Jointly published by Elsevier Science Ltd, Oxford and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 40, No. 1 (1997) 103–138

DATABASE TOMOGRAPHY FOR TECHNICAL INTELLIGENCE: COMPARATIVE ROADMAPS OF THE RESEARCH IMPACT ASSESSMENT LITERATURE AND THE JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

R. N. KOSTOFF,* H. J. EBERHART,** D. R. TOOTHMAN,*** R. PELLENBARG****

**Office of Naval Research, 800 North Quincy Street, Arlington, VA 22217–5660 (USA)*
***Naval Air Warforce Center, China Lake (USA)*
****DSTI, Inc. (USA)*
*****Naval Research Laboratory (USA)*

(Received April 25, 1997)

This paper shows how Database Tomography can be used to derive technical intelligence from the published literature. Database Tomography is a patented system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and performing phrase proximity analyses. Phrase frequency analysis provides the pervasive themes of a database, and the phrase proximity analysis provides the relationships among the pervasive themes, and between the pervasive themes and sub-themes. One potential application of Database Tomography is to obtain the thrusts and interrelationships of a technical field from papers published in the literature within that field. This paper provides applications of Database Tomography to analyses of both the non-technical field of Research Impact Assessment (RIA) and the technical field of Chemistry. A database of relevant RIA articles was analyzed to produce characteristics and key features of the RIA field. The recent prolific RIA authors, the journals prolific in RIA papers, the prolific institutions in RIA, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically as well as the particular papers/journals/institutions cited most prolifically, are identified. The pervasive themes of RIA are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and subthemes. A similar process was applied to Chemistry, with the exception that the database was limited to one year's issues of the Journal of the American Chemical Society. Wherever possible, the RIA and Chemistry results were compared. Finally, the conceptual use of Database Tomography to help identify promising research directions was discussed.

0138–9130/97/US $ 15.00

Background

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a concommittent increase in the need for scientific and technical intelligence to insure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, there have become available many techniques which can support and complement direct human intelligence gathering. In particular, techniques which identify, select, gather, cull, and interpret large amounts of technological information semi-autonomously can expand greatly the capabilities of human beings for performing technical intelligence.

This paper shows how Database Tomography[1–5] can be used to derive technical intelligence from the published literature. Database Tomography is a patented system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequency analysis and performing phrase proximity analyses. The phrase frequency analysis provides the pervasive themes of a database, and the phrase proximity analysis provides the relationships among the pervasive themes, and between the pervasive themes and sub-themes.

One potential application of Database Tomography is to obtain the thrusts and interrelationships of a technical field from papers published in the literature within that field. This paper originated with a benchmark application of Database Tomography to analysis of the field of Research Impact Assessment (RIA). RIA was selected for this benchmark because of the author's familiarity with the field[6–8] and subsequent ability to validate and verify the results from the computerized analysis.

RIA uses combinations of methodologies to ascertain the impact of research on the same field of research, on allied research fields, on technology, on systems, and on operations. The main approaches employed in RIA[8] include qualitative (e.g., Peer Review), semi-quantitative (e.g., .Retrospective Studies), and quantitative (e.g., Bibliometrics).

To execute the study reported in this paper, a database of relevant RIA articles is generated using a unique search approach,[9] and the database is analyzed to produce characteristics and key features of the RIA field. The recent prolific RIA authors, the journals prolific in RIA papers, the prolific institutions in RIA, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically as well

---

The views in this paper are solely those of the authors and do not represent the views of the Department of the Navy, any of its components, or DSTI, Inc.

as the particular papers cited most prolifically, are identified. In addition, the most highly cited years, journals, and countries are also shown. The pervasive themes of RIA are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and subthemes.

Based on the positive benchmark results for RIA, the application of Database Tomography to a technical field, Chemistry, was then performed, and the results from the two studies are compared where practical. To execute the Chemistry study, a database of all papers published in the 1994 edition of a leading Chemistry journal, the *Journal of the American Chemical Society* (JACS), as abstracted in the *Science Citation Index* (SCI) is generated, and the database is analyzed to produce characteristics and key features of the Chemistry field as reflected in JACS. The recent prolific JACS authors, the prolific institutions in JACS, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically as well as the particular papers cited most prolifically, are identified. In addition, the most highly cited years, journals, and countries are also shown. The pervasive themes of JACS are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and subthemes.

In the *Appendices*, selected results from other Database Tomography studies are shown to display further capabilities of this system. One form of taxonomy from a Near-Earth Space study is shown; another type of taxonomy from a Former Soviet Union applied research study is presented; and a method to help identify promising research directions from computerized analysis of the published literature is discussed.

What is the importance of applying Database Tomography to a non-physical science field such as RIA, or a physical science field such as Chemistry? Database Tomography provides a map of the field of interest and, analogous to ordinary roadmaps, serves as a structured guide to reach a specific destination efficiently. Suppose one wants to understand the limitations of the major RIA techniques, and perhaps identify promising avenues for improving these techniques. One could start with hit-or-miss literature searches or randomized personal contacts, or one could start with Database Tomography.

Database Tomography would identify the main intellectual thrust areas in RIA or Chemistry, and the relationships among those thrust areas. As part of the analysis output, the main RIA or Chemistry techniques conceptualized and employed would be identified. The major journals associated with each thrust area and technique would be identified, the major authors for each technique and thrust area would be identified, and

the major institutions and countries associated with each technique and thrust area would be identified. The ancillary techniques and the science and technology areas which could support and improve a technique or thrust area would be identified, and conversely techniques or thrust areas which could be impacted by a given technique would be identified.

The map, then, provides a comprehensive overview of the full picture, and allows specific starting points to be chosen rationally for more detailed investigations into a topic of interest. It does not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but allows these detailed efforts to be executed more efficiently.

Database generation

The key step in the RIA literature analysis is the generation of the database. For the present study, the database consists of selected journal abstracts (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the *Science Citation Index* (SCI) and the *Social Sciences Citation Index* (SSCI). The SCI accesses about 3000 journals (mainly in the physical sciences) and the SSCI accesses about half that amount (mainly in the social sciences). In the SCI and SSCI, the title, keyword, and abstract fields were searched using keywords relevant to RIA. The resultant abstracts were culled to those relevant to RIA.

The search was performed with the recently developed technique of Simulated Nucleation,[9] which includes two powerful Database Tomography tools: multi-word phrase frequency analysis and phrase proximity analysis. An initial database of titles, keywords, and abstracts was created from a core of papers known to be highly relevant to RIA. A phrase frequency analysis was performed on this textual database. The high frequency single, double, and triple word phrases obviously relevant to RIA were then used as search terms in the SCI and SSCI databases. The process was repeated on the new database of titles, keywords, and abstracts which was found. A few more iterations were performed until convergence was obtained. Before the final iteration, a phrase proximity analysis was performed on the database in addition to the phrase frequency analysis. This additional analysis provided relevant phrases closely related to the main themes which may not have had high frequency occurrence. The value of this search approach is that the search terms are obtained from the authors in the SCI and SSCI databases, not by guessing on the part of the searcher. The resulting final database may

be the most complete RIA journal database in existence. The titles of the papers in the final RIA database are listed at the end of Ref. 8.

As stated in the background section, the JACS database consisted of SCI abstractions of all the papers contained in the 1994 issues of JACS.

Prolific authors

In both RIA and JACS, the author field was separated from the database, and a frequency count of author appearances was made. The most prolific authors follow, in order of decreasing publications. Two caveats are in order here.

For RIA, the journals searched were limited to those in the SCI and SSCI. Relevant articles in other journals were not included. Books or major reports were not included. The keywords used were a finite set of the author's discretion, and undoubtedly overlooked some relevant articles in RIA. The time frame of the articles was 1991–early 1995. Thus, there may be excellent researchers writing in the field of RIA who were omitted from the following list due to the finite selection process, and the author's apologies are extended to anyone who falls into this category. In particular, those authors whose work has been referenced in the main body of Ref. 8, and who do not appear on the following list, should be considered as an ex officio part of the list.

For the Chemistry component of the study, only JACS was used. The time frame of the study is 1994. Relevant Chemistry articles in other journals were not included. Books or major reports were not included. Thus, there are undoubtedly excellent researchers writing in the field of Chemistry who were omitted from the following list due to the finite selection process, and the authors' apologies are extended to anyone who falls into this category.

There were approximately 2300 RIA papers retrieved and approximately 2150 JACS papers. There were approximately 2975 RIA authors, and approximately 6535 JACS authors, which average to 1.3 authors per RIA paper, and 3 authors per JACS paper. The ratio of JACS authors per paper does not differ appreciably from the 3.37 authors per paper obtained in a recent study of the near-earth space literature. It appears that the RIA papers tend to be individual efforts, while the JACS (and space) papers tend to be team efforts. The JACS (and space) studies could involve multiple disciplines and potentially large experiments (certainly true for the space studies), which would account for the difference in authors per paper.

87.3% of the RIA authors produced one paper and 7.3% produced two papers, while 84.3% of the JACS authors produced one paper and 10.7% produced two papers. Thus, in both cases, about 5% of the authors produced three or more papers, although in each

case the mode author produced one paper. However, as Table 1 shows, a few authors in each field produced an order of magnitude more papers than the average or mode author. While the RIA numbers are spread over four years, the JACS numbers are for a single year, and the top JACS numbers are quite impressive.

Table 1
Most prolific authors – RIA

GARFIELD-E 91;	SCHUBERT-A 18;	VANRAAN-AFJ 17;	GLÄNZEL-W 14;
BRAUN-T 13;	GRILICHES-Z 11;	MCCAIN-KW 10;	LEYDESDORFF-L 10;
NARIN-F 9;	KOSTOFF-RN 9;	COURTIAL-JP 9;	BONITZ-M 9;
VINKLER-P 8;	NEDERHOF-AJ 8;	MOED-HF 8;	EGGHE-L 8;
ROUSSEAU-R 7;	WELLJAMSDOROF-A 6;	TIJSSEN-RJW 6;	TERRADA-ML 6;
PINERO-JML 6;	PETERS-HPF 6;	PERITZ-BC 6;	PAO-ML 6;
MENDEZ-A 6;	MACZELKA-H 6;	LANCASTER-FW 6;	

Table 1a
Most prolific authors – JACS

SCHLEYER-PV 13,	RHEINGOLD-AL 13,	BOGER-DL 13,		TROST-BM 10,
PAQUETTE-LA 10,	WHITESIDES-GM 9,	SPIRO-TG 9,		REBEK-J 9,
MOROKUMA-K 8,	LIPPARD-SJ 8,	HROVAT-DA 8,	HAW-JF 8,	DIXON-DA 8,
BUCHWALD-SL 8,	BORDEN-WT 8,	ADAM-W 8,		KITAGAWA-T 7,
HOUK-KN 7,	GELLMAN-SH 7,	BRAUMAN-JI 7,		WILLNER-I 6,
SQUIRES-RR 6,	SCHREIBER-SL 6,	ROBB-MA 6,		OLIVUCCI-M 6,
NICOLAOU-KC 6,	INGOLD-KU 6,	ECHEGOYEN-L 6,		CLARDY-J 6,
BORDWELL-FG 6,	BERNARDI-F 6,	BERGMAN-RG 6,		ARDUENGO-AJ 6,

Code: The number following each author's name represents the number of papers authored or co-authored in the literature database.

Prolific journals

A similar process was used to develop a frequency count of journal appearances for RIA. Similar limitations to those mentioned above apply to the journals, and similar apologies are extended to journals not listed. The most prolific journals follow in order of decreasing frequency. While many disciplines are represented in the RIA table, there seems to be large representation from the Medical/Psychological Sciences field and the Information/Library Sciences field. There are 645 separate journals listed for RIA. While the average number of papers per journal is 3.57, the most prolific journals contain one to two orders of magnitude more RIA papers.

Table 2
Most prolific journals – RIA

SCIENTOMETRICS 336; CURRENT CONTENTS/LIFE SCIENCES 139; CURRENT CONTENTS 86; CURRENT CONTENTS/SOCIAL & BEHAVIORAL SCIENCES 68; CURRENT CONTENTS/CLINICAL MEDICINE 68; CURRENT CONTENTS/PHYSICAL CHEMICAL & EARTH SCIENCES 44; CURRENT CONTENTS/ENGINEERING TECHNOLOGY & APPLIED SCIENCES 41; SCIENCE 40; NATURE 34; JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 33; BRITISH MEDICAL JOURNAL 31; JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION 26; BEHAVIORAL AND BRAIN SCIENCES 25; SCIENTIST 20; SCIENCES 20; CURRENT CONTENTS /AGRICULTURE BIOLOGY & ENVIRONMENTAL 20; INFORMATION PROCESSING & MANAGEMENT 19; BULLETIN OF THE MEDICAL LIBRARY ASSOCIATION 17; JOURNAL OF INFORMATION SCIENCE 16; AMERICAN PSYCHOLOGIST 16; LIBRARY & INFORMATION SCIENCE RESEARCH 15; HIGHER EDUCATION 15;

Code: The number following each journal represents the number of papers in the literature database published in the journal.

Prolific institutions

A similar process was used to develop a frequency count of institutional address appearances, and similar apologies are extended to institutions not listed. The most prolific institutions follow in order of decreasing frequency. It should be noted, especially with regard to the universities, that many different organizational components may be included under the single organizational heading. Lack of space precluded printing out the components under the organizational heading.

For RIA, 1125 institutions are represented (average 2 papers per institution, and 2.64 authors per institution), and for JACS, 750 institutions are represented (average 2.9 papers per institution, and 8.7 authors per institution). The most prolific RIA institutions are almost two orders of magnitude above the average in papers generated, while the most prolific JACS institutions are an order of magnitude above the average. These differences reflect the more concentrated nature of JACS papers in teams and institutions relative to those of RIA papers. Interestingly, even though the RIA and JACS subject matter are very different, a number of institutions rank as the most prolific in both fields (HARVARD UNIV, UNIV OF ILLINOIS, YALE UNIV, UNIV OF PENN, UNIV OF MINNESOTA, UNIV OF TEXAS, UNIV OF WISCONSIN).

Table 3
Most prolific institutions – RIA

INST SCI INFORMAT 109; HARVARD UNIV 61; UNIV OF ILLINOIS 39; HUNGARIAN ACAD SCI 35; LEIDEN UNIV 32; INDIANA UNIV 32; UNIV OF MICHIGAN 31; YALE UNIV 25; UNIV OF PENN 23; UNIV OF N CAROLINA 22; UNIV OF MINNESOTA 21; UNIV OF TEXAS 21; UNIV OF LONDON 20; JOHNS HOPKINS UNIV 20; UNIV OF WISCONSIN 19; PENN STATE UNIV 19; CSIC 19; UNIV OF SUSSEX 18; OHIO STATE UNIV 17; CORNELL UNIV 17; UNIV OF PITTSBURGH 16; UNIV OF CAMBRIDGE 16; STANFORD UNIV 16; UNIV OF MARYLAND 15; UNIV OF CALIF SAN FRANCISCO 15; UNIV OF CALIF DAVIS 14; DREXEL UNIV 14; UNIV OF IOWA 13; UNIV OF SO CALIF 13; UNIV OF INSTELLING ANTWERP 13; UNIV OF CALIF BERKELEY 12; UNIV OF CALIF LOS ANGELES 12;

Table 3a
Most prolific institutions – JACS

MIT 67; UNIV-ILLINOIS 56; UNIV-TEXAS 51; UNIV-CALIF-BERKELEY 51; SCRIPPS-CLIN-&-RES-INST 49; STANFORD-UNIV 47; CALTECH 46; HARVARD-UNIV 43; NORTHWESTERN-UNIV 39; UNIV-WISCONSIN 38; DUPONT-CO-INC 37; UNIV-MINNESOTA 35; EMORY-UNIV 35; UNIV-TORONTO 32; UNIV-PENN 32; PURDUE-UNIV 31; CORNELL-UNIV 30; YALE-UNIV 30; PRINCETON-UNIV 29; TEXAS-A&M-UNIV 29; COLUMBIA-UNIV 27; OHIO-STATE-UNIV 27; MICHIGAN-STATE-UNIV 27; UNIV-GEORGIA 25; INDIANA-UNIV 24; UNIV-PITTSBURGH 23; HEBREW-UNIV-JERUSALEM 23; UNIV-CALIF-SAN-DIEGO 22; UNIV-TOKYO 22; UNIV-WASHINGTON 22; UNIV-ROCHESTER 22; UNIV-DELAWARE 21; TOKYO-INST-TECHNOL 21; PENN-STATE-UNIV 20; UNIV-N-CAROLINA 20; OSAKA-UNIV 19; KYOTO-UNIV 19; CNRS 18; RUTGERS-STATE-UNIV 18; IOWA-STATE-UNIV-SCI-&-TECHNOL 17; UNIV-MICHIGAN 17; UNIV-CALIF-IRVINE 17; UNIV-VIRGINIA 17; UNIV-CALIF-SANTA-BARBARA 16; UNIV-ERLANGEN-NURNBERG 16; NAGOYA-UNIV 16; UNIV-CALIF-DAVIS 16; UNIV-CALIF-LOS-ANGELES 16; UNIV-FLORIDA 15; UNIV-ALBERTA 15; UNIV-BRITISH-COLUMBIA 15; NATL-RES-COUNCIL-CANADA 15;

Code: The number following each institution represents the number of times a name of a representative from that institution appears as an author or co-author in the literature database.

Prolific countries

A similar process was used to develop a frequency count of institutional address appearances, and similar apologies are extended to institutions not listed. The most prolific countries follow in order of decreasing frequency.

For RIA, 56 countries are represented, and for JACS, 44 countries are represented. The United States is about an order of magnitude more prolific than its nearest competitor, and is as prolific as its major competitors combined. In the four studies performed so far using the present approach (RIA, Chemistry [JACS], Near-Earth Space, Hypersonic-Supersonic Flow), this dominant relationship between the United

States and its nearest competitors is observed. Generically, the western democracies tend to be the most prolific. In addition, Japan is in the first JACS tier and second RIA tier; Hungary is high in RIA; and India and Russia are both well into the second RIA and JACS tiers.

Table 4
Most prolific countries – RIA

USA, 1595; UK, 279; CANADA, 138; NETHERLANDS, 8O; GERMANY, 79; FRANCE, 71; AUSTRALIA, 69; SPAIN, 58; HUNGARY, 46; BELGIUM, 45; INDIA, 32; ISRAEL, 3O; RUSSIA, 29; NORWAY, 25; JAPAN, 23; ITALY, 22; SWEDEN, 21; DENMARK, 16; SOUTH-AFRICA, 16; MEXICO, 15;

Table 4a
Most prolific countries – JACS

USA 2040; JAPAN 276; CANADA 168; GERMANY 148; FRANCE 116; UK 109; ITALY 97; SPAIN 58; SWITZERLAND 53; ISRAEL 48; NETHERLANDS 43; SWEDEN 40; AUSTRALIA 35; BELGIUM 18; DENMARK 18; SOUTH-KOREA 18; INDIA 12; RUSSIA 12; TAIWAN 8;

Code: The number following each country represents the number of times a name of a representative from that country appears as an author or co-author in the literature database.

Prolific citations

The citations in all 2300 RIA papers were aggregated into a file of over 37000 entries, and the citations in all 2154 JACS papers were aggregated into a file of over 85000 entries. The authors most frequently cited, the specific papers most frequently cited, the journals most frequently cited, and the years most frequently cited were identified. The highly cited authors, papers, journals, and years are presented in order of decreasing frequency.

While the numbers of RIA and JACS papers are about the same, there are more than twice as many citations per paper on average in JACS relative to RIA. However, many of the RIA articles were editorials or editorial-like, and did not contain references, and therefore no conclusions should be drawn about differences in numbers of citations per journal research article based on these data.

For RIA, there are 30400 papers and 18140 authors cited (average of 1.68 papers per author), and for JACS, there are 64800 papers and 32450 authors cited (average of 2 papers per author). Therefore, those RIA authors that do cite draw from a modestly wider group of authors than the JACS authors that cite. For RIA, 72% of authors cited are cited once and 14.5% are cited twice, while for JACS 60% of authors cited are cited

once and 16.7% are cited twice. For RIA, 89.7% of the papers that are cited are cited once and 6.5% are cited twice, while for JACS, 83% of the papers that are cited are cited once and 11% are cited twice. Thus, the authors cited distribution seems to follow the more classic inverse hyperbolic Lotka's Law at low citations, while the paper cited distribution follows a somewhat sharper trajectory closer to a cubed law.

For RIA, a number of the most highly cited authors are also the most prolific (Garfield, Narin, Braun, Schubert). These particular authors are recognized leaders in the RIA field, and their work also focuses on the quantitative aspects of RIA. Because of the time lag between papers and citations, differences should be expected between the most prolific authors and the most cited authors. Authors who are new to the field and are prolific may have relatively few citations. Also, some established authors who are highly cited may require substantial time to produce seminal papers.

For JACS, some of the most highly cited authors are also the most prolific (Boger, Trost). However, some of the prolific authors could have been highly cited in other journals, which would not have been reflected in this single journal study. Also, some of the highly cited authors could have been prolific in other journals.

For RIA, the first tier of highly cited papers represents many of the seminal quantitative approaches (Garfield, Schubert, Small, Lotka), while the second tier reflects the more qualitative approaches (Kuhn, Price, Cole). This should not be surprising, since with the advent of fast high-storage computers and massive databases, technology enables the shifting of focus to more quantitative data-intensive studies.

For the JACS database, the most highly cited papers reflect the evolution of metal-complex chemistry, with a continuing focus on transition metals (d-shell especially) reactions. There is a clear, continued emphasis on the synthesis (i.e. first reported formation) of a great variety of such complexes. Also reported are new and novel applications of instrumental techniques to characterize the new complexes, especially those involving organic moieties as ligands, especially application of such techniques as nuclear magnetic resources (NMR), X-ray diffraction, and mass spectrometry to determine the structure of new transition metal complexes. The body of literature analyzed (1994 JACS) clearly shows an increasing utilization of computer-based techniques as ab initio molecular orbital calculations, and molecular orbital calculations, and molecular mechanistic approaches to elucidate structure, and provide guidance in understanding mechanism of formation and catalytic pathways mediated by an increasing body of complexes.

For RIA, the most highly cited journals are congruent with the most prolific journals. The top five cited journals (*Scientometrics*, *JASIS*, *Science*, *Nature*, *JAMA*) are within the top seven prolific journals (if Current Contents is treated as a single journal).

One would expect more congruence between the highly cited and highly prolific journals (and most highly cited and prolific institutions, if the data were available) than between the highly cited and prolific authors. The time lags between publication and citation are not insignificant relative to the span of an author's productive career, whereas the time lags for journals (and institutions) are relatively smaller compared to the period over which a journal (or institution) has established a reputation for publishing quality in given fields.

The JACS authors cited 6725 different journals and other sources, with an average of over 12.6 citations per journal. However, the most highly cited journal by far is JACS, receiving 25% of total citations, or three orders of magnitude higher citations than average. Its citations equal those of the next seven most cited journals combined.

Table 5
Most cited authors – RIA

GARFIELD-E 870; NARIN-F 181; PRICE-DD 159; BRAUN-T 142; SMALL-H 141; SCHUBERT-A 139; MORAVCSIK-MJ 105; EGGHE-L 90; MERTON-RK 90; MOED-HF 82; MCCAIN-KW 78; COLE-S 77; LEYDESDORFF-L 77; ZUCKERMAN-H 77; BROOKES-BC 72; CALLON-M 71; GRILICHES-Z 70; ARUNACHALAM-S 69; COLE-JR 66; NEDERHOF-AJ 65; SMALL-HG 65; MARTIN-BR 64; LINDSEY-D 61; KOSTOFF-RN 60; CRANE-D 58; CRONIN-B 57; ALLISON-PD 56; FRAME-JD 54; CHUBIN-DE 53; MACROBERTS-MH 53; LINE-MB 52; PAO-ML 52; CICCHETTI-DV 51; IRVINE-J 51; VINKLER-P 51; KUHN-TS 50; VANRAAN-AFJ 50; LONG-JS 49; CARPENTER-MP 48; ABT-HA 47; PERITZ-BC 46; PRICE-DJD 46; VLACHY-J 46; HARGENS-LL 45; HAMILTON-DP 44; NALIMOV-VV 43; WHITE-HD 43; COURTIAL-JP 42; LOTKA-AJ 40;

Table 5a
Most cited authors – JACS

BOGER-DL 307; FRISCH-MJ 225; TROST-BM 175; DEWAR-MJS 171; COREY-EJ 154; COLLMAN-JP 127; EVANS-DA 120; HEHRE-WJ 119; BORDWELL-FG 116; WIBERG-KB 116; OLAH-GA 114; JORGENSEN-WL 108; COTTON-FA 106; POPLE-JA 102; NICOLAOU-KC 99; ADAM-W 95; LIAS-SG 87; LEHN-JM 86; MOSS-RA 86; BAX-A 82; PAQUETTE-LA 82; MARCUS-RA 73; EVANS-WJ 71; HOFFMANN-R 71; ALLINGER-NL 64; CURRAN-DP 64; BROWN-HC 63; DUNNING-TH 62; BECKWITH-ALJ 60; CRABTREE-RH 60; SHELDRICK-GM 60; BROOKHART-M 59; TURRO-NJ 59; DENMARK-SE 58; GOULD-IR 58; REED-AE 58; STILL-WC 58; BERNARDI-F 56; CRAM-DJ 56; NEGISHI-E 56; NEWCOMB-M 56; PAULING-L 56; BALDWIN-JE 55; KUBAS-GJ 55; HOUK-KN 54; YAMAMOTO-Y 54; BARTON-DHR 53; JENCKS-WP 53; BECKE-AD 52; DOYLE-MP 52; GROVES-JT 52; ARDUENGO-AJ 51;

Code: The number following each author's name represents the number of times this person was first author of a reference cited in the literature database.

Table 6
Most cited papers – RIA

GARFIELD-E-1979-CITATION-INDEXING 55
SCHUBERT-A-1989-SCIENTOMETRICS-V16-P3 40
GARFIELD-E-1972-SCIENCE-V178-P471 40
SMALL-H-1973-J-AM-SOC-INFORM-SCI-V24-P265 35
LOTKA-AJ-1926-J-WASHINGTON-ACADEMY-V16-P317 35
KUHN-TS-1970-STRUCTURE-SCI-REVOLU 33
PRICE-DD-1963-LITTLE-SCI-BIG-SCI 32
COLE-JR-1973-SOCIAL-STRATIFICATIO 29
NARIN-F-1976-EVALUATIVE-BIBLIOMET 27
SMITH-LC-1981-LIBR-TRENDS-V30-P83 25
CRANE-D-1972-INVISIBLE-COLLEGES 24
PETERS-DP-1982-BEHAVIORAL-BRAIN-SCI-V5-P187 22
MERTON-RK-1973-SOCIOLOGY-SCI 22
MARTIN-BR-1983-RES-POLICY-V12-P61 22
SMALL-HG-1974-SCI-STUD-V4-P17 21
HAMILTON-DP-1990-SCIENCE-V250-P1331 20
MORAVCSIK-MJ-1975-SOC-STUD-SCI-V5-P86 19
KING-J-1987-J-INFORM-SCI-V13-P261 19
HOWARD-GS-1987-AM-PSYCHOL-V42-P975 19

Table 6a
Most cited papers – JACS

FRISCH-MJ-1992-GAUSSIAN-92, 90
HEHRE-WJ-1986-AB-INITIO-MOL-ORBITA, 65
DEWAR-MJS-1985-J-AM-CHEM-SOC-V107-P3902, 50
FRISCH-MJ-1990-GAUSSIAN-90, 39
HARIHARAN-PC-1973-THEOR-CHIM-ACTA-V28-P213, 39
LIAS-SG-1988-J-PHYS-CHEM-REF-D-S1-V17, 38
MOLLER-C-1934-PHYS-REV-V46-P618, 38
STILL-WC-1978-J-ORG-CHEM-V43-P2923, 28
HEHRE-WJ-1972-J-CHEM-PHYS-V56-P2257, 24
LEHN-JM-1988-ANGEW-CHEM-INT-EDIT-V27-P89, 24
MCMILLEN-DF-1982-ANNU-REV-PHYS-CHEM-V33-P493, 23
REED-AE-1988-CHEM-REV-V88-P899, 23
BECRE-AD-1988-PHYS-REV-A-V38-P3098, 22
WEINER-SJ-1984-J-AM-CHEM-SOC-V106-P765, 21
BONDI-A-1964-J-PHYS-CHEM-US-V68-P441, 20
MOHAMADI-F-1990-J-COMPUT-CHEM-V11-P440, 20
VOSKO-SH-1980-CAN-J-PHYS-V58-P1200, 20
FRISCH-MJ-1992-GAUSSIAN-92-REVISION, 19
JORGENSEN-WL-1983-J-CHEM-PHYS-V79-P926, 19
POPLE-JA-1976-INT-J-QUANTUM-CHEM-S-V10-P1, 19
WUTHRICH-K-1986-NMR-PROTEINS-NUCLEIC, 19
HAY-PJ-1985-J-CHEM-PHYS-V82-P299, 18
MARCUS-RA-1985-BIOCHIM-BIOPHYS-ACTA-V811-P265, 18
PARR-RG-1989-DENSITY-FUNCTIONAL-T, 18

Table 7
Most cited journals – RIA

SCIENTOMETRICS,	1343;	J-AM-SOC-INFORM-SCI,	679;	SCIENCE,	646;	NATURE,	388;
JAMA-J-AM-MED-ASSOC,	387;	AM-PSYCHOL,	346;	SOC-STUD-SCI,	324;	J-DOC,	276;
NEW-ENGL-J-MED,	268;	RES-POLICY,	251;	CURR-CONTENTS,	245;	AM-SOCIOL-REV,	222;
J-INFORM-SCI,	183;	COLL-RES-LIBR,	141;	LANCET,	138;	AM-ECON-REV,	123;
ANN-INTERN-MED,	115;	ESSAYS-INFORMATION-S,	114;	BRIT-MED-J,	113;
J-PERS-SOC-PSYCHOL,	113;	J-APPL-PSYCHOL,	109;	INFORM-PROCESS-MANAG,	98;
PSYCHOL-BULL, 98;									

Table 7a
Most cited journals – JACS

J-AM-CHEM-SOC	17883;	J-ORG-CHEM	3257;	J-CHEM-PHYS	2916;	TETRAHEDRON-LETT	2593;
J-PHYS-CHEM-US	2496;	INORG-CHEM	2204;	BIOCHEMISTRY-US	1799;
ANGEW-CHEM-INT-EDIT	1795;	J-CHEM-SOC-CHEM-COMM	1568;	ORGANOMETALLICS	1312;
SCIENCE	1226;	CHEM-PHYS-LETT	1051;	CHEM-REV	1039;	TETRAHEDRON	997;
ACCOUNTS-CHEM-RES	985;	P-NATL-ACAD-SCI-USA	858;	J-BIOL-CHEM	813;	NATURE	800;
J-ORGANOMET-CHEM	721;	UNPUB	681;	J-CHEM-SOC	612;	J-MOL-BIOL	525;	CAN-J-CHEM	507;
CHEM-BER	472;	J-MAGN-RESON	470;	J-COMPUT-CHEM	418;	BIOCHIM-BIOPHYS-ACTA	379;
ACTA-CRYSTALLOGR-B	361;	B-CHEM-SOC-JPN	359;	HELV-CHIM-ACTA	346;
PURE-APPL-CHEM	342;	CHEM-LETT	334;	SYNTHESIS-STUTTGART	328;	CHEM-PHYS	283;
MACROMOLECULES	278;	J-ANTIBIOT	277;	ANGEW-CHEM	255;	J-MED-CHEM	250;
BIOPOLYMERS	242;	LANGMUIR	239;	MOL-PHYS	233;	PHYS-REV-B	232;	ANAL-CHEM	225;
INT-J-MASS-SPECTROM	222;	NUCLEIC-ACIDS-RES	222;	J-CHEM-SOC-DALTON	215;
J-CHEM-SOC-DA 209; BIOCHEM-BIOPH-RES-CO 204; THEOR-CHIM-ACTA 202;									

Table 8
Most cited years – RIA

1990,3092;	1989,2826;	1991,2726;	1988,2580;	1987,2177;
1992,2094;	1986,1942;	1985,1773;	1984,1436;	1983,1288;
1982,1217;	1993,1122;	1981,1092;	1979,1023;	1980,981;

Table 8a
Most cited years – JACS

1992	8297;	1993	7764;	1991	7470;	1990	6265;	1989	5282;	1988	4742;	1987	4072;	1986	3499;
1985	3299;	1984	2757;	1983	2445;	1982	2372;	1980	1991;	1981	1874;	0	1711;	1994	1669;
1978	1625;	1979	1537;	1977	1380;	1976	1343;								

Code: The number following each paper represents the number of times the paper was cited in the literature database.

Prolific keywords

A similar process was used to obtain prolific keyword appearances. The paucity of RIA keywords is due to the fact that relatively few authors submitted keywords to the database. There are approximately an order of magnitude more keywords from JACS.

For RIA, the keywords, when viewed as an integral whole, describe the following RIA scenario: Use of Peer Review and quantitative Performance Indicators such as Citation Analysis and Bibliometrics for the purpose of Quality Assurance of University Publications from Medical and Educational Research.

For JACS, the keywords, when viewed as an integrated whole, describe the following scenario of chemistry as reflected in JACS: a continued focus on the synthesis of transition and heavy-metal complexes, and the elucidation of formation pathways (mechanisms) and structure of the various complexes. There is a continued emphasis on possible catalytic activity (especially redox reactions) associated with the complexes, and an increasing examination of the biological aspects at transition metal complex chemistry. Indeed, some cited work clearly examines the interactions of such bio-molecules as proteins and metals, both as metals catalyzing protein formation and/or controlling protein conformations. Also, the cited papers deal at length with instrumental techniques associated with metal-complex structure elucidation. As only one metal-complex structure out of many possible may prove to be active, structure elnsidation is clearly of interest within the research community.

Table 9
Most prolific keywords – RIA

PEER REVIEW 19; RESEARCH 13; CITATION 7; CITATION ANALYSIS 7; CITATIONS 6; PUBLICATION 4; PERFORMANCE INDICATORS 4; BIBLIOMETRICS 4; UNIVERSITIES 3; QUALITY ASSURANCE 3; PUBLISHING 3; PUBLICATIONS 3; PREVENTION 3; PERFORMANCE 3; MEDICAL RESEARCH 3; ITALY 3; EDUCATIONAL RESEARCH 3; EDUCATION 3; DECISION SUPPORT SYSTEMS 3;

Table 9a
Most prolific keywords – JACS

COMPLEXES 220; CHEMISTRY 146; DERIVATIVES 120; SPECTROSCOPY 110; MECHANISM 108; MOLECULES 80; CRYSTAL-STRUCTURE 77; BINDING 68; ABINITIO 64; REACTIVITY 63; SPECTRA 61; PROTEINS 59; COMPLEX 56; LIGANDS 56; GAS-PHASE 54; ACID 53; 1 51; ENERGIES 47; WATER 46; MODEL 43; ORGANIC-SYNTHESIS 42; RESOLUTION 40; SYSTEMS 40; NMR 40; BOND 38; STRUCTURE 37; NUCLEAR-MAGNETIC-RESONANCE 37; RECOGNITION 37; CLEAVAGE 37; OXIDATION 37; MOLECULAR-STRUCTURE 36; PROTEIN 35; IONS 35; ALCOHOLS 35; GENERATION 35; DESIGN 35; DYNAMICS 33; CARBON 32; KETONES 32; DNA 31; RESONANCE 31; KINETICS 31; ESTERS 30; ACTIVATION 30; ELECTRON-TRANSFER 30; ELECTRONIC-STRUCTURE 30; AQUEOUS-SOLUTION 30;

NUCLEAR MAGNETIC-RESONANCE 29; STEREOCHEMISTRY 29; REDUCTION 29; STATE 28; EXCHANGE 28; ANALOGS 27; CRYSTAL 27; HYDROGEN 27; PHOTOCHEMISTRY 26; LIGAND 26; REACTIONS 26; COORDINATION 25; DEPENDENCE 25;

Code: The number after each keyword represents the number of times the keyword appeared in the papers of the literature database.

Pervasive themes

To obtain pervasive themes, single, double, and triple word phrases from the text of the database were identified, and the high frequency high technical content phrases were identified as the pervasive themes. In this particular exercise, the databases for RIA and JACS were each split into two parts (titles and abstracts), and the analysis was done on each part. The titles of the papers were put into a separate database, and the multiword frequency analysis was performed. The abstracts of the papers constituted a separate database as well.

Following are the raw data outputs from these two subdatabases for both RIA and JACS. The number preceding the phrase is the frequency of appearance of the phrase in the database.

Those phrases in RIA which are relatively specific are underlined, and will be used for future literature searches as keywords. The major themes include quantitative RIA approaches such as BIBLIOMETRICS/SCIENTOMETRICS/CITATIONS, qualitative approaches such as PEER REVIEW, and more generic terms such as (RESEARCH or SCIENCE) PRODUCTIVITY/OUTPUT/PERFORMANCE/BENEFIT/IMPACT.

The major Chemistry themes as reflected in JACS include study of Reactions (RATE CONSTANTS, TRANSITION STATE, ELECTRON TRANSFER, DIELS-ALDER) and Complexes (SPACE GROUP, TRANSITION-METAL, MOLECULAR-HYDROGEN, CRYSTAL STRUCTURE) using both experimental approaches (X-RAY DIFFRACTION, NMR SPECTROSCOPY, MASS SPECTROMETRY) and computational approaches (COMPUTATIONAL QUANTUM CHEMISTRY, AB INITIO MOLECULAR ORBITAL METHODS, MOLECULAR MECHANICS CALCULATIONS).

Table 10
Title double word frequencies – RIA

315 CITATION-CLASSIC COMMENTARY 116 CITATION- CLASSIC 115 CLASSIC COMMENTARY 43 CITATION ANALYSIS 35 PERFORMANCE INDICATORS 24 RESEARCH PRODUCTIVITY 22 EVALUATION RESEARCH 20 BIBLIOMETRIC ANALYSIS 16 RESEARCH PERFORMANCE 14 SCIENTIFIC PRODUCTIVITY 13 PEER-REVIEW PROCESS 13 SCIENTIFIC LITERATURE 12 LITTLE

SCIENTOMETRICS 11 BIBLIOMETRIC STUDY 11 SCIENTIFIC PRODUCTION 10 CITATION IMPACT 10 PUBLICATION PRODUCTIVITY 10 RESEARCH IMPACT 9 BIBLIOMETRIC INDICATORS 9 CHOLESTEROL LOWERING 9 CITATION INDEX 9 CITATION INDEXES 9 CITATION PATTERNS 9 LOWERING TRIALS 9 PEER REVIEWERS 9 REFORM OPTIONS 9 RESEARCH ASSESSMENT 9 SCIENCE CITATION 9 SCIENTIFIC PERFORMANCE 8 BIG SCIENTOMETRICS 8 CITATION RATES 8 INTERNATIONAL SCIENTIFIC 8 QUALITATIVE EVALUATION 8 RESEARCH METHODS 8 SCIENTOMETRICS BIG 7 ASSESSMENT EXERCISE 7 CITATION DATA 7 PEER- REVIEW 7 RESEARCH BENEFITS 7 RESEARCH EVALUATION 7 SCIENCE POLICY 7 SCIENTIFIC COLLABORATION 7 UNITED-STATES SCIENCE 6 CITATION COUNTS 6 CONSUMER RESEARCH 6 EDITORIAL PEER-REVIEW 6 IMPACT ASSESSMENT 6 JOURNAL ARTICLES 6 LEDERBERG JOSHUA 6 MEDICINE VOL 6 NOBEL CLASS 6 PEER-REVIEWED JOURNALS 6 PEERLESS SCIENCE 6 QUANTITATIVE INDICATORS

Code: The number following each word pair represents the number of times the word pair appeared in all the titles of the literature database.

Table 11
Title triple word frequencies – RIA

115 CITATION- CLASSIC COMMENTARY 17 RESEARCH AND EVALUATION 11 EVALUATION AND RESEARCH 10 EVALUATION OF RESEARCH 9 CHOLESTEROL LOWERING TRIALS 9 CITATION AND OUTCOME 9 FREQUENCY OF CITATION 8 LIBRARY AND INFORMATION-SCIENCE 8 LITTLE SCIENTOMETRICS BIG 8 OPTIONS FOR PEER-REVIEW 8 OUTCOME OF CHOLESTEROL 8 SCIENCE AND TECHNOLOGY 8 SCIENTOMETRICS BIG SCIENTOMETRICS 7 INDICATORS IN HIGHER-EDUCATION 7 RESEARCH ASSESSMENT EXERCISE 6 INTENT OF PEER-REVIEWED 6 PEER-REVIEW AND UNITED-STATES 6 RELIABILITY OF PEER-REVIEW 6 REPRINTED FROM SCIENCE 6 RESEARCH IMPACT ASSESSMENT 6 SCIENTOMETRICS AND BEYOND 6 UNITED-STATES SCIENCE POLICY 5 APPLICATIONS FOR RESEARCH 5 COMMENTARY ON STUDIES 5 COMMUNICATION AND BIBLIOMETRICS 5 EVALUATION AND TEACHING 5 INQUIRY FOR LIBRARY-SCIENCE 5 INTERNATIONAL SCIENTIFIC COLLABORATION 5 METHODS AND APPLICATIONS 5 QUALITY OF CARE 5 REPRINTED FROM THEORETICAL 5 THEORETICAL MEDICINE VOL

Table 12
Title single word frequencies – RIA

432 COMMENTARY 390 RESEARCH 317 CITATION-CLASSIC 273 PEER-REVIEW 258 CITATION 158 SCIENCE 153 ANALYSIS 151 SCIENTIFIC 137 EVALUATION 121 CLASSIC 117 CITATION- 105 PERFORMANCE 87 INDICATORS 80 JOURNALS 72 BIBLIOMETRIC 72 CITATIONS 71 PRODUCTIVITY 70 IMPACT 66 LITERATURE 66 STUDY 61 ASSESSMENT 61 JOURNAL 54 DEVELOPMENT 54 PUBLICATION 53 REVIEW 52 QUALITY 49 INTRODUCTION 45 STUDIES 44 SCIENTOMETRICS 41 REPRINTED 41 VOL 39 INTERNATIONAL 38 METHOD 38 METHODS 37 PG 35 DATA 35 INFORMATION 35 PAPERS 33 STRUCTURE 33 THEORY 31 PATTERNS 31 POLICY 31 PROCESS 31 PUBLICATIONS 30 PSYCHOLOGY 30 SYSTEM 29 CASE 29 PRODUCTION 28 COMMUNICATION 28 EVALUATING 28 INFLUENCE 28 REPLY 28 SYSTEMS 28 TECHNOLOGY 27 BEHAVIOR 27 BIBLIOMETRICS 27 COMPARISON 27 SCIENTOMETRIC 26 CLINICAL 26 MEDICAL 25 ARTICLES 25 EFFECTS 25 HUMAN 25

Table 13
Abstract double word frequencies – RIA

152 PEER REVIEW 54 EVALUATION RESEARCH 52 CITATION INDEX 44 HEALTH CARE 44 PERFORMANCE INDICATORS 38 CITATION ANALYSIS 38 SCIENCE CITATION 34 UNITED STATES 30 SOCIAL SCIENCES 29 REVIEW PROCESS 26 ARTICLES PUBLISHED 25 INFORMATION SCIENCE 25 RESEARCH PRODUCTIVITY 23 IMPACT FACTOR 21 PAPERS PUBLISHED 21 SCIENTIFIC RESEARCH 20 RESEARCH PERFORMANCE 19 JOURNAL ARTICLES 19 SOCIAL SCIENCE 18 HIGHLY CITED 18 RESEARCH ASSESSMENT 18 TOTAL NUMBER 17 CITATION RATES 17 PAPER PRESENTS 17 RESEARCH OUTPUT 17 SCIENTIFIC PRODUCTIVITY 16 CITATION PATTERNS 16 EVALUATIVE RESEARCH 16 MENTAL HEALTH 15 CHEMICAL ENGINEERING 15 HEALTH PROMOTION 15 INFORMATION RETRIEVAL 15 PAPER DESCRIBES 15 SCIENTIFIC COMMUNITY 15 SCIENTIFIC LITERATURE 14

Code: The number following each word pair represents the number of times the word pair appeared in all the abstracts of the literature database.

Table 14
Abstract triple word frequencies – RIA

36 SCIENCE CITATION INDEX 31 QUALITY OF CARE 24 NUMBER OF CITATIONS 23 SCIENCE AND TECHNOLOGY 18 LIBRARY AND INFORMATION 18 RESEARCH AND EVALUATION 16 PEER REVIEW PROCESS 13 NUMBER OF PUBLICATIONS 12 EVALUATION OF RESEARCH 12 NUMBER OF PAPERS 11 NUMBER OF AUTHORS 10 RESEARCH AND DEVELOPMENT 9 RESEARCH ASSESSMENT EXERCISE 8 EVALUATION AND RESEARCH 8 JOURNAL CITATION REPORTS 8 MAIN OUTCOME MEASURES 8 NUMBER OF ARTICLES 8 QUALITY OF LIFE 8 SCIENCES CITATION INDEX 8 SOCIAL SCIENCES CITATION 7 CITATION INDEX SCI 7 CITING AND CITED 7 PUBLISHED IN JOURNALS 7 QUANTITATIVE AND QUALITATIVE 7 SCIENTIFIC AND TECHNOLOGICAL 7 SCIENTISTS AND ENGINEERS 7 SOCIAL WORK JOURNALS 6 CORONARY HEART DISEASE 6 INSTITUTES OF HEALTH 6 JOURNAL OF CLINICAL 6 NATURE OF SCIENCE 6 PUBLICATION AND CITATION 6 RESEARCH AND ASSESSMENT

Table 15
Abstract single word frequencies – RIA

1189 RESEARCH 386 CITATION 368 JOURNALS 359 STUDY 343 ANALYSIS 338 DATA 319 SCIENTIFIC 313 SCIENCE 296 REVIEW 269 ARTICLES 268 JOURNAL 262 INFORMATION 252 PAPER 246 CITATIONS 239 AUTHORS 238 QUALITY 236 PAPERS 232 PERFORMANCE 228 NUMBER 226 EVALUATION 203 PUBLISHED 200 ARTICLE 196 STUDIES 195 SOCIAL 193 PEER 190 TWO 188 HEALTH 185 IMPACT 185 LITERATURE 182 BASED 165 PROCESS 161 FIELD 160 CARE 154 INDICATORS 152 SYSTEM 151 DEVELOPMENT 151 MODEL 148 PRODUCTIVITY 146 YEARS 145 CITED 143 PUBLICATIONS 135 ASSESSMENT 133 METHODS 128 MEDICAL 124 PUBLICATION 120 POLICY 119 COUNTRIES 115 FOUND 115 INDEX 114 AREAS 111 CLINICAL 110 FINDINGS 109 GROUP 109 TECHNOLOGY 105 DIFFERENCES 104 FACULTY 103 MEASURES 100 LEVEL 100

Table 10a
Title double word frequencies – JACS

56 TOTAL SYNTHESIS; 52 CHEM ENGN; MOLECULAR RECOGNITION; 38 PHYS CHEM; 35 RATE CONSTANTS; 34 MOLEC BIOL; 31 NUCLEAR MAGNETIC-RESONANCE; 28 STRUCTURAL CHARACTERIZATION; 28 THEORET CHEM; 26 EXPTL STN; 26 TRANSITION-METAL COMPLEXES; 24 PHARMACEUT SCI; 24 RAY CRYSTAL-STRUCTURE; 24 USA TRANSITION-METAL; 23 DIELS-ALDER REACTIONS; 22 RADICAL CATIONS; 22 X-RAY STRUCTURE; 21 MOLECULAR-ORBITAL METHODS; 21 RESONANCE RAMAN; 20 ENANTIOSELECTIVE SYNTHESIS; 20 STEREOSELECTIVE SYNTHESIS; 19 AB-INITIO STUDY; 19 ANORGAN CHEM; 19 BOND ACTIVATION; 19 IRON III; 19 MOLECULAR MECHANICS; 19 USA NUCLEAR-MAGNETIC-RESONANC; 18 REDUCTIVE ELIMINATION; 17 OXIDATIVE ADDITION; 16 ANTITUMOR ANTIBIOTICS; 16 CARBENE COMPLEXES; 16 II COMPLEXES; 16 MOLECULAR-STRUCTURE; 16 POTENTIAL-ENERGY SURFACE; 15 DIELS-ALDER REACTION; 15 ISOTOPE EFFECTS; 15 RUTHENIUM II; 14 CRYSTAL-STRUCTURE; 14 ELECTRON-TRANSFER REACTIONS; 14 III COMPLEXES; 14 PHOTOINDUCED ELECTRON-TRANSFER; 14 SELF-ASSEMBLED MONOLAYERS; 13 MOLECULAR CALCULATIONS; 13 RIBONUCLEOTIDE REDUCTASE; 13 SOLID-STATE NMR;

Table 11a
Title triple word frequencies – JACS

13 USA NUCLEAR MAGNETIC-RESONANCE; 13 USA TRANSITION-METAL COMPLEXES; 11 COMPUTAT QUANTUM CHEM; 11 USA DIELS-ALDER REACTIONS; 10 ABSOLUTE RATE CONSTANTS; 10 SYNTHESIS AND CHARACTERIZATION; 10 USA CONVERGENT FUNCTIONAL-GROUPS; 9 EFFECTIVE CORE POTENTIALS; 9 SYNTHESIS AND STRUCTURE; 9 USA MOLECULAR-ORBITAL METHODS; 8 PREPARATION AND CHARACTERIZATION; 7 KINETIC ISOTOPE EFFECTS; 7 POTENT ANTITUMOR ANTIBIOTICS; 6 C-H BOND ACTIVATION; 6 ENHANCED FUNCTIONAL ANALOGS; 6 USA METAL-PROMOTED CYCLIZATION; 6 USA MOLECULAR MECHANICS; 6 USA RAY CRYSTAL-STRUCTURE; 5 ATOMIC BASIS SETS; 5 BASIS SETS FIRST-ROW; 5 GAUSSIAN BASIS FUNCTIONS; 5 IRON III COMPLEXES; 5 MARCUS INVERTED REGION; 5 MOLECULAR MECHANICS CALCULATIONS; 5 NONCOVALENT BINDING SELECTIVITY; 5 PREPARATION AND PROPERTIES; 5 SETS FIRST-ROW ATOMS; 5 STRUCTURE AND REACTIVITY; 5 SYNTHESIS AND REACTIVITY; 5 USA MOLECULAR-HYDROGEN COMPLEXES; 4 AB-INITIO CALCULATIONS; 4 AB-INITIO MOLECULAR-ORBITAL STUDY; 4 ALPHA BETA-UNSATURATED; 4 ANTITUMOR ANTIBIOTIC CC-1065; 4 ASYMMETRIC TOTAL SYNTHESIS; 4 BRIDGED TETRAHYDROINDENYL LIGANDS; 4 CHIRAL TITANOCENE CATALYST; 4 CONICAL INTERSECTIONS IDENTICAL; 4 DENSITY FUNCTIONAL THEORY; 4 ELECTROCHEMISTRY OF SPONTANEOUSLY; 4 ENGLAND POTENTIAL-ENERGY SURFACES; 4 ESCHERICHIA-COLI RIBONUCLEOTIDE REDUCTASE; 4 EXPERIMENTAL AND THEORETICAL; 4 EXPERIMENTAL AND THEORETICAL-STUDY; 4 INTERSECTIONS IDENTICAL NUCLEI; 4 KINETIC AND MECHANISTIC; 4 MECHANISM OF ASSEMBLY; 4 MOLECULAR- STRUCTURE CRYSTAL-STRUCTURE; 4 OPENING METATHESIS POLYMERIZATION; 4 OXYGEN ATOM TRANSFER; 4 PHOTOINDUCED CHARGE TRANSFER; 4 PHOTOSYNTHETIC REACTION CENTER; 4 PLATINUM II COMPLEXES; 4 SCANNING TUNNELING MICROSCOPY; 4 SOLIDE INORGAN MOLEC; 4 SPONTANEOUSLY ADSORBED MONOLAYERS;

Table 12a
Title single word frequencies – JACS

2218 CHEM; 2042 USA; 613 COMPLEXES; 393 SYNTHESIS; 274 JAPAN; 274 REACTIONS; 237 CHEMISTRY; 213 BIOCHEM; 195 STRUCTURE; 189 DERIVATIVES; 183 COMPLEX; 182 REACTION; 177 SPECTROSCOPY; 168 CANADA; 166 NY; 165 MECHANISM; 159 NMR; 153 BINDING; 153 MOLECULAR; 150 MA; 148 GERMANY; 146 MOLEC; 145 ORGAN; 138 ACID; 136 II; 132 IL; 129 GAS-PHASE; 127 CAMBRIDGE; 127 FAC; 126 BOND; 126 PHYS; 124 DNA; 123 MOLECULES; 121 LIGANDS; 120 MODEL; 120 RESONANCE; 116 REACTIVITY; 115 FRANCE; 115 TX; 115 VOL; 114 FORMATION; 114 IONS; 113 CO; 113 CRYSTAL-STRUCTURE; 111 STUDY; 110 WATER; 107 RECOGNITION; 107 SCH; 105 CHARACTERIZATION; 104 PROTEINS; 101 KU;

Table 13a
Abstract double word frequencies – JACS

356 KCAL MOL; 165 AB INITIO; 139 SPACE GROUP; 117 RATE CONSTANTS; 84 TRANSITION STATE; 81 H-1 NMR; 80 KJ MOL; 73 ELECTRON TRANSFER; 71 ANGSTROM BETA; 67 X-RAY DIFFRACTION; 64 NMR SPECTROSCOPY; 64 ROOM TEMPERATURE; 63 GROUND STATE; 62 AQUEOUS SOLUTION; 57 GROUP P2; 56 FREE ENERGY; 55 HYDROGEN BONDS; 55 INITIO CALCULATIONS; 55 MOLECULAR ORBITAL; 53 CHEMICAL SHIFT; 53 CRYSTAL STRUCTURE; 53 POTENTIAL ENERGY; 53 PROTON TRANSFER; 53 RATS CONSTANT; 52 ANGSTROM ALPHA; 52 DOUBLE BOND; 51 HYDROGEN BONDING; 50 DOUBLE DAGGER; 49 DEGREES BETA; 49 GAS PHASE; 47 DEGREES GAMMA; 46 ISOTOPE EFFECTS; 44 EXCITED STATE; 43 CRYSTAL STRUCTURES; 42 HYDROGEN BOND; 42 RADICAL CATION; 40 ACTIVE SITE; 40 SOLID STATE; 40 TRANSITION STATES; 39 FE III; 39 GOOD AGREEMENT; 39 MASS SPECTROMETRY; 39 MONOCLINIC SPACE; 39 NMR SPECTRA; 38 CHEMICAL SHIFTS; 38 FORCE FIELD; 38 MOLECULAR MECHANICS; 35 ISOTOPE EFFECT; 35 TEMPERATURE DEPENDENCE; 34 BASE PAIRS; 34 C-13 NMR; 34 HYDROGEN ATOM; 33 ENERGY SURFACE; 33 EXPERIMENTAL DATA; 33 RADICAL CATIONS; 32 ACTIVATION ENERGY; 32 AMINO ACID; 31 BASIS SETS; 31 DNA CLEAVAGE; 31 ELECTRONIC STRUCTURE; 31 ET AL; 31 MOLECULAR DYNAMICS; 31 RING OPENING; 30 IRON III; 30 KINETIC ISOTOPE; 30 PREVIOUSLY REPORTED; 30 X-RAY CRYSTALLOGRAPHY; 29 METAL IONS; 28 DELTA DELTA; 28 EPR SPECTRA; 28 SIDE CHAIN; 27 ELECTRON DENSITY; 27 EXCITED STATES; 27 ORBITAL CALCULATIONS; 27 RESONANCE RAMAN; 26 BASIS SET; 26 CRYSTAL DATA; 26 FREE ENERGIES; 26 SIDE CHAINS; 26 VIBRATIONAL FREQUENCIES; 25 FE II; 25 INITIO MOLECULAR; 25 INTERSYSTEM CROSSING;

Table 14a
Abstract triple word frequencies – JACS

57 SPACE GROUP P2; 53 AB INITIO CALCULATIONS; 38 MONOCLINIC SPACE GROUP; 35 LEVEL OF THEORY; 29 POTENTIAL ENERGY SURFACE; 27 MOLECULAR ORBITAL CALCULATIONS; 25 AB INITIO MOLECULAR; 23 KINETIC ISOTOPE EFFECTS; 22 H-1 NMR SPECTROSCOPY; 22 INITIO MOLECULAR ORBITAL; 21 TRICLINIC SPACE GROUP; 17 ION CYCLOTRON RESONANCE; 17 MOLECULAR MECHANICS CALCULATIONS; 17 VAN DER WAALS; 15 AGREEMENT WITH EXPERIMENT; 15 H-1 NMR SPECTRA; 15 INTERPRETED IN TERMS; 14 AB INITIO METHODS; 14 DETERMINED BY X-RAY; 14 ELECTRON PARAMAGNETIC RESONANCE; 14 EXPLAINED IN TERMS; 13 KCAL MOL RESPECTIVELY; 13 SINGLE-CRYSTAL X-RAY DIFFRACTION; 13 SPACE GROUP C2; 12 AGREEMENT WITH EXPERIMENTAL; 12 CP RH

CO; 12 DENSITY FUNCTIONAL THEORY; 12 DISCUSSED IN TERMS; 12 LASER FLASH PHOTOLYSIS; 12 LEVELS OF THEORY; 12 NUCLEAR MAGNETIC RESONANCE; 12 SECOND-ORDER RATE CONSTANTS; 11 DELTAH DOUBLE DAGGER; 11 H–1 AND C–13; 11 HEATS OF FORMATION; 11 ORDERS OF MAGNITUDE; 11 ORTHORHOMBIC SPACE GROUP; 11 SYSTEM SPACE GROUP; 10 AB INITIO QUANTUM; 10 AMINO ACID RESIDUES; 10 CALF THYMUS DNA; 10 CHARACTERIZED BY X-RAY; 10 DELTAS DOUBLE DAGGER; 10 FOURIER TRANSFORM ION; 10 HYDROGEN ATOM TRANSFER; 10 MOLECULAR DYNAMICS SIMULATIONS; 10 PREPARED AND CHARACTERIZED; 10 SINGLE AND DOUBLE; 10 SINGLE CRYSTAL X-RAY; 10 TRANSFORM ION CYCLOTRON; 10 X-RAY CRYSTAL STRUCTURES;

Table 15a
Abstract single word frequencies – JACS

792 REACTION; 710 ANGSTROM; 620 TWO; 617 DEGREES; 583 COMPLEXES; 526 BOND; 506 STRUCTURE; 500 ENERGY; 498 COMPLEX; 485 MOL; 479 OBSERVED; 465 CO; 444 GROUP; 424 STATE; 416 FOUND; 412 FORMATION; 398 REACTIONS; 371 KCAL; 367 NMR; 354 CALCULATIONS; 354 MOLECULAR; 346 BINDING; 344 DATA; 339 RATE; 332 ELECTRON; 331 ACID; 327 FORM; 321 II; 319 STRUCTURES; 306 ION; 297 RING; 297 TRANSFER; 293 RADICAL; 292 HYDROGEN; 290 EFFECTS; 288 DETERMINED; 288 SOLUTION; 287 SIMILAR; 285 SPECTRA; 283 DELTA; 278 MODEL; 267 TEMPERATURE; 264 ADDITION; 262 MOLECULES; 259 SPECIES; 258 DNA; 253 COMPOUNDS; 253 METAL; 251 TRANSITION; 250 BETA; 247 IONS; 246 ANALYSIS; 246 SURFACE; 237 VALUES; 229 CONSTANTS; 229 LIGAND; 228 SOLVENT; 227 WATER; 226 EFFECT; 226 PRODUCTS; 225 PH; 223 GROUPS; 222 MECHANISM; 221 CRYSTAL; 220 FE; 220 R-RAY; 217 STUDIED; 216 INTERACTIONS; 215 ENERGIES; 215 STUDIES; 214 CHEMICAL; 214 FORMED; 212 HIGH; 211 RESPECTIVELY; 210 EXPERIMENTAL; 209 INTERMEDIATE; 207 CALCULATED; 204 RELATIVE; 203 CORRESPONDING; 200 ALPHA;

Theme relationships

To obtain the theme and subtheme relationships, a phrase proximity analysis is performed about each theme phrase. Typically, forty to sixty multi-word phrase themes are selected from a multi-word phrase analysis of the type shown above. For each theme phrase, the frequencies of words within +–50 words of the theme phrase for every occurrence in the full text are computed. A phrase frequency dictionary is constructed which shows the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those subthemes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the most closely related phrases to the cluster theme (e.g., see the example

(Table 16 – Citation – Abstract Database) following this section for part of a typical filtered cluster from the study).

Because of space limitations in this document, only two themes were chosen for the RIA phrase proximity analysis, and one theme for the JACS phrase proximity analysis. Peer review was one obvious high frequency RIA theme. Citation was chosen as the other RIA theme because of its high frequency, although Bibliometrics could have been an appropriate alternate theme. Complexes was chosen as the JACS theme, while Reaction could have been an equally appropriate theme.

The full text database was split into two databases. One was the abstract narrative, and it was hoped that performing the phrase proximity analysis on this database would yield mainly topical theme relationships. The other database consisted of records (one for each published paper) containing four fields: author(s), title, journal name, author(s) institutional address(es). It was hoped that performing the phrase proximity analysis on this database would yield not only topical theme relationships from the proximal title phrases, but also relationships between technical themes and authors, journals, and institutions.

Table 16
Theme phrase "Citation" – Abstract Database – Sort by Eij

Cij	Ci	Ii (Cij/Ci)	Ij (Cij/Cj)	Eij (Ii*Ij)	Cluster member
150	386	0.389	0.389	0.1510	CITATION
137	368	0.372	0.355	0.1321	JOURNALS
106	246	0.431	0.275	0.1183	CITATIONS
107	268	0.399	0.277	0.1107	JOURNAL
94	236	0.398	0.244	0.0970	PAPERS
65	115	0.565	0.168	0.0952	INDEX
70	145	0.483	0.181	0.0875	CITED
91	269	0.338	0.236	0.0798	ARTICLES
93	313	0.297	0.241	0.0716	SCIENCE

Code: Cij is co-occurrence frequency, or number of times cluster member appears within +–50 words of cluster theme in total text;
Ci is absolute occurrence frequency of cluster member;
Cj is absolute occurrence frequency of cluster theme;
Ii, the cluster member inclusion index, is ratio of Cij to Ci;
Ij, the cluster theme inclusion index, is ratio of Cij to Cj,
Eij, the equivalence index, is product of inclusion index based on cluster member Ii (Cij/Ci) and inclusion index based on cluster theme Ij (Cij/Cj).

In the following figures, the underlined topic is the cluster theme. The cluster members were segregated by their values of Inclusion Indices (Ij and Ii), but due to

space limitations, only the summary relational results are presented. Ij is the ratio of Cij to Cj, and is the Inclusion Index based on the theme phrase. Ii is the ratio of Cij to Ci, and is the Inclusion Index based on the cluster member. The dividing points between high and low Ij and Ii are the middle of the "knee" of the distribution functions of numbers of cluster members vs. values of Ij and Ii. All cluster members with Ij greater than or equal to 0.1 were defined as having high Ij. All cluster members with Ii greater than or equal to 0.5 were defined as having high Ii.

A high value of Ij means that, whenever the theme phrase appears in the text, there is a high probability that the cluster member will appear within +–50 words of the theme phrase. A high value of Ii means that, whenever the cluster member appears in the text, there is a high probability that the theme phrase will appear within +–50 words of the cluster member.

Phrases in the category HIGH Ij HIGH Ii are coupled very strongly to the theme phrase. Whenever the theme phrase appears, there is a high probability that the cluster member will be physically close. Whenever the cluster member appears, there is a high probability that the theme phrase will be physically close. Whenever either word appears in the text, the other will be physically close.

Consider phrases located under the heading HIGH Ij LOW Ii in Tables 17 and 18. Whenever the cluster member appears in the text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially larger than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning.

Single word phrases have absolute frequencies of an order of magnitude higher than double word phrases. Thus, the phrases under the heading HIGH Ij LOW Ii are typically high frequency single words. They are related to the theme phrase but much broader in meaning than the theme phrase. A small fraction of the time that these broad single words appear, the more narrowly defined double word phrase theme will appear physically close. However, whenever the narrowly defined double word phrase theme appears, the broader related single word cluster member will appear. The phrases under this heading can also be viewed as a higher level taxonomy of technical disciplines related to the theme.

Consider phrases located under the heading LOW Ij HIGH Ii. Whenever the cluster member appears in the text, there is a high probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the text, there is a low probability that it will be physically close to the cluster member. This type of situation

occurs when the frequency of occurrence of the cluster member Ci is substantially smaller than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Thus, the phrases under the heading LOW Ij HIGH Ii tend to be low frequency double and triple word phrases, related to the theme phrase but very narrowly defined.

A large fraction of the time that these very narrow double and triple word phrases appear, the relatively broader double word phrase theme will appear physically close. However, a small fraction of the time that the relatively broad double word phrase theme appears, the more narrow double and triple word phrase cluster member will appear. This grouping has the potential for identifying "needle-in-a-haystack" type thrusts which occur infrequently but strongly support the theme when they do occur. One of many advantages of full text over key or index words is this illustrated ability to retain low frequency but highly important phrases, since the key word approach ignores the low frequency phrases.

Table 17 – RIA

PEER REVIEW

The first grouping analyzed is the BLOCK database; low Ii high Ij. The words describe the more generic associations with PEER REVIEW. The major journals whose RIA articles tend to focus on peer review are shown to include SCIENCE, NATURE, and BEHAVIORAL AND BRAIN SCIENCES. The major countries associated with peer review are USA and ENGLAND. The major users of peer review in this database tend to represent the medical community (MEDICAL; MEDICAL ASSOCIATION; SCH MED; MD). In summary, peer review has major emphasis in America and England, is featured in the major journals of Science, Nature, and Behavioral and Brain Sciences, and is employed widely in the medical community.

The second grouping analyzed is the BLOCK database; high Ii low Ij. The words describe the more specific associations with PEER REVIEW. Authors who focus on peer review are shown to include CHUBIN, HACKETT, CICCHETTI, RUBIN, TRACEY, LOCK, and DICKSON. Journals closely associated with peer review in this database include JOURNAL OF CHILD NEUROLOGY, TECHNOLOGY REVIEW, JOURNAL OF PSYCHIATRY, ANGEWANDTE CHEMIE INTERNATIONAL, and BEHAVIORAL AND BRAIN SCIENCES. Institutions which appear often with peer review include JOHNS HOPKINS UNIV, YALE UNIV, SUNY-STONY BROOK, and NEW ZEALAND UNIV. Subthemes related to peer review include REFORM OPTIONS, MANUSCRIPT AND GRANT SUBMISSIONS, INTERNAL AND EXTERNAL STANDARDS, SCIENCE POLICY, PERFORMANCE REVIEW, REFEREES, QUALITY ASSESSMENT, QUALITY ASSURANCE, and RELIABILITY.

The third grouping analyzed is the ABSTRACT database; low Ii high Ij. The generic related themes from this database include the validity of the peer review process (PROCESS, CRITERIA, QUALITY, OBJECTIVE, RELIABILITY), the journal focus of peer review (MANUSCRIPTS, AUTHORS, JOURNALS, ARTICLES, EVALUATION), and the medical focus of peer review (HOSPITAL, HEALTH, MEDICAL, CLINICAL).

The fourth grouping analyzed is the ABSTRACT database; high Ii, low Ij. Specific themes include those related to process performance and quality (DEFICIENCIES, GRIEVANCES, BLINDED PEER REVIEW, NON-BLINDED PEER REVIEW, FOG INDEX, CONTROL GROUP, SHORTCOMINGS, READIBILITY), those related to the uses and purposes of peer review (RESEARCH SELECTION, IMPACT EVALUATION, QUALITY ASSESSMENT, OVERSIGHT, AUDIT, RESEARCH IMPACT), those related to the focus on selecting journal publications (EDITORIAL PROCESSES, SELECTION REVIEW, PUBLISHED IN JOURNALS, MANUSCRIPTS), and those related to the medical focus (TRAUMA CENTER, AMBULATORY CARE, CAESAREAN SECTIONS, MEDICARE, PRIMARY CARE, PERINATAL).

CITATION

The fifth grouping analyzed is the BLOCK database; low Ii high Ij. The words describe the more generic associations with CITATION. The major countries appear again to be the USA and ENGLAND; The major journal appears to be CURRENT CONTENTS, the major author appears to be GARFIELD, and the major institution appears to be INST-SCI-INFORMAT. These results show the sensitivity of the conclusions to the theme phrases chosen for the proximity analysis. The inclusion of citation classic commentaries in the database gave heavy weighting to CURRENT CONTENTS in which they appeared. Had BIBLIOMETRICS been chosen as a theme word, then in addition journals such as SCIENTOMETRICS would have appeared prominently, as would institutions such as HUNGARIAN ACADEMY OF SCIENCES and CHI-RES-INC, and authors such as NARIN and BRAUN.

The sixth grouping analyzed is the BLOCK database; high Ii low Ij. The words describe the more specific associations with CITATION. The authors closely associated with citations include GARFIELD, BURCHINSKY, DUPLENKP, HARGENS, WELLJAMSDOROF, and BOTT. The journals associated with citations include AMERICAN PSYCHOLOGIST, METEORITICS, CHEMICKE LISTY, SOUTH AFRICAN JOURNAL OF SCIENCE, AMERICAN JOURNAL OF ROENTGENOLOGY, SCIENCE TECHNOLOGY AND HUMAN VALUES. Institutions associated with citations include INST-SCI-INFORMAT, INST GERONTOL-KIEV, UNIV OF ILLINOIS, and UNIV OF MICHIGAN. Subthemes related to citation include COUNTS, RATES FREQUENCY, RANKINGS, INDEXES, LINKS, HIGH IMPACT RESEARCH, IMPACT FACTOR, JOURNAL ARTICLES, PUBLICATIONS, CHAPTERS.

The seventh grouping analyzed is the ABSTRACT database; low Ii high Ij. The generic related themes from this database include types of documents cited (PAPERS, ARTICLES, PUBLICATIONS), characterization of material cited (RESEARCH, SCIENCE, LITERATURE), and yields from citations (ANALYSIS, PATTERNS, INFORMATION, DATA).

The eighth grouping analyzed is the ABSTRACT database; high Ii, low Ij. Specific themes include those related to citation focus areas (CITATION INDEX DATABASE, CITATION MATRIX, CITATION STUDIES, CITATION RATE, CITATION COUNTS, JOURNAL CITATION REPORTS, MEDIAN CITATION, CITATIONS PER ARTICLE, CITATION HISTORY, CITATION PROCESS, CITATION RETRIEVAL, CITATION IMPACT, CITATION FREQUENCY), citation analysis techniques (MEDIAN CITATION, MEAN CITATION, AVERAGE CITATION, MEAN VALUE FUNCTION, BIBLIOGRAPHIC COUPLING, ANALYSIS OF CITATIONS, LOGLINEAR, POISSON PROCESS, RELATIVE INDICATORS, COUNTS, COCITATION), outputs of citation techniques (RESEARCH FRONTS, HIGHLY CITED PAPERS, MAPPINGS), and specific technical areas analyzed (DERMATOLOGY, RADIOLOGY, HEART DISEASE, MARINE BIOLOGY, SAFETY SEATS, CAPITAL PUNISHMENT, AND ASTRONOMERS).

Table 18 – JACS

COMPLEXES

The first grouping analyzed is the BLOCK database; low Ii high Ij. The words describe the more generic associations with COMPLEXES. The major countries associated with research into COMPLEXES are USA, JAPAN, CANADA, ITALY, FRANCE, GERMANY, SPAIN, ENGLAND, and SWITZERLAND. The major states in the US associated with research into COMPLEXES are NY, MA, CA, IL, MO, DE, GA, PA, TX, NJ, MI, FL, and MN. The major research institutions associated with COMPLEXES are STANFORD, BERKELEY, EMORY, CALTECH, DELAWARE, DUPONT, and NORTHWESTERN. The major types of COMPLEXES researched include TRANSITION-METAL, IRON, RUTHENIUM, MOLYBDENUM, RHODIUM, TUNGSTEN, and PALLADIUM. The major analytical techniques associated with COMPLEXES include X-RAY, SPECTROSCOPY, NMR, and MASS-SPECTROMETRY. The major phenomena researched associated with COMPLEXES include SYNTHESIS, REACTIONS, CRYSTAL STRUCTURE, REACTIVITY, ELECTRON TRANSFER, ACTIVATION, POLYMERIZATION, CLUSTERS, CATALYSIS, OXIDATION, BINDING, and INSERTION.

The second grouping analyzed is the BLOCK database; high Ii low Ij. The words describe the more specific associations with COMPLEXES. Organizations closely associated with COMPLEXES research include SEARLE, HOKKAIDO-UNIV, KYOTO-UNIV, UNIV-PARMA, MERCK-SHARP, LOS-ALAMOS-NATL-LAB, UNIV-LAUSANNE, EMORY-UNIV, UNIV-DELAWARE, BERKELEY, UNIV-BARCELONA, UNIV-STRASBOURG, UNIV-SYDNEY, UNIV-MISSOURI, UNIV-ZARAGOZA, TEXAS A&M, UNIV-CHICAGO, UNIV-FLORIDA, AND BROOKHAVEN-NATL-LAB. Authors closely associated with COMPLEXES include SOLARI-E, FLORIANI-C, HEINEKEY-DM, COLLMAN-JP, and GOULD-IR. This grouping clearly emphasizes an institutional focus of where research is conducted. Both industrial concerns and academic facilities are emphasized, roughly equally. Significant themes seem to be, as expected, synthesis and characterization of complexes, but with a curious attention to fixing gases (MOLECULAR OXYGEN, HYDROGEN, OR CARBON MONOXIDE) within the complex, perhaps as one step in a catalysis reaction. Indeed, several of the papers in this grouping focus on 'reactive' complexes which could clearly be related to catalytic activity. It is likely that this group focuses heavily on catalysis as an overall theme.

The third grouping analyzed is the ABSTRACT database; low Ii high Ij. The generic related themes from this data base include understanding the actual structure of complexes, often by application of instrumental techniques (NUCLEAR MAGNETIC RESONANCE, X-RAY DIFFRACTION, ULTRA VIOLET OR INFRARED SPECTROSCOPY, others), an apparent extended examination of copper and iron complexes, and a weak reference to potential catalysis. There seems to be less emphasis on the actual formation (synthesis) of the complexes in this grouping.

The fourth grouping analyzed is the ABSTRACT database; high Ii, low Ij. Specific themes include those related to formation (synthesis) of a broad spectrum of metal complexes (focus on metals such as the platinum group, iron, nickel, copper) many of which appear to include multi-metal atom centers, (e.g. Pt-Pt) and even mixed multi-metal atom centers (e.g. Pt-Ir), and an emphasis on metal complexes involving carbon monoxide as a ligand, as well as some emphasis on unusual carbon-based ligands (e.g. per fluorinated species). This version of the data base clearly seems to focus on the chemistry (esp. synthesis) of metal complexes.

The data base shows that the most prolific JACS authors were Schleyer, Rheingold, Boger and Trost, who published a total of 49 papers in 1994. These authors published extensively on focused themes, research topics their groups likely have pursued for several years before, and after, 1994. Specifically, Schleyer examined in depth synthesis complexes of alkali metals (e.g. sodium), a very unusual topic as alkali metals in general form complexes only rarely, as well applied computer based technology to elucidate the structure of such comples. Rheingold's group published extensively on complexes involving metal-rutal

bonds, and multi-metal atom clusters in catatopic systems. Boger, alone among the prolific authors, focused on bio-active molecules and their synthesis and reactivity as a function of structure. Trost, and his associates, appeared to examine transition metal catalysis of traditional, well characterized organic system reaction such as the Diehls-Alder reaction (which involves no metal species). In general, it is clear that the four authors are publishing heavily in broad areas of contemporary organic metallic chemistry: synthesis catalysis, structure and mechanism determination, and metal-mediation of bio-active molecules. Indeed, it is clear that these authors are defining the direction of these themes by their prolific research and publication programs.

Conclusions and applications

This paper has provided maps of the RIA and JACS Chemistry fields, although only a small fraction of the raw data has been presented. A Competitive Intelligence (CI) professional who has interest in these fields has many options for proceeding further from the map, depending on this person's specific interests. For example, if the analyst wanted to understand the intellectual foundations of RIA or JACS Chemistry, then a reading of the most highly cited papers would be an excellent starting point. If the analyst wanted to overview the current literature, then two approaches are available. The comprehensive literature survey used as the database for the RIA analysis and reproduced in the back of Ref. 8 is one avenue. Another is to peruse the journals which contain the highest frequency of recent publications. This latter approach is worthwhile since computerized search approaches don't always identify the full scope of related articles to the topic of interest, and journals which focus on such a topical area could yield a cornucopia of useful information through browsing.

If the analyst wants to contact experts in a particular thrust area or technique, then contact could be made with the specific individuals or the institutions identified with given techniques in the theme relationships section. If the analyst wants to generate a taxonomy of the S&T field based on the technical relationships used by the research performers, then the approaches described in Appendix I might prove helpful. If the analyst wants to utilize the literature to help identify promising research directions, then the approach described in Appendix II might prove useful. The key conclusion is that, starting from the raw data, the analyst can generate any cross-cutting relationships desired to proceed further in specific directions of personal interest.

References

1. R. N. KOSTOFF et al, "System and Method for Database Tomography", U.S. Patent Number 5440481, August 8, 1995.
2. R. N. KOSTOFF, Database tomography: Multidisciplinary research thrusts from co-word analysis, *Proceedings: Portland International Conference on Management of Engineering and Technology,* October 27–31, 1991.
3. R. N. KOSTOFF, Database tomography for technical intelligence, *Proceedings: Eighth Annual Conference of the Society for Competitive Intelligence Professionals*, Los Angeles, CA, 1993.
4. R. N. KOSTOFF, Database tomography for technical intelligence, *Competitive Intelligence Review*, 4:1, Spring 1993.
5. R. N. KOSTOFF, Database tomography: Origins and applications, *Competitive Intelligence Review*, Special Issue on Technology, 5:1, Spring 1994.
6. R. N. KOSTOFF, Research performance effectiveness and impact, Chapter 31, In: G. H. GAYNOR, (Ed.), *Handbook of Technology Management*, McGraw-Hill, Inc., 1996.
7. R. N. KOSTOFF, Federal research impact assessment: Axioms, approaches, applications, *Scientometrics*, 34 (1995) 163–206.
8. R. N. KOSTOFF, *The Handbook of Research Impact Assessment*, Fifth Edition, Summer 1995, DTIC Report Number ADA296021.
9. R. N. KOSTOFF, H. J. EBERHART, D. R. TOOTHMAN, Database tomography for information retrieval, *Journal of Information Science*, 23:4, 1997.
10. D. R. SWANSON, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and Medicine*, 30:1, 1986.
11. M. D. GORDON, R. K. LINDSAY, Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil, *JASIS*, 47:2, February 1996.

Appendix I - Generation Of Taxonomies

TAXONOMIES

The different types of Database Tomography outputs allow different types of taxonomies, or classifications into component categories, to be generated. Such categorizations, anologous to the independent axes of a mathematical coordinate system, allow the underlying structure of a field to be portrayed more clearly, leading to more focused analytical and management analyses. There is a major difference between the taxonomy obtained by this approach and other taxonomies. The present taxonomy derives from the language and natural divisions of the database, and therefore database entries are easily categorized. Other taxonomies are usually generated top-down and usually attempt to force-fit database subjects into pre-determined categories.

One of the advantages of the present full text approach, relative to the index or key word approach, is that many types of taxonomies can be generated: i.e., science, technology, institution, journal, person name, etc. Even within one of these categories, such as science, many types of taxonomies can be developed, depending on the interests of the analyst and the reason for the taxonomy. Two separate types of taxonomies will be discussed here.

I – PHRASE FREQUENCY TAXONOMY

The first type of taxonomy derives from the phrase frequencies. The authors examined the phrase frequency outputs, then arbitrarily grouped the high frequency phrases into different, relatively independent, categories for which all remaining terms would be accounted. Two examples of taxonomies are presented: the first is from a study of research papers related to the utilization of near-earth space, and the second is from a study of reports from the Foreign Applied Sciences Assessment Center (FASAC) assessing different areas of applied research in the former Soviet Union.

EXAMPLE 1 – NEAR EARTH SPACE RESEARCH TAXONOMY

About 5500 research papers relating to utilization of near earth space were drawn from the SCI. Phrase frequencies were generated from the abstracts, and the high frequency phrases were arbitrarily categorized. These relatively independent categories consist of Space Platform (E.G., SATELLITE, SPACECRAFT), Satellite Function

(E.G., MAPPING, TRACKING), Satellite Type (E.G., GEOSAT, LANDSAT), Measuring Instrument (E.G., RADIOMETER, MICROWAVE LIMB SOUNDER), Region Examined (E.G., SEA, UPPER ATMOSPHERE), Location Examined (E.G., NORTH ATLANTIC, SOUTHERN HEMISPHERE), Variable Measured (E.G., TEMPERATURE, SOIL MOISTURE CONTENT), Variable Derived (E.G., RADIATION BUDGET, GENERAL CIRCULATION PATTERN), Analytical Tool (E.G., DATA PROCESSING, LEAST SQUARES), Products (E.G., TIME SERIES, TOTAL OZONE MAPPING), Space Environment (E.G., SOLAR WIND, MAGNETIC FIELD).

EXAMPLE 2 – FORMER SOVIET UNION APPLIED RESEARCH

About 35 full-length reports on the status of different areas of applied research in the Former Soviet Union were used as the database. Phrase frequencies were generated from the reports, and the high frequency phrases were arbitrarily categorized. An applied research taxonomy was generated. It consists of Information (IMAGE PROCESSING, PATTERN RECOGNITION, SIGNAL PROCESSING, ARTIFICIAL INTELLIGENCE, ETC.), Physics (SHOCK WAVES, RADIO WAVES, QUANTUM ELECTRON, MAGNETIC FIELD, CHARGED PARTICLE ACCELERATORS, OPTICAL PHASE CONJUGATION, ETC.), Environment (INTERNAL WAVES, OCEANIC PHYSICS, SEA SURFACE, IONOSPHERIC MODIFICATION, RADIO WAVE PROPAGATION, ETC.), and Materials (THIN FILM, COMPOSITE MATERIALS, FRACTURE MECHANICS, SOLID FUEL CHEMISTRY, STRENGTH MATERIAL, ETC.).

II – PHRASE PROXIMITY TAXONOMY

The second type of taxonomy derives from the phrase frequency and proximity analysis. From the phrase frequency analyses, fifty or sixty high frequency technical phrases were identified as pervasive themes. The next step was to group these high frequency phrases into categories of related themes. A proximity analysis was done for each of these high frequency phrases. A phrase frequency dictionary, or cluster, was generated for each phrase. This cluster contained those phrases which were in close physical proximity to the pervasive theme throughout the text. The degree of overlap among clusters was computed. Clusters which shared more than a threshold number of common phrases were viewed as overlapping. These overlapping clusters were viewed

as links in a chain, with the different chains being relatively independent. Each chain was then defined as a category of the larger taxonomy.

For the study of applied research in the Former Soviet Union, the following taxonomy, or megacluster grouping, was generated. The numbered themes (e.g., 1. IONOSPHERIC HEATING/ MODIFICATION) are the categories, or megaclusters. The component themes (e.g., *RADIO WAVE), preceded by an asterisk (*), are the clusters, or pervasive themes from the phrase frequency analysis.

1. IONOSPHERIC HEATING/ MODIFICATION
*RADIO WAVE
*WAVE PROPAGATION
*QUANTUM ELECTRON
*IONOSPHERIC MODIFICATION
*PHASE CONJUGATION

2. IMAGE/ OPTICAL PROCESSING
*PARALLEL PROCESSING
*PATTERN RECOGNITION
*IMAGE PROCESSING
*COMPUTER VISION
*DIGITAL COMPUTER
*ARTIFICIAL INTELLIGENCE
*DATA PROCESSING
*COMPUTER SCIENCE
*OPTICAL PROCESSING
*SPATIAL LIGHT MODULATOR
*SIGNAL PROCESSING
*LIQUID CRYSTAL
*LIGHT MODULATOR
*PROGRAMMING LANGUAGES
*INTEGRAL EQUATIONS

3. AIR-SEA INTERFACE

*SURFACE WAVE
*OCEANIC PHYSICS
*INTERNAL WAVE
*SEA SURFACE
*BOUNDARY LAYER

*ATMOS OCEANIC PHYS
*REMOTE SENSING

4. LOW OBSERVABLE
*LOW OBSERVABLE
*THIN FILM

5. EXPLOSIVE COMBUSTION
*KINETICS AND CATALYSIS
*SOLID FUEL
*EXPLOSION AND SHOCK
*SHOCK WAVE
*CHEMICAL PHYSICS
*EXPLOS SHOCK WAVE
*STRENGTH MATER
*FRACTURE MECHANICS
*COMPOSITE MATERIALS
6. PARTICLE BEAMS
*NEUTRAL BEAM
*PARTICLE ACCELERATOR
*ATOMIC ENERGY
*PLASMA PHYSICS
*ELECTRON BEAM
*CHARGED PARTICLE ACCELERATOR
*CHARGED PARTICLE

7. AUTOMATIC/ REMOTE CONTROL
*AUTOMATIC CONTROL
*REMOTE CONTROL

8. FREQUENCY STANDARDS
*FREQUENCY STANDARD
*HYDROGEN MASER

9. RADAR CROSS SECTION
*CROSS SECTION
*ELECTROMAGNETIC WAVE
*RADIO ENGINEERING

From the multiword frequency analysis, the science discipline taxonomy for the FASAC database was defined as Information, Physics, Environment, and Materials. In terms of the megaclusters, Information would encompass IMAGE/ OPTICAL PROCESSING and AUTOMATIC/ REMOTE CONTROL; Physics would encompass IONOSPHERIC HEATING/ MODIFICATION, PARTICLE BEAMS, FREQUENCY STANDARDS, and RADAR CROSS SECTION; Environment would encompass AIR-SEA INTERFACE; and Materials would encompass EXPLOSIVE COMBUSTION and LOW OBSERVABLE. Categorizing the database with the megacluster subcategories allows a re-interpretation of the FASAC database. FASAC can be viewed as a compendium of those aspects of FSU science of interest to the U. S. for strategic and military purposes rather than viewed as a microcasm of all of FSU science.

APPENDIX II – IDENTIFICATION OF PROMISING RESEARCH DIRECTIONS

INTRODUCTION

This Appendix describes a literature-based approach to identifying opportunity-driven promising directions in science and technology. The method is generic to all fields of endeavor for which a literature exists, is dual use in the broadest sense, and has the potential to revolutionize how promising directions are identified. The approach is a computer-based analysis of the desired literatures using appropriate experts for data interpretation. The proposed procedure offers a potential quantum improvement over earlier related research efforts in the medical literature (10, 11). The technique would use the Database Tomography system described in this paper.

BACKGROUND

In the mid-1980s, Don Swanson showed that logical connections in the existing medical literature can be integrated to help identify promising medical research directions (10). His three literature-based investigations have hypothesized that 1) dietary fish oil would be helpful in treating Raynaud's Disease; 2) magnesium is important to migraine; and 3) there is a relationship between arginine and Somatomedin C. There has been medical corroboration of Swanson's discoveries (11).

Gordon and Lindsay used computer-based tools to replicate and extend Swanson's work (11). A more detailed summary of their work, as well as additional improvements possible with the authors' approach, is in the Procedure section which follows. Basically, they used word frequency analysis to examine the literature of interest, they used the high frequency words or phrases to identify related intermediate literatures, and then used a combination of high frequency phrases and weak relations between the phrases to identify the promising research directions from the related literatures.

For example, they performed a phrase frequency analysis of the Raynaud's Disease (RD) literature, and found that BLOOD VISCOSITY was a crucial element in RD. They then performed a phrase frequency and weak phrase proximity (ratio of phrase appearance in BLOOD VISCOSITY literature to appearance in total medical literature) analysis of the BLOOD VISCOSITY literature. Their analyses confirmed Swanson's results, and showed that FISH OIL and EICOSAPENTAENOIC ACID (one of fish oil's main chemical constituents) offerred substantial promise as research directions. Experiments performed subsequent to Swanson's findings have confirmed these predicitions.

The authors believe this strong dependence on high frequency phrases and only latter stage employment of the weak proximity condition severely constrains the technique's potential. Based on the authors' database analyses of the past five years, it was found that the strong physical proximity of phrases in text is of equal importance to the occurrence frequency of those phrases when constructing structural maps of science and technology. In fact, for identifying promising research and technology directions, strong phrase proximity may be far more important than phrase frequency. High frequency phrases tend to reflect both the obvious and the mainstream efforts, while low frequency phrases located in close proximity to phrases of topical interest have much greater chance of uncovering 'needles-in-a-haystack'. In addition, as was shown in a recent paper, the full power of the authors' analytic approach requires the use of both phrase frequency and strong phrase proximity at every iterative step in the analysis (9).

The authors' approach uses the Database Tomography tools of phrase frequency analysis in conjunction with strong phrase proximity analysis. This allows identification not only the mainstream high-frequency relationships, but the less-explored low-frequency high-proximity relationships as well. This provides the capability to identify the most promising science and technology directions with the least restrictions.

PROCEDURE

This section summarizes Gordon and Lindsay's work on literature-based discovery, and shows how the combination of Database Tomography and their approach would eliminate the major deficiencies in their present approach. This combined approach could have tremendous payoff in many technical and non-technical fields.

The initial summary of Gordon and Lindsay's work will focus on their example of Raynaud's Disease (RD). The objective of their approach is to find something in the published literature that will point to new directions for treating/ curing, etc. RD. They use the following approach. Search the literature (MEDLINE, in their particular case) to retrieve all documents which contain Raynaud* in the appropriate fields (560 documents). Using word frequency analysis (including different types of word frequency analysis statistics), identify high frequency terms related to RD.

For example, they find BLOOD is such a term. They then identify the subset of the Raynaud documents which contain blood-related terms (BLOOD FLOW, BLOOD VISCOSITY, PLATELET AGGREGATION, ETC.), and repeat the word frequency analysis on this subset (232 documents). They find that ideas related to BLOOD FLOW should be pursued further. In particular, they find that BLOOD VISCOSITY is related to BLOOD FLOW, is a possible cause of impaired flow, and is statistically prominent in its own right.

Here comes a crucial part of their approach. They go back into the literature, and search for all records related to BLOOD VISCOSITY, whether or not they are related to RD. After performing a word frequency analysis and a weak proximity analysis on this information retrieved, they prune the list of terms to 115 which they judge to be initial candidates for discovery. The details of the pruning are not relevant for what follows here. Of the 115 terms, they find that only 34 did not appear in the list of the original 560 Raynauds records. These 34 terms are what they call disjoint from Raynauds, and are therefore true candidates for discovery. They finally arrive at FISH OIL, and EICOSAPENTAENOIC ACID (one of fish oil's main chemical constituents) as the discovery items.

The purpose of their study was to replicate Swanson's approach for identifying promising directions in medical research, done without computerized information retrieval techniques, ten years earlier. They did replicate, and they also show that follow-up medical research has corroborated Swanson's discoveries. Thus, their method and Swanson's appear to have great promise in mining the medical literature for promising new directions. What, then, are the deficiencies?

Their approach is based mainly on word frequency analysis, and the use of high frequency terms to guide promising directions. Only in the last step of their analysis do they employ a weak proximity analysis condition. Based on the authors' experience, high word frequencies tend to reflect mainstream research approaches heavily published in the literature. Use of high frequency terms at most stages of the analysis will effectively eliminate concepts, accepted or alternative, which have received little support in the past and are lightly represented in the literature.

What is required for a more complete computer-based analytical tool is a method that gives equal emphasis to low frequency terms as well as high frequency terms. In practice, the low frequency term analyzer would probably be more valuable for identifying promising opportunities. High frequency relationships tend to be more obvious, and probably many of these types of relationships are known without use of the computerized analysis. According to *Gordon* and *Lindsay*, *Swanson* was able to hypothesize the promising opportunities without the use of the computerized analysis. While high frequency relationships are useful in mapping structural relationships among science and technology disciplines, as has been shown with the Database Tomography efforts, it is the low frequency relationships which have the greater potential of finding the 'needles in a haystack'.

However, while there are relatively few high frequency relationships, and the analytical problem is relatively bounded, there are very large numbers of low frequency relationships. The problem becomes pragmatically intractable if no further conditions

are placed on the low frequency relationships. The additional conditions on the low frequency relationships required to make the problem tractable derive from the word proximity analyses. Examine only those low frequency terms which are also strongly related to the dominant themes of the problem. In other words, examine those low frequency terms which have high inclusion indices (number of appearances within some domain around the dominant term/ number of appearances in the total text) relative to the dominant terms. Thus, whenever these low frequency terms appear in the text, they are located physically close to the dominant themes.

The Raynaud example will now be used to show how Database Tomography in conjunction with Gordon and Lindsay's method could have worked. Using DT, two major pathways could have been examined, where Gordon and Lindsay examined only one. For the first pathway, use Gordon and Lindsay's database and replicate, using word frequency analysis, that BLOOD VISCOSITY appears important. Examine the BLOOD VISCOSITY literature further, as they did. Then, do a word frequency analysis of the BLOOD VISCOSITY literature, and identify the high frequency terms.

At this point, perform a strong word proximity analysis for BLOOD VISCOSITY on the retrieved blood viscosity literature. Identify (using the numerical indicators from the proximity analysis) those terms which, when they appear in the blood viscosity literature, are located physically close to BLOOD VISCOSITY. Thus, for argument's sake, FISH OIL may appear 100 times in the blood viscosity literature (and not in the RAYNAUD* literature; keep the requirement of disjointness), but in only 30 of those times does it appear physically close to BLOOD VISCOSITY. It would have an inclusion index of 30/100=.3. However, a potential low frequency term like VISUALIZATION may appear only 5 times in the BLOOD VISCOSITY literature (again, not in the RAYNAUD* literature), but in 4 of those times it appears physically close to BLOOD VISCOSITY. It would have an inclusion index of 4/5=.8.

Then, investigate both FISH OIL (high frequency and low inclusion) and VISUALIZATION (high inclusion and low frequency) further, with the use of the medical experts, for promising research directions.

For the second pathway, perform a strong word proximity analysis on the initial RD literature. Based on the results of this analysis, define a promising intermediate literature, analogous to the BLOOD VISCOSITY literature on the first pathway. Perform word frequency and strong proximity analyses on this intermediate literature, and interpret the data with the support of medical experts to arrive at (hopefully) further promising research directions.