ELSEVIER

Research Policy 24 (1995) 869-882

research policy

Research requirements for research impact assessment ☆

Ronald N. Kostoff *

Office of Naval Research, 800 N. Quincy St., Arlington, VA 22217, USA

Final version received July 1994

Abstract

This paper describes research required to advance the state of research impact assessment. Generic research requirements, such as certification, quality, motivation, and review frequency are discussed initially. Then, research requirements for retrospective methods (such as projects Hindsight and TRACES), qualitative methods (such as peer review), and quantitative methods (such as cost-benefit analysis and bibliometrics) are described.

1. Background

In research sponsoring organizations, the selection and continuation of research programs must be made on the basis of outstanding science and potential contribution to the organization's mission. Recently, there have been increasing pressures to link science and technology programs and goals even more closely and clearly to organizational as well as broader societal goals. This is reflected in a number of studies (Brown, 1992; Carnegie, 1992; National Academy of Sciences (NAS), 1992), in the controversial National Institutes of Health strategic planning process, in the controversial statements by the previous National Science Foundation director about closer alignment with industry and other government agencies, and in conversations with numerous government officials.

In tandem with the pressures for more strategic research goals are motivations to increase research assessments and reporting requirements to insure that the increasingly strategic research goals are being pursued by proposed and existing research programs. The 1992 Congressional Task Force report on the health of research (Brown, 1992) stated, as one of its two recommendations: "Integrate performance assessment mechanisms into the research process using legislative mandates and other measures, to help measure the effectiveness of federally funded research programs".

According to the statement of Genevieve Knezo, a Congressional Research Service representative, at a 1993 research assessment colloguium:

> The House Science Committee has asked the Congressional Research Service to develop some options for legislative language that might be included in the mandates of the agencies

☆ The views expressed in this paper are solely those of the author and do not represent the views of the Office of Naval Research.

* Tel. 703-696-4198; Fax: 703-696-4274; Internet: kostofr@onrhq.onr.navy.mil

SSDI 0048-7333(94)00798-5

> that they have responsibility for, which would require or in some way discuss the need for R&D to be evaluated. We are exploring that right now. We have a task force in Congressional Research Service composed of about twelve people who are surveying the agencies for which the Committee has responsibility. We are also surveying agencies outside the jurisdiction of the House Science Committee, DOD and NIH specifically. (Knezo, 1993)

The Government Results and Performance Act of 1993 (Public Law 103–62) was passed on 3 August 1993. This Act provides for the establishment of strategic planning and performance measurement in the Federal government, and for other purposes. Not only will the Federal agencies be required to establish performance goals for program activities, but as the law states, they will be required to establish performance indicators to be used in measuring or assessing the relevant outputs, service levels, and outcomes of each program activity.

Due to increased world competition, and the trends toward corporate downsizing, parallel pressures exist for industrial research organizations to link research programs more closely with strategic corporate goals and to increase research performance and productivity. In tandem with the increasing governmental interests in research assessment stated above, there is considerable industrial interest in research assessment as well. As an example, the Industrial Research Institute (IRI), whose 260 member companies invest over $55 billion annually in R&D, has shown intense interest in measuring research performance and effectiveness. The IRI has commissioned one of its internal panels (headed by Dr. James W. Tipping) to research the field and write a position paper on measuring and improving effectiveness of R&D on company performance. According to Dr. Tipping, two roundtables on this subject have been held. They have been oversubscribed but limited to 50 companies (Tipping, 1993).

When the above activities are integrated and a mosaic is constructed, the inescapeable trend for the future becomes clear. The research sponsoring agencies will become more accountable to the Administration and Congress on the relationship between sponsored programs and strategic goals, and soon thereafter the research performers will become more accountable to the sponsoring agencies. In addition, the accountability of industrial research to the broader corporate goals will increase (as has been observed over the past decade), and improved methods of measuring research performance and productivity will be sought continually by industrial research organizations. It is therefore important that research managers and administrators in government, industry, and academia understand the assessment approaches which could be utilized to evaluate research quality and goal relevance, and that researchers gain an understanding of these evaluation approaches as well.

In the Congressional Task Force report on the health of research (Brown, 1992) mentioned above, the authors recognized the difficulty of integrating performance assessment mechanisms into the research process. In addressing the difficulty of implementing this recommendation, the report stated further:

> More daunting than political resistance to performance assessment are the technical obstacles. Because policy-oriented assessment has not been a part of the research process in the past, its implementation must be both gradual and flexible. There are some initial efforts underway.

The reference in the Task Force report for these 'initial efforts' (Kostoff, 1992) is the text of a presentation at a management conference.

Since Kostoff (1992), a number of studies have been performed by the author on different aspects of research impact assessment (RIA) (see, for example, Kostoff (1993a,b,c)). One observation from these studies and the accompanying literature surveys is that very little research assessment activity is reported in the literature by the Federal agencies. This is despite the intense interest by the Administration and Congress in research assessment activities. There are two main reasons for this finding. One has been stated eloquently by Averch (1990):

Since 1985, no breakthrough methods of any variety have been invented that more definitively reveal the ex-post scientific or social value of past research investments. While the computer permits greater use of ex-ante project selection methods and expost evaluation methods, the evidence is sparse that there is much payoff to public or private sector R&D administrators from making greater use of them.

Consequently, R&D administrators have little incentive to use current evaluation technologies for ex-ante decisions about R&D level or allocations. This lack of incentive is certainly consistent with the published evidence on their use. R&D administrators do use ex-post evaluations for political and organizational purposes, for example, to convince sponsors that they are interested in rational decision processes and that they are funding good work. However, the research evaluation literature between 1985–1990 contains very few demonstrations that evaluation makes any difference at all to the critical decisions about the level and allocation of scarce scientific and technical resources.

Or, stated in another way by Kerpelman and Fitzsimmons (1985):

> ... Based upon our review of the literature, it would appear that formal, strategic evaluation of research programs is not performed on a regular basis in either government or industrial laboratories. Government funding programs are evaluated on an irregular basis as well. We surmise that much evaluation is informal and non-technique oriented and hence not reported outside of the organization which conducts it.

The other reason for little reported assessment activity is that the research assessment techniques available have a number of inherent deficiencies, and their reliability, validity, and credibility leave much to be desired. The agencies are not fully motivated to publish evaluation information obtained with such problematic techniques which could impact their funding status. Thus, there is a need for more research, development, and pilot studies into RIA techniques to improve their credibility and eventual acceptability by the agencies and industry. The need for this R&D is time critical because of the present and projected increasing reporting pressures being brought upon the agencies and industry as described above.

2. Contents of present paper

The present paper examines a number of different RIA approaches in use or proposed, and recommends research which could address some of the deficiencies of these approaches. Following an initial section on generic research requirements, this paper is divided into three segments, which range from research requirements for qualitative to quantitative approaches. The first segment deals with retrospective approaches. These methods make little use of mathematical tools but attempt to draw on documented approaches and results wherever possible. In practice, some of these approaches (namely, studies of accomplishments resulting from sponsored research programs, or studies of systems and the research products which were eventually converted and incorporated into those systems) are widely used by the research sponsoring organizations.

The second segment deals with qualitative approaches to RIA. Foremost among these are variants on the common theme of peer review. While peer review (evaluation of research and its consequences by 'peers', or experts on the different facets of research and its impacts) is the method used most widely to evaluate research, it has its detractors.

The third segment deals with the quantitative and fiscal approaches to RIA. These approaches make heavy use of mathematical and analytic tools, and utilize computer capabilities extensively. Probably the heaviest concentration of literature papers today are in this category. It should be noted that there are hybrid techniques which span more than one of the three categories. For example, a recent retrospective study of significant events in cancer research (Narin, 1989) included a bibliometric component (citation and co-citation analyses).

3. Recommended areas for research in RIA

3.1. General

This initial section discusses research required for RIA which transcends any particular technique. The issues addressed are those which have hindered the acceptability of the RIA product for decades.

The first issue addressed, certification of RIA managers, is as much an education and training issue as a research issue. Successful resolution of this issue would, in the author's estimation, result in a major advance in the profession of RIA. In the author's experience, most of the people responsible for RIA in the technical agencies and high-tech industries are engineers and scientists who have converted from performing engineering and science to assessing engineering and science. Their training in assessment techniques ranges from minimal to non-existant. Their knowledge of the breadth of available techniques, and when to apply these techniques, is, except for a few notable cases, very limited.

Yet, the tools available for research assessment, and the conditions under which these tools should be applied, are no less complex than the analogous diagnostic tools and application conditions available to an M.D. Internist. In fact, the research assessor's operating conditions may be more complex. The Internist typically has a series of standard protocols to follow in arriving at a diagnosis. No suite of standard protocols is available to the research assessor today. How much credibility would the diagnosis of an Internist have if the Internist had training in his discipline equivalent to that of the average research assessor? The conclusion drawn here is that in order for research assessment to progress from today's practice of random application of a few well-known techniques to tomorrow's application of a suite of more sophisticated approaches tailored to specific problems, the people responsible for research assessment must have appropriate training.

Research should be addressed to the types of training which would offer preparation for assessing research from many perspectives. What are the elements of successful research assessment, and what are the educational requirements which would lead to successful research assessment? What would be the contents of the curricula; where would they be offered? For many fields, such as Airline Pilot, Brain Surgeon, there are aptitude and personality prerequisites. Are there similar prerequisites for a potentially successful research assessor? Finally, how should certification of research assessors be done, and enforced?

The second issue addressed is research assessment quality. In many fields, such as construction, surgery, music, quality of the product can be ascertained readily upon inspection. Yet how is quality of a RIA ascertained? One reads papers and reports which summarize RIAs, including procedures and results. From these, it is almost impossible to differentiate high from medium or low quality RIAs. How much preparation was done by the members of an evaluation panel before the actual meeting? How much background work did their leader do, and how intense was his probing, and consequently that of the panel, during the evaluation process? Was free discourse during the proceedings encouraged, or suppressed?

More research is needed into what constitutes a quality assessment. It is important to understand how these factors can be communicated in a report, and how they can be identified by independent readers.

The third generic issue is that of motivation *and associated incentives*. This issue has some overlap with the previous issue of quality. The research managers and administrators, and those with responsibility for higher level oversight, have to be convinced of the value of RIA to their organizations for the improved allocation of research resources. More important than any evaluation criteria selected is the dedication of an organization's management to the highest quality objective review, and the associated emplacement of rewards and incentives to encourage quality reviews. The team assigned responsibility to carry out RIA must be motivated to generate the highest quality product, not just 'answer the mail', as is done in many organizations today. This means selecting the best suite of methods available to accomplish organizational objectives, and selecting the most competent and objective individuals to participate in the RIA. The RIA managers must be motivated to examine the impact from as many perspectives as possible, to gain the most complete understanding. Finally, the objectives, importance, and benefits of RIA must be articulated and communicated to the researchers and research managers at the initiation of RIA, so that the reviewees will participate in the RIA as fully and as cooperatively as possible.

What are the best motivating factors for producing quality research assessments? What are the best incentives? How does one insure that the range of individuals from upper management to the person conducting the details of the assessment remain motivated throughout the assessment process to provide the highest quality product?

The final generic issue addressed is frequency and level of detail of RIA. How frequently should research be reviewed from a cost-effectiveness viewpoint? The more frequently research is reviewed, the more chances exist to identify wayward research and redirect the efforts. However, as was shown in Kostoff (1994c) in determining an estimate of peer review costs, costs of research reviews are not negligible. The main conclusion of the cost study was that for serious panel-type peer reviews, where sufficient expertise is represented on the panels, *total real costs will dominate direct costs*. This conclusion would also be true for mail-type peer reviews. While the total costs of mail-type peer reviews would be less than those of panel-type peer reviews due to the absence of travel costs, the ratio of total costs to direct costs for mail-type peer reviews would be very high. The major contributor to total costs for either type of review is the time of all the players involved in executing the review. With high quality performers and reviewers, time costs are high, and the total review costs can be a non-negligible fraction of total program costs, especially for programs that are people intensive rather than hardware intensive.

Thus, there is some sort of optimum point where the costs of performing the review balance the probability of achieving cost savings by identifying and redirecting or terminating wayward research. Research is required to determine this review frequency, as a function of discipline, organization, level of basic or applied, type of performer, and other key parameters.

At what level of organization (i.e. Principal Investigator, program, Division, Discipline, etc.) should reviews be performed, and at what frequency? Should the same RIA approach, or combinations of RIA approaches, be applied at each level of organization with the same degree of intensity and effort? Or, should the suites of RIA techniques and review frequencies be functions of the level of organization being reviewed? These are key issues of practical importance on which negligible amounts of research have been performed.

3.2. Retrospective methods

Over the past 20 years, a number of retrospective studies have been performed for the purpose of shedding light on what conditions lead to successful conversion of R&D to technology and systems. The studies referenced here are described and critiqued in more detail in Kostoff (1993b,1994b). Hindsight (Director of Defense, Research and Engineering (DOD), 1969) was a retrospective study of R&D events which impacted selected military systems. TRACES refers to three studies which used a similar retrospective approach for selected technology innovations (Illinois Institute of Technology Research Institute (IITRI), 1968; Battelle, 1973; Narin, 1989). ARPA refers to a study performed by the Institute for Defense Analysis (IDA, 1991) which analyzed retrospectively 49 projects funded by the Advanced Research Projects Agency (ARPA).

The Hindsight, TRACES, and ARPA studies provided valuable insight into the parameters which affect the quality and productivity of research. These types of studies should be expanded. More organizations, such as in the ARPA study, should be examined from a retrospective viewpoint. More technologies and systems, as in the TRACES and Hindsight studies, should be examined.

One shortcoming of all these studies is that the indirect impacts characteristic of basic research did not receive a proper accounting. Mainly the direct impacts of research on the final product are identified. In the expanded studies recommended here, more emphasis should be expended on identifying and tracing the pathways of the indirect impacts of research. Especially for basic research, the research products are disseminated broadly, impacting eventually not only the sponsoring organization's goals, but the broader societal goals as well. These broader impacts should be captured within the studies.

The latest technologies, such as information processing and computer hardware and software, should be employed in these retrospective approaches. In the recent TRACES study (Narin, 1989), whose goal was to determine the effectiveness of different research settings or support mechanisms in bringing about important advances in cancer research, citation and co-citation analysis were used to help identify the direct impact of other science fields on the cancer fields of interest. If citation and co-citation analysis were combined with word frequency and co-word analysis (an analytic approach which assumes that words which occur together frequently in some domain are related, and the strength of that relationship is proportional to the co-occurrence frequency—see Kostoff (1993c, 1994a) for more details), it would be possible to trace some of the indirect impact pathways. Citations of successive generations of papers, for example, could document the diffusion and dissemination of the products of research.

Alternatively, network approaches (where research and technology sub-areas are represented as network nodes, and the links represent the strengths of impact of the nodes on each other—See Kostoff (1994d)) could explore the information flow among research, technology, and downstream mission areas. Combined with co-nomination techniques (a co-occurrence approach which concentrates on nomination by experts of other experts in a field, and allows the construction of communication linkages—see Kostoff (1992) for more details), these approaches could not only shed light on information dissemination, but on the people involved in the information diffusion process as well.

Central to credible work in tracking the diffusion of information from research is a database of research products at various evolutionary stages which can feed the models. Since the research product evolutionary pathways transcend the research originating organization, and can intersect all societal sectors, the cooperation of many public and private organizations would be required to develop a database of research products in their evolutionary stages. Development and construction of such a database should start in the near future.

3.3. Peer review

Peer review of research represents evaluation by experts in the field, and is the method of choice in practice in the US (e.g. Salasin et al., 1980; Logsdon and Rubin, 1985; Chubin and Hackett, 1990, Chubin, 1994; Kostoff, 1994c). Its objectives range from being an efficient resource allocation mechanism to a credible predictor of research impact.

One of the central problems in peer review is lack of credibility in its predictive reliability. For an organization conducting peer review of research, it would be desirable to relate the reviewers' scores to downstream impacts on the organization's mission. A few studies have been done relating reviewers' scores on component evaluation criteria to proposal or project review outcomes (e.g. Department of Energy (DOE), 1982; Kostoff, 1992, appendix II). Some studies have been done in which reviewers' ratings of research papers have been compared to the numbers of citations received by these papers over time (Bornstein, 1991a). Correlations between reviewers' estimates of manuscript quality and impact and the number of citations received by the paper over time were relatively low. Bornstein concludes, after an extensive survey of peer review reliability and validity, that:

> If one attempted to publish research involving an assessment tool whose reliability and valid-

ity data were as weak as that of the peer review process, there is no question that studies involving this psychometrically flawed instrument would be deemed unacceptable for publication. (Bornstein, 1991b)

In any case, the author is not aware of reported studies, singly or in tandem, that have related peer review scores/rankings of proposals to *downstream impacts* of the research on technology, systems, and operations. This type of study would require an elaborate data tracking system over lengthy time periods which does not exist today. More studies are necessary to relate evaluations by peers of research proposals and existing research programs to future impacts of this research. Presently, the data to validate different predictive models does not exist. As stated above and reiterated later, what is required is a database which allows tracking of the evolution of products of research in their various metamorphosed stages. Having such a database would allow not only validation of peer review predictive models, but bibliometric predictive models and other quantitative predictive models as well. The database would allow predictive reliability to be determined for a number of different types of impact. These would include impact on the research area of interest, impact on allied research areas, impact on technology, impact on systems, impact on operations, etc. The research product evolution database concept is described in more detail at the end of this paper.

There are very few comparative studies of different types of peer groupings and the quality of the peer review product. Studies should be done varying mail vs. panel review, British model vs. standard model (peer review using professionals instead of eminent persons), panel size, types of reviewer expertise, time expended by the reviewers and reviewees on the process, combining or separating the quality and relevance components of a review, and correlating these variables with the quality of the product. Central to the result would be how cost of the review varies with quality of the product and is affected by the different variables.

A continuing problem in peer review is the validity and reliability of the review results. An excellent discussion of this problem can be found in Cicchetti (1991), as well as in other commentary in the journal issue in which Cicchetti's article appears. To improve validity and reliability, research needs to be done on optimal numbers of reviewers utilized, ascertaining whether author anonymity impacts the results, and ascertaining whether training people to perform peer reviews would increase review quality as well as reliability and validity.

While Kostoff (1994c) included a very approximate estimation of total peer review time and dollar costs for one peer review scenario, more accurate time and cost estimates would be required when comparing different types of peer review scenarios. Extensive data taking would be necessary, because of the many different types of peer reviews in existence. However, since total peer review costs can be substantial, and since cost reduction with consistent quality would be one of the goals of these different types of suggested studies, both the extensive data taking and development of improved peer review cost estimating procedures would be well justified from an economic viewpoint.

The application of expert systems and knowledge-based systems for proposal evaluation and program review could supplement peer review. Few studies have been done along these lines, but a recent dissertation (Odeyale, 1993) and follow-on studies (Odeyale and Kostoff, 1994a,b,c) address this problem in detail. Much more work would be required to validate the application of these advanced technologies as useful supplements to peer review, but more research in this direction could determine whether there is potential for real payoff.

One of the potential benefits resulting from a peer review is constructive feedback to the reviewee(s) followed by an improvement in the reviewee's conduct of research. Studies should be done to ascertain reviewees' perceptions of the peer review and the review's value in improving the conduct of research. A recent study (Luukkonen and Stahle, 1993) addresses peer review from the reviewee's perspective, but much more can be done to improve the information transfer from the reviewers to the reviewee, and to insure that the review's recommendations were translated into improved research.

3.4. Quantitative methods

Quantitative approaches are a relatively inexpensive (compared to high quality peer review or retrospective studies) method for obtaining measures of science quality. Because of potential multiple interpretations of the numbers obtained, quantitative approaches should be coupled with peer review to enhance the value of each approach. Bibliometrics, especially evaluative bibliometrics, uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. A recent comprehensive review of bibliometrics (White and McCain, 1989) shows the sparsity of bibliometric studies for research impact evaluation reported by the Federal government. Macroscale bibliometric studies characterize science activity at the national (e.g. Hicks et al., 1986; Braun et al., 1989), international, and discipline level. The biennial *Science and Engineering Indicators* report (National Science Foundation (NSF), 1989)) tabulates data on characteristics of personnel in science, funds spent, publications and citations by country and field, and many other bibliometric indicators. There have been numerous microscale bibliometric studies reported in the literature (e.g. Frame, 1983; McAllister et al., 1983; Mullins, 1987; Mullins et al., 1988; Moed and Van Raan 1988; Irvine, 1989; Van Raan et al., 1989; Luukkonen, 1990a; Luukkonen and Stahle, 1990; Luukkonen et al., 1992).

In the practical use of bibliometrics, one of the problems which arises is cross-discipline comparisons of outputs. For example, how should the paper or citation output of a program in Solid-State Physics be compared to that of Shallow Water Acoustics? What types of normalizations are required to allow comparisons among these different types of programs and fields? Is there a threshold for disaggregation below which the normalization factors apply to all the subfields? For example, can the normalization factor for Acoustics be applied to a program in High Frequency Shallow Water Acoustics, or can the normalization factor for Shallow Water Acoustics be applied to the program in High Frequency Shallow Water Acoustics?

Or, is credible normalization not possible? The choice of important bibliometric indicators to use for research performance measurement may not be straightforward. A recent study surveyed about 4000 researchers to identify appropriate bibliometric indicators for their particular disciplines (Australia, 1993). The respondents were grouped in major discipline categories across a broad spectrum of research areas. While the major discipline categories agreed on the importance of publications in refereed journals as a performance indicator, there was not agreement about the relative values of the remaining 19 indicators provided to the respondents. For the respondents in total, the important performance indicators were:

- Publications (publication of research results in refereed journals);
- Peer Reviewed Books (research results published as commercial books reviewed by peers);
- Keynote Addresses (invitations to deliver keynote addresses, or present refereed papers and other refereed presentations at major conferences related to one's profession);
- Conference Proceedings (publication of research results in refereed conference proceedings);
- Citation Impact (publication of research results in journals weighted by citation impact);
- Chapters in Books (research results published as chapters in commercial books reviewed by peers);
- Competitive Grants (ability to attract competitive, peer reviewed grants from the ARC, NH&MRC, rural R&D corporations and similar government agencies).

Thus, the survey results indicated that the important performance indicators may rank differently for different disciplines. This suggests that multiple indicators would be required for any cross-field comparisons. Under these circumstances, cross-discipline comparisons would require not only normalizations for the same indicators, but some type of weighting correction to account for the different relative importance of the indicators on different disciplines. More research on these issues needs to be done to make cross-discipline comparisons using bibliometrics more acceptable.

An area of bibliometrics which has been gaining in popularity over the past decade has been that of partial/multiple indicators (e.g. Martin and Irvine, 1983; Rubenstein and Geisler, 1988, 1991). In some applications, different partial indicators are combined to give an overall figure of merit. A number of research issues need to be addressed here. If the indicators do not form an orthogonal set, there will be multiple counting, and the results will be skewed. As a hypothetical example, if it were shown that publications were strongly correlated with awards, then including publications and awards in the figure of merit would be a double counting of publications. There needs to be research showing how the different leading indicators are related to each other, and the degree to which they overlap.

Typically, the indicators are combined in a linear manner to arrive at the figure of merit. In addition to the problem that the weighting factors may be field-dependent, as discussed in the section on cross-discipline comparison above, the linear assumption may be invalid over the full range of the indicators. For example, marginal utility theory would suggest that while it might be twice as valuable for a researcher to publish two papers per year compared to one paper, it would probably not be twice as valuable if the researcher were to publish 40 papers per year as opposed to 20. Research needs to be done to identify the utility functions for these indicators, and identify the regions where the linear assumption is valid.

One rapidly emerging area, for which substantial databases are in existence, is patent citation analyses (e.g. Carpenter et al., 1981; Carpenter, 1982; Carpenter and Narin, 1983; Narin et al., 1984; Wallmark and Sedig, 1986; Collins and Wyatt, 1988; Narin and Olivastro, 1988a,b, 1991, 1992; Narin et al., 1988; Van Vianen et al., 1990). Yet there has been negligible use of this capability by the Federal government for research impact assessment, and assessment of the conversion of science to technology. Studies should be done to ascertain the regions of validity of patent citation analysis, and the constraints and limitations of the technique. For those technologies and research disciplines where the technique has validity, studies should be done using patent citation analysis to track the diffusion of research information. Perhaps the technique could be used in tandem with the other citation approaches in supplementing the retrospective approaches suggested in the section on proposed retrospective studies. It would be valuable to understand the parameters which influence the successful conversion of science to technology.

A number of specific studies are suggested for large multi-spectrum federally supported laboratories, to ascertain whether these organizations are making effective and efficient use of their multi-discipline capabilities:

- Examine distribution of disciplines in coauthored papers, to see whether the multidisciplinary strengths of the lab are being utilized fully;
- Examine distribution of organizations in coauthored papers, to determine the extent of lab collaboration with universities/industry/other labs and countries;
- Examine nature (basic/applied) of citing journals and other media (patents), to ascertain whether lab's products are reaching the intended customer(s);
- Determine whether the lab has its share of high impact (heavily cited) papers and patents, viewed by some analysts as a requirement for technical leadership;
- Determine which countries are citing the lab's papers and patents, to see whether there is foreign exploitation of technology and in which disciplines;
- Identify papers and patents cited by the lab's papers and patents, to ascertain degree of lab's exploitation of foreign and other domestic technology;

- Compare the lab's output (papers/citations normalized over disciplines) with that of other similar institutions, taking into account the concerns expressed above on cross-discipline normalization.

An interesting production function approach to cost-efficiency of basic research (Averch, 1987, 1989) essentially regresses desirable research outputs (citations per dollar, etc.) against research inputs (quality of the investigator's department, etc.). One potential application is prediction of high output proposals based on prior knowledge of the investigator and proposal characteristics (the research inputs). This could be a useful supplement for proposal peer reviews, especially in those cases where quality differences among different proposals are not large, and use of prior knowledge could impact the outcome. Studies should be done to:

- Identify the appropriate output measures;
- Identify the appropriate input measures;
- Estimate the production functions for different disciplines;
- Provide some understanding of the predictive reliability of the approach.

A 1991 study (Mansfield, 1991) weighed the costs of academic research against the benefits realized from the earlier introduction of innovative products and processes due to the academic research. For agencies which sponsor some accelerated research programs, or which have the charter of funding accelerated research programs to hasten transitions, marginal cost-benefit studies of the type used by Mansfield should be made to study the research impacts. Applications of these approaches to the early stages of basic research should be evaluated, such that the indirect impacts of basic research are given appropriate credit in an economic sense.

Modern quantitative techniques utilize computer technology extensively, usually supplemented by network analytic approaches, and attempt to integrate disparate fields of research. One class of techniques which tends to focus more on macroscale impacts of research exploits the use of co-occurrence phenomena. In co-occurrence analysis, *phenomena that occur together frequently in some domain are assumed to be related, and the strength of that relationship is assumed to be related to the co-occurrence frequency*. Networks of these co-occurring phenomena are constructed, and then maps of evolving scientific fields are generated using the link-node values of the networks. Using these maps of science structure and evolution, the research policy analyst can develop a deeper understanding of the interrelationships among the different research fields and the impacts of external intervention, and can recommend new directions for more desirable research portfolios.

Little evidence of Federal use of these techniques (co-citation, co-word, co-nomination, and co-classification analysis) has been reported in the open literature. However, as computerized databases get larger, and more powerful computer software and hardware become readily available, their utilization in assessing research impact should increase substantially. These techniques are discussed in more detail in Kostoff (1992, appendix III, 1993c,d, 1994a; Tijssen and Van Raan, 1994). The Tijssen paper contains an excellent exposition on mapping techniques for displaying the structure of related science and technology fields.

For mapping the structures of different fields of science and technology, comparative studies should be done of co-word, co-citation, and co-nomination approaches, and hybrid combinations of these co-occurrence techniques. There should be synergistic benefits from the hybrid approaches, since different complementary data are used in each approach.

Recently, a new co-word approach that deals directly with text and requires no indexing or key words was developed (Kostoff, 1993d, 1994a). The methodology can be applied to any text database, consisting of published papers, reports, memos, etc., which can be placed on computer storage media. This approach has been used to identify pervasive thrust areas of science and technology, the connectivity among these areas, and sub-thrust areas closely related to and supportive of the pervasive thrust areas. The approach utilizes a computer-based algorithm to extract and order data from a large body of textual material which, for example, may describe a broad spectrum of science. The algorithm extracts words and word phrases which are repeated throughout this large database, and allows the user to create a taxonomy of pervasive research thrusts from this extracted data. The algorithm then extracts words and phrases which occur *physically close* to the pervasive research thrusts throughout the text, and allows the user to determine interconnectivity among the research thrusts, as well as determine research sub-thrusts strongly related to the pervasive thrusts. While the focus of applications has been to identify *technical* thrusts and their interrelationships, the raw data obtained by the extraction algorithms allows the user to relate *technical thrusts to institutions, journals, people, geographical locations, and other categories.*

For the full-text co-word analysis, automated data analysis and interpretation techniques should be developed to reduce the labor intensity of the process. The full-text technique should be applied to technical journals to identify emerging research and technology areas, as well as the evolving structure of the technical discipline. For example, with present desktop computer memory capabilities, full-text co-word analysis could be applied to one or more year's issues of the *Journal of the American Chemical Society* to identify the emerging research areas in chemistry, and to provide some understanding of the interrelationships among the different areas in chemistry (and perhaps among chemistry and other discipline areas as well).

3.5. Database infrastructure development

To fully understand a research program, especially in the assessment of that program, evaluators must be cognizant of the large body of research being conducted throughout the world. In addition, to fully understand the impacts of research on different technologies, evaluators must be cognizant of the large body of existing technology and technology being conducted throughout the world, and the existing and potential shortcomings in those technologies.

With the advent of high speed and high storage capacity computers, and advances in database software packages, the capability exists now to make large amounts of information available to researchers and evaluators. In particular, the capability exists to provide information about funded research and technology development programs being conducted throughout the world, as well as information about existing technologies. The author has developed a multiagency research database which describes programs being funded by defense and non-defense Federal agencies. This database has been of immense help in assessing research programs, as well as helping to plan research programs. *However, a much larger and more comprehensive database, covering not only research but technology as described above, would be of substantial benefit to the research and technology performer community, the research and technology evaluation community, and the research and technology user community.* Such a database would involve the cooperation of many government agencies, and a number of industrial organizations as well. The requirements of, and planning for, such a database should be started in the near future.

As stated previously, central to credible work in predicting and tracking the diffusion of information from research is a database of research products at various evolutionary stages which can feed the predictive models. This database of research products could be linked in part with the above-proposed database of research and technology. Since the research product evolutionary pathways transcend the research originating organization, and can intersect all societal sectors, the cooperation of many public and private organizations would be required to develop a database of research products in their evolutionary stages. Development and construction of such a database should start in the near future.

One approach to constructing this research product evolution database has its conceptual heritage in Kostoff (1994d). The products of research and technology development programs would be entered into a database on a periodic basis. The research and technology product antecedents which led to these latest products would be identified. Linkages would be constructed to show the evolution of the research products over time, with appropriate credit given to the programs which spawned the initial research products.

As a particular example of an entry in the proposed database, assume that research program P1 has a number of products. These products could include papers, patents, reports, presentations, graduate students, etc. The various products would be entered into the database, and their ties with P1 and its input characteristics (evaluation scores, funding, etc.) would be retained. These products would be related to their antecedents, and these antecedents would be part of the database after the initial transient start-up period. For example, a paper which resulted from P1 would be linked through its references to research and development products (other papers, patents, presentations, etc.) resulting from other programs. A patent resulting from P1 would have similar linkages. For those products whose antecedent research and development products cannot be traced as easily as papers or patents (such as devices that are developed and not published in the literature), the program manager of P1 would enter the product and its antecedents in the database. In technology development and engineering development programs, there tends to be less of a readily available documentary trail of program products, and the program manager of these types of programs would have to supply more of the product and antecedent information than the nominal research program manager.

Included with the entry of an antecedent in the database would be some measure of its relative importance in the generation of the research product resulting from P1. Thus, for a patent which resulted from P1 and referred to five papers, some measure of importance of the impact of each of the five papers on the successful development of the patent of interest should be provided by the program manager. If providing an importance measure proves to be infeasible in practice because of sheer data volume limitations, then all antecedents could be assumed to have equal importance. Provision of an importance measure should not be ruled out at present, since visionary approaches could conceivably overcome this problem.

Thus, in its steady state operation mode, the database would consist of large amounts of research and technology development products with quantitative measures of the strength of their linkages. If it were desired to examine the multiple impacts of a given research (or technology) program on downstream 'products', then the total output of the program could be integrated forward in time over the linkages. The downstream impact could then be related to the program inputs (evaluation scores, funding, etc.) to arrive at the desired information. Programs with little downstream impact would be identified as well as those with high downstream impact. If it were desired to start with a given downstream impact (say, a successfully developed system) and identify those research and development programs which contributed to successful development of the system (as well as the strength of their contribution) this could be done as well. The integration would be performed backwards in time over the linkages to arrive at the various research and technology development products which spawned the successful impact, and the research and technology development programs could then be identified from their products.

There may be other valid approaches to developing such a product tracking database and, at this early conceptual stage, all approaches should be considered. The most important factor is for government and private organizations to start serious planning of this database in the near future.

References

Australia, 1993, *Research Performance Indicators Survey*, National Board of Employment, Education and Training, Commissioned Report No. 21, (Australian Government Publishing Service, Canberra, Australia).

Averch, H., 1987, Measuring the Cost-Efficiency of Basic Research Investment: Input-Output Approaches, *Journal of Policy Analysis and Management*, 6, 3.

Averch, H., 1989, Exploring the Cost-Efficiency of Basic Research Funding in Chemistry, *Research Policy*, 18.

Averch, H., 1990, *Policy Uses of "Evaluation of Research" Literature*, OTA Contractor Report.

Battelle Columbus Laboratories, 1973, *Interactions of Science and Technology in the Innovative Process: Some Case Studies*, Final Report, Prepared for the National Science Foundation, Contract NSF-C 667, 19 March 1973.

Bornstein, R.F., 1991a, The Predictive Validity of Peer Review: A Neglected Issue, *Behavioral and Brain Sciences*, 14, 1.

Bornstein, R.F., 1991b, Manuscript Review in Psychology: Psychometrics, Demand Characteristics, and an Alternative Model *Journal of Mind and Behaviour*, 12.

Braun, T., Glanzel, W. and Schubert, A., 1989, An Alternative Quantitative Approach to the Assessment of National Performance in Basic Research, in: David Evered and Sara Harnett (Eds.), *The Evaluation of Scientific Research*, (John Wiley and Sons, Chichester, UK).

Brown, G.E., 1992, *Report of the Task Force on the Health of Research*, Chairman's Report to the Committee on Science, Space, and Technology, US House of Representatives, No. 56-819, (US Government Printing Office, Washington).

Carnegie, 1992, Commission *Enabling the Future: Linking Science and Technology to Societal Goals*, Carnegie Commission on Science, Technology, and Government, (Carnegie Commission, New York, NY).

Carpenter, M.P., Narin, F., and Woolf, P., 1981, Citation Rates to Technologically Important Patents, *World Patent Information*, 3, 4.

Carpenter, M.P., 1982, *Assessment of the Linkages Between Patents and Fundamental Research*, Presented at the OECD Patents and Innovation Statistics Seminar, Paris, France.

Carpenter, M.P. and Narin, F., 1983, Validation Study: Patent Citations as Indicators of Science and Foreign Dependence, *World Patent Information*, 5, 3.

Chubin, D.E. and Hackett, E.J., 1990, *Peerless Science: Peer Review and US Science Policy*, (State University of New York Press, Albany, NY).

Chubin, D.E., R.N. Kostoff, 1994, Grants Peer Review in Theory and Practice, in: (Ed.), Special Issue on Research Impact Assessment, *Evaluation Review*, 18, 1.

Cicchetti, D.V., 1991, The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation, *Behavioral and Brain Sciences*, 14, 1.

Collins, P. and Wyatt, S., 1988, Citations in Patents to the Basic Research Literature, *Research Policy*, 17.

Director of Defense, Research and Engineering, 1969, *Project Hindsight*, Office of the Director of Defense Research and Engineering, Washington, DC, DTIC No. AD495905.

Department of Environment, 1982, *An Assessment of the Basic Energy Sciences Program*, Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0123.

Frame, J.D., 1983, Quantitative Indicators for Evaluation of Basic Research Programs/Projects, *IEEE Transactions on Engineering Management*, EM-30, 3.

Hicks, D., Martin, B. and Irvine, J., 1986, Bibliometric Techniques for Monitoring Performance in Technologically Oriented Research: The Case of Integrated Optics, *R&D Management*, 16, 3.

IDA, *DARPA Technical Accomplishments*, Volume I, IDA Paper P-2192, February 1990; Volume II, IDA Paper P-2429, April 1991; Volume III, IDA Paper P-2538, July 1991 (Institute for Defense Analysis).

Illinois Institute of Technology Research Institute IITRI, 1968, *Technology in Retrospect and Critical Events in Science*, Report, December.

Irvine, J., 1989, Evaluation of Scientific Institutions: Lessons from a Bibliometric Study of UK Technical Universities, in: David Evered and Sara Harnett (Eds.), *The Evaluation of Scientific Research*, (John Wiley and Sons, Chichester, UK.

Kerpelman, L.C. and Fitzsimmons, S.J., 1985, *Methods for the Strategic Evaluation of Research Programs: The State-of-the Art*, and *Annotated Bibliography*, NSF Contract No. PRA 8400688 (Abt Associates Inc., Cambridge, MA).

Knezo, G., 1993, Congressional Research Service, at a colloquium on Effective Research Assessment: Criteria and Models for Success, Washington DC, 16 December 1993, organized by Susan E. Cozzens, Rensselaer Polytechnic Institute, under NSF Grant No. SBE-9220059.

Kostoff, R.N., 1992, "*Research Impact Assessment*", Presented at the Third International Conference on Management of Technology, Miami, FL, 17-21 February, 1992. Larger text available from author.

Kostoff, R.N., 1993a, Evaluating Federal R&D in the US, in: B. Bozeman and J. Melkers (Eds.), *Assessing R&D Impacts: Method and Practice*, (Kluwer Academic Publishers, Norwell, MA).

Kostoff, R.N., 1993b, Semi-Quantitative Methods for Research Impact Assessment, *Technological Forecasting and Social Change*, 44, 3.

Kostoff, R.N., 1993c, Co-Word Analysis, in: B. Bozeman and J. Melkers (Eds.), *Assesing R&D Impacts: Method and Practice*, (Kluwer Academic Publishers, Norwell, MA).

Kostoff, R.N., 1993d, Database Tomography for Technical Intelligence, *Competitive Intelligence Review*, 4, 1.

Kostoff, R.N., 1994a, Database Tomography: Origins and Applications, Special Issue on Technology, *Competitive Intelligence Review*, 5, 1.

Kostoff, R.N., 1994b, "Research Impact Assessment: Semi-Quantitative Methods", in: R.N. Kostoff (Ed.), Special Issue on Research Impact Assessment, *Evaluation Review*, 18, 1.

Kostoff, R.N., 1994c, Quantitative/Qualitative Federal Research Impact Evaluation Practices, *Technological Forecasting and Social Change*, 45, 2.

Kostoff, R.N., 1994d, Research Impact Quantification, *R&D Management*, 24, 3.

Logsdon, J.M. and Rubin, C.B., 1985, *An Overview of Federal Research Evaluation Activities*, Report, The George Washington University, Washington DC. See also J.M. Logsdon and C.B. Rubin, 1985, *Federal Research Evaluation Activities*, (Abt Associates, Cambridge, MA).

Luukkonen, T., 1990, Bibliometrics and Evaluation of Research Performance, *Annals of Medicine*, 22, 3.

Luukkonen, T. and Stahle, B., 1990, Quality Evaluations in the Management of Basic and Applied Research, *Research Policy*, 19.

Luukkonen, T., Persson, O. and Sivertsen, G., 1992, Understanding Patterns of International Scientific Collaboration, *Science, Technology, and Human Values*, 17, 1.

Luukkonen, T., and Stahle, B., 1993, Evaluation of Research Fields-Scientists' Views", *Nord 1993, 15*, (Nordic Council of Ministers, Copenhagen).

Mansfield, E., 1991, Academic Research and Industrial Innovation, *Research Policy*, 20.

Martin, B. and Irvine, J., 1983, Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio-Astronomy, *Research Policy*, 12.

McAllister, P.R., Narin, F. and Corrigan, J.G., 1983, Programmatic Evaluation and Comparison Based on Standardized Citation Scores, *IEEE Transactions on Engineering Management*, EM-30, 4.

Moed, H.F. and Van Raan, A.F.J., 1988, Indicators of Research Performance: Applications in University Research Policy, in: A.F.J. Van Raan, (Ed.), *Handbook of Quantitative Studies of Science and Technology*, (North Holland, Amsterdam).

Mullins, N., 1987, Evaluating Research Programs: Measurement and Data Sources, *Science and Public Policy*, 14, 2.

Mullins, N., Snizek, W. and Oehler, K., 1988, The Structural Analysis of a Scientific Paper, in: A.F.J. Van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology*, (North Holland, Amsterdam).

Narin, F., 1989, The Impact of Different Modes of Research Funding, in: David Evered and Sara Harnett (Eds.), *The Evaluation of Scientific Research*, (John Wiley and Sons, Chichester, UK).

Narin, F. and Olivastro, D., 1988a, *Identifying Areas of Leading Edge Japanese Science and Technology*, Final Report, NSF Contract No. SRS-8507306.

Narin, F. and Olivastro, D., 1988b, Technology Indicators Based on Patents and Patent Citations, in: A.F.J. Van Raan (Ed.), *Handbook of Quantitative Studies in Science and Technology*, (Elsevier Science Publisher, Amsterdam).

Narin, F. and Olivastro, D., 1991, *The Technological Utilization of European Science*, Presented at Joint EC-Leiden Conference on Science and Technology Indicators, Leiden, 23-25 October 1991.

Narin, F. and Olivastro, D., 1992, Status Report-Linkage between Technology and Science", *Research Policy*, 21.

Narin, F., Carpenter, M.P. and Woolf, P., 1984, Technological Performance Assessments Based on Patents and Patent Citations, *IEEE Transactions on Engineering Management*, EM-31, 4.

Narin, F., Rosen, M. and Olivastro, D., 1988, Patent Citation Analysis: New Validation Studies and Linkage Statistics, in: A.F.J. Van Raan, A.J. Nederhof and H.F. Moed, (Eds.), *Science Indicators: Their Use in Science Policy and Their Role in Science Studies*, (DSWO Press, the Netherlands).

National Academy of Sciences, *The Government Role in Civilian Technology: Building a New Alliance*, 1992, Committee on Science, Engineering, and Public Policy, National Academy of Sciences, (National Academy Press).

NSF, 1989, *Science and Engineering Indicators — 1989*, National Science Board Report NSB 89-1, GPO, Washington DC.

Odeyale, C.O., 1993, *Design and Development of a Knowledge-Based System for Effective and Unbiased Military Biomedical R&D Management*, Ph. D. Dissertation, Walden University.

Odeyale, C.O. and Kostoff, R.N., 1994a, R&D Management Expert Networks: I. Knowledge Representation and Inference Strategies, *HEURISTICS, the Journal of Knowledge Engineering and Technology*, 7, 1.

Odeyale, C.O. and Kostoff, R.N., 1994b, R&D Management Expert Networks: II. Prototype Construction and Validation, *HEURISTICS, the Journal of Knowledge Engineering and Technology*, 7, 1.

Odeyale, C.O. and Kostoff, R.N., 1994c, R&D Management Expert Networks: III. A Graph Theoretical Approach to Knowledge Engineering, *HEURISTICS, the Journal of Knowledge Engineering and Technology*, in press.

Rubenstein, A.H. and Geisler, E., 1988, The Use of Indicators and Measures of the R&D Process in Evaluating Science and Technology Programs, in: J. Roessner (Ed.), *Government Innovation Policy*, (St. Martins Press).

Rubenstein, A.H. and Geisler, E., 1991, Evaluating the Outputs and Impacts of R&D/Innovation, *International Journal of Technology Management*, 6, 3/4.

Salasin, J. et al., 1980, *The Evaluation of Federal Research Programs*, MITRE Technical Report MTR-80W123.

Tijssen, R. and Van Raan, A., 1994, Mapping Changes in Science and Technology, in: R.N. Kostoff (Ed.), Special Issue on Research Impact Assessment, *Evaluation Review*, 18, 1.

Tipping, J.W., 1993, Industrial Research Institute, Telephone conversation.

Van Raan, A.F.J., 1989, Evaluation of Research Groups, in: David Evered and Sara Harnett (Eds.), *The Evaluation of Scientific Research*, (John Wiley and Sons, UK).

Van Vianen, B.G., Moed, H.F. and Van Raan, A.F.J., 1990, An Exploration of the Science Base of Recent Technology, *Science Policy*, 19.

Wallmark, J.T. and Sedig, K.G., 1986, Quality of Research Measured by Citation Method and by Peer Review—A Comparison, *IEEE Transactions on Engineering Management*, EM-33, 4.

White, H.D. and McCain, K.W., 1989, "Bibliometrics", in: M.E. Williams (Ed.), *Annual Review of Information Science and Technology*, 24.