SANDIA REPORT

SAND2005-7538
Unlimited Release
Printed November 2005

Generation of Large-Scale Maps of Science and Associated Indicators

Kevin W. Boyack and Richard Klavans

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.osti.gov/bridge>

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>

SAND2005-7538
Unlimited Release
Printed December 2005

Generation of Large-Scale Maps of Science and Associated Indicators

Kevin W. Boyack
Computational Biology Dept. 1412
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-0310

Richard Klavans
SciTech Strategies, Inc.
Berwyn, PA 19312

Abstract

Over the past several years, techniques have been developed for clustering very large segments of the technical literature using sources such as Thomson ISI's Science Citation Index. The primary objective of this work has been to develop indicators of potential impact at the paper level to enhance planning and evaluation of research. These indicators can also be aggregated at different levels to enable profiling of departments, institutions, agencies, etc. Results of this work are presented as maps of science and technology with various overlays corresponding to the indicators associated with a particular search or question.

Contents

1. Introduction	6
1.1. Need for an Evaluation System	6
1.2. Definitions	7
1.3. Chronological History	7
1.4. Report Structure	9
2. Journal-level Mapping	10
2.1. Multi-step Mapping of Science	11
2.2. Mapping of Science and Technology	14
3. Paper-level Mapping	16
3.1. Paper-level Map of Science	17
3.2. Paper-level Map of Science and Technology	22
4. Applications	25
5. References	29

Figures

- Figure 1. Map of science generated using the IC-Jaccard similarity measure [3]. The map is comprised of 7,121 journals from year 2000. Large font size labels identify major areas of science. Small labels denote the disciplinary topics of nearby large clusters of journals. 10
- Figure 2. Map of the backbone of science with 212 clusters comprising 7,000 journals [3]. Clusters are denoted by circles that are labeled with their dominant ISI category names. Circle sizes (area) denote the number of journals in each cluster. Circle color depicts the independence of each cluster, with darker colors depicting greater independence. Dominant cluster-to-cluster citing patterns are indicated by arrows. Arrows show all relationships where the citing cluster gives more than 7.5% of its total citations to the cited cluster, with darker arrows indicating a greater fraction of citations given by the citing cluster. 11
- Figure 3. Structural map of science generated using a multi-step bibliographic coupling process. The map is comprised of 671 clusters representing 7,351 journals from year 2001. Dominant linkages between journal clusters are shown. 13
- Figure 4. Structural map of science and technology from the 2003 combined dataset. The map contains 852 clusters representing 7,445 journals and 1,198 conference proceedings. Dominant linkages between clusters are shown. 15
- Figure 5. Map of current papers from 2002. Five major fields are shown by different colored markers. The map has been rotated to place Mathematics (not marked) at the top and physics at the right, which is our standard convention. The inset shows

the detail in a small portion of the map including how linkages form research communities.....	18
Figure 6. Comparison of the current (left) and reference (right) maps of papers from 2002.....	19
Figure 7. Membership of current community 73336 from the 2002 map of science. Common terms are highlighted to show the topical focus of the community.	19
Figure 8. Structural (left) and detailed (right) maps of science from 2002. In the detailed map, research communities have been positioned on the structural map.....	20
Figure 9. Left half of the poster “The Structure of Science” by Kevin Boyack and Richard Klavans. Colored flags and text correspond to sections of the galaxy-like map using the same colored labels. Each white dot represents one research community.....	21
Figure 10. Right half of the poster “The Structure of Science” showing overlays of three different broad topics within science and their disciplinary distribution.....	22
Figure 11. Combined structural and detailed map of science and technology (S&T) from 2003. Large circles indicate subdisciplines (clusters of journals) while the small dots are for research communities (clusters of papers).	24
Figure 12. Structural map of science and technology showing the publication activity of Sandia National Laboratories for 2003.	27
Figure 13. Structural map of science and technology showing the combined publication activity of the US Department of Energy National Laboratories for 2003.	28

Tables

Table 1. Steps for generating current and reference maps of science at the paper level.	16
---	----

1. Introduction

This document comprises the final report for an LDRD project entitled “A method of evaluating research using new innovation, risk, and impact indicators.” The overall goal of this project has been to create a literature-based system that would aid Sandia managers and researchers in evaluating both current and proposed research by generating very detailed large-scale maps of science with research indicators at the individual paper level. This goal has been met, although not without significant course changes along the way.

1.1. Need for an Evaluation System

In this age of modern information availability, organizations are searching for ways to capture tacit knowledge and use that knowledge to evaluate and select between new ideas. We are continually falling behind in our retention of knowledge due to the rapidity with which new information becomes available, and from a lack of good tools to help us retain what we have learned. This retention of knowledge is critical as it pertains to the management of R&D resources. Lack of knowledge can result in a misallocation of R&D dollars, an action that can have serious economic, social, technological, and political consequences. It is the consequences of these decisions that drive us to develop a literature-based system of research indicators.

We propose a method of evaluating research ideas by creating new indicators that, in essence, provide a numerical history of accumulated knowledge. These are research community-based indicators that can give the researcher, administrator, or evaluator a means to quickly evaluate an idea within the context of the history of similar ideas. A research community is a group of papers that are written on the same research topic, and is thus a unit of production that has meaning. Indicators will not only help in the evaluation process, but will also facilitate focused access to pertinent literature for those individual researchers with a desire to gain more detailed knowledge and thereby improve their ideas.

For such a system and its associated indicators to be valid, it must be based on “all of science.” The key underlying technology to enable this goal is therefore a model of all of science. Thus, the project has focused on generating large scale maps of science that include all individual papers. We have also focused extensively on the accuracy of the mapping process, as the accuracy of the maps determine the accuracy of the associated indicators.

The map of science can be used as tool for science strategy. This is the terrain in which organizations, institutions, and individual researchers locate their scientific capabilities. Indicators that describe the potential scientific and economic impact of each research community allow people from policy makers to individual to decide which areas to explore, exploit, abandon or ignore.

1.2. Definitions

A few definitions are given here to clarify concepts that will be used in the balance of the report.

Paper – a published article, appearing in either a journal or conference proceedings.

Keyword – a key term or phrase assigned to a paper, typically by authors or publishers.

Journal – a periodical (typically peer-reviewed) that publishes papers on a specific set of topics.

Citation – the act of one paper citing another paper (in its reference list).

Co-citation – two papers are co-cited if they are both cited by the same paper.

Bibliographic coupling – two papers are bibliographically coupled if they both cite the same paper.

Similarity (or relatedness) measure – an association strength between two objects (e.g., papers), typically ranging between 0 and 1.

Graph – consists of objects (commonly called “nodes”) and the linkages between those objects (commonly called “edges”).

Graph layout – the visual positioning of nodes in a graph in the x,y plane.

Clustering – assignment of nodes to groups based on criteria such as distance, linkages, etc.

Research community – a cluster of papers on the same topic, published during a given year.

Indicator – a metric describing some feature of a research community.

Subdiscipline – a cluster of journals that are strongly linked to each other, and that cover a particular topic space.

1.3. Chronological History

A brief chronological history of the project is given here, while more detail on specific parts of the project and associated technologies are given in subsequent sections of the report.

This project was initially funded as a late-start LDRD during FY03. The initial idea was to map the usage of keywords and phrases using power law characteristics, and to base indicators on various properties including citations statistics associated with keyword

usages over time. We recognized that there would be difficulties with this approach due to the natural ambiguity of language and word usage, and proposed to construct ontologies that would reduce those difficulties.

The scope and underpinnings of the project changed almost immediately when the Principal Investigator, Kevin Boyack, met Richard Klavans at an academic workshop at which both were speaking on different methods for science mapping. Klavans presented an extremely large scale map of science that was based on measures and methods that had not been published in the open literature. He also had worked for 10 years with Fortune 500 firms using indicators based on scientometric analysis. Klavans used co-citation analysis rather than keyword based methods for his indicators, and although he had published little due to the proprietary nature of the work, it was clear that his knowledge and techniques easily exceeded anything in the open literature. After many discussions, and after presentation of Klavans past work to decision makers in the LDRD program at Sandia, the project was expanded to include Klavans as a consultant. In addition, the basis for mapping was changed from keywords to citations in order to build on Mr. Klavans' know-how.

Since that point, work has progressed steadily. The key decision points and milestones are as follows:

- A project plan was created in which Klavans' previous methods would be tested at the journal level, and then expanded to the paper level.
- A decision was made to perform detailed accuracy studies of journal level maps to 1) determine the best similarity measure, 2) determine the effects of dimension reduction, and 3) determine the effects of inclusion of multidisciplinary journals.
- Klavans originally used a hierarchical clustering method and multi-dimensional scaling (MDS) to generate journal maps. Boyack proposed the use of VxOrd (a graph layout tool developed at Sandia as part of a previous LDRD) because of its scalability to very large datasets. VxOrd was adopted as the layout/clustering engine for mapping.
- Inter-citation measures were found to be more accurate than co-citation measures, and dimension reduction using VxOrd was found to improve the accuracy of most journal maps.
- A new similarity measure (K50) proposed by Klavans was used and found to be the most accurate similarity measure in most cases. K50 is simply a modified cosine measure which subtracts the expected value from the full cosine value.
- Detailed accuracy studies of journal maps were published in two journal articles.
- The journal mapping process was further refined to use bibliographic coupling (co-occurrence of references) as the similarity type. The process was deemed sufficiently accurate to be expanded to the mapping of papers.

- Mapping at the paper level proceeded with duplicate and parallel efforts by Boyack and Klavans. Coding was done independently. Cross-checking allowed us to fix errors in both processes to ensure the accuracy of the overall method and process.
- Now that maps existed at two levels (journals and papers), the process was updated to superimpose the two maps for visualization purposes. A poster “The Structure of Science” was created, and is now part of a traveling exhibit called “Places & Spaces”, which shows some of the best examples of mapping from the cartographic (places) to the abstract (spaces).
- Accuracy of our maps at the paper level was calculated in several different ways, and submitted as a journal article.
- Indicators were calculated at the research community level.
- Due to a request by the Computation and Information Sciences LDRD investment area team, mapping was expanded to include Conference Proceedings literature. The effect of adding this literature was characterized.
- Maps from multiple years were linked and integrated to enable calculation of dynamic indicators.

1.4. Report Structure

The balance of the report consists of three main sections: journal-level mapping, paper-level mapping, and application of the technology. Specific issues associated with each are reported on here. These sections consist primarily of work that has not been published or submitted elsewhere. Previous work and papers that were written as a part of this project are referenced [1-5], and are not included in the text.

2. Journal-level Mapping

Our early efforts at mapping journals are detailed in two papers that have been published in the open literature [2, 3]. Background materials, motivations, detailed literature searches, and results of our studies can be found in those papers. Those studies presented, for the first time, detailed accuracy results related to various similarity measures and their effects on mapping of journals. The accuracy studies represent a major contribution to the scientometrics and science mapping fields, and represent a major part of the overall technical advances achieved by this project. In addition, a detailed map of the structure of science was published with commentary. That map of science, in scatterplot (Figure 1) and journal cluster forms (Figure 2), is reproduced here to allow comparison with our most recent work on journal mapping, which is presented below.

Figure 1. Map of science generated using the IC-Jaccard similarity measure [3]. The map is comprised of 7,121 journals from year 2000. Large font size labels identify major areas of science. Small labels denote the disciplinary topics of nearby large clusters of journals.

of Figure 2 gives the impression that there is a center to science, and that science is evenly spread. Yet, our experience, and other publications, would suggest otherwise. Moya finds a ring structure [6] while Small finds a linear structure to science [7]. VxOrd is a space filling algorithm, and thus does not leave white space in a layout if there are enough nodes to fill in a space. VxOrd is an interactive layout routine, and allows the user to watch the progression of the layout. When running VxOrd on these journal maps, we note that the layout originally forms a loose ring with white space in the center, and that the layout then relaxes to fill the middle space.

We hypothesized that if we could recursively cluster journals, that is if we would cluster the clusters, the layout of the clusters would form a more accurate structural layout both in terms of position of the journal clusters, and in terms of the distance between them.

This new idea was tested on a new set of data. While the previous study, and our published results were done using the 2000 dataset, we moved to a more recent set of data (from 2002) whose results would be more current.

For these new calculations, we used the combined Science Citation Index Expanded (SCIE) / Social Science Citation Index (SSCI) data from fileyear 2002. The data consisted of 1.07M records of individual papers from over 7,351 separate journals. There were a total of 24.5 million references from the 1.07 million records.

Our previous work had shown that inter-citation measures were more accurate than co-citation measures. But given that one of our desires was to use the same process for mapping journals and for mapping papers, use of inter-citation is impractical given that papers typically do not cite papers published in the current year. A one-year paper-level map generated from inter-citation would leave out the majority of papers. The only way to generate paper-level maps of current papers from citation data is to use bibliographic coupling (where papers are related if they common entries in their reference lists). Thus, we decided to use bibliographic coupling for the journal-level maps as well.

The process for generating the map was as follows. First bibliographic coupling (BC) counts were computed for each pair of papers. The counts were then summed up to the journal level to get the bibliographic coupling counts between pairs of journals. Similarity between journals was calculated as the K50 coefficient where

$$K50_{i,j} = K50_{j,i} = \max \left[\frac{(RAW_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(RAW_{i,j} - E_{j,i})}{\sqrt{S_i S_j}} \right],$$

the expected value of the cosine $E_{i,j} = \frac{S_i S_j}{(SS - S_i)}$,

$$\text{and } SS = \sum_{i=1}^n S_i.$$

The K50 coefficient is essentially a cosine minus its expected value. Properties of the K50 coefficient are detailed in Klavans and Boyack [2].

Multiple-step graph layout is performed with VxOrd, resulting in clusters of journals that form the disciplinary structure of science. The first level layout and clustering were performed using VxOrd with the cutting parameter set to its maximum. This cuts a majority of edges and results in a graph layout with very distinct clusters. The clusters were determined automatically using a modified single-link algorithm that uses distances and uncut edges as input parameters. 671 clusters of journals were found using this method.

The second-level layout was done by once again aggregating coupling counts, this time to the journal cluster level, and then recalculating K50 values. VxOrd was run once again with the 671 clusters as input, and with the cutting parameter set very low. The resulting structural map is shown in Figure 3.

Figure 3. Structural map of science generated using a multi-step bibliographic coupling process. The map is comprised of 671 clusters representing 7,351 journals from year 2001. Dominant linkages between journal clusters are shown.

This new process has the desired effect of maintaining white space within the structural map of science. Most disciplines are comprised of several different clusters of journals, each of which contains several journals. The clusters are called subdisciplines in that they each represent a topical section of one of the larger disciplines within science.

It is also interesting to note that the ordering of disciplines in the maps from Figures 2 and 3 are very similar, even though different data (2002 vs. 2000), a different similarity

process (bibliographic coupling vs. inter-citation), and a different process (multi-step vs. single step) were used. In each case we have rotated the map so that Math is at the top, and Physics is at the right. The ordering of the major disciplines is the same in each case as one moves clockwise around the maps: Mathematics, Computer Science, Physics, Chemistry, Life Sciences, Medical Sciences, Social Sciences, and back to Mathematics. This shows the robustness of the mapping process in general.

2.2. Mapping of Science and Technology

During the second year review of this LDRD project, the Computation and Information Sciences (CIS) investment area team made a request that this mapping and indicators effort include more of the technology literature. This is very important in the computer science field, as a large fraction of the computer science literature is published in conference proceedings rather than in journals. Accordingly, we updated our process to include the ISI Proceedings database, which Sandia had recently subscribed to. We had also recently received the 2003 fileyear data, and so a new map was generated using these data, and using a process identical to that used for the 2002 science map.

The 2003 fileyear dataset consisted of 1.35 million records (papers) from the combined SCIE/SSCI/ ISI Proceedings databases. These records came from 7,445 journals and 1,198 conference proceedings, and contained a total of 29.23 million references.

The multi-step journal mapping procedure produced 852 separate journal/proceedings clusters in a layout as shown in Figure 4, which also shows the effects of adding the proceedings into the dataset. The sizes of the clusters indicate the number of papers represented by each (using cubed root scaling). The color indicates the prevalence of conference papers, where the lightest blue indicates clusters with a low percentage (<25%) of conference papers, and the darkest blue indicates clusters with a very high percentage (>75%) of conference papers. As expected, computer science is very heavily influenced by conference papers, and benefits tremendously by their inclusion in the map. Some subdisciplines within physics and engineering also have a large fraction of their work in conferences. By contrast, the life sciences, medical sciences, and social sciences are influenced very little by conference papers.

We also note that a comparison of the maps in Figures 3 and 4 shows that while there is a considerably larger (~20%) number of clusters in the S&T map of Figure 4, and particularly in computer science and engineering, the overall structure and ordering of the disciplines remains the same. Thus, while inclusion of conferences may add significant detail to the map, they do not have a large effect on the overall structure of science and technology.

We consider this map as shown in Figure 4 to be the current definitive map of science and technology for several reasons. First, it represents the largest literature set ever mapped, and thus is the most comprehensive map available. Second, it does include proceedings, and thus covers more technology than a pure science (journals) map. Third, the consistency of the maps regardless of the particular fileyear mapped or the details of

the mapping process suggest that the structure is robust. We propose to use this map as a template for presenting other results, as will be detailed in Section 4.

Figure 4. Structural map of science and technology from the 2003 combined dataset. The map contains 852 clusters representing 7,445 journals and 1,198 conference proceedings. Dominant linkages between clusters are shown.

3. Paper-level Mapping

Our ultimate goal was to produce paper-level maps of science and to calculate indicators at the level of clusters of papers. The journal-level mapping described previously was beneficial in multiple regards. First, the journal set is a much smaller set, and allowed us to tune and refine our approaches before using them on the much larger paper sets. Second, the journal map gives us a framework for presenting results at an aggregated disciplinary level that is intuitive for most technical people.

After each stage of journal mapping, the associated paper mapping was done. We note that for papers, two maps were calculated rather than just one. Current paper maps, using the papers from the current fileyear, were calculated using bibliographic coupling. Reference paper maps, consisting of the prior literature that was referenced by the current papers, were also calculated. Both the current and reference maps can be calculated using the same process. Although both maps use the same list of citing:cited paper pairs as input, they differ in the initial calculation of the number of coupling counts between papers.

The process used to calculate paper maps is very similar to that used to calculate the journal maps. The only major difference is that some thresholding levels were applied to the processing of papers that were not necessary for journals. The purpose of thresholding is to remove papers that have very weak relationships to other papers. The process is as follows for current and reference papers:

Table 1. Steps for generating current and reference maps of science at the paper level.

Step	Current paper map	Reference paper map
1	Calculate bibliographic coupling counts between pairs of current papers	Calculate co-citation counts between pairs of reference papers
2	Apply thresholding – remove any papers that have a maximum co-occurrence count of 1, or that only co-occur with one other paper This defines the Set-A current papers Removed papers are Set-B	Apply thresholding – remove any papers that have a maximum co-occurrence count of 3 or less, or that only co-occur with one other paper This defines the Set-A reference papers Removed papers are Set-B
3	Calculate K50 coefficients for Set-A papers	Calculate K50 coefficients for Set-A papers
4	Limit the set of coefficients to the top 10 per Set-A paper and ordinate using the VxOrd graph layout routine using maximum cutting	Limit the set of coefficients to the top 10 per Set-A paper and ordinate using the VxOrd graph layout routine using maximum cutting
5	Cluster the Set-A papers using Klavans' modified single-link clustering algorithm and the results of the VxOrd run (coordinates and remaining edges) This results in current communities	Cluster the Set-A papers using Klavans' modified single-link clustering algorithm and the results of the VxOrd run (coordinates and remaining edges). This results in reference communities
6	Assign Set-B papers to communities if	Assign Set-B papers to communities if

	they can be unambiguously assigned using numbers of potential partners and/or distances <i>At this point we have a list of current papers and their current community assignments</i>	they can be unambiguously assigned using numbers of potential partners and/or distances <i>At this point we have a list of reference papers and their reference community assignments</i>
7	Assign reference papers to current communities based on numbers of potential partners <i>At this point we have current communities containing both current papers and reference papers</i>	Assign current papers to reference communities based on numbers of potential partners <i>At this point we have reference communities containing both reference papers and current papers</i>
8	Link all communities, both current and reference communities, based on overlap between their reference paper members	
9	Calculate indicators for each community	

Steps 1-7 were repeated several times over the course of this project, each time using data from different fileyears and different database combinations.

3.1. Paper-level Map of Science

Using the distinction introduced in the journal-level mapping discussion, we denote a map of science as one that does not include conference proceedings, and a map of science and technology as one that does include conference proceedings. Here we detail our first attempts to map science at the paper level.

We used the combined ISI databases from the 2002 fileyear in this study (2002 SCIE and SSCI databases). There were 1,069,764 current papers and 10,911,939 unique reference papers in the dataset. Of the over 1 million current records, we limited the number of “mappable” records to the 833,307 that are bibliographically coupled to another current record within the set. This excludes the majority of editorials, book reviews, and similar items that are indexed by ISI, and limits the dataset primarily to records containing publishable technical advances. This parallels the 10,911,039 reference papers used in the set in that, since they were referenced, they can be considered to be publishable technical advances that were worth referring to.

Our work with the 2002 paper-level maps focused on ways to establish the accuracy of these very comprehensive and detailed maps. Details of the accuracy studies are available in two papers [4, 5], and will not be reproduced here. In total, we generated eight different maps using the 2002 dataset, and ran various statistical measures against each of the maps to establish relative accuracies and biases. In the end we determined that the K50 measure generated better maps than raw frequency measures (which have been historically used in large maps) because they give higher local accuracy, reduce disciplinary biases at subsampling levels, and generate very few large clusters [4]. Large clusters are considered problematic in such mapping efforts. It was also shown that

running VxOrd at maximum cutting values was advantageous in that it reduced cluster sizes and biases as well.

The 2002 map of current papers generated using the K50 measure and maximum cutting within VxOrd is described here. Of the 833,307 mappable records, only 731,289 were actually included in the map due to the application of the coupling thresholds mentioned in Table 1. After ordination and clustering, these papers generated 96,500 separate clusters, or current research communities. The 2002 current papers map, as generated by VxOrd, is shown in Figure 5.

Figure 5. Map of current papers from 2002. Five major fields are shown by different colored markers. The map has been rotated to place Mathematics (not marked) at the top and physics at the right, which is our standard convention. The inset shows the detail in a small portion of the map including how linkages form research communities.

Figure 6 shows the current and reference maps side by side. While there are differences, the overall structure and distribution of fields between the two maps is quite similar. The biggest difference is that biology (proxy for all life sciences in this context) forms more of an interface between chemistry and medicine in the reference map than it does in the current map.

Figure 6. Comparison of the current (left) and reference (right) maps of papers from 2002.

In addition to the accuracy studies reported in [4, 5], we did some spot checking of abstracts and titles in order to see if the research communities had the type of topical focus that we expected. The membership of a representative community is listed in Figure 7, and shows that the community does indeed have a strong topical focus. The only paper that appears not to fit is the sixth paper from the journal *Obes Surg*. Spot checking of another 50 communities gave similar results.

Journal	Paper Title	Weight
<i>Scientometrics</i>	Reflections on scientific collaboration , (and its study): past, present, and future	4)
<i>Scientometrics</i>	Continuity and discontinuity of collaboration behaviour since 1800	4)
<i>Scientometrics</i>	Elite researchers in ophthalmology: Aspects of publishing strategies, collaboration and multi-disciplinarity	3)
<i>Scientometrics</i>	The effect of team consolidation on research collaboration and performance of scientists. Case study of Spanish university researchers in Geology	3)
<i>Scientometrics</i>	Recognition and international collaboration : the Brazilian case	2)
<i>Obes Surg</i>	Progress of the International Federation for the Surgery of Obesity	2)
<i>Scientometrics</i>	A. H. Zewail: Research collaborator par excellence	1)
<i>Scientometrics</i>	Age structures of scientific collaboration in Chinese computer science	1)
<i>Philos Sci</i>	The epistemic significance of collaborative research	1)
<i>Braz J Med Biol Res</i>	The Brazilian investment in science and technology	1)
<i>Region Anesth Pain M</i>	What do we measure by co-authorships ?	1)
<i>Scientometrics</i>	Authorship patterns in agricultural sciences in Egypt	1)

Figure 7. Membership of current community 73336 from the 2002 map of science. Common terms are highlighted to show the topical focus of the community.

We note that, although the maps of Figures 5 and 6 are “pretty pictures”, they do not produce the white space and differentiation between disciplines that we desire for purposes of planning and evaluation. In these maps of science, the lines between disciplines are somewhat blurred, and neighboring communities are more topical than disciplinary. In other words, two research communities that focus on very similar topics will be close to each other on the maps of Figures 5 and 6 even though they may be from different disciplines. We made the decision that it would be more advantageous to map research communities closer to their dominant disciplines or subdisciplines. In that way, similar topics might appear in different sections of a map, and would show that there either exists or should exist a pathway for interdisciplinarity. Such a pathway would not show up on the topical map of Figure 5.

Thus, we took the additional step of mapping the research communities from Figure 5 onto the structural map shown in Figure 3. The position of each community was calculated by triangulating its constituent journal positions from the structural map. For instance, if a research community had 10 members, 8 of which were from a journal in subdiscipline 1, and 2 of which were from a journal in subdiscipline 2, the x position of the research community was calculated as $x_c = [0.8^2(x_1) + 0.2^2(x_2)] / [0.8^2 + 0.2^2]$. Use of squares in this formula places the community closest to its dominant subdiscipline. The placement of research communities on the structural map is shown in Figure 8.

Figure 8. Structural (left) and detailed (right) maps of science from 2002. In the detailed map, research communities have been positioned on the structural map.

The detailed community map meets our needs in that it shows great detail, but maintains the context of a discipline-level map. This map also shows the density of topics within various areas of science, as well as the topic-based interdisciplinary pathways between disciplines that are not visible on the structural map alone. For example, Biochemistry is not linked to just other biological sciences. Distinct pathways between biochemistry several other areas including Brain Science, Chemistry, Microbiology, Geosciences, and several areas in Medicine can be seen.

This particular map was turned into a poster that is now a part of the *Places & Spaces: Cartography of the Physical and Abstract* (<http://vw.indiana.edu/places&spaces/>)

traveling exhibit. Portions of the framed poster are reproduced in Figures 9 and 10, and tell the story of the map and of some of its uses.

Figure 9. Left half of the poster “The Structure of Science” by Kevin Boyack and Richard Klavans. Colored flags and text correspond to sections of the galaxy-like map using the same colored labels. Each white dot represents one research community.

Figure 10. Right half of the poster “The Structure of Science” showing overlays of three different broad topics within science and their disciplinary distribution.

3.2. Paper-level Map of Science and Technology

As mentioned in section 2.2, a request was made that we incorporate proceedings literature into our maps. Thus, after generating the structural S&T map using the 2003 dataset, as shown in Figure 4, we proceeded to generate paper-level maps using the same data. New current and reference paper maps were generated using these data, and using a process identical to that used for the 2002 paper-level maps.

The 2003 fileyear dataset consisted of 1.35 million records (papers) from the combined SCIE/SSCI/ ISI Proceedings databases. These records came from 7,445 journals and 1,198 conference proceedings, and contained a total of 29.23 million references. Of the 1.35 million paper, a total of 997,775 current papers were ultimately mapped because of the thresholds that were applied. The ordination and clustering steps identified a total of 117,433 current communities within the map. Given that accuracy studies were done for the 2002 science map, and that our experience is that the mapping process is robust, we did not feel the need to repeat the accuracy studies for these 2003 S&T maps.

For the 2003 reference papers map, there were originally 12,570,587 unique references available. The majority of these papers were referred to only once, and were thus removed from consideration by our thresholding criteria. A total of 871,038 papers were mapped into 76,376 separate reference research communities.

Current communities were mapped onto the structural map as noted previously. The resulting combined map is shown in Figure 11. This map has much the same character as those shown in Figures 8-10 in that the densities of topics in various fields of science are depicted. In addition, the relative interdisciplinarity of different fields are easily seen.

Additional processing has been done to calculate S&T maps (including proceedings) for the 2002 and 2004 dataset to allow linking of communities from year to year. The structural map is robust from year to year over short periods of time, and thus the structural map from 2003 is used for the 2002 and 2004 fileyears as well. Journals and proceedings that are unique to either the 2002 or 2004 fileyears are assigned to subdisciplines based on their dominant linkages to journals within the 2003 structural map.

Figure 11. Combined structural and detailed map of science and technology (S&T) from 2003. Large circles indicate subdisciplines (clusters of journals) while the small dots are for research communities (clusters of papers).

4. Applications

Once maps such as those shown in Figure 11 are generated, various metadata and metrics can then be calculated for each research community. For metadata, we generate community summaries, key phrases, and coherence values from the abstracts of the papers in the community using routines obtained from the Computational Research Laboratory at New Mexico State University. The key phrases in particular give us a set of terms that describe the particular research question that is the focus of the research community. The key phrases from several communities are shown in the table below as examples.

C_95080	C_11965	C_103409	C_92818
North Korea South Asia arms transfers foreign policy	quiescent H-mode H-mode operation edge harmonic MHD oscillation harmonic oscillation	crystalline phases liquid crystalline aspect ratio electron microscopy silver colloids	molecular beam beam epitaxy output power GaAsN layers PL intensity

The primary indicators that we use are size and vitality. Size is, of course, the number of papers in the community. Although it is a very simple indicator, it does reflect to some degree the importance of a topic in the current research environment. Larger topics are those receiving more attention and more funding, and are thus more important to those sponsoring research.

Vitality is a measure that is related to the *average age of all cited references* from the papers in a community. The vitality for community c is calculated as

$$V_c = \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{Age_j + 1} \right),$$

where n is the number of references from all current papers assigned to a community, and Age_j is the age of reference j in years. Vitality is thus bounded between zero and one. Communities that refer to more recent research (in the former of younger papers) have a higher vitality. The research in those topics is updating itself more quickly. Highly vital topics are, thus, fast moving areas of research. Vitality is one metric by which we can compare different communities within a subdiscipline. A high vitality does not necessarily mean that the research in one community is better than that in another, but merely that it is in a faster moving, or more vital, topic. We do not compare vitalities between subdisciplines because different subdisciplines in science and technology have different citing cultures, and thus different natural vitalities. Rather, we calculate the mean vitality for each subdiscipline, and use it as a reference vitality.

Maps of science with indicators at the research community level have many potential applications related to aiding the planning and evaluation of research, which was the original goal for this project.

At the individual staff level, these data would allow a staff member to find a community of interest, perhaps one closely related to current research, or perhaps one that is related to a new idea, see the papers and researchers that are part of the community along with the indicators associated with the community. The data could also highlight related communities, those that are strongly linked through citation, that have strong indicators, thus alerting the staff member to information that he or she might not already know. Such information could be very useful in directing new research or writing proposals. This particular application has not yet been implemented. We hope to implement this as a system that could be accessed by all Sandia staff during FY06.

In terms of evaluation at the organizational level, the map of science and technology can be used as a template upon which to overlay other information. For instance, Figure 12 shows the publication activity for all of Sandia National Laboratories during 2003. To generate this overlay, all papers authored by Sandia were identified, which gave us a list of all of the research communities in which Sandia was an active participant. We were then able to count the number of research communities in each subdiscipline. In many cases, Sandia was active in ten or more communities within a particular subdiscipline.

Figure 12 shows both the number of communities, and the relative vitality of the communities in which Sandia was active by subdiscipline. The size of each circle shows the number of communities, or the relative levels of Sandia activity, and the color of each circle shows the relative vitality for those communities in which Sandia participated. An example of a relative vitality calculation follows:

Suppose that Sandia was active in three communities in a particular subdiscipline, with vitalities of 0.20, 0.23, and 0.26, and suppose that the base vitality for that subdiscipline has a vitality of 0.20. The relative vitality is then calculated as

$$\text{Avg}(0.20, 0.23, 0.26) / 0.20 = 1.15$$

In this case, the vitality of the Sandia activity would be 15% greater than that of the world at large for the particular subdiscipline. Using our color scale from Figure 12, this vitality would merit a red dot. Figure 12 shows that Sandia has a higher than average vitality (red and orange circles) in many areas of science and technology such as physics, engineering, and computer science. Sandia has a lower than average vitality (yellow and white dots) in fewer areas, such as some areas between physics and computer science.

Overlays such as this can be done for nearly any type of rollup imaginable. They can be done for companies, agencies, countries, etc. Side-by-side comparison of multiple overlays (e.g. for different companies or laboratories) can be very informative. For instance, the map of DOE activity at all of its national laboratories is shown in Figure 13. In addition, other metrics, such as citation counts, linked patent counts, or community continuity, can be defined and used as circle colors.

Figure 12. Structural map of science and technology showing the publication activity of Sandia National Laboratories for 2003.

Figure 10 shows overlays on the structural map of various research topics, such as nanotechnology, and can show the footprint of a particular technology across different fields of science. Such overlays can be done for nearly any technology or topic, and can show the spread of a technology from one field to another with the time-sequencing of data. Thus, our maps of science and technology are suitable for dynamic, as well as static studies.

Figure 13. Structural map of science and technology showing the combined publication activity of the US Department of Energy National Laboratories for 2003.

5. References

1. Boyack, K.W., *Mapping knowledge domains: Characterizing PNAS*. Proceedings of the National Academy of Sciences, 2004. **101**(Suppl. 1): p. 5192-5199.
2. Klavans, R. and K.W. Boyack, *Identifying a better measure of relatedness for mapping science*. Journal of the American Society for Information Science and Technology, 2006, in press.
3. Boyack, K.W., R. Klavans, and K. Börner, *Mapping the backbone of science*. Scientometrics, 2005. **64**(3): p. 351-374.
4. Klavans, R. and K.W. Boyack, *Quantitative evaluation of large maps of science*. Scientometrics, 2006, in press.
5. Klavans, R. and K.W. Boyack. *Mapping world-wide science at the paper level*. in *10th International Conference of the International Society for Scientometrics and Informetrics*. 2005. Stockholm, Sweden: Karolinska University Press.
6. Moya-Anegón, F., et al., *A new technique for building maps of large scientific domains based on the cocitation of classes and categories*. Scientometrics, 2004. **61**(1): p. 129-145.
7. Small, H., *Visualizing science by citation mapping*. Journal of the American Society for Information Science, 1999. **50**(9): p. 799-813.

Distribution:

1	MS0114	N. M. Rahal	5925
1	MS0123	H. R. Westrich	1011
1	MS0123	D. Chavez, LDRD Office	1010
1	MS0145	C. E. Meyers	0220
2	MS0310	K. W. Boyack	1412
1	MS0310	M. D. Rintoul	1410
1	MS0310	G. S. Davidson	1400
1	MS0511	R. D. Watkins	1010
1	MS0511	W. R. Cieslak	1010
2	MS0899	Technical Library	9616
1	MS0915	N. P. Searls	5900
2	MS9018	Central Technical Files	8945-1
2	MS0899	Technical Library	4536