examine a mid-1990s study evaluating both outputs and outcomes that clearly illustrated the usefulness of an agency's database. Well-funded in-house bibliometrics to support ongoing decision-making is examined next. We finish with two recent studies employing innovative newer techniques that point the way to the future.

State of the art in the 1980s

To provide methodological perspective, we step back in time to the early 1980s when the National Institutes of Health commissioned a series of bibliometric studies from CHI Research. This work featured prominently in reviews written in the 1980s, including the 1986 Office of Technology Assessment report, because the analysis was innovative and comprehensive even by today's standards. The studies built upon a specialized bibliometric database CHI constructed for NIH containing funding acknowledgements for papers in leading biomedical journals as well as citation and institutional information.

In one typical study, CHI reported on the National Institute of Child Health and Human Development (NICHD). The results included:

- NICHD's major impact was in the subfields of endocrinology, obstetrics and gynecology, pediatrics, anatomy and morphology and embryology. Although support of biochemistry and molecular biology was emphasized (17% of NICHD-sponsored papers) the Institute accounts for only 1.6% of the papers in that subfield.
- NICHD's 'activity index' — a measure of its publications in a subfield in relation to all publications in that subfield — was highest for embryology at 7.2 times the expected level.
- NICHD supported more papers than any other NIH institute in the subfields of fertility, obstetrics–gynecology, pediatrics, anatomy–morphology and embryology.
- Subfields exhibiting large changes in NICHD supported papers were: nutrition and dietetics — 33% increase; cell biology/cytology/histology — 31% increase; and fertility — 19% decrease. Research in geriatrics declined with the creation of the National Institute on Aging in 1974.
- NICHD's record of supporting outstanding investigators is manifest in citations to its publications. In 15 of the 16 subfields in which it was most

active, 10% or more of the NICHD's papers were in the top decile of cited papers. This indicates that the quality of its programs rose or had been maintained despite funding restrictions in place at the time.
- That NIH institutes are interdependent in their research was clear from interaction in citation. Only 20% of references in NICHD-supported papers were to research supported by that institute; the rest were to papers supported by others. Conversely, 84% of the citations to NICHD research were in papers with other support. Interdependence was strong among NICHD, NIADDK, NCI and NIGMS.

This study was one in an ongoing series providing comparable profiles across NIH institutes in the early 1980s. The comprehensiveness and methodological sophistication of this work remains unmatched in recent years in the USA.

Tension

COSEPUP (2000) benchmarked US research in a report that revealed tension between peer review and bibliometric method. International benchmarking compares the quality and impact of research in one country (or region) with world standards. This report sought to experiment with a benchmarking methodology to evaluate the research-leadership status of the United States.[3] Although it could have built upon the bibliometric resources in NSF's *Science and Engineering Indicators*, this report did not. Indeed, the report denounces bibliometric indicators.[4] The report states that such indicators are useful, but by themselves inadequate because (and this is a full recount of the reasoning relevant to bibliometrics) 'for example a paper that describes truly innovative research may receive few citations if no one else is doing comparable work'. Therefore, 'expert judgment of panel members afforded the most effective means for assessing research' (pages 6–7). Benchmarking, however, is commonly thought of as a quantitative exercise. Thus a certain tension is discernable in the work of the scientists on the panels.

COSEPUP chose three areas for study: mathematics, immunology, and materials science and engineering. In each area COSEPUP appointed a panel of eminent scientists to produce a report. The methods used by the panels were:

- The virtual congress — panel members called leading experts in subfields to ask who were the five to 20 best people in the world and constructed an indicator from their country affiliations.
- Esteem indicators — the panels counted numbers of prize recipients and representation among conference speakers by country.
- Journal publication analysis — the panels scanned five journals and tabulated the locations of principal investigators and their subfields.
- Citation analysis — the panels of scientists turned to an existing British bibliometric analysis. The immunology panel also bought a 'high impact' immunology database from Thomson-ISI, classified prolific authors by country and ranked them by citation counts.

Each panel concluded that the USA was at least among the world leaders in its field. However, each panel also identified subfields in which the USA lagged the world leaders. Each panel identified key infrastructure concerns.

The panels used bibliometrics; if they had not the report would hardly be credible as a benchmarking exercise. However, the panels lacked access to or information on existing US data sources or the expertise to produce a bibliometric study tailored for their needs. Scanning five journals is a limited and crude technique, needless to say far behind the methodological frontier as it stood in the late 1990s.

As a methodological experiment the report also commented on methodological issues. The report recognized and wrestled with the inherent subjectivity involved in asking very senior US scientists to decide whether their life's work had led to US leadership in their field. The work also was said to be cost-effective, which is important in a method proposed for routine use. However, a number of very senior scientists both domestic and foreign made substantial time commitments to produce the report. Since they were not paid for their time, the cost of their participation was implicitly born by their institutions and the agencies awarding them their grants. Thus, the procedure's true costs were understated by including only direct billings to the Academy. Kostoff's handbook of evaluation (1997) includes a calculation of full costs for studies of this type.

Harmonious combination

A more harmonious combination of qualitative and bibliometric methodology was illustrated by two studies conducted by Roessner *et al* (1997, 1998) for the National Science Foundation. These probed how NSF support for engineering, especially research

A number of very senior scientists made substantial time commitments to produce the report. Since they were not paid for their time, the cost of their participation was implicitly born by their institutions and funding agencies

and related activities, contributed to the development and commercialization of recent, significant innovations. To trace the impact of NSF research on technology, these studies used the retrospective case study approach pioneered in the Hindsight and TRACES studies of the late 1960s and early 1970s (IITRI, 1968; DoD, 1969). The innovations studied here were: Internet, magnetic resonance imaging (MRI), reaction injection molding (RIM), computer-aided design applied to electronic circuits (CAD/EC), optical fiber for telecommunications and analog cellular phones.

The method involved:

1. Identifying the technologies that underpinned each innovation and deciding which were unique to the innovation and which were supporting technologies that already existed.
2. Library search of online databases to find all major works describing the development of the technologies.
3. Institutional analysis — identifying major organizations that developed the technologies from the library search, discussions with NSF staff, interviews and searches of NSF's awards database.
4. Personal and phone interviews — those identified using the above techniques were interviewed about the history of the technology and NSF's role.
5. Bibliometrics — described in detail below.

In nearly all six cases, Department of Defense and other government agency support for R&D was important. Without exception the cases revealed the essential role that government support of education and training, especially graduate education, had on engineering innovation. Indeed, if a consistent pattern stood out across all six cases, it was the critical role played by human capital — individual inventors, technical entrepreneurs and students trained in the state-of-the-art. Regulatory policy shaped the course of innovation in the cellular phone and RIM cases. Fundamental research was found to play a supportive rather than central role in the six cases of engineering innovation.

NSF emerged consistently as a major, often the major, source of support for education and training of the PhD scientists and engineers who went on to make major contributions to each innovation. NSF's support of university research infrastructure emerged as the likely second most influential activity. NSF's direct research support was key to successful innovation in just one case: CAD/EC. NSF's research support produced knowledge essential to the evolution of all cases. NSF's organizational leadership was commanding, highly visible and unique in the Internet case. In about half the cases, NSF shaped the evolution of research areas by encouraging university researchers to address problems relevant to industry using workshops and symposia, which brought university and industry people together to discuss promising areas.

The bibliometric results by area were:

- RIM — 294 patents were found using keywords 'reaction injection molding'. Patents cited seven or more times were virtually all related to RIM and were issued to private firms. The percentage of patents with references to scientific literature suggested the area was science-linked when compared against Narin's published data.
- NMR — Thomson-ISI's research front database was used to produce a map of specialty clusters in the 1990 Research Front Database. The map showed two sets of interrelated specialties related to imaging and spectroscopy respectively. The strong link between the two areas confirmed findings from interviews. One tentative conclusion from this experiment was that the bibliometric mapping technique represents a potentially helpful tool for bounding the research themes relevant to an innovation.
- Cellular telephony — papers — five names were found as citing or cited authors in the 1988, 1989 and 1990 Research Front Databases. Their number of papers in the Research Front Database was counted. One of the 46 clusters associated with the papers contained papers by a number of pioneers in the field, suggesting that searches of this type would prove a valuable way of validating and extending the range of knowledge at the inception of a case study.
- Cellular telephony — patents — 10 highly cited patents were identified in online searching. Patents cited by these 10 and cited by those and so on were retrieved. A co-citation analysis of these patents was conducted and a map was drawn confirming the centrality of AT&T and Motorola in the field. The map contained distortions as all multi-dimensional scaling maps do (the two most closely linked patents were not the closest together on the map), but nevertheless was thought to provide a useful overview of the patented technologies underlying cellular telephony, including an intimation of the digital technology to come.

The cases offer rich detail and a subtle understanding of NSF's role in enabling technological innovation that can serve to enhance decision-making and further research. The depth and richness of the qualitative work contrasts with the bibliometrics which seemed not to get very far, producing largely what a filter of patents would have done.[5] The step of bounding the area of study with a filter or a map would ideally be the first step, not the end point. Lack of normalizations hampered interpretation of data produced.

Filtering and normalized indicators

The method of starting with field definition and filters and using normalized indicators to produce

strong interpretations was illustrated in an unpublished study by Albert (1996). This study evaluated the contribution of a federally maintained database to research developments in a particular field. The database analyzed was the standard reference database Thermodynamic Properties of Refrigerants and Refrigerant Mixtures (REFPROP) maintained by the National Institute of Standards and Technology (NIST). The research field examined was CFC (chlorofluorocarbon) replacement.

The study sought to evaluate the impact of the REFPROP database on publishing and patenting in CFC replacement. It was therefore necessary to define the research field of CFC replacement. This was achieved using a technology filter comprising keywords and, in the case of patents, International Patent Classifications. An additional stage was added to the process of identifying relevant patents. This involved examining all of the patents that cited to the patents identified by the initial filter and determining which of them should be included. These additional patents were building on CFC-replacement technology and may represent the next generation of refrigeration technologies that do not mention CFCs explicitly. The patents and papers were then split into two groups according to whether or not the organizations producing them subscribed to the REFPROP database. This grouping allowed for a comparative analysis of the impact of the REFPROP database.

The study showed that REFPROP subscribers held just over half of the US patents and just under half of the scientific and technical papers in CFC-replacement technology. The influence of REFPROP was particularly strong among the large patenting and publishing organizations. Of the top 10 patenting organizations in this field, seven were subscribers to REFPROP, including four of the top five. Similarly, five of the top 10 publishing organizations in this field were subscribers, including all three commercial organizations in the top 10.

REFPROP subscribers were also responsible for almost two-thirds of the highly cited US patents in CFC-replacement. This was a higher percentage than expected, given that REFPROP subscribers accounted for just over half of the US patents in this area. It suggested that REFPROP subscribers tend to produce research whose impact was greater than average.

In this study, bibliometric method was used to assess the impact of a database on research and technological development. Bibliometric analysis might be assumed to be ineffective in assessing the impact of such resources. However, this bibliometric study made visible the link between the database and high-performing research organizations. The study would have been stronger if more methods had been brought to bear. For example, the contents of papers and patents could have been analyzed for explicit mention of REFPROP information, or interviews could have probed how often researchers referred to the database in their daily work. Best practice would recommend combining qualitative and bibliometric method.

Bibliometrics to evaluate societal outcomes

Research agencies in recent years have sought to achieve with their funding more complex, outcome-related goals including fostering the participation of underrepresented minorities in research. An unpublished study by Thomas (1999) used bibliometrics to assess such a program. The bibliometric methodology was combined with qualitative analysis in a final report to the sponsor. The National Institutes of Health program entitled Research Centers in Minority Institutions (RCMI) was set up to enhance the biomedical and behavioral research capabilities of academic institutions in which a large proportion of students come from minority groups. RCMI funding was first awarded to institutions in 1985. The purpose of the bibliometric evaluation was to establish whether the RCMI funding had led to an increase in the quality and quantity of research publications produced by these minority institutions.

The evaluation was based on analysis of the publication records of institutions that had received RCMI funding for at least ten years. Two time periods were selected for the analysis — the period immediately before RCMI funding was awarded (1981–84) and a period after institutions had received funding for a number of years (1993–97).

Many characteristics of institutions' publications were analyzed. These include: number of publications; quality of journals in which papers were published; citation impact of papers; and percentage of papers co-authored with other institutions.

The institutions were separated into three groups depending on the extent of their research experience prior to the receipt of RCMI funding. The results showed that:

- the least experienced institutions had not increased the quantity or impact of their publications. However, they had increased their level of co-authorship, suggesting that RCMI funding had allowed them to extend their links within the scientific community.
- RCMI funding had the greatest impact on institutions with a modest level of previous research experience. Institutions with modest levels of output in the first period wrote more papers, published these papers in higher-quality journals, and received more citations from later research in the second period.
- The publication records of institutions with the highest level of prior research experience did not change markedly after RCMI funding was awarded. This may be because RCMI funding represented a smaller percentage of their overall research budget.

In establishing that prior research experience may affect the impact of research funding, this bibliometric study provided information useful to managers designing future funding programs. In future to achieve maximum impact on research output, it may be best to target this type of money on institutions with modest levels of research experience. The least experienced institutions may need another type of program.

The bibliometric evaluation formed part of a larger study into the impact of the RCMI program. The larger study included analysis of RCMI-funded institutions' success in competing for grants, their number of research staff and research students, and their development of infrastructures to support research. Thus the full study illustrates best practice of combining sophisticated, normalized bibliometric analysis with qualitative method.

Well-supported in-house bibliometrics

A somewhat different thread in bibliometric work is represented by methods supporting routine in-house decision-making. The early work of CHI for the NIH, although it was contracted out, exemplifies this. More recently, Kostoff (2000) has championed this type of analysis, developing bibliometric software tools to support routine agency decision-making. Kostoff was for many years director of the Office of Technical Assessment at the US Navy Office of Naval Research and in his work used quantitative evaluation techniques. Recently he has applied techniques of text mining to management of science and technology. These techniques are often used in commercial competitive intelligence work, but are often too micro-level for government evaluation studies.

In 1998 Kostoff conducted a prototype implementation of a text mining approach. This involved several steps:

1. The iterative development of a filter to identify papers associated with a technical theme (such as Fullerenes or ship hydrodynamics). Six databases were used as sources.
2. The frequency with which all words and phrases appeared in the documents was computed. Topical experts selected the useful phrases.
3. For each useful phrase, a dictionary of closely associated phrases was constructed by counting the number of times all other phrases occurred in close proximity to each useful phrase. Each associated phrase was assigned a measure of the strength of its association to the useful phrase. A threshold was used to filter out the most closely linked phrases. Topic experts identified the themes and their conceptual relationships.
4. Analysis identified pervasive technical themes in the database, the relationship among the themes and the relationship of supporting sub areas.
5. Bibliometric analysis using authors, journals, addresses, etc. related the themes to performers.

Bibliometrics identified the location of critical infrastructure in each technical area. This was useful for finding experts for workshops and review panels and for planning visits

Bibliometrics identified the location of critical infrastructure in each technical area. This was useful for finding experts for workshops and review panels and for planning visits. Bibliometrics also tracked productivity, impact and the critical intellectual heritage. Because no norms were available, it was important to compare bibliometrics across disciplines so that anomalies in any one could be spotted and universal trends identified. The resolution of the categories was an important parameter in the study. Finer categories ('welded titanium alloys' rather than 'materials', for example) are more useful, but are more expensive and time-consuming to construct.

The process centered on the subject experts and so the conclusions reflected experts' biases and limitations. For a credible analysis that detects the maximum number of data anomalies, experts with diverse knowledge are required and a generalist is needed to identify unique patterns in a technical domain that the domain expert might not recognize as unique. From an organization's long-range strategic viewpoint the main output is not the documents generated but rather the broadening of the experts' perspectives. There was a steep learning curve for the experts, who had to learn how to use the tools to address the study's objectives and how to analyze and interpret the information produced.

Kostoff believes that all S&T management decision aids are inter-related and need to be integrated to support S&T strategic management. Thus a program peer review should be accompanied by metrics to gauge progress toward strategic goals; should have roadmaps to place the program under review in its larger spatial and temporal context; should have text mining to insure roadmap comprehensiveness and so on.

Recent innovations

Finally, we examine two recent studies that employ innovative new techniques and highlight likely future methodological directions. In an unpublished study, Hicks (2000) applied new techniques of Internet-based analysis. This study used the Internet

to attempt to assess an outcome, namely whether an institute's research was influential in environmental policy circles. Referencing patterns on the web were used to identify authoritative websites, i.e. those receiving many references from websites that were themselves authoritative. The algorithm used has a precursor in CHI's influence methodology developed by Pinski and Narin (1976) in the mid-1970s.[6] The technique begins with pages found in a standard search engine search and augments them with pages that link to and from these pages. Several attempts were made and it was discovered that the most favorable results were obtained when precisely targeted search terms were available (for example 'transboundary air pollution' rather than 'environmental problem') and the terms were related to pages on the institute's website that offered resources — such as software or data — as opposed to brochure pages.

The analysis of authoritativeness of websites suggested two things; first that the institute's resources were used in the policy sphere, and second that there were few if any comparable institutes. The institute's resources were found to be explicitly referenced on the web, hence we know that they were influential. The institute's resources were used in the international policy sphere, as evidenced by co-referencing of the institute's web pages with those from international policy-making bodies. That the institute occupied a unique position in the world was deduced from the observation that there were few research organizations on the lists of web authorities in areas of international environmental policy-making. These authority lists comprised government agencies and non-governmental organizations.

It is fortuitous for outcome assessment, particularly for those aiming to influence public discussion, that these techniques are becoming available. In this example, they seemed to confirm the institute's policy influence, something that has been impossible to validate quantitatively before. However, without normalization it is unclear whether the results obtained were evidence of strong, normal or weak performance. The study also showed that the results obtained were extremely sensitive to the search phrase used. Furthermore, it was possible to influence the results by structuring a website with many pages that referenced each other. Finally, search engine coverage is unstable day-to-day, which is not conducive to high-quality evaluation.

Boyack and Borner (2003) provided an example of evaluation employing three-dimensional landscape software. This study sought to analyze and visualize the impact of government funding on the amount and citation counts of research publications. The study examined an extramural program of the National Institute on Aging. Publication and grant data were compiled and connected. In addition, citation information was available for some particularly highly cited authors. Tables were produced reporting trends in funding and publication patterns by topic area and journal. In addition, the publications and grants were clustered by assessing the similarity of the words in their titles. The results were portrayed on three-dimensional landscape maps. The peaks on the maps were labeled with research topic areas such as: Alzheimer's disease, nursing homes, etc. The size of peaks corresponded to number of publications. In addition funding data were portrayed using two types of markers, one for big grants and one for small. Finally arrows were used to indicate whether number of publications was growing. The maps were examined to find correlations and deduce the impact of funding.

The authors conclude that their results were inconclusive. Their underlying data was not comprehensive, which limited the analysis. They suggest that grant agencies require principal investigators to fill out forms with reams of information on publishing, co-authorship patenting, etc. to build clean databases for this type of work. Though it would solve the software engineer's problem, this approach is likely to be unpopular with the principal investigators. In addition, the technique of trying to read correlations between data sets from a 3-D map seems sub-optimal. The results could have been better presented using mature, low-overhead techniques such as properly organized tables or charts plotting grant size against growth in number of publications. The analytical added value of the 3-D picture remains unclear.

Conclusion

Bibliometric techniques have found application in research evaluation in the USA, although historically peer review played a more prominent role. Interest in quantitative methods including bibliometrics has grown over the past decade under pressure from the GPRA and PART initiatives. In addition, increasing interest in evaluation of outcomes accompanied by the enhanced visibility of links between science and technology has encouraged development of new bibliometric methods. The increasingly complex goals of recent federal R&D funding programs have encouraged collaboration among evaluation providers to deliver evaluations at the state-of-the-art in both bibliometric and qualitative techniques. These trends were illustrated by examining a variety of evaluations undertaken in the past decade that incorporated bibliometric technique. These demonstrated that bibliometric method continues to evolve in response to changing circumstances and therefore remains essential to evaluation of federal agency research.

Acknowledgements

We are grateful to an anonymous referee for extremely useful comments. This paper is drawn in large part from a report by the Japanese National Institute of Science and Technology Policy (NISTEP) entitled: *Quantitative Methods of Research Evaluation Used by the US Federal Government* (Hicks *et al*, 2002). The

authors gratefully acknowledge the support provided by NISTEP and MEXT for the work that underpins this analysis.

Notes

1. Not infrequently, the National Academies were contracted to perform peer reviews.
2. In the USA there has always been an interest in bibliometrics from academic information scientists and librarians. But the evaluative use of the techniques were innovated by the companies.
3. Although it was described as experimental, the Royal Society had conducted such an exercise a decade earlier (Advisory Board for the Research Councils, 1986).
4. Routine in National Academy publications.
5. A filter being a protocol for delineating patents in a technology of interest using available information such as words, classification codes and citation relationships.
6. The Pinski and Narin work is a precursor of Google and is cited on their methodology page.

References

Advisory Board for the Research Councils (1996),*Evaluation of National Performance in Basic Research* (Department of Education and Science, London).

M Albert (1996), *Association between NIST's REFPROP Database and Innovativeness in CFC-Replacement Technology*, Final Report to NIST on CHI-9621 (CHI Research Inc., Haddon Heights, NJ), December.

H Averch (1993), 'Annotated bibliography on evaluation of research, 1985–1990', in B Bozeman and J Melkers, *Evaluating R&D Impacts: Methods and Practice* (Kluwer, Dordrecht), pages 279-300.

K W Boyack and K Borner (2003), 'Indicator-assisted evaluation and funding of research: visualizing the influence of grants on the number and citation counts of research papers', *Journal of the American Society for Information Science and Technology*, 54(5), pages 447–461.

Committee on Science, Engineering, and Public Policy (COSEPUP) (2000), *Experiments in International Benchmarking of US Research Fields*, (National Academy Press, Washington DC). Available at: <http://books.nap.edu/html/exp_in_bench/pdf/>

S Cozzens, S Popper, J Bonomo, K Koizumi and A Flanagan (1994), *Methods for Evaluating Fundamental Science*, RAND/CTI DRU-875/2-CTI, Washington, DC.

Department of Defense (1969), *Project Hindsight*, Office of the Director of Defense Research and Engineering, DTIC Report No. AD495905, October. See also C W Sherwin and R S Isenson (1967), 'Project hindsight: defense department study of the utility of research', *Science*, 156, pages 1571–1577.

L Georghiou and D Roessner (2000), 'Evaluating technology programs: tools and methods', *Research Policy*, 29, pages 657–678.

D Hicks (2000), Final report to Abt Associates, CHI-9912 (CHI Research, Inc., Haddon Heights, NJ), January.

D Hicks, P Kroll, F Narin, P Thomas, R Ruegg, H Tomizawa, Y Saitoh and S Kobayashi (2002), *Quantitative Methods of Research Evaluation Used by the US Federal Government*, NISTEP Study Material No. 86, Second Theory-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP) (Ministry of Education, Culture, Sports, Science and Technology, Japan),153 pages, May. Also annotated and translated into Japanese.
Available at: <http://www.nistep.go.jp/achiev/ftx/eng/mat086e/idx086e.html>

Illinois Institute of Technology Research Institute (IITRI) (1968), *Technology in Retrospect and Critical Events in Science* (National Science Foundation, Washington DC).

R N Kostoff (1997),*The Handbook of Research Impact Assessment*, Seventh Edition, DTIC Report Number ADA296021 (Office of Naval Research: Arlington, VA), summer. Available at: <http://www.dtic.mil/dtic/kostoff/index.html>

R N Kostoff (2000), *Implementation of Textual Data Mining in Government Organizations*, presented at the Federal Data Mining Symposium and Exposition, Washington DC, March.

J M Logsdon and C B Rubin (1988), 'Research evaluation activities of ten federal agencies', *Evaluation and Program Planning*, 11, pages 1–11.

J Melkers and D Roessner (1997), 'Politics and the political setting as an influence on evaluation activities: national research and technology policy programs in the United States and Canada', *Evaluation and Program Planning*, 20(1), pages 57–75.

F Narin and E Noma (1985), 'Is technology becoming science?', *Scientometrics*, 7(3), pages 369–381.

Office of Technology Assessment (1986), *Research Funding as an Investment: Can We Measure the Returns? — A Technical Memorandum* (Washington, DC, US Congress, Office of Technology Assessment, OTA-TM-SET-36), April. Available at: <http://www.wws.princeton.edu/~ota/disk2/1986/8622_n.html>

Office of Technology Assessment (1991), *Federally Funded Research: Decisions for a Decade*, OTA-SET-490 (Washington, DC, US Government Printing Office), May.
Available at: <http://www.wws.princeton.edu/~ota/>

G Pinski and F Narin (1976), 'Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics', *Information Processing and Management*, 12(5), pages 297–312.

D Roessner, B Bozeman, I Feller, C Hill and N Newman (1997), *The Role of NSF's Support of Engineering in Enabling Technological Innovation*, first year final report for National Science Foundation (SRI International, Arlington, VA), January. Available at: <http://www.sri.com/policy/stp/techin/>

D Roessner, R Carr, I Feller, M McGeary and N Newman (1998), *The Role of NSF's Support of Engineering in Enabling Technological Innovation: Phase II*, final report to National Science Foundation (SRI International, Arlington, VA) May. Available at: <http://www.sri.com/policy/stp/techin2/>

D Stokes (1997), *Pasteur's Quadrant: Basic Science and Technological Innovation* (Brookings Institution Press, Washington DC).

P Thomas (1999), *Bibliometric Evaluation of the Research Centers in Minority Institutions (RCMI) Program*, report to Quantum Research Corporation, CHI-9814 (CHI Research, Inc., Haddon Heights, NJ), November.