

## **SCIENCE AND TECHNOLOGY TEXT MINING: STRATEGIC MANAGEMENT AND IMPLEMENTATION IN GOVERNMENT ORGANIZATIONS**

by

**Ronald N. Kostoff<sup>(1)</sup> and Eliezer Geisler<sup>(2)</sup>**

### **ABSTRACT**

This report focuses on the strategic role and the implementation of textual data mining (TDM) in government organizations, with special emphasis on TDM to support the management of science and technology (S&T). It begins by defining TDM, and discussing the strategic management process in federal government organizations and the role of TDM as an integral part of this process. The report then proceeds to describe some of the uses and applications of TDM. The results of a demonstration program by the U.S. Office of Naval Research show some potential benefits from TDM: (1) integration of national and multi-national S&T databases; (2) supporting strategic decisions on the direction and funding of government S&T; and (3) creation of usable S&T databases to support strategic decisions in other areas of government. Implications of the demonstration program relative to larger scale implementation of TDM are discussed. The report ends with a description of the principles and requirements of higher quality TDM studies. The appendix describes conceptually how a college-based TDM training program could be implemented.

**KEYWORDS:** text mining; textual data mining; strategic management; decision aids; science and technology; data warehouse; unstructured free text; resource allocation; strategic options; strategy formulation; strategy implementation; roadmaps; metrics; peer review; information retrieval; bibliometrics.

| Report Documentation Page                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                    |                                     | Form Approved<br>OMB No. 0704-0188       |                                  |                                 |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------|-------------------------------------|------------------------------------------|----------------------------------|---------------------------------|
| Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.                                                                                                                                                                                                                                                                                                                                                     |                                    |                                     |                                          |                                  |                                 |
| 1. REPORT DATE<br><b>01 MAR 2004</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                    | 2. REPORT TYPE                      |                                          | 3. DATES COVERED<br>-            |                                 |
| 4. TITLE AND SUBTITLE<br><b>SCIENCE AND TECHNOLOGY TEXT MINING STRATEGIC MANAGEMENT AND IMPLEMENTATION IN GOVERNMENT ORGANIZATIONS</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                    |                                     | 5a. CONTRACT NUMBER                      |                                  |                                 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 5b. GRANT NUMBER                         |                                  |                                 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 5c. PROGRAM ELEMENT NUMBER               |                                  |                                 |
| 6. AUTHOR(S)<br><b>Ronald Kostoff; Eliezer Geisler</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                    |                                     | 5d. PROJECT NUMBER                       |                                  |                                 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 5e. TASK NUMBER                          |                                  |                                 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 5f. WORK UNIT NUMBER                     |                                  |                                 |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br><b>Office of Naval Research, 800 N. Quincy St., Arlington, VA, 22217</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                    |                                     | 8. PERFORMING ORGANIZATION REPORT NUMBER |                                  |                                 |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                    |                                     | 10. SPONSOR/MONITOR'S ACRONYM(S)         |                                  |                                 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 11. SPONSOR/MONITOR'S REPORT NUMBER(S)   |                                  |                                 |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT<br><b>Approved for public release; distribution unlimited</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                    |                                     |                                          |                                  |                                 |
| 13. SUPPLEMENTARY NOTES                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                    |                                     |                                          |                                  |                                 |
| 14. ABSTRACT<br><b>This report focuses on the strategic role and the implementation of textual data mining (TDM) in government organizations, with special emphasis on TDM to support the management of science and technology (S&amp;T). It begins by defining TDM, and discussing the strategic management process in federal government organizations and the role of TDM as an integral part of this process. The report then proceeds to describe some of the uses and applications of TDM. The results of a demonstration program by the U.S. Office of Naval Research show some potential benefits from TDM: (1) integration of national and multi-national S&amp;T databases; (2) supporting strategic decisions on the direction and funding of government S&amp;T; and (3) creation of usable S&amp;T databases to support strategic decisions in other areas of government. Implications of the demonstration program relative to larger scale implementation of TDM are discussed. The report ends with a description of the principles and requirements of higher quality TDM studies. The appendix describes conceptually how a college-based TDM training program could be implemented.</b> |                                    |                                     |                                          |                                  |                                 |
| 15. SUBJECT TERMS                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                    |                                     |                                          |                                  |                                 |
| 16. SECURITY CLASSIFICATION OF:                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                    |                                     | 17. LIMITATION OF ABSTRACT               | 18. NUMBER OF PAGES<br><b>67</b> | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT<br><b>unclassified</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | b. ABSTRACT<br><b>unclassified</b> | c. THIS PAGE<br><b>unclassified</b> |                                          |                                  |                                 |

---

***NOTE: The views in this report are solely those of the authors and do not represent the views of the Department of the Navy, or the Illinois Institute of Technology.***

- (1) Office of Naval Research, 800 N. Quincy Street, Arlington, VA 22217, U.S.A. E-mail: [kostofr@onr.navy.mil](mailto:kostofr@onr.navy.mil)
- (2) Stuart School of Business, Illinois Institute of Technology, Chicago, IL.

## INTRODUCTION

This report focuses on the strategic role and implementation of textual data mining (TDM) in government organizations. In the past several years, there has been a surge in studies on the roles that knowledge and science and technology (S&T) have in the successful operation of business as well as government organizations. Text mining has emerged as one of the more powerful techniques to extract useful information from complex databases and data warehouses (Westphal and Blaxton, 1998; Thuraisingham, 1999). Because much of the S&T management language is unstructured free text, TDM is critical for large-scale analysis of this textual component, hence TDM is the algorithmic focal point of this report.

In recent years, some commercial applications of text mining have rapidly emerged.

By using a confluence of techniques such as rule induction, artificial intelligence, and relational databases, these ‘search and discover’ systems have scanned large commercial data warehouses and yielded useful patterns of consumer behavior and similar findings (Berry and Linoff, 1997). In health care, for example, hospital chains are using text mining in their databases on patients, illnesses, and medical experience to identify patterns in resources utilization and patient behavior. These patterns are then used to improve decisions on the allocation of scarce resources for delivery of medical services (Borok, 1997).

Data mining has long been a valuable tool in scientific research in those areas where massive data are collected, such as in astronomy, biotechnology (bio-sequencing), and geo-sciences (Geisler, 2000). As these tools became more sophisticated, and the use of TDM more prevalent, the potential value of this method - beyond discovering patterns in science - soon became evident; in particular, the value to strategic management of the S&T organizations in industry and in government. In the latter case, the convergence of the movement to commercialize public S&T and the potential gains from knowledge on patterns and relations that emerged from TDM have created a fertile background for the utilization of TDM in government S&T (Geisler, 1995; Geisler and Frey, 1997; Kostoff, 1992).

Government S&T organizations are especially suited to TDM applications. The network of US Federal S&T institutions has accumulated very large S&T data warehouses from the outputs of both public and private government sponsored programs, and is also mandated to streamline and strategically chart its future activities to benefit the public-at-large, in addition to the parent agencies.

To place TDM in its role as an S&T strategic management decision aid, this report first overviews S&T strategic management issues, including modern decision aids and their integrated role in enhancing the strategic management process,. Then, the report focuses on one of the critical decision aids, TDM, and the role that TDM has in the strategic management of government S&T. The report ends with a detailed description of how TDM can be implemented, and the problems, processes, and requirements for successful implementation.

## **STRATEGIC MANAGEMENT OF S&T IN GOVERNMENT ORGANIZATIONS**

In general, strategic management is a process that starts with the establishment of the overall direction of the organization through a broad vision and achievable objectives. This is followed by the identification and mobilization of capabilities and resources that enable the organization to accomplish its objectives (Allison and Kaye, 1997). In government organizations, strategic management involves the establishment of objectives that are consonant with the overall mission of the parent agency and the expressed concerns and expectations of the various constituencies of the organization (Bryson, 1995; Koteen, 1997).

S&T in industrial organizations is viewed, strategically, as an instrument or dimension of the overall strategy of the firm (Burgelman and Rosenbloom, 1997).

The outcomes from S&T, in the form of knowledge, skills, and technical competencies can be applied in the competitive positioning of the firm. S&T thus benefits the firm by its contributions to improved flexibility, better performance of units and processes, and to the firm's stock of skills and capabilities. These are all utilized in meeting the strategic objectives of competition, such as timely reaction to environmental changes, or cost-savings that lead to cost-leadership in the marketplace (Geisler, 1999a).

The link between the strategy process and S&T can also be viewed in terms of the contributions of S&T to the needs of this process. Strategic management relies on the identification of trends, concepts, and configurations of competencies that are, or will be, needed to define the strategic options and to attain the goals embedded in them. S&T helps in the identification of these trends and competencies.

In government organizations, this link is even more powerful. Strategic management in this sector is based on achieving goals that are interpretations of needs and expectations of the public and other constituencies in the government. Performance targets are identified, and these are related to the longer-term concepts

and requirements. The S&T programs in the government organization must therefore be able to supply the needed support and skills necessary to achieve the goals (Moore, 1995). The strategic S&T output is not only the improved underlying technology base needed for long-term advancement, but most importantly the advanced human resource capabilities required to develop and exploit this technology.

The strategy process is composed of two major parts: strategy formulation and strategy implementation. The first includes the selection of the longer-term objectives and the ways to achieve them. Implementation includes program identification and selection, program management and review, program evaluation, program transition, and productivity and impact tracking. In both parts of the strategy process, text mining is useful. Strategic planning is enhanced substantially by comprehensive knowledge of what S&T has been, and is being, conducted on a *global* scale. In the implementation stage, the knowledge extracted from textual databases is equally useful in program identification, management, and evaluation.

Text mining provides S&T and other government managers a powerful set of decision-aids, by offering a systematic view of the state-of-the-art and the trends, concepts, and relations that populate it. These reflect the wishes, desires, and expectations of the relevant constituencies (Chelsey and Wenger, 1999; Wise, 1995).

The link between strategic goals and S&T outcomes in Government organizations is also dependent on or at least strongly influenced by the GPRA (Government Performance and Results Act). This act requires the evaluation of major S&T programs in light of the Government's strategic goals, as they are translated into each Government agency's own goals and objectives (Government Executive, February 2002). Compliance with this and similar legislation, such as the technology transfer acts, calls for assessment of long-term (strategic) and short-term (tactical) goals and activities.

Performance management and metrics are essential elements of this assessment effort. In the strategic component of the evaluation of government S&T programs, textual data mining is a powerful tool. It allows the evaluator to trace the progression of certain activities of S&T—and their outputs—downstream from the strategy formulation to the tactical aspects of strategy implementation. For example, if a government agency has among its strategic goals to have its S&T “contribute to national priorities and social welfare” and to “promote national prominence in S&T”, TDM may be used to track the process by which

these strategic goals are translated into operational and tactical outputs from S&T programs and activities. The metrics for such performance that links strategy to tactics contains such illustrative measures as partnerships with industry, patents, bibliometrics and similar outputs from government S&T. TDM provides a means to identify shared attributes of technology transfer and technology commercialization from strategic to tactical.

In the implementation portion of the tactical aspects of government S&T, TDM is a decision aid that allows for managers of government S&T to assess the various outputs and their metrics and to link them to the strategic goals and the mission of the government S&T enterprise. This process complements the “top-down” assessment that tracks government S&T outputs from the strategic to the tactical. TDM is a decision aid that allows S&T managers to assess linkages among the various tactical outputs (e.g., patents, bibliometrics, partnerships, and other measures of diffusion of technology). Both the “top-down” and the tactical linkages are approaches that provide managers with a measure of value accrued from the S&T in their organizations—hence the contributions of such S&T to the mission of the agency (organizational strategic goals), and to higher-order national goals (determinants of the organization’s strategic goals).

## **INTEGRATED DECISION-AIDS FOR S&T MANAGEMENT SUPPORT**

The growth in available databases, and information storage and processing capabilities, has resulted in an attendant proliferation of computer-based management decision aids to support the strategic management process. These management support techniques include road-maps, metrics, peer review, text mining, information retrieval, bibliometrics, and retrospective studies. Each of these techniques has its own nomenclature and literature, and superficially is treated as an independent process. In reality all these techniques are inter-related and are valuable to the degree that they synergistically support the strategic management process. For example, road-maps require metrics for goal setting and progress tracking and text mining for placing the defined S&T program in its larger national and global context. But road-maps also support strategic planning and program reviews. Text mining requires information retrieval for source material and bibliometrics for interpretation, but literature text mining also supports planning and review by identifying the published state-of-the-art. In reality, all these seemingly diverse techniques support not only the strategic management process in aggregate, but all support each of the process stages.

The potential benefits to S&T from the integrated use of these techniques may be substantial, but the benefits realized so far have been minimal. There are two central reasons for this deficiency. First, there has been little understanding of, and little attention paid to, the intrinsic quality of these decision aids. Second, these decision aids have not been implemented properly into the overall S&T management process.

The following two sections examine the implementation-related problems, and requirements for high quality decision aids, respectively.

### **Implementation-Related Problems**

There are three major implementation-related problems with management decision aids, both in practice and in the published literature. These problems are: 1) The management support techniques tend to be treated as add-ons; 2) The management support techniques tend to be treated independently; and 3) There is a major mismatch between the developers of the (especially literature-based) management support techniques and the users of these techniques. The first two of these problems stem from the same fundamental cause, namely, that advanced computerized management support techniques are not conceptualized and implemented as an organic component of the management structure. The third problem arises from the separation of the contributors to the published literature from the implementing practitioners.

#### **(1) Techniques Treated as Add-ons**

The various decision aid tools and procedures are not incorporated into the structure of the organization, but are treated as add-ons. For example, management/technology metrics are generally not imbedded as an integral part of an organization's intrinsic operating structure. They tend to be employed on a fragmented basis in response to external pressures. They tend to make use of whatever data is available as a result of ordinary business practices, and not the desired type of focused data that would address progress toward corporate strategic goals if the use of metrics were an integral organizational component. This metrics example can be extrapolated generically to other management science techniques as stated previously; they all tend to be used on a sporadic basis. This fragmented approach makes little use of the full power available from integrating the existing management science tools.

#### **(2) Techniques Treated Independently**



Generally, the various management science techniques, if used at all within an organization, are employed independently. One person or group may be doing metrics, another person or group peer review, a third person or group road-maps, a fourth person or group text mining, and so on. The synergies that can be exploited by employing these tools in a unified approach are never realized. Kostoff (1997e, 2002c, 2003b, 2004c) presents an example of promoting and stimulating innovation through a combination of workshop-based and literature-based approaches; this example illustrates some of the synergistic benefits possible from accessing multiple management science tools. In the complex systems of management science, as in the complex systems of physical/ biological/ engineering sciences, the whole is indeed greater than the sum of its parts. In all these complex multi-component systems with highly interactive elements, the intelligence that links the components and allows communication and control provides the benefits from the synergy.

### ( 3) Mismatch Between Performers and Users

Over the past few years, the first author has conducted a number of literature surveys and subsequent studies in fields that can be loosely called 'management science', including research assessment (Kostoff, 1997a, 1997j, 1997k, 2001h), peer review (Kostoff, 1997c, 1997h, 1998d, 2001f), metrics (Kostoff, 1997i, 1998c, 1998e, 2001e, 2001e, 2002b, 2004i), text mining (Kostoff et al., 1997j-k, 1998a, 1999a, 1999f, 2000a, 2000b, 2001d, 2001g, 2002a, 2003a, c-g, 2004a-h, j-o), information retrieval (Kostoff et al., 1997f, 2001b), resource allocation and project selection (Kostoff, 1997a), technology transfer (Kostoff, 1997g), and road-maps (Kostoff, 1997d, 2001a). The specific conclusions from the metrics survey will be described, and then generalized to cover all the areas surveyed.

Most of the documents retrieved in the metrics survey described the generation of a multitude of metrics of large data aggregates, with no indication of the relevance of these metrics to any questions or decisions supporting S&T evaluations. The foundation of this problem is the strong dichotomy between the researchers who publish metrics studies in the literature, and the managers who use metrics to support budgetary allocation and other management decisions. Most of the people who employ metrics for management purposes do not document their experiences and approaches in the literature. Most of the principle and concept and (potential) application papers in the metrics literature are written by people who have never used or applied metrics for management decision-making purposes. In addition, many of the researchers who perform metrics studies focus on single approaches or single approach applications, in order to promote the concepts that they have

developed. Conversely, the managers who use metrics have very eclectic requirements. They need suites of metrics, or suites of metrics combined with other evaluation approaches and decision aids, in order to perform comprehensive multi-faceted S&T evaluations. Thus, there is a serious schism between the incentives and products of the metrics researchers (suppliers) and the incentives and requirements of the metrics users (customers).

Consequently, there are two major gaps in the literature on S&T metrics. First, there are few relevant papers published. Second, most of the concept and principle and (potential) application papers that do exist bear little relation to the reality of what is required to quantitatively support science and technology assessments and evaluations for decision-making. Because of the deficiency of metrics studies relevant to S&T applications, it is difficult to extract the conditions for high quality metrics-based evaluations solely from the open literature. Drastic alterations in this overall situation are required if metrics are going to support future government and industry business requirements in any credible manner.

While there are some minor differences among the diverse management decision aid domains surveyed, the following observation generally appears to transcend disciplines, and can be considered universal and invariant. Most of the people who conduct program evaluations/ assessments/ plans (including practitioners who use the management science tools listed above in their repertoire) do not document their studies and/ or approaches/ techniques in the literature, and most of the management science papers in the literature are written by people who have never conducted program evaluations/ assessments/ plans. Consequently, there is a major gap in the management science literatures, that is reflected as a major split between the theory and the practice of management science.

Consider, for example, the advanced operations research (and other) techniques available in the literature for resource allocation applications (Hall, 1990; Kostoff, 1997a), and then observe how resources are allocated in practice. Or, as another example, consider the esoteric literature publications on information retrieval techniques (Greengrass, 1997; TREC, 2002), and contrast those with methods actually used by librarians and other information resource personnel to retrieve information. Or, as a third example, consider the sophisticated methods on TDM in the literature, and contrast this with how the majority of R&D people actually perform TDM (i.e., reading technical papers with no computer-based support).

Many of the papers in the management science literature are very sophisticated, while most of the techniques actually used by the practitioners are very primitive.

While the literature papers may have substantial academic merit, many bear little relation to the reality of conducting program evaluations/ assessments/ plans. The practice of management science lags far behind what the technology of management science can offer (Geisler, 1997).

The proposed TDM implementation process, and its precursor development programs, described in the remainder of this report, were developed to overcome the limitations imposed by condition 3) above (mismatch between performer and user). The performer and the user were unified, and the continual interplay between satisfying user requirements and performer opportunities resulted in a TDM process that was maximized from the combination of both perspectives.

Overcoming the real limitations imposed by conditions 1) and 2) above is not within the province of the TDM developer, but rather is a function of how the implementing organization chooses to integrate text mining with the other decision aid techniques to support its cohesive strategic management process.

## **REQUIREMENTS FOR HIGH QUALITY MANAGEMENT DECISION AIDS**

Before the applications, implementation, and the benefits from TDM as a management decision aid are discussed, the meaning of the quality aspect of such tools will be reviewed. Quality will be described in the context of management decision support, rules for high-quality management support procedures using these aids, and criteria for more effective implementation of these decision aids into the larger management process. To provide tangible demonstration of the decision aid quality problem, and set the stage for the more universal conclusions which follow, two illustrative examples will be presented. The first concerns quality issues related to S&T road-maps, and the second concerns the meaning of quality in the context of information retrieval for text mining.

## **QUALITY ISSUES RELATED TO S&T ROAD-MAPS**

A 1997 web document on road-maps (Kostoff, 1997d), and an updated journal paper (Kostoff and Schaller, 2001a) focused on principles of high quality road-maps, different classifications of road-maps, and specific examples of many different types of road-maps. As shown by the Bibliography in Kostoff (1997d),

there are hundreds of documents that come under the broad umbrella of S&T road-maps. One major problem in interpreting and drawing credible conclusions from these documents is the inability to ascertain the quality of any given road-map. There are no independent tests of quality. Unlike the physical and engineering sciences, there are no primary physical reference standards against which one can benchmark the road-map product.

Even the metrics of road-map quality are unclear. Road-map (and other decision aids) quality is a very subjective term, and has intrinsic and extrinsic components.

Quality depends not only on the technical construction of the road-map (the intrinsic component), but depends on the objectives of the road-map application as well (the extrinsic component). If the objective of the application is to attract investor interest in a technology/ system, then the quality metric would relate to dollars invested subsequent to the road-map. How well the road-map represented the state or potential of S&T is of little consequence, as long as the major objective of capital attraction was achieved. Alternatively, if the objective of the application is to reflect the state and potential of S&T fully, then this becomes the metric of quality. The latter concept/ application of road-map quality is the one used in the remainder of this section.

To illustrate the road-map metrics quality problem further, consider the following example. Suppose a prospective technology-push road-map has been constructed for high energy-density batteries. Suppose further that fifteen years after the road-map was developed, an assessment was performed of the road-map's predictions or targets relative to the battery state-of-the-art. Suppose even further that the assessment showed the road-map development plan was followed religiously by the technical community, and the long-range technical goals were achieved exactly as predicted by the road-map. Does that mean the road-map was of high quality; i.e., that it reflected the state and potential of battery S&T fully?

Not necessarily. The road-map developers may have been very conservative in their targets, and did not 'push the envelope' to develop the field as vigorously as technology would have allowed. The developers may also have been very narrow in their outlook, and may not have drawn from other disciplines sufficiently to develop the batteries to the greatest extent. It could be stated that the road-map was precise (in predicting the goals that were actually achieved), but was not accurate (the most visionary goals were not predicted).

On the other hand, the road-map in this case may have been of the highest quality. The developers may well have had very ambitious targets, and may have drawn from other disciplines to the maximum extent possible. The point to be made here is that the concepts of road-map quality, and its associated metrics, are very complex and diffuse, yet very important if road-maps are to become useful operational tools.

A high quality S&T road-map that integrates all temporal stages of development requires the following conditions: (1) the retrospective component must be a comprehensive reflection of the evolution and relation of all critical sciences and technologies that resulted in the technology of present interest; (2) the present time component must be a comprehensive reflection of all critical science and technology related to the technology of interest; and (3) the prospective component should reflect some degree of vision by the planners and should incorporate all the critical science and technology areas that relate to the technology of interest and to the projected targets. The road-map's utility is enhanced substantially if some intrinsic processing capability is present; i.e., if the quantitative relationships between the road-map's component elements can be incorporated in functional form, and sensitivity or tradeoff studies can then be done. Its utility is enhanced further if critical attributes (cost, schedule, risk, performance targets) can be displayed throughout (Zurcher and Kostoff, 1997). Thus, a high quality S&T road-map is analogous to a high resolution picture of the evolving/ changing relationships among science and technology areas related critically to the focal road-map technology, and incorporates especially the concepts of awareness, coordination, vision, completeness, and risk (Kostoff and Schaller, 2001a).

## **QUALITY ISSUES RELATED TO INFORMATION RETRIEVAL FOR TEXT MINING**

A 1997 article on information retrieval (Kostoff et al., 1997f) focused on the use of computational linguistics imbedded in an iterative relevance feedback procedure. In this approach, a database query is expanded by incorporating phrase patterns from relevant documents, and the query is contracted by subtracting phrase patterns obtained from non-relevant documents. The final product is a query that will retrieve documents with two aggregate characteristics; the maximum number of relevant documents will be retrieved, and the ratio of relevant to non-relevant documents will be very large; i.e., the signal-to-noise ratio will be large.

Quality in the context of information retrieval requires three conditions. Two of these conditions are the aggregate characteristics mentioned above. The third condition derives from the definition of 'relevant', and requires the desired definition of 'relevant' to be incorporated into the query development process. As in the previous road-map example, the operational meaning of 'relevant' depends on the objectives of the query. Is the purpose of the query to retrieve all the papers in: (1) a very narrowly focused target technical field, (2) in allied technical fields as well, (3) and/ or in very disparate technical fields that have the potential to provide innovative new insights to advance the target technical field (Kostoff, 1997e, 2003b)

Each of these purposes defines a very different concept of 'relevant', and would result in very different numbers of 'relevant' documents being retrieved. The operational definition of relevance will be the major determinant of the volume of papers retrieved.

Typical S&T literature surveys have none of these three quality conditions. Most queries consist of a few key words fairly closely associated with the desired narrow target literature, with minimal (if any) iterative steps. The results will either contain substantial noise if the search terms are relatively broad, or will be very limited if the search terms are narrowly focused. Some iterative approaches will provide substantial numbers of records with high signal-to-noise ratio using a constrained definition of relevant; i.e., not accessing the disparate literatures from which innovative ideas could potentially flow. Rarely, if ever, are all three necessary conditions for a high quality information retrieval fulfilled. Why is this?

Probably the main reason is time and cost. Information retrieval/ text mining efforts (e.g., Kostoff et al., 1999a, 2000a-b) have shown that an iterative process that incorporates a broad scope of 'relevant' disciplines to the target discipline requires the participation of a technical domain expert(s) and a computational linguistics expert(s) (or at least a documented procedure using computational linguistics tools).

There is substantial judgement and interpretation required by at least one expert at each iterative step, and this effort directly translates into significant resource expenditures. The downside of not expending sufficient resources to obtain a high quality product is that allied and related literatures that could serve as the engines of innovation are not accessed.

As an example of the level of effort required for a reasonable quality query, the first author, in conjunction with two technical domain experts, developed a query related to the hydrodynamic flow over solid bodies (for examining flow around ships). Three iterative steps were required; each step required the technical expert(s) to read

many hundreds of the retrieved records in order to identify those that were relevant and not relevant. Then, computational linguistics analyses (Kostoff *et al.*, 1997f) were performed on both the relevant and non-relevant records to identify phrase patterns and relationships characteristic of the relevant records and the non-relevant records. Substantial time and judgement were required to select the appropriate phrases unique to the relevant records and the non-relevant records, and then modify the query accordingly using the key phrases identified. Approximately 200 terms were contained in the final query. Even then, the process could have continued for more iterations, but it was not considered cost-effective given the time and resource constraints of the specific study.

## **GENERALIZED CONCLUSIONS ON DECISION AIDS QUALITY**

Conclusions of quality drawn from the above two specific examples, as well as from myriad examples over many decision aid applications, can be generalized to many other S&T management decision aids. For example, a high quality peer review provides a comprehensive picture of the intrinsic evolution and status of S&T, and its inter-relationships with other S&T and with potential end-use applications. High quality text mining provides a comprehensive picture of the global S&T trends and status, and their inter-relationship with other S&T and with potential end-use applications. Quality applications of all these decision aids reflect most accurately the history, status, and potential of the S&T area(s) being examined, relate these S&T areas to allied S&T areas and draw insights from disparate S&T disciplines, and incorporate challenges to the frontiers of S&T through a vision of their implementation. Since many of the differences between high and low quality decision aids applications revolve around what could have been included as opposed to what was actually included in the application (projects, papers, patents), and since what could have been included is a highly subjective topic, the metrics of evaluating decision aid product quality are very difficult to quantify.

Thus, since quality cannot be ascertained or measured easily from examination of the final decision-aid output product, then the focus for evaluating quality must be shifted from the decision aid product to the decision-aid application process. The next section addresses the process requirements for insuring that the decision aids applications are of high quality.

## **HIGH QUALITY DECISION AID APPLICATIONS REQUIREMENTS**

Successful applications of high quality decision-aids depend on the following twelve requirements.

(1) Senior Management Commitment

The most important factor in a high-quality S&T evaluation is the serious commitment from the organization's most senior management with evaluation decision authority, and the associated emplacement of rewards and incentives to encourage such evaluations. Incorporated in senior management's commitment to quality evaluations is the assurance that a credible need for the evaluation exists, as well as a strong desire that the evaluation be structured to address that need as directly and completely as possible.

(2) Evaluation Manager Motivation

The *second* important factor is the operational evaluation manager's motivation to perform a technically credible evaluation. The manager:

- (1) sets the boundary conditions and constraints on the evaluation's scope;
- (2) selects the final specific evaluation techniques used;
- (3) selects the methodologies for how these techniques will be combined/ integrated/ interpreted; and
- (4) selects the experts who will perform the interpretation of the data output from these techniques.

In particular, if the evaluation manager does not follow, either consciously or subconsciously, the highest standards in selecting these experts, the evaluation's final conclusions could be substantially determined even before the evaluation process begins. Experts are required for all the evaluation processes considered (peer review, retrospective studies, metrics/ economic studies, road-maps, and text mining), and this conclusion about expert selection transcends any of these specific applications.

(3) Statement of Objectives

The *third* most important factor is the transmission of a clear, unambiguous statement of the S&T evaluation's objectives (and conduct) and potential impact/ consequences to all participants at the initiation of the process. Participants are usually more motivated to contribute when they understand the importance of the evaluation to the achievement of the organization's goals, and understand in particular how they and the organization will be potentially impacted by the outcome.



Clear objectives and goals tend to derive from the seamless integration of evaluation processes in general into the organization's business operations. Evaluation processes should not be incorporated in the management tools as an afterthought, as is the case in practice today, but should be part of the organization's front-end design. This allows optimal matching between data generating/ gathering and evaluation requirements, not the present procedure of force fitting evaluation criteria and processes to whatever data is produced from non-evaluation requirements. When the evaluation processes are integrated with the organization's strategic management, the objectives drive the metrics that in turn determine what data should be gathered. Ad hoc evaluation processes tend to let the available data drive the metrics and the quantifiable goals.

#### (4) Competency of Technical Evaluators

The *fourth* important factor is the role and competency of technical experts in any S&T evaluation. While the requirements for experts in peer review, retrospective studies, road-maps, and text mining are somewhat obvious, there are equally compelling reasons for using experts in metrics-based evaluations. Metrics should not be used as a stand-alone diagnostic instrument. Analogous to a medical exam, even quantitative metric results from suites of instruments require expert interpretation to be placed into proper context and gain credibility. The metrics results should contribute to, and be subordinate to, an effective peer review of the technical area being examined.

Thus, this fourth critical factor consists of the evaluation experts' competence and objectivity. Each expert should be technically competent in his/ her subject area, and the competence of the total evaluation team should cover the multiple science and technology areas critically related to the science or technology area of present interest. In addition, the team's focus should not be limited to disciplines related only to the present technology area (which tends to reinforce the status quo and provide conclusions along very narrow lines). It should be broadened to disciplines and technologies that have the potential to impact the overall evaluation's highest-level objectives (which would be more likely to provide equitable consideration to revolutionary new paradigms).

#### (5) Relevance of Evaluation Criteria to Future Action

The *fifth* important factor is one that has been violated in almost every use of metrics by government agencies, industrial organizations, and academic institutions. In general, this factor tends to be violated for the evaluation criteria used in any of

the evaluation approaches under the decision aids umbrella. The factor will be stated in terms of a metrics-based evaluation, but it should be considered as applicable to all evaluation techniques:

*Every S&T metric, and associated data, presented in a study or briefing should have a decision focus; it should contribute to the answer of a question that would then be the basis of a recommendation for future action.*

Metrics and associated data that do not perform this function become an end in themselves, offer no insight to the central focus of the study or briefing, and provide no contribution to decision-making. They dilute the theme of the study, and, over time, tend to devalue the worth of metrics in credible S&T evaluations. Because of the political popularity and subsequent proliferation of S&T metrics, the widespread availability of data, and the ease with which this data can be electronically gathered/ aggregated/ displayed, most S&T metrics briefings and studies are immersed in data geared to impress rather than inform. While metrics studies provide the most obvious examples, this conclusion can be easily generalized to any of the evaluation methods.

#### (6) Selection of Evaluation Criteria

The *sixth* important factor, evaluation criteria, will depend on:

- the interests of the audience for the evaluation,
- the nature of the benefits and impacts,
- the availability and quality of the underlying data,
- the accuracy and quality of results desired,
- the complementary criteria available and suites of diagnostic techniques desired for the complete analysis,
- the status of algorithms and analysis techniques, and
- the capabilities of the evaluation team.

#### (7) Reliability of Evaluation

The *seventh* important factor is reliability or repeatability. To what degree would an S&T evaluation be replicated if a completely different team were involved in selection, analysis, and interpretation of the basic data? If each evaluation team were to generate different evaluation criteria, and in particular, generate far different interpretations of these criteria for the same topic, then what meaning or credibility or value can be assigned to any S&T evaluation? To minimize repeatability

problems, a diverse and representative segment of the overall competent technical community should be involved in the construction and execution of the evaluation.

#### (8) Evaluation Integration

The *eighth* important factor is the seamless integration of evaluation processes in general into the organization's business operations. Evaluation processes should not be incorporated in the management tools as an afterthought, as is generally the case in many organizations, but should be part of the organization's front-end design. This allows optimal matching between data generating/ gathering and evaluation requirements, not the present procedure of force-fitting evaluation criteria and processes to whatever data is produced from non-evaluation requirements.

#### (9) Normalization Across Technical Disciplines

For evaluations that will be used as a basis for comparison of science and technology programs or projects, the *ninth* important factor is normalization and standardization across different science and technology areas. For science and technology areas that have some similarity, use of common experts (on the evaluation teams) with broad backgrounds that overlap the disciplines can provide some degree of standardization. For very disparate science and technology areas, some allowances need to be made for the relative strategic value of each discipline to the organization, and arbitrary corrections applied for benefit estimation differences and biases. Even in this case of disparate disciplines, some normalization is possible by having some common team members with broad backgrounds contributing to the evaluations for diverse programs and projects. However, normalization of the criteria interpretation for each science or technology area's unique characteristics is a fundamental requirement. Because credible normalization requires substantial time and judgement, it tends to be an operational area where quality is sacrificed for expediency.

#### (10) Global Data Awareness

The *tenth* important factor, of equal importance to reliability and normalization, is global data awareness. What S&T projects, developed systems or operations, or events, that exist globally are in any way supportive of, related to, or impacted by, the S&T programs under review. This factor is foundational to S&T investment strategy, and how a program or body of S&T is planned, selected, managed, coordinated, integrated, and transitioned. It is imperative that the latest information technology resources be used to the greatest extent possible during the complete

S&T evaluation process to insure that global S&T resources are being exploited to the maximum extent.

#### (11) Cost of S&T Evaluations

The *eleventh* important factor is the cost of S&T evaluations. The true total costs of developing a high quality evaluation using sophisticated normalization techniques and diverse experts for analyses and interpretation can be considerable, but tend to be understated. For high quality evaluations, where sufficient expertise is represented on the evaluation team, the major contributor to total costs is the time of all the individuals involved in gathering, presenting, normalizing, and interpreting the data. With high quality personnel involved in the evaluation process, time costs are high, and the total evaluation costs can be non-negligible. Especially when suites of diagnostics are combined, as when a metrics-based evaluation is performed in tandem to a qualitative peer-review process (Kostoff, 1997b), the real costs of these experts could be substantial. Costs should not be neglected in designing a high quality S&T evaluation process.

#### (12) Maintenance of High Ethical Standards

The *twelfth*, and final, critical factor, and perhaps a foundational factor, in any high quality S&T evaluation is the maintenance of high ethical standards throughout the process. There is a plethora of potential ethical issues, including technical fraud, technical misconduct, betraying confidential information, and unduly profiting from access to privileged information, because there is an inherent bias/ conflict of interest in the process when real experts are desired to design, analyze, and interpret an S&T evaluation. The evaluation managers need to be vigilant for undue signs of distortion aimed at personal gain.

In summary, for management decision-aids to gain wider acceptance, more attention needs to be paid to quality. This includes both intrinsic/ extrinsic quality, and implementation quality. The quality metrics need to be sharpened for specific applications, the requirements for high quality applications have to be considered carefully, and the decision aids need to be integrated into an organization's overall management processes.

### **TEXTUAL DATA MINING**

Now that the role of decision aids has been examined in the context of their support of the strategic management process, the focus of the report sharpens to address TDM specifically. Data mining in general, and TDM in particular, are defined. The

impact of TDM on strategic management is described. The background, structure and objectives, and lessons learned, of a prototype TDM implementation program are discussed, including the finding that the need and technology exist for large-scale implementation of TDM. A proposed TDM implementation process is outlined, based on the prototype program. Because of the foundational role of text S&T databases in TDM support of S&T strategic management, and the perceived present deficiencies of global S&T text databases to support the text mining process, actions critical to upgrading the quantity and quality of global S&T databases are specified.

## **DEFINITIONS**

Generically, data mining is the extraction of useful information from data. Conceptually, data mining can be divided into two major categories, non-textual (structured) and textual (unstructured).

Non-textual data mining focuses on data within a structured context, such as numbers, images, and words as data. Its main use has been in the classification, correlation, and clustering of data to identify patterns and relationships of interest. It is especially valuable for physical data analysis, and analyses of other types of multi-attribute systems.

Textual data mining focuses on words and phrases within an unstructured context (i.e., free text). It also has been used for classification, correlation, and clustering of data to identify patterns and relationships of interest. However, these relationships and patterns are in the realm of concepts rather than attributes. Since words and phrases are the projections of concepts onto the communication plane of the communicator, and since each communicator uses a unique plane of communication, the de-convolution of the phrase and word data back to the concepts that generated them is a complex and non-unique mapping process. Because much of the language of S&T is unstructured free text, TDM is valuable for analyses of this textual component. The remainder of this discussion focuses on TDM.

TDM can be subdivided further into two categories, non-computer assisted and computer assisted. Non-computer assisted TDM represents the bulk of TDM today. Experts in the subject area of the text read and analyze the literature of interest without the assistance of any computer analytic tools. Computer assisted TDM incorporates sophisticated information technology tools and techniques to augment the experts' analyses of the literature. To extract useful information from

the large volumes of text that are available today in electronic form, computer assisted TDM is a necessity. Computer-assisted TDM is at the earliest stage of incubation, and is ripe for advancement and exploitation (Losiewicz, Oard, & Kostoff, 2000).

## **TEXTUAL DATA MINING AND S&T STRATEGIC MANAGEMENT**

As a management decision-aid, TDM must be a quality tool that benefits the organization in its strategic evaluation of S&T. As stated earlier, the TDM implementation process proposed below in this report is designed to overcome the mismatch between performers and users. By generating valuable information on S&T, both at the level of the organization and the discipline(s) or topics, TDM helps answer the following critical questions:

- (1) *What* S&T is being performed (globally and by the organization, its industry, or its parent agency)?
- (2) *Who* is performing this work?
- (3) *Where* is it being performed?
- (4) What *messages, patterns, and relationships* can be extrapolated from the databases mined? and
- (5) What is *not* being performed (globally or at the level of the organization, industry, or patent agency)?

Answers to these questions, albeit perhaps partial or incomplete, may contribute to the following components of the strategic management of S&T. First, the long term *planning* of S&T benefits from a more precise view of what S&T is and is not being performed. Formulation of strategic goals is a process that depends on background knowledge of current S&T achievements and S&T directions in which these topics may progress. A critical component of the S&T strategic planning is the projection of specific human resource skills necessary for achieving the desired S&T goals.

Second, the identification and selection of management procedures may greatly benefit from the generation of knowledge about patterns, trends, messages, and relationships in S&T performed by the organization and by other entities, as well as globally. Such knowledge would indicate gaps in the organization's S&T (*vis-a-vis* its goals, needs, and the requirements of the parent agency), and in the skills and competencies that the organization requires to perform and to survive. These competencies depend on inputs from S&T, hence the gaps between what S&T is

globally performed and what is done within the organization are strong indications of strategic deficiencies (Fraser and Sibley, 1998).

Third, the selection of review, oversight, and evaluation - imposed by the sponsoring agency and by senior management - will greatly benefit from answers to the five questions above. TDM may yield knowledge about the state-of-the-art and how the organization compares with similar S&T institutions. Such knowledge may provide background for the selection and application of standards and review benchmarks.

Finally, TDM may yield knowledge about relationships and patterns in global S&T that allow the organization to develop a strategic view of the S&T environment it faces. This in turn can be translated by management to chart its interactions with the environment and to introduce necessary changes in the general as well as tactical direction of S&T.

TDM may also be used to generate knowledge on commercialization, technology transfer, and the link between strategic goals and the mission of government organizations—and their S&T programs. TDM aids in establishing the link between the national strategic goals and the strategic goals of the government agency. By extension, TDM also helps in establishing the link between the strategic goals of the agency, and the goals for its S&T programs and activities. Then, TDM may also provide the tool to link the strategic goals of agency and S&T, to the outputs and tactical management objectives of the S&T programs.

Government agencies and their S&T programs are subjected to evaluation requirements from Congress and the Administration. They must show credible linkage between national strategic goals, the mission of the agency, and the ways and means by which their S&T contributes to such strategic goals. Managers of these agencies and S&T programs need decision aids that allow them to establish such linkages and to adequately measure them. TDM is such a decision-aid that can monitor and help in the assessment of the strategic *and* the tactical components of S&T, its generation, and its transfer and commercialization, so as to achieve the contributions from government S&T to higher national and strategic goals.

However, these potential benefits will not materialize unless the organization is able to appropriately implement TDM. Thus, implementation (as complex and difficult as it may be) is a crucial element of successful TDM (App, 1997). A 1998 prototype TDM program at the Office of Naval Research showed that successful

implementation requires three conditions: establishing the TDM resource infrastructure; providing a level of training that will result in high quality output; providing incentives and motivation for using TDM in strategic and tactical applications. In this report, a framework for implementation is proposed based on the findings of the prototype program.

## **PROTOTYPE IMPLEMENTATION PROGRAM**

### **Background**

The prototype TDM implementation program conducted in FY98 at the Office of Naval Research was the culmination of seven years of prior preparatory efforts. These prior efforts were conducted to remedy deficiencies in the then-existing TDM approaches designed in particular to support research evaluation. Some of this history that describes the evolution of co-word-based TDM in 1991 to its use in the prototype implementation program will now be summarized. A much more detailed description can be found in Kostoff et al. (1997f, 1998a).

Much of the forefront TDM used for research evaluation prior to the initial efforts in the early 1990s was centered about the use of co-word analysis. This technique is based on analyzing the co-occurrence frequency of words or phrases in the same syntactic domain. Modern development of co-word analysis for purposes of evaluating research originated in the mid-1970s (Callon *et al.*, 1979; Callon *et al.*, 1983; Callon, 1986). The method developed initially by Callon focused on analyzing the content of articles and reports. In one of the first descriptions and applications of the method (Callon *et al.*, 1979), the impact of French government intervention in the field of macromolecular chemistry was examined. A database of over 4,000 articles covering the field of interest was generated. Key or index words were assigned to each article in the database. A basic assumption was that the key words describing an article had some linkages in the author's mind, and the different fields or functions represented by these words had some relation.

Each time a pair of words occurred together in the key word list of an article, it was counted as a co-occurrence of the pair. The number of co-occurrences for each pair was calculated for all the articles in the database. A co-occurrence matrix was constructed whose axes were the index words in the database and whose elements were the number of pair co-occurrences of the index words. A two-dimensional map was constructed that would display visually the positions of



the key words relative to each other based on their co-occurrence values from the matrix. While different maps had different axes pairs, the central features of the maps appeared to be display of the relationship structures, and the strength of the relationships, between the words.

There were at least two major problems with this approach: (1) the text was not analyzed directly; and (2) the analysis was performed solely on the key words. The bias and error introduced from key word analysis was unknown, but use of key words continued to affect the credibility of the technique for years (Healey, Rothman, & Hoch, 1986; Leydesdorff, 1987).

Subsequent co-word studies focused on: biotechnology (Rip & Courtial, 1984); aquaculture (Bauin, 1986); patents (Callon, 1986); industrial ceramics (Turner *et al.*, 1988); polymer science (Callon, Courtial, & Laville, 1991); neural networks (Van Raan & Tijssen, 1991); chemical engineering (Peters & Van Raan, 1991); combined word frequency analysis of citing articles with co-citation analysis (Braam & Van Raan, 1991a; 1991b); and material science (Van Raan, 1996). All of these reported studies used key words or index words, not full text.

Callon's classical co-word analysis did not allow the richness of the semantic relationships in full text to be exploited, and it was restricted to formally published papers. In order to allow any form of free text to be used, Database Tomography (DT) was developed (Kostoff, 1991a, 1995).

In 1990-1991, experiments were performed at the Office of Naval Research (Kostoff, 1991b) that showed the frequency with which phrases appeared in full text narrative technical documents was related to the main themes of the text. The phrases with the highest frequencies of appearance represented the main, 'pervasive' themes of the text. In addition, the experiments showed that the physical proximity of the phrases was related to the thematic proximity. These experiments formed the basis of DT.

The DT method in its entirety requires generically three distinct steps. The first step is identification of the main themes of the text being analyzed. The second step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps will be summarized below. Time evolutions of themes have not yet been performed.

First, the frequencies of appearance in the total text of all single word phrases (e.g., matrix), adjacent double word phrases (e.g., metal matrix), and adjacent triple word phrases (e.g., metal matrix composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database.

Second, for each theme phrase, the frequencies of phrases within  $\pm M$  (nominally 50) words of the theme phrase for every occurrence in the full text are computed, and a phrase frequency dictionary is constructed. This dictionary contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster) yielding, among many results, those sub-themes closely related to and supportive of the main cluster theme.

Third, threshold values are assigned to the numerical indices, and these indices are used to filter out the most closely related phrases to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allows an understanding of the theme interrelationships not heretofore possible with previous text abstraction techniques (using index words, key words, etc.).

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles (Kostoff, 1991a, 1992, 1993, 1994), the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated infrastructure/ bibliometric information (authors, journals, addresses, etc), the final results have also included relationships among the technical themes and authors, journals, institutions, etc (Kostoff et al., 1997j-k, 1998a, 1999a, 1999f, 2000a, 2000b, 2001d, 2001g, 2002a, 2003a, c-g, 2004a-h, j-o).

These more recent journal-abstract-based DT processes performed represent the framework of a TDM approach that couples the TDM research and associated computer technology processes closely with the TDM user. Strategic database maps are developed on the front end of the process using bibliometrics and DT,

with heavy involvement from topical domain experts (either users or their proxies) in the DT component of strategic map generation. The strategic maps themselves are then used as guidelines for detailed expert analysis of segments of the total database. The authors believe that this is the proper use of automated techniques for TDM, to augment and amplify the capabilities of the expert by providing insights to the database structure and contents, not to replace the experts by a combination of machines and non-experts.

## **Objectives and Structure**

The confluence of: (1) the rapid expansion of information technology hardware and software in recent years; (2) the rapid expansion of massive S&T electronic databases with wide availability; (3) the technical results and knowledge gained from the recent DT studies; and (4) the perceived strategic need of the naval forces for TDM augmentation of their capabilities as they evolve toward network-centric information technology dominated operation, resulted in a 1997 proposal by the first author to establish a prototype program for implementation and integration of TDM. The program's two specific objectives were: (1) demonstrate feasibility and added value of employing topical area experts on the TDM studies; and (2) understand how to apply TDM to a broad spectrum of databases.

The approved FY98 program contained the same four basic building blocks of the prior DT-based activities: (1) information retrieval using an iterative query approach with relevance feedback and term expansion; (2) bibliometric studies of retrieved records; (3) computational linguistics studies of retrieved records; and (4) interpretation and analysis of retrieved records and computer output. Unlike the previous DT studies, where the majority of the funds went toward process development, the majority of the funding for studies performed in the FY98 program was allocated to topical area experts. Three studies were performed [ship hydrodynamics S&T, aircraft S&T (Kostoff *et al.*, 2000b), fullerenes S&T (Kostoff *et al.*, 2000a)], using a total of five topical area experts. Six source databases were examined in the course of the three studies. These included databases of technical papers and reports (Science Citation Index, Engineering Compindex, National Technical Information Service Technical Reports), databases of government and industry project narratives (RADIUS, IR&D), and a web-based patent database. Time and resource limitations permitted only the technical papers and reports databases to be used in the studies.

## LESSONS LEARNED FROM DEMONSTRATION PROGRAM

A number of valuable lessons were learned from the FY98 program, and were incorporated in future efforts. These lessons are now summarized. The first four lessons discussed relate to the four basic building blocks of the FY98 program described above.

### **(1) Value of Iterative Query Reformulation**

The iterative query approach of simulated nucleation (Kostoff *et al.*, 1997f) was used for all studies. Typical queries required about three iterations until convergence was obtained, and ranged in size from a dozen terms to a couple of hundred terms.

The query size depended on the objectives of the study, and the contents of the database relative to the study objectives.

For example, one of the studies focused on the S&T of the aircraft platform (Kostoff *et al.*, 1999c). The query philosophy was to start with AIRCRAFT, then add terms which would expand the aircraft S&T papers retrieved and would eliminate papers not relevant to aircraft S&T. The SCI query required 207 terms and 3 iterations, while the EC query required 13 terms and one iteration. Because of the technology focus of the EC, most of the papers retrieved using an *aircraft* or *helicopter* type query term focused on the S&T of the platform itself, and were aligned with the study goals. Because of the research focus of the SCI, many of the papers retrieved focused on the science that could be performed from the aircraft platform, rather than the S&T of the platform, and were not aligned with the study goals. Therefore, no adjustments were required to the EC query, whereas many *not* Boolean terms were required to eliminate aircraft papers not aligned with the main study objectives. It is analogous to the selection of a mathematical coordinate system for solving a physical problem. If the grid lines are well aligned with the physical problem to be solved, the equations will be relatively simple. If the grid lines are not well aligned, the equations will contain a large number of terms required to translate between the geometry of the physical problem and the geometry of the coordinate system.

The iterative query approach provided an increased ratio of relevant to non-relevant papers; it provided an increased signal-to-noise ratio. The approach allowed more records in the specific targeted field to be retrieved; it provided an increased signal.

The approach allowed more records in allied S&T fields to be retrieved, and in some cases allowed relevant records in disparate fields to be retrieved. The latter

capability has high potential for generating innovation and discovery from disparate disciplines (Kostoff, 1997e, 2003b).

## **(2) Value of Bibliometrics**

Frequency lists were generated (in highest frequency order) of authors, journals, organizations, countries, cited authors, cited papers, and cited journals. Bibliometric analyses were then performed on the retrieved records, and comparisons were made among the diverse disciplines studied (Kostoff *et al.*, 2000a-b, 2001d).

The frequency lists, and the subsequent analyses, allowed the critical infrastructure in each field to be identified. This is useful for identifying credible experts for workshops and review panels, and for planning itineraries of productive individuals and organizations to be visited. For assessment purposes, the bibliometrics allowed productivity and impact of specific papers/ authors/ organizations to be tracked and estimated. For further analyses, the bibliometrics allowed the critical intellectual heritage in each field (highly cited authors/ papers /journals) to be identified. For perspective and context, it is important to compare bibliometrics across disciplines, so that anomalies in any one discipline can be spotted more easily, and universal trends can also be identified.

## **(3) Value of Computational Linguistics**

### *3.1 Phrase Frequency Analysis*

Single word, adjacent double word, and adjacent triple word phrases were extracted from the abstracts of the retrieved papers, and their frequencies of occurrence in the text were computed. A taxonomy was generated (top-down, bottom-up, or some hybrid) whose categories covered the technical scope of the phrases, and the phrases and their associated occurrence frequencies were placed in the appropriate categories. The frequencies in each category were summed, thereby providing an estimate of levels of category technical emphasis on a global basis.

This proved to be a very useful approach for estimating levels of emphasis ('Emphasis' is used rather than 'effort', since *phrases* rather than funding were being computed). It allowed judgements of *adequacy* and *deficiency* in selected S&T categories to be made. However, in order for these judgements to be made, some additional context was necessary. Either requirements-driven levels of emphasis for the different categories needed to be provided, and/ or opportunity-driven levels of emphasis for the different categories needed to be provided. For the specific areas

studied, phrase frequency analyses of requirements/ guidance documents were performed to obtain quantitative estimates of levels of emphasis for context, and the phrase frequency results from the S&T documents were then matrixed against the phrase frequency results from the requirements /guidance documents. Judgements of adequacy and deficiency of technical emphasis in the different categories could be estimated. The opportunity-driven levels of emphasis, which are a statement of what could be done in the categories with state-of-the-art S&T, were estimated based on intuition and judgement of the technical experts, and more softly matrixed against the phrase frequency results from the S&T documents to provide further judgements of adequacy and deficiency.

Obviously, more hierarchical levels in the taxonomy lead to greater resolution of the subcategories, and thereby to greater specificity of judgements of adequacy and deficiency that can be made. For example, if the lowest level materials category in a taxonomy of ship subsystems is 'materials', then a gross judgement of adequacy or deficiency of technical emphasis in 'materials' is all that can be made. This does not help guide decisions because of the lack of specificity. If, however, the lowest level materials category includes subcategories such as 'welded titanium alloys', then judgements as to the adequacy or deficiency of technical emphasis in 'welded titanium alloys' can be made. The more detailed the category, the more useful the result from a programmatic viewpoint, and the greater are the numbers of adequacy or deficiency judgements that can be made. However, the greater the number of categories, the lower the frequencies of the phrases required for statistical significance, the greater the amount of work required, and the more expensive and time consuming is the study. Thus, a tradeoff between study time and costs, and quality of results required, must be performed.

It was also found useful to apply phrase frequency analysis to multiple database fields to gain different perspectives. The fields (keywords, titles, abstracts) are used by their originators for different purposes, and the phrase frequency results can provide a different picture of the overall discipline studied based on which field was examined. For example, in the aircraft study (Kostoff *et al.*, 2000a), a high frequency keyword focal area was concentrated on the mature technology issues of longevity and maintenance; this view of the aircraft literature was not evident from the high frequency abstract phrases. The lower frequency abstract phrases had to be accessed to identify thrusts in this mature technology/ longevity/maintenance area.

Keywords are author/ indexer summary judgements of the main focus of the paper's contents, and represent a higher level description of the contents than the actual words in the paper/ abstract. Thus, one explanation for the difference between the conclusions from the high frequency keywords and abstract phrases is that the body of non-maintenance abstract phrases, when considered in aggregate from a gestalt viewpoint, are viewed by the author/ indexer as maintenance/ longevity oriented. However, while there may be a difference in high frequency phrases between the two data sources, there may be far less of a difference when both high and low frequency phrases are considered. Thus, a second possible explanation is that, in some technical areas in different databases, there is more diversity in descriptive language employed. Rather than a few high frequency phrases to describe the area, many diverse low frequency phrases are used. This could result from the research encompassing a wider spectrum of smaller effort topics. It could also result from the absence of a recognized discipline, with its accepted associated language. This would reflect the arbitrary combination of a number of diverse fields to produce the technical area, with the associated numerous but low frequency thrusts. Another explanation is that maintenance and longevity issues are politically popular now, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

Also, the abstract phrases from the aircraft study contain heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the keywords. Again, the indexers may view much of the testing as a means to the end, and their keywords reflect the ultimate objectives or applications rather than the detailed approaches for reaching these objectives. However, there was also emphasis on high performance in the abstract phrases, a category conspicuously absent from the keywords. In fact, the presence of mature technology and longevity descriptors in the keywords, coupled with the absence of high performance descriptors, provides a very different picture of aircraft literature research from the presence of high performance descriptors in the abstract phrases, coupled with the absence of mature technology and longevity/maintenance descriptors.

This analytical procedure, and subsequent analytical procedures based on the phrase proximity results (described later), are not independent of the analyst's domain knowledge; they are, in fact, expert-centric. The computer techniques play a strong supporting role, but they are subservient to the expert, and not vice versa. The computer-derived results help guide and structure the expert's analytical processes; the computer output provides a framework upon which the expert can construct a comprehensive story. The conclusions, however, will reflect the biases and

limitations of the expert(s). Thus, a fully credible analysis requires not only domain knowledge by the analyst(s), but probably domain knowledge representing a diversity of backgrounds. It was also found useful in each study to incorporate a generalist with substantial experience in analyzing different technical domains, who could identify unique patterns for that technical domain not evident to the domain experts.

### *3.2 Phrase Proximity Analysis*

After the high frequency phrases had been identified from the phrase frequency analysis, phrases of particular interest to the objectives of the study were selected.

The algorithms then constructed frequency dictionaries of phrases in the text located in close physical proximity to the phrase of interest, with numerical indicators accompanying each dictionary phrase. The indicators served as filters, and allowed only those phrases most closely associated with the theme phrase to be selected finally. The process was applied to different database fields/ combination of fields to generate a variety of association results.

Applied to the infrastructure component (title, author, journal, organization, country), the proximity analysis identified the key authors/ journals/ organizations closely related to specific technical areas of interest. This is particularly useful when attempting to define the infrastructure for an unfamiliar area. Applied to the abstract component, proximity analysis allowed closely related themes to be identified. This may be of particular value in identifying low frequency phrases closely associated with high frequency themes; the so-called 'needle-in-a haystack'.

In this application, however, the background and perspective of the technical expert were extremely important, since the core requirement is to recognize signal from a substantial amount of clutter.

Further applied to the abstract, proximity analysis allows taxonomies with relatively independent categories to be generated using a 'bottom-up' approach. This is a potentially powerful capability, since taxonomies are used in all phases of S&T performance and management, and a technique that can generate credible taxonomies semi-autonomously in relatively undeveloped disciplines has many applications. In the present DT approach, the taxonomies are generated by selecting many high frequency themes from the phrase frequency analysis, constructing a phrase frequency dictionary for each theme of phrases located physically close to the theme in the text, and then grouping related themes whose dictionaries contain more than a threshold number of shared phrases into categories. The process is somewhat labor intensive at present, but has the potential for substantial automation with time and labor reduction.



Applied to different record fields as part of the query process, proximity analysis allows complementary and disparate literatures that contain themes related to the target literature to be accessed. This approach has a high potential for innovation and discovery from other disciplines (Kostoff et al., 1997f, 2003b). Finally, proximity analysis has proven to be useful for estimating levels of technical emphasis closely associated with specific technical sub-areas.

#### **(4) Value of Technical Domain Expertise: The Learning Curve**

The FY98 experience showed conclusively that, for a high quality text mining study, close involvement of the technical domain expert is required in all stages where the computational linguistics component was used (information retrieval, phrase frequency and proximity analyses, and integrated interpretation and analysis). To insure that multiple perspectives are incorporated into the study, such that maximum data anomalies will be detected, multiple domain experts with diverse backgrounds and text mining experts who have analyzed many different disciplines are required.

From an organization's long-range strategic viewpoint, the main output from a text mining study is not necessarily the documents or files of data generated. The main output is the technical expert(s) who has had his/ her horizons and perspectives broadened substantially as a result of participating in the full text mining process.

The text mining tools/ techniques/ tangible products are of secondary importance to the organization's long-term strategic health relative to the expert with advanced capabilities. There was a steep learning curve required to integrate the domain expert with the computational tools. The operational mechanics were not the problem; the major roadblock was the time required for the expert to understand how the tools should be applied to address the study's specific objectives, and how their products should be analyzed and interpreted. The problem stems from the fact that text mining requires additional skills beyond traditional science and engineering training and experience, and technical domain experts do not necessarily develop such skills in the traditional technical specialty career. Due to the learning curve problem, substantial time was required to train the expert how to use and interpret computational tools.

#### **(5) Update of Text Mining Lessons Learned**

The initial version of this report was published in 1999 (Kostoff and Geisler, 1999b).

Since that time, additional text mining studies have been performed, as mentioned previously. These studies have strengthened the validity of the above lessons

learned from the FY98 pilot program, and have provided additional insights as well. These post-1999 studies have continued the discipline-oriented studies published in the 1997-1999 period, and have focused on aircraft technology, ship hydrodynamics, fullerenes research, analytical chemistry, electric power sources, electrochemical power sources, fractals, nonlinear dynamics, wireless, LANs, abrupt wing stall, neuroscience, and fullerenes applications.

Additionally, the technique of Citation Mining was developed. In Citation Mining, one or more published research papers is selected as the unit of analysis. All the papers that cite the unit of analysis are retrieved, and text mining is performed on the citing papers. Two Citation Mining studies have been completed. One focused on granular system dynamics (Kostoff et al, 2001c), and the other focused on macromolecular mass spectrometry (Kostoff et al, 2004d).

Also, the technique of literature-based asymmetry detection was developed. In literature-based asymmetry detection, text mining of a target literature is used to identify potential asymmetries where none would be expected. Then, subsets of the target literatures that consist of only the asymmetric categories are retrieved, and the ratios of these retrievals are predicted to reflect the ratios of the actual asymmetries.

So far, one study of literature-based asymmetry detection has been performed (Kostoff, 2003c). The ratios of bilateral cancer incidence have been predicted for four organs, and have shown to exhibit excellent agreement with actual patient incidence.

Underlying the development and demonstration of these advanced text mining applications has been the development of computational linguistic processes for text clustering and interpretation. While most text mining researchers have focused on algorithm development, the ONR text mining effort has continued to concentrate on developing the processes in which the algorithms are imbedded. In a surgical analogy, the mainstream text mining community has developed advanced scalpels, with few surgical objectives and processes. The ONR effort has developed the surgical objectives and techniques, with scalpels as required. The supporting evidence for these assertions and conclusions is that most of the ONR technical specialty studies are published in the technical specialty literatures, while the rest of the text mining community is limited to publishing in the information technology literature only, due to the community's algorithmic focus and process de-focus.

Most of the ONR upgrades since the 1999 paper was published have been in the area of concept and document clustering. The 2000-2002 published studies added

statistical concept clustering (both single word and multi-word phrases), and began to eliminate manual clustering. The gains were savings in time and additional perspectives offered by the statistical groupings. The losses were the perspectives that only a human expert can provide when assigning text to different categories. On balance, the trade-off is viewed as highly cost-effective.

The post-2002 studies have added different types of document clustering. Assignment of document clusters to categories defined by the concept (words/phrases) clusters, and counting the number of documents in each category, has supplanted the assignment of words/ phrases to these categories, and the subsequent counting of their frequencies, as a method for estimating levels of emphasis.

As the concept and document clustering capabilities have expanded, the study scopes have evolved and transitioned from a single discipline focus to a multiple discipline focus. The latest multi-discipline studies are along two lines. One is country and regional studies, where the emphases are identification of core technology competencies and the supporting infrastructure for each competency. The other is literature-based discovery, where the concept is to identify disparate and disjoint literatures where advances in one or more disciplines could be extractor to other disciplines for innovation and discovery. This multi-discipline text mining study trend has increased the requirement for performers who are cognizant of multiple disparate disciplines. It has narrowed the pool of candidates with sufficient expertise and breadth, and surfaced dramatically the immediate need for educating and developing this cadre of broadly-based experts.

Further, recent experience has shown that even multi-disciplinary efforts are limited when addressing the newest types of text mining challenges. Inter-disciplinary capabilities are required to provide the most comprehensive results. Here, performers are required to be:

individually skilled in the information technology processes and tools,  
expert in the central technology themes, and  
highly knowledgeable in science and technology areas that could contribute to knowledge advances in the central technology themes.

The paradox is that the specific expertise requirements in the disparate technologies may not be known when the studies are initiated, but surface as the studies proceed and related disciplines are uncovered. Compounding the challenge imposed by the limited human resource pool is the additional challenge of inducing such people to

work on truly inter-disciplinary efforts. As a recent study has shown (Kostoff, 2002c), true interdisciplinary research faces the barriers of Culture, Time, Evaluation, Publication, Employment, Funding, Promotion, and Recognition. Each of these barriers provides a dis-incentive for active participation in inter-disciplinary efforts. Substantial changes in the incentives and rewards for performing true interdisciplinary research will be required to develop the cadre necessary to work on multi-discipline text mining studies.

As will be discussed in more detail in a later section, an equally, if not more, serious barrier to credible text mining studies is the quality of the base text data itself. Even using the most advanced information extraction techniques coupled with trained interdisciplinary personnel, the best text processing and interpretation cannot compensate for lack of text data or poor quality text data. The fundamental text data is severely deficient for the following reasons:

- relatively small fraction of S&T performed globally documented;
- many more dis-incentives than incentives to publish;
- small fraction of documented S&T reaches widely disseminated databases;
- volume and quality of database content varies widely.

When the data deficiency is combined with the deficiency of information extraction processes, the resource infrastructure available to address the challenging multi-discipline text mining problems described above is highly sub-standard. The problem needs to be addressed on multiple fronts in parallel before serious inroads in the text mining challenges can be made.

## **(6) Cost and Time Estimates**

Because of the start-up costs associated with the learning curve, long-range involvement of the expert(s) with text mining of the program/ topic area is cost-effective. For TDM studies that are not overly time-intensive, the Program Officers in an S&T sponsoring organization could serve as the technical experts. For more detailed time-intensive TDM studies, the Program Officers might require support from contractors, or might want the contractors to perform the complete TDM study. To insure that long-range involvement is executed appropriately, a strategic plan showing how text mining is integrated into an organization's business operations is required. Such a plan would address the role of textual data mining in the context of overall data mining, the role of the organization's overall data mining in the context of allied organizations' data mining in similar technical areas, and how

different types/ classes of technical experts should be integrated most efficiently into the data mining process.

Text mining cannot be used sporadically to realize its full benefits, but must become an integral part of any S&T sponsor's business operations. A strategic plan that presents TDM in this larger context is required to insure that text mining integration is implemented in a cost-effective manner. Such a plan would identify the different ways text mining would support the S&T sponsor's operations, such as planning, reviews, assessments, metrics, oversight response, etc. Each of these applications has different objectives, metrics to address those specific objectives, data requirements for each metric, different types of experts required, and different suites of text mining tools required. A strategic plan allows a top-down driven approach to text mining, in which the desired objectives are the starting point, and the data required to satisfy the objectives can be identified, and planned for, in advance. Without such a plan, the organization is constrained by whatever data exists and has been gathered for other purposes. This bottom-up approach forces the organization to use whatever metrics the existing data will support, whether or not these metrics are most appropriate to satisfying the overall objectives of the application.

Since there is a wide range of text mining studies that can be performed. The cost and time of each study will depend on the scope of the study and quality of the final product desired. For a text mining study that consists of the four building blocks of the FY98 studies, the cost and time will depend on:

- (a) The complexity of the query, the number of query iterations, and the level of analysis effort applied to each iteration
- (b) The number of bibliometric quantities examined, and the complexity of analysis applied to each metric.
- (c) The number of computational linguistics algorithms employed, the number of different applications of each algorithms desired, and the level of analytical detail associated with the application of each computational linguistics algorithm.
- (d) The number and sophistication of other text mining techniques and tools used, such as clustering, strategic analysis, visualization, or other.
- (e) The complexity of integration and interpretation of results from the above analytical components.

A textual data mining study could range from a simple query of a focused technical field by a Program Officer with little or no analysis to a complex query with complex analyses by an external contractor. The costs associated with these studies could range from no out-of-pocket costs for the simple queries to six figures for complex queries with sophisticated analyses. The times required for such studies could range from minutes to months.

### **(7) Need for Large-Scale Implementation**

As time has evolved since the FY98 pilot program, it has become clear that the TDM pilot program and follow-on efforts could serve as prototypes for how TDM should be implemented on a larger scale. Further, it has become clear that there is sufficient information about existing TDM tools and processes that implementation could provide useful results now, and thereby increase the customer base and support for TDM now. It also has become clear that TDM has substantial value for supporting strategic management of S&T, and that the S&T community would benefit from accelerated introduction of TDM to a wide variety of S&T-related users.

Two major avenues have to be pursued simultaneously if there are to be any chances of a high-quality powerful TDM process achieving wide applicability within the potential user community. First, a broad segment of the S&T-related user community needs to gain understanding of, and experience with, TDM. This is an absolute necessity for converting TDM from a literature-based phenomenon to a working support system. Second, a number of specific tools and especially process development techniques have to be developed and/ or refined. This would increase the quality and power of TDM. It appears that the most prudent way to accomplish both objectives would be to train a wide segment of interested users with the techniques and processes available today, while at the same time upgrading and refining these tools and processes. Once the *cultural* roadblock of using TDM is removed through positive application experiences, acquiring and learning new tools and processes would be reduced to a secondary problem.

A process that would accomplish these two objectives was developed, and is presented in the next section. It could be applied at any organizational level, including that of the total Federal government. It applies to industry or academia as well.

## **PROPOSED IMPLEMENTATION PROCESS**

## Objective

The objective of this proposed process is to augment the capabilities of S&T managers for covering their technical spheres of responsibility by providing them in-house training for using TDM. This would provide the managers with both the tools, and the understanding of how to apply the tools to achieve the desired enhanced awareness of their technical fields. It would be a major first step in integrating the text mining capability with the management of S&T.

## Approach

The approach proposed is learning by doing, and is based mainly on the experiences from the FY98 TDM pilot program, and subsequent follow-on efforts. The pilot program showed the necessity for a 'hands-on' experiential approach with close supervision to provide technical experts a basic understanding of TDM principles and applications.

The initial pilot program employed five separate technical domain expert contractors in FY98, none of whom had any real text mining experience. They were initially provided with papers outlining the techniques to be used, and were provided briefings on the details of the approach. However, the contractors made little progress until they started working the assigned problem 'hands-on', accompanied by very close supervision. In the initial phases of the project, they were able to become familiar with the mechanics of text mining (i.e., operating the algorithms).

It was not until the final interpretive and integrative phases that they developed some understanding of where and how to apply the techniques and algorithms, and how to perform the analyses to extract substantial information from the data. In other words, the contractors had to ***experience the complete process*** (i.e., perform a full study) before they developed a minimal understanding of what text mining could offer and how to proceed to obtain its full benefits. At the end of the study, they had developed a much greater understanding of, and appreciation for, the benefits that TDM could offer, and could start to perform studies on their own. They could perform the types of extensive queries and bibliometrics studies that would be of central interest to Program Officers and other S&T program-related personnel in obtaining a more integrated and expanded view of their technical fields.

Subsequent experience has shown that the quality of the final product is increased further when the performers go beyond multi-disciplinary operational modes to inter-disciplinary operational modes. They become versed not only in their own

technical specialty, but the requisite information technology, and other supporting technologies as well.

The approach proposed here would start with an initial two-week full time text mining effort. A group of perhaps three people would constitute the class. It would be conducted in a dedicated on-site text mining facility, so that office responsibilities would not interfere with the training. Each student would bring a problem(s) of work-related interest, and the problem selected for each student to data-mine would be negotiated with the instructor. After some instruction, the students would proceed to work the problem. There would be close supervision by the instructor to enhance the knowledge transfer process. Because of the sequential nature of the text mining approach, almost all of the two weeks would be spent on query development and generation of the related bibliometrics. Operationally, the students would focus almost exclusively on reading abstracts of the literature in their chosen work-related area, deciding on the relevance of specific documents to the central focus of their topical area. After the two-week period, the students would return to their regular activities, and complete the projects on a part time basis. The complete approach and results would be documented and placed on the Web, to allow access to the findings by the larger technical community.

Past experience has shown that when the technical experts complete the full cycle described here, and understand the enhanced capabilities that text mining can provide to support their job responsibilities, they are much more motivated to use text mining as an integral part of their daily activities. Thus, the important output of long-term benefit is the ingraining of the text mining process into the student's psyche. Once they understand the process, they can apply its principles to any database, and easily incorporate new analysis tools as they become available.

## **Resources**

The following resources are proposed.

### *(1)Time*

The time required for each student to complete the initial full cycle is not negligible.

However, experience has shown that substantial time and effort are required to achieve the full benefits of what TDM has to offer. TDM is deceptively simple; many people believe that a straight-forward extrapolation from familiar literature searches is all that is required to achieve TDM-based insights and understanding.

Moreover, there is a widespread belief in the 'magic bullet' approach to TDM. Many people believe that there exist one or a few stand-alone TDM tools that can



be applied easily to various literatures to produce highly useful results to support S&T management.

In fact, TDM is inherently complex. Serious gaps in the base text data exist, due to lack of documentation and poor quality of much of the existing documentation.

The human mind is required to identify these gaps, and compensate for them in the analysis to the greatest extent possible. Further, to assemble concepts and their inter-relations from verbal fragments of even reasonable quality documentation is an extremely difficult task, and is not fully amenable to the mechanistic approaches of the software tools. The context of the verbal fragments is the important driver to enhanced understanding, and assembling the verbal fragments requires the contextual oversight that only the human expert can provide. In order for the human expert to understand how and when to apply the context, substantial learning time is required. The main value of the software tools is not to do TDM, but to organize the input data such that the human expert can place it into its proper context more easily.

### *(2)Instruction*

Quality instruction will be the critical path to wide scale dissemination of high quality TDM techniques. There are very few people with appropriate demonstrated experience in all the important aspects of TDM (query development and information retrieval, bibliometrics, computational linguistics, cross-literature discovery) related to supporting S&T management to insure that high quality training results. There is substantial evidence that people who have been trained poorly in TDM develop negative attitudes about the potential of TDM. This is not due to any intrinsic defects in TDM, but rather because the full capabilities of TDM were not communicated to the student due to instructor deficiencies. Many high quality TDM instructors need to be trained starting now, and the time for training is substantial.

### *(3)Software Tools*

There are many software tools available that could support TDM now. For the initial training, tools selected should have the capability to support iterative query development, bibliometrics, and computational linguistics (phrase frequency and some type of phrase proximity/ clustering to relate concepts from one literature or many literatures). These algorithms should preferably be integrated into one seamless tool, operable on a single platform. As time proceeds and the community's understanding of text mining improves, the tools would continually be updated, but tool development should not hinder the initiation of the implementation. Tool

development, and process improvement, should be funded in parallel with the implementation procedure.

### **Logistics**

Each full-time class would last for four weeks, and contain four students. There would be a one-week interval between classes to allow the instructor to provide support to the students who are completing their project on a part-time basis. This would result in about 40 students per instructor produced annually. This number could be increased or decreased, depending on evolving experiences with the class and its products. Because of the sparse number of qualified instructors presently available relative to what is needed for full-scale implementation, some targeted pilot demonstration programs in a few agencies would be required initially, until adequate competent instructors become available for larger-scale implementation. In fact, this need to train qualified instructors is as important a driver for immediate implementation initiation as the other more technical reasons presented above.

### **Student Personnel**

The initial student pool would include S&T Program Officers, field management and liaison personnel, acquisition personnel, and personnel involved in planning and oversight of S&T programs. The latter group of members may find the class particularly valuable, since it would allow them to access a literature that they may not have the opportunity to access often. Given their projected responsibilities, it would be very useful for them to at least be familiar with the S&T component of this literature. Other types of people could also be added to this class.

### **Student Teams**

One assumption above is that each student works on a separate problem. There may be cases where it is desirable to have all the students in a class work on the same problem. For example, assume one class consists of the principals on a newly-created Integrated Product Team (IPT), whose central focus is developing the S&T for a broad-based operational theme. Hypothesize a theme such as Autonomous Flying Systems, the focal point of an interdisciplinary workshop conducted by the first author in December 1997 (Kostoff, 1997e). The class members could all address the problem of text mining the multi-disciplined literature on autonomous operations, or autonomous systems. Much of the time in the four-week class would be spent on reading this literature, and sharpening the focus of the central theme.. The students could divide the task of reading the literature and deciding on applicability of the records to the central theme, or all the students

could read all the records and ascertain whether they have consensus as to applicability of individual records to the central theme.

The lessons from the FY98 pilot TDM program and its aftermath have shown there is no better way of sharpening the focus of the central theme than this relevance selection procedure. In fact, this four-week process would be an excellent method of having the IPT initiate work as a team on the central technical problem of interest. The combination of the succeeding bibliometrics component and technical thrust analysis component (derived from a combination of research and development databases) would allow the IPT to define an appropriate mix of inter-disciplinary and multi-category topics and personnel for a workshop on the central theme. What better way to initiate the technical definition component of an IPT than this?

### **Quality Control**

It is of utmost importance that a high level of quality be maintained throughout the text mining process. Experience has shown that low-quality text mining will disinterest potential users, whereas high quality text mining will motivate people to incorporate it into their daily work activities. To insure that the students are motivated to perform at their best, they would be evaluated on their performance, and this would contribute to their annual performance review. If such discipline is not invoked, there is the danger that the text mining class could be treated in the cavalier way other types of required briefings are treated. If complete attention is not given to text mining, the results will be shallow, and the intrinsic value of the process will not be evident to the students.

## **DEVELOPING GLOBAL S&T DATABASES**

The prototype implementation program and its aftermath have yielded lessons regarding the nature of the databases necessary for effective implementation and useful application. As a result of these studies, it has become clear that: (1) insufficient S&T results (both foreign and domestic) are being documented, (2) those that are documented are incomplete from the perspective of potential sponsor and user applications, (3) many of those that are documented and incomplete are relatively inaccessible to a wide variety of potential users, and (4) the technology and need now exist to correct this situation on a global basis.

### **Background**

Science and technology have become global, as the world has effectively contracted due to the Internet and other high-speed forms of travel and communications. From an S&T agency sponsor's perspective, there are many applications where knowledge of past, present, and future global S&T products/ programs/ proposals would be of immense value. These applications include both tactical and strategic program planning, program selection and termination, program management and review, program transition and product utilization, product/ program impact and productivity tracking, response to oversight organizations, and public relations. All of these applications have their unique goals and objectives. Each of the goals and objectives for each of the applications has its own unique metrics. Each of these specific metrics has its own unique data requirements.

For strategically managed organizations, the logic flow in developing data for specific objectives should be from goals to metrics to data. The data generated by an organization either internally or from external sources would be that targeted data derived from specific goals and objectives. The situation today is completely convoluted; the S&T organizations are at the mercy of the large database providers (e.g., SCI, RADIUS, Engineering Compendex, NTIS Technical Reports, etc) for the input S&T data. A very simple example for the case of the SCI (Kostoff, 1998b) showed (for example) the absence of sponsor fields, and the inability to distinguish among reference authors with the same name, are major obstacles to productivity and impact tracking. Thus, the present situation is backwards; the available S&T data drives the studies that can be done and the objectives that can be addressed, rather than the objectives driving the data.

## **Requirements**

To overcome these limitations, a series of databases must be developed on a global scale with multi-national support. Three steps are required to generate a useful product.

### *(1) Documenting S&T*

The foundational requirement upon which high quality text mining rests is that past, present, and future S&T that has been, is being, and will be performed should be documented. Contrary to present thinking, where the belief exists that there is too much data being placed in the literature, there is actually a very modest amount of S&T that is documented relative to what could, and should, be documented. Except for unclassified academic research, motivations for the remainder of S&T performers for documenting their output are not high. For truly breakthrough research, from which the performer would be able to profit substantially, the

incentives are to conceal rather than reveal. For research that aims to uncover product problems, there is little motivation (from the vendor or the sponsor or the developer) to advertise or amplify the mistakes that were made or the shortcuts that were taken. For very focused S&T, the objective is to transition to a salable product as quickly as possible; no rewards are forthcoming for documentation, and the time required for documentation reduces the time available for development. Insufficient documentation is not an academic issue; in a variety of ways, it retards the progress of future S&T.

A similar situation occurs in industrial companies. Although there has been a dramatic growth in knowledge management systems installed by many firms, there is a limited amount of technical and managerial data that is deposited in these systems. This phenomenon may be explained by the following: (1) the reluctance of managers to part with knowledge they believe is a component of their power and standing in the organization; (2) the amount of additional work it takes to convert information into a format acceptable to the knowledge system; (3) lack of an orderly procedure for such deposits and withdrawals; and (4) lack of adequate incentive for managers to undertake such additional tasks (Geisler, 1999(a)).

Thus, the first step in the development of the series of multi-national databases envisioned here is to set requirements and procedures for insuring that as much as possible of the S&T that is being performed will be documented. In addition, where feasible, the documentation should be targeted for the highest quality database. As an example, a completed S&T project could be documented for the performer's personal records, as an internal organizational memo, as a limited distribution technical report, as a widely distributed technical report, as a paper in a conference proceeding, and/or as a paper in a peer-reviewed journal.

There are at least two characteristics that distinguish these forms of documentation: (1) the level of expert quality control increases roughly in the order shown, and (2) the pro-active intrinsic quality of the document tends to increase as a result of the knowledge that the quality control of the first item will be enforced. Thus, a requirement that every S&T output document be submitted for journal publication would have the effect of raising the quality bar on what is already substantially documented. The benefits of such improved documents to subsequent S&T would be enormous, and would go a long way toward eliminating the repetition of mistakes. Finally, requirements may have to be established to insure that the documentation contains a) the full scope of information needed to address the objectives and goals discussed above (e.g., every document should contain sponsor information, if appropriate, etc), and b) a comprehensive entry in each covered field.

In particular, with massive electronic databases of journal papers, the Abstract

becomes the operational proxy for the full paper, and the wide spectrum of Abstract content quantity and quality provides a major impediment to uniform database interpretation..

### *(2) Creating S&T Databases*

The next step is to develop different types of comprehensive multi-national S&T databases for different applications. For example, there could be databases of published research papers (e.g., expanded and more complete versions of the SCI), databases of published technology papers (e.g., Engineering Compendex), databases of published technical reports (e.g., NTIS), databases of conference proceedings (some are in EC, but many of the smaller ones never translate into databases), databases of program narratives (e.g., RADIUS), databases of patents (e.g., IBM patent database), databases of new S&T concepts (e.g., electronic agency proposals), databases of off-the-shelf technology, and many other types as well.

The diversity of databases, and the specific fields to be contained in each, would be determined by the different types of applications envisioned by the multi-national sponsors, and the specific data types and formats required for each application. This step has to be closely linked to the first step. The databases have to draw upon and include as much of the S&T documentation as is possible. For example, the SCI could probably double or triple the number of journals it includes with present or near-future technology. It should be the multi-national S&T sponsor agencies, and other entities, along with the state of information technology, that determine the breadth and depth of the different database contents, not the developers.

### *(3) Broad Database Availability*

The final step is to make the databases friendly and readily available to a wide variety of users. Presently, there are many incomplete fragmented databases, each with its own field structures and formats, and with its own unique query and output protocols. Access to many of these databases is very difficult, very limited, and many of these databases are not widely known. For all practical purposes, if a database is not widely known, readily available, and easy to use, it may as well not exist.

In industrial firms, for example, integration of databases is a major challenge for knowledge systems. Diverse databases exacerbate the problems of poor participation by managers. Some attempts at integration, centralization, and streamlining of databases have led to resistance on the part of managers who considered such effort as the unwelcome interference from senior management designed to complicate established work processes (Leon, 1999).

### **Why Not Depend on the Web Alone**

At this point, the question could be raised: Why not depend on the Web for global S&T data? Why do a parallel development, and spend unnecessary funds to replicate what exists on the Web? This is a valid question, and the response derives from the experiences gained from studies of S&T text mining.

One of the key findings from the FY98 text mining pilot program was that, in general, a separate query had to be developed for each database examined. Each database accesses a particular culture, with its unique language and unique types of documentation and expression. A query that optimizes for one database may be very inadequate for another database. The example in the lessons learned section of this report on the Aircraft S&T study, where a query of 207 terms was required to obtain acceptable signal-to-noise ratio for the SCI, while a query of 13 terms produced even a better S/N for the EC, validates this statement. The conclusion reached was that the effort required for an acceptable query depended on the objectives of the overall study, and the relationship of the contents of the database to the objectives of the study.

In addition, the pilot program showed that a query optimized for one field of a database could be very inadequate when applied to another field. Again in the aircraft study, the picture of the total technical discipline derived from a database of record Abstracts was in some cases very different from the picture derived from a database of record keywords. Abstracts, keywords, and titles have different structures, and are generated by authors or indexers for different purposes.

The bottom line that resulted from the pilot program is that developing an S&T database query that will retrieve sufficient technical documents to be of operational use is not a simple procedure. It requires close interaction with technical experts, an in-depth understanding of the contents and structure of the potential databases to be queried, the relation of these database contents to the problem of interest, and substantial time and effort on the part of the technical expert and supporting information technologist. This runs counter to the unfounded assertions being promulgated by the algorithm developers and vendors in the information technology community: sophisticated tools exist that will allow low-cost non-experts to perform comprehensive and useful data retrieval and analysis with minimal expenditures of time and resources.

The Web is a conglomeration of many types of data, with no central structure to the records, with data of widely varying contents and quality and verification, and unknown completeness and coverage. Given the experiences from the pilot program, there is no evidence that a rigorous query of high quality and utility could be made of the Web as it exists now and in the foreseeable future. Even with the more uniform multi-national databases proposed here, serious queries and subsequent text mining will not be simple or easy, and will require substantial time and effort. There are no 'magic bullets' for text mining.

The Web does have its use for simple non-rigorous queries. Some information can be obtained that would not be available from the structured databases envisioned in this memo. The Web could complement the proposed multi-national S&T databases, but its utility for S&T sponsors would be of a far different nature from that of the proposed multi-national databases.

### **Multi-National Database Implementation**

The first step is to identify the total concept. The approach would be to specify the different types of management support studies and operations that require global S&T data. Then, metrics that gauge progress for these studies and operations would be identified, and the data necessary to feed and satisfy these metrics would be specified. The types, and field structures, of databases that would contain this data would be developed. An estimate of the costs, times, and general levels of effort involved in developing, maintaining, and quality controlling these databases would be computed.

The major prospective international partners would then be contacted, and their input would be requested. Larger issues would be included in these discussions (legal, political, economic, etc), and especially cost-sharing and management issues would be addressed. At some point, the State Department, and possibly the Commerce Department, would probably have to be involved.

### **BENEFITS FROM TDM: A MANAGERIAL PERSPECTIVE**

When properly implemented (as suggested in this report), TDM may be more apt to yield strategic benefits to S&T managers, at all levels of government organizations. The existence of global databases that are routinely and systematically mined for useful information represents a very valuable strategic tool. Experience with many industrial companies has shown that managers are generally



unaware of the many merits and the richness of potential benefits that such knowledge systems can offer (Fleisher, 1999).

From the S&T manager's perspective, it is immensely valuable to have insights into the state-of-the-art, trends, emerging topics, and other relationships in a given S&T topic. Such knowledge means much more than a decision-aid. It represents a solid background for strategic formulation and its implementation regarding basic questions: (1) the direction in which S&T in the organization will follow and (2) the composition (topics and resources) of such future S&T effort (Poister and Streib, 1999).

By making certain that these questions are answered, S&T managers have now the capability to place their organization within the existing and the foreseeable state of science and technology. This is akin to a firm's ability to competitively place itself in a winning position in its industry, so that it can sustain such a position in the longer term (Elfring & de Man, 1998).

The prize to be achieved by S&T managers is a competence to be not only competitive but in relative control of their destiny. This is the true essence of strategic management. TDM can contribute to accomplishing this goal and to winning the prize.

## **CONCLUSIONS**

In a competitive global environment, the fundamental limitation to the quality of S&T strategic management is understanding the S&T that has been performed, is being performed, and is planned to be performed. Limits on this S&T understanding constrain the utility of any modern decision aids that support strategic management.

Three avenues exist to enhance this S&T understanding: direct personal knowledge transfer, analysis of tangible physical systems; and analysis of documented results from the management and performance of S&T. Personal knowledge transfer is slow and very limited in scope. Analysis and reverse engineering of tangible physical systems is slow and incomplete, and the S&T understanding obtained is limited.

Analysis of documented results was the focus of the present report. It offers access to the widest amount of information. However, as this report has shown, the existing S&T databases are very incomplete and limited relative to what could be generated with global S&T sponsor agreements. Further, the methods for extracting information from existing databases are very inefficient, and their algorithmic components have not been integrated very well with their human interpretive components.

This report has presented an approach to improving the database limitation problem through the joint construction of multi-national S&T databases whose core data entries are more complete. These databases would be end-use customer requirements driven, rather than database vendor driven. This report has also presented an approach to improving the information extraction process, and has proposed a plan to implement this approach. The implementation plan, like the information extraction approach, requires time and adequate effort. The approach and plan are based on the authors' belief that high quality textual text mining is inherently complex, and no simple 'magic bullets' or other 'quick fixes' will produce the technical intelligence of which text mining is capable. The relatively extensive training proposed and required is probably most cost-effective for organizations that want to develop a long-term strategic TDM capability.

## REFERENCES

- Allison, M. and J. Kaye, *Strategic Planning for Non-Profit Organizations* (New York: John Wiley & Sons, 1997).
- Anwar, M. A., and A. B. Abu Bakar, Current State of Science and Technology in the Muslim World, *Scientometrics*, 40(1), 1997.
- App, S., From Performance Measurement to Performance Management, *Public Manager*, 26(3), 1997, pp. 29-31.
- Bauin, S., Aquaculture: A Field by Bureaucratic Fiat In: Callon, M., J. Law, and A. Rip, (Eds.), *Mapping the Dynamics of Science and Technology* (London: Macmillan Press Ltd., 1986).
- Berry, M., and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support* (New York: John Wiley & Sons, 1997).
- Borok, L., Data Mining: Sophisticated Forms of Managed Care Modeling Through Artificial Intelligence, *Journal of Health Care Finance*, 23(3), 1997, pp. 20-36.
- Braam, R. H. Moed, and A. Van Raan, Mapping of Science by Combined Co-Citation and Word Analysis. 1. Structural Aspects, *Journal of the American Society for Information Science*, 42(4), 1991a and Mapping of Science by Combined Co-Citation and Word Analysis. 2. Dynamical Aspects, *Journal of the American Society for Information Science*, 42(4), 1991b.
- Braun, T., Schubert, A. P., and Kostoff, R. N. "Growth and Trends of Fullerene Research as Reflected in its Journal Literature." *Chemical Reviews*. 100:1. 23-27. January 2000.
- Braun, T., Schubert, A., and Kostoff, R. N. "A Chemistry Field in Search of Applications: Statistical Analysis of U. S. Fullerene Patents". *Journal of Chemical Information and Computer Science*. 42:5. 1011-1015. 2002.

- Bryson, J., *Strategic Planning for Public and Nonprofit Organizations: A Guide to Strengthening and Sustaining Organizational Achievement* (San Francisco: Jossey-Bass Publishers, 1995).
- Burgelman, R., and R. Rosenbloom, Technology Strategy: An Evolutionary Process Perspective, In: M. Tushman and P. Anderson, *Managing Strategic Innovation and Change* (New York: Oxford University Press, 1997) pages 273-286.
- Callon, M., Pinpointing Industrial Invention: An Exploration of Quantitative Methods for the Analysis of Patents In: Callon, M., J. Law, and A. Rip, (Eds.), *Mapping the Dynamics of Science and Technology* (London: Macmillan Press Ltd., 1986).
- Callon-M , J. P. Courtial, and F. Laville, Co-Word Analysis As a Tool for Describing the Network of Interactions Between Basic and Technological ResearchCThe Case of Polymer Chemistry, *Scientometrics*, 22(1), 1991, pp 155-205.
- Callon M., J. P. Courtial, and W. A. Turner, APROXAN: Visual Display Technique for Scientific and Technical Problem Networks, Second Workshop on the Measurement of R&D Output, Paris, France, December 5-6, 1979.
- Callon, M., J. P. Courtial, W. A. Turner, and S. Bauin, From Translations to Problematic Networks: An Introduction to Co-word Analysis, *Social Science Information* 22, 1983.
- Chelsey, J., and M. Wenger, Transforming an Organization: Using Models to Foster a Strategic Conversion, *California Management Review*, 41(3), 1999, pp. 54-73.
- Del Rio, J. A., Kostoff, R. N., Garcia, E. O., Ramirez, A. M., and Humenik, J. A. "Phenomenological Approach to Profile Impact of Scientific Research: Citation Mining." *Advances in Complex Systems*. 5:1. 19-42. 2002j.
- Elfring, T., and A. deMan, Theories of the Firm, Competitive Advantage, and Government Policy, *Technology Analysis & Strategic Management*, 10(3), 1998, pp. 283-293.

- Fleisher, C., Public Policy Competitive Intelligence, *Competitive Intelligence Review*, 10(2), 1999, pp. 23-36.
- Fraser, M., and E. Sibley, Strategic IT Alignment and Managing the Use of Very Large Databases, *Public Manager*, 27(1), 1998, pp. 39-42.
- Geisler, E., Industry-University Cooperation: A Theory of Inter-Organizational Relations, *Technology Analysis & Strategic Management*, 7(2), 1995, pp. 217-229.
- Geisler, E., *Managing the Aftermath of Radical Corporate Change: Reengineering, Restructuring, and Reinvention* (Westport, CT: Quorum Books, 1997) pp. 46-50.
- Geisler, E., Harnessing the Value of Experience in the Knowledge-Driven Firm, *Business Horizons*, May-June, 1999(a), pp. 18-26.
- Geisler, E., Strategic Management of Information Technology: Empirical Findings in Three Sectors, *Proceedings of the Second Portland International Conference on the Management of Technology (PICMET)*, Portland, OR, July 25-29, 1999(b).
- Geisler, E., *The Metrics of Science and Technology: Evaluation and Measurement of Research, Development, and Innovation* (Westport, CT: Quorum Books, 2000).
- Geisler, E., and D. Frey, Commercialization of Energy Related Technology to Industry: The Case of the U.S. National Energy Laboratories, *International Journal of Global Energy Issues*, 9(1-2), 1997, pp. 16-23.
- Greengrass, E., *Information Retrieval: An Overview*, TR-R52-02-96, NSA, 28 February 1997.
- Hall, D., and Nauda, A., An Interactive Approach for Selecting IR&D Projects, *IEEE Transactions on Engineering Management*, (37:2), 1990.
- Healey, P., H. Rothman, and P. Hoch, An Experiment in Science Mapping for Research Planning, *Research Policy*, 15, 1986.

- Kostoff, R. N., *Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis*, Proceedings, Portland International Conference on Management of Engineering and Technology, 1991a.
- Kostoff, R. N., *Word Frequency Analysis of Text Databases*, ONR Memorandum, 5000 Ser 10P4/1443, April 12, 1991b.
- Kostoff, R. N., *Research Impact Assessment*, Proceedings, Third International Conference on Management of Technology, Miami, FL, 1992. Larger text available from author.
- Kostoff, R. N., Database Tomography for Technical Intelligence, *Competitive Intelligence Eview*, 4(1), 1993.
- Kostoff, R.N., 1994, Database Tomography: Origins and Applications, *Competitive Intelligence Review. Special Issue on Technology*, 5(1), 1994.
- Kostoff, R. N. *et al.*, *System and Method for Database Tomography*, U.S. Patent Number 5440481, 1995.
- Kostoff, R. N., "*The Handbook of Research Impact Assessment*", Seventh Edition, Summer 1997, DTIC Report Number ADA-296021. Also, see [http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/), 1997a.
- Kostoff, R. N., "Peer Review: The Appropriate GPRA Metric for Research", *Science*, Volume 277, 1 August 1997b.
- Kostoff, R. N., "Research Program Peer Review: Principles, Practices, Protocols", [http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/), 1997c.
- Kostoff, R. N., "Science and Technology Roadmaps", [http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/), 1997d.
- Kostoff, R. N., *Science and Technology Innovation*, [http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/), 1997e. Also, Kostoff, R. N. "Science and Technology Innovation". *Technovation*. 19:10. 593-604. October 1999.

- Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", *Journal of Information Science*, 23:4, 1997f.
- Kostoff, R. N., "Accelerating the Conversion of Science to Technology: Introduction and Overview", *Journal of Technology Transfer*, Special Issue on Accelerating the Conversion of Science to Technology, 22:3, Fall 1997g.
- Kostoff, R. N., "The Principles and Practices of Peer Review", *Science and Engineering Ethics*, Special Issue on Peer Review, 3:1, 1997h.
- Kostoff, R. N., "Use and Misuse of Metrics in Research Evaluation", *Science and Engineering Ethics*, 3:2, 1997i.
- Kostoff, R. N., "Database Tomography for Technical Intelligence: Analysis of the Research Impact Assessment Literature", *Competitive Intelligence Review*, 8:2, Summer 1997j.
- Kostoff, R. N., Eberhart, H. J., Toothman, D. R., and Pellenbarg, R. "Database Tomography for Technical Intelligence: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society:", *Scientometrics*, 40:1, 1997k.
- Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature", *Information Processing and Management*, 34:1, 1998a.
- Kostoff, R. N., "The Under-reporting of Research Impact", *The Scientist*, September 14, 1998b.
- Kostoff, R. N., "Metrics for Planning and Evaluating Science and Technology", *R&D Enterprise - Asia Pacific*, 1:2-3, 1998c.
- Kostoff, R. N., "GPRA Science and Technology Peer Review", SciCentral, [www.scicentral.com](http://www.scicentral.com), 1998d. Also available at Kostoff, R. N. "Science and Technology Peer Review: GPRA". DTIC Technical Report Number ADA418868.

- Kostoff, R. N., "Science and Technology Metrics",  
[http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/), 1998e.
- Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Road-maps Using Bibliometrics and Database Tomography". *Journal of the American Society for Information Science*. 15 April 1999a.
- Kostoff, R. N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations". *Technology Analysis and Strategic Management*. 11:4. 1999b.
- Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". *Journal of Chemical Information and Computer Science*. 40:1. 19-39. Jan-Feb 2000a.
- Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". *Journal of Aircraft*, 37:4. 727-730. July-August 2000b.
- Kostoff, R. N., and Schaller, R. R. "Science and Technology Roadmaps". *IEEE Transactions on Engineering Management*. 48:2. 132-143. May 2001a.
- Kostoff, R. N. "The Extraction of Useful Information from the BioMedical Literature". *Academic Medicine*. 76:12. December 2001b.
- Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". *JASIST*. 52:13. 1148-1156. 52:13. November 2001c.
- Kostoff, R. N., Toothman, D. R., Eberhart, H. J., and Humenik, J. A. "Text Mining Using Database Tomography and Bibliometrics: A Review". *Technology Forecasting and Social Change*. 68:3. November 2001d.
- Kostoff, R. N. "Normalization for Citation Analysis". *Cortex*. 37. 604-606. September 2001e.



Kostoff, R. N., Miller, R., Tshiteya, R. "Advanced Technology Development Program Review – A US Department of the Navy Case Study". *R&D Management*. 31:3. 287-298. July 2001f.

Kostoff, R. N., and DeMarco, R. A. "Science and Technology Text Mining". *Analytical Chemistry*. 73:13. 370-378A. 1 July 2001g.

Kostoff, R. N., and Del Rio, J. A. "Physics Research Impact Assessment". *Physics World*. 14:6. 47-52. June 2001h.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography". *Journal of Power Sources*. 110:1. 163-176. 2002a.

Kostoff, R. N. "Citation Analysis for Research Performer Quality". *Scientometrics*. 53:1. 49-71. 2002b.

Kostoff, R. N. "Overcoming Specialization." *BioScience*. 52:10. 937-941. 2002c.

Kostoff, R. N. "Text Mining for Global Technology Watch". In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. 2003. Vol. 4. 2789-2799. 2003a.

Kostoff, R. N. "Stimulating Innovation". International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003b.

Kostoff, R. N. "Bilateral Asymmetry Prediction". *Medical Hypotheses*. August 2003c.

Kostoff, R.N. "Role of Technical Literature in Science and Technology Development." *Journal of Information Science*. 29:3. 223-228. 2003d.

Kostoff, R. N. "The Practice and Malpractice of Stemming". *JASIST*. 54: 10. June 2003e.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". SSC San Diego SDONR 477, Space and Naval Warfare Systems Center. San Diego, CA. June 2003f.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". TR NAWCAD PAX/RTR-2003/164 Naval Air Warfare Center, Aircraft Division, Patuxent River, MD. 2003g.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography". *International Journal of Bifurcation and Chaos*. January. 2004a.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". *Fractals*. March 2004b.

Kostoff, R. N., Boylan, R., and Simons, G. R. "Disruptive Technology Roadmaps". *Technology Forecasting and Social Change*. 71:1-2. 141-159. January-February 2004c.

Kostoff, R.N., Del Rio, J. A., Bedford, C.W., Garcia, E.O., and Ramirez, A.M. "Macromolecule Mass Spectrometry-Citation Mining of User Documents". *Journal of the American Society for Mass Spectrometry*. March 2004d.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". *Journal of Aircraft*. March 2004e.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power: Military Requirements and Literature Structure." *Academic and Applied Research in Military Science*. In Press. 2004f.

Kostoff, R. N. "Data – A Strategic Resource for National Security". *Academic and Applied Research in Military Science*. In Press. 2004g.

Kostoff, R. N., Block, J. A., and Pfeil, K. M. "Information Content in Medline Record Fields". *International Journal of Medical Informatics*. In Press. 2004h.

Kostoff, R. N. "*Science and Technology Transition Metrics*". DTIC Technical Report. In Press. 2004i.

Kostoff, R. N., Andrews, J., Buchtel, H., Pfeil, K., Tshiteya, R., and Humenik, J. A. "Text Mining and Bibliometrics of the Journal Cortex". *Cortex*. Invited for Publication. 2004j.

Kostoff, R. N., Tshiteya, R., and Stump, J. "Wireless LAN Roadmaps using Bibliometrics and Database Tomography". Submitted for Publication. 2004k.

Kostoff, R. N., and Block, J. A. "Factor Matrix Text Filtering and Clustering." Submitted for Publication. 2004l.

Kostoff, R. N., Tshiteya, R., Humenik, J. A., and Pfeil, K M. "Power Source Roadmaps Using Database Tomography and Bibliometrics". Submitted for Publication. 2004m.

Kostoff, R. N., Del Rio, J. A., Smith, C., and Malpohl, G. "Mexico Technology Assessment using Text Mining." To be Submitted for Publication. 2004n.

Kostoff, R. N., Del Rio, J. A., Briggs, M., and Malpohl, G. "China Technology Assessment using Text Mining." To be Submitted for Publication. 2004o.

Koteen, J., *Strategic Management in Public and Nonprofit Organizations* (New York: Praeger Publishers, 1997, 2<sup>nd</sup> Edition).

Leon, M., *Integration: Consultants Who Practice and Preach, Knowledge Management*, June, 1999, p. 90.

Leydesdorff, L., Various Methods for the Mapping of Science, *Scientometrics*, 11, 1987.

Losiewicz, P., Oard, D., and Kostoff, R. N. "Textual Data Mining to Support Science and Technology Management". *Journal of Intelligent Information Systems*. 15. 99-119. 2000.

Moore, M., *Creating Public Value: Strategic Management in Government* (Cambridge, MA: Harvard University Press, 1995).

Peters, H. and A. Van Raan, A., Co-Word Based Science Maps of Chemical Engineering, Research Report to the Netherlands, 1991.

Poister, T., and G. Streib, Strategic Management in the Public Sector, *Public Productivity & Management Review*, 22(3), 1999, pp. 308-325.

Rip, A. and J. P. Courtial, Co-word Maps of Biotechnology: An Example of Cognitive Scientometrics, *Scientometrics*, 6(6), 1984.

Thuraisingham, B., *Data Mining: Technologies, Techniques, Tools, and Trends* (New York: CRC Press, 1999).

TREC, NIST Special Publication: SP 500-251, The Eleventh Text Retrieval Conference (TREC 2002), Department of Commerce, National Institute of Standards and Technology.

Turner, W.A., G. Chartron, F. Laville, and B. Michelet, Packaging Information for Peer Review: New Co-Word Analysis Techniques In: Van Raan, A. F. J. (Ed.), *Handbook of Quantitative Studies of Science and Technology* (North Holland, 1988).

Van Raan, A. and R. Tijssen, R., "The Neural Net of Neural Network Research: An Exercise in Bibliometric Mapping," Centre for Science and Technology Studies, University of Leiden, 1991.

Van Raan, A. (1996). Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises. *Scientometrics*. 36:3.

Westphal, C. and T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems* (New York: John Wiley & Sons, 1998).

Wise, R., The Balanced Scoreboard Approach to Strategy Management, *Public Manager*, 26(3), 1995, pp. 47-50.

Zurcher, R.J., and Kostoff, R.N., "Modeling Technology Roadmaps", in Kostoff, R. N., (ed), *Journal of Technology Transfer*, Special Issue on Accelerating the Conversion of Science to Technology, 22:3, Fall 1997.

## **APPENDIX-COLLEGE STUDENT TRAINING PROGRAM**

### Background

This report has shown the need for using text mining in support of all aspects of S&T, and has demonstrated its present under-utilization. One reason for its under-utilization is that text mining has not been ingrained into technical professionals from the earliest stages of their careers. This appendix presents a proposal for both educating/ training prospective technical professionals (i.e., college students) in the value and approaches of incorporating text mining into their professional activities, and performing useful text mining studies in support of sponsor agencies.

### Objective

The objective of this program is to train college students majoring in technical specialties how to use text mining. At the same time, these students would generate text mining products of value to sponsor organizations.

### Conceptual Approach

The overall approach is to train a group of students in the fundamental processes and tools of text mining for a full time period, then have them perform specific text mining studies on a part-time basis while in school.

The initial full time period would be for a summer, on location at a sponsor's organization. To insure that the student has sufficient technical training for text mining, the work period would be preferably at the end of the sophomore year. This would provide a reasonable balance between technical adequacy and future college productivity, although exceptionally talented students could start at the end of the freshman year.

The organization of students into groups, and the types of students in each group, would depend upon the text mining problem type chosen. If the problem is text mining of a single discipline (e.g., electrochemical power sources, fullerenes, high speed flow), then one student with interest in the discipline could be assigned to the full study, or a group of students with interest in the discipline could form a team, where each student focused on a different component of the study (e.g., query development, background, bibliometrics, computational linguistics/ taxonomy). If the problem is text mining of a

multi-disciplinary problem (e.g., assessment of a country's technology core competencies, literature-based discovery), then the students would have to be organized into an inter-disciplinary team. Each team member would focus on a complementary aspect of the total problem, and students from a variety of disciplines would be required. The following task example applies to an individual student working on a complete single discipline study.

Each student would be given a technical discipline to 'mine', selected in concert with his/ her interests and background. The student would generate a comprehensive query using the information retrieval and clustering processes identified in this report and other text mining documents (see Kostoff references), and would use the marginal utility approach (Kostoff et al, 2004a) to insure the query is efficient.

The student would then apply this query, or portions thereof, to a variety of databases, to gain an understanding of the types of literature available. Once the desired literature sub-sets have been retrieved, the student would apply a combination of manually and computer-intensive techniques to analyze the retrieved literature, and gain multiple perspectives on its structure.

The student would read a sample of the retrieved literature, and record a number of judgements and metrics on each document. This would provide the student with an in-depth understanding of the discipline, and provide a benchmark of metrics and judgements against which the computer-intensive technique results would be compared.

The student would then generate bibliometrics both computer-intensively as well as judgementally. This would offer insight into the infrastructure of the discipline, as well as its origins and evolution. The student would then perform a variety of computational linguistics analyses, including manual word/ phrase and document clustering, and statistical word/ phrase and document clustering. This would provide insight into the pervasive thrusts in the discipline, as well as the relationships among the thrusts. Many different clustering approaches would be examined, to show how the multiple attributes of a project or technical concept could translate to multiple clustering perspectives. The student could then apply these results to potential discovery in the discipline.

The student would then return to school, and work about twenty hours per week on the text mining. Support for both the summer employment and part-time school employment would come from a sponsor agency. The agency would make available to the student the databases necessary to support the student's activities.

The student would be given a combination of longer and nearer-term projects while in school. The student would be expected to pursue the long-term project, with interruptions from nearer-term projects. An agency contact, or mentor, would be responsible for the student's assignments. The agency contact, in agreement with the student, would identify the longer term project. Other agency personnel would identify the shorter term projects, and provide them to the student through the agency point of contact. These shorter-term projects would not be 'fire-drill' types of activities, but would be consonant with academic time scales.

On the college side, a monitor would oversee the text mining efforts of all the students in the program. The monitor would insure quality control, and insure productivity as well. Each student would submit a report at the completion of each project, and would receive class credits as well. It is important not to neglect the socialization aspects of the work. Student motivation will be high if students function as part of a group dedicated to text mining. Therefore, each participating institution would have a core of at least four to six students contributing to the text mining effort. The group would meet periodically, including with the institution monitor and perhaps with the sponsoring agency contact point as well. Group members would provide technical and motivational support to each other.

### Conceptual Implementation Prototype

The Illinois Institute of Technology (IIT) has a didactic mechanism used for projects of the type described above. It is called an IPRO (Interprofessional Projects). These are full length courses, led by an advisor, offered in all semesters and in the summer. They are offered to juniors and seniors. The students form a team of 3-10. They receive a project-challenge: to solve a technical problem, to design a machine, software, etc. The team thus formed is inter-disciplinary, including students from all areas of engineering and the sciences, as well as from business.

One can create an IPRO-like project of such a small group, under the supervision of one or two advisors. The challenge will be to learn TDM and to apply it to a specific situation or problem. In the example above, the students would be given a technical discipline to mine. Even though it's a given discipline, probably an inter-disciplinary group of students would bring more benefits to the group, rather than having an individual student work on the DTM project. The text mining experience has shown that no disciplinary area is really "pure", to the extent that other disciplines cannot contribute to DTM of the selected disciplinary area.

IPROs at IIT have been supported by various industrial and government organizations.

When an industrial company supports it, usually the problem on which the students work is of interest to the company. For example, students in some IPROs designed a new garage-door opener, using a revolutionary approach. The project was supported by a company that manufactures these door openers. The support, however, is quite reasonable, paying for incidentals, as students still pay tuition and receive credit hours. IPROs are obligatory for undergraduates in engineering. The outcome suggested above is very suitable

The following elements are needed for adequate training in TDM: 1) the problem or challenge, defined in operational terms so that students can tackle it within the timeframe of an academic semester. 2) An academic advisor or even two advisors and monitors who guide the students; and 3) a strong relationship to the curriculum, with adequate rules and regulations and academic credits offered for the effort. The project must be integrated into the university's on-going processes, to allow the students a strong sense of identification and a feeling that this a worthwhile effort with full academic credentials and performance appraisal.

The emphasis in such training is not on the individual student, but on the context within which such training takes place. It goes beyond the realm of internship—as a continuing training program with different agencies spearheading it and covering diverse numbers and types of disciplines and segments of such disciplines as problems and challenges to students. Furthermore, the cooperative effort between software engineering, programming, and systems analysis, TDM expertise, and scientific and engineering expertise needed to accomplish a truly outstanding TDM effort makes this an outstanding candidate for an interdisciplinary approach with full academic sponsorship.

IPRO students meet with their advisor once a week (or more if he/she so desires). Midway through the course they present their initial findings to a select audience. At the end of the semester, each IPRO group presents its final findings in a full day event, where students, faculty, and outside guests are invited to listen to the presentations. Each IPRO team prepares a visual presentation (poster presentation) in the lobby, and also gives a 20 minutes oral and audiovisual presentation.