PERSPECTIVES

Refining search terms for nanotechnology

Alan L. Porter · Jan Youtie · Philip Shapira · David J. Schoeneck

Received: 30 April 2007 / Accepted: 29 June 2007 / Published online: 3 August 2007

Abstract The ability to delineate the boundaries of an emerging technology is central to obtaining an understanding of the technology's research paths and commercialization prospects. Nowhere is this more relevant than in the case of nanotechnology (hereafter identified as "nano") given its current rapid growth and multidisciplinary nature. (Under the rubric of nanotechnology, we also include nanoscience and nanoengineering.) Past efforts have utilized several strategies, including simple term search for the prefix nano, complex lexical and citation-based approaches, and bootstrapping techniques. This research introduces a modularized Boolean approach to defining nanotechnology which has been applied to several research and patenting databases. We explain our approach to downloading and cleaning data, and report initial results. Comparisons of this approach with other nanotechnology search formulations are presented. Implications for search strategy development and profiling of the nanotechnology field are discussed.

Keywords Bibliometric analysis · Nanoscience and engineering · Nanotechnology publication · Nanopatenting · Research profiling · Search strategies · Nanoinformatics

Introduction

There are many ongoing efforts to assess the evolving nature of nanotechnology research and innovation systems in the US and internationally. A fundamental building block of this work involves development of an operational definition of nanotechnology *in specific bibliometric terms.*

Nanotechnology is held to be the manipulation of molecular-sized materials to create new products and processes.[1] It encompasses contributions from fields

A. L. Porter · P. Shapira
Georgia Institute of Technology, Atlanta, USA

J. Youtie (✉)
Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, GA, 30332-0640, USA
e-mail: jan.youtie@innovate.gatech.edu

P. Shapira
Center for Nanotechnology in Society (CNS-ASU), Program in Nanotechnology Research and Innovation Systems Analysis, Tempe, AZ, USA

D. J. Schoeneck
Search Technology, Inc., Norcross, GA, USA

[1] Here, we follow the definition developed by the US National Nanotechnology Initiative (NNI) which defines nanotechnology as "encompassing the science, engineering, and technology related to the understanding and control of matter at the length scale of approximately 1–100 nanometers." Importantly, NNI adds that "nanotechnology is not merely working with matter at the nanoscale, but also research and development of materials, devices and systems that have novel properties and functions due to their nanoscale dimensions and components" (PCAST 2005).

such as physics, chemistry and biochemistry, molecular biology, and engineering, with potential applications in areas as diverse as drug delivery and discovery, environmental sensing, manufacturing, and quantum computing. However, to robustly track the development of research and commercialization in nanotechnology, there is a need to define in greater detail the multiple sub-fields within the nanotechnology domain. This will make possible the ability to search large-scale and multiple databases to retrieve relevant research articles, patents applications and awards, and other information types to map and assess nanotechnology research and commercialization trajectories.

In this paper, we provide an overview of the method and process we are using to develop refined nanotechnology search terms. We also compare with other nanotechnology search definitions, discuss our approach to downloading and cleaning data, and report initial results. The paper concludes with reflections on our search process and planned and potential analyses of the resulting databases.

Background

A brief evolutionary history of the authors' nanotechnology data interests underlies the development of the approach used to delineate the nanotechnology domain described herein. The work reported in this paper was undertaken primarily to develop real-time databases of research activity (publications) and innovation (patents) to map, analyze and model nano research and innovation systems in the US and globally with colleagues in the Center for Nanotechnology and Society at Arizona State University (CNS-ASU).[2] However, the genesis of our approach to nano profiling has its origins in an earlier project on "Creative Capabilities and the Promotion of Highly Innovative Research in Europe and the United States" (CREA).[3] This project involves analysis of creative research in the domains of nanotechnology and human genetics. In 2005, for the CREA project, researchers in Georgia Tech's Technology Policy and Assessment Center (TPAC) identified over 100,000 Web of Science (WOS) records and over 10,000 US patents relating to nanoscience and engineering (NSE). The bibliometric definition to search for these records was developed by the Fraunhofer Institute for Systems and Innovations Research (2002). These research publication and patent abstract records were imported into *VantagePoint* text mining software for analyses.[4] In the CREA project, the bibliometric records were used to identify academic, government, and corporate researchers publishing in nanotechnology fields between 1995 and 2004.

Also leveraged were activities undertaken through the Partnership for Innovation project at North Carolina State University (NCSU) on nanotechnology (sponsored by the National Science Foundation). This project seeks to foster knowledge transfer from non-industrial research to promote industrial innovation. With guidance especially from Professor Angus Kingon (a nanoscientist in the Department of Materials Science and Engineering at NCSU), we explored various nanotechnology search algorithms. The approach in this project called for the compilation of a relatively encompassing set of nanotechnology research publication and patent records from which we could then extract records that would subsequently inform a particular technology transfer endeavor. For instance, in early 2007, the NCSU Center for Innovation Management Studies (CIMS) team organized a workshop at Purdue University. As background to that workshop, we created a profile of nanotechnology research activity at Purdue based on more than 2000 nanotechnology publications authored by researchers at that institution since 1990. This information was made available to CIMS to enhance work with Purdue colleagues to identify the key thrusts and leading researchers, and to identify industrial counterparts apt to be especially interested in those thrust areas.

To advance these activities, a variety of nanotechnology profiling efforts were reviewed. These include Kostoff et al. (2006a, b), a Brazilian nanopatent search (Alencar et al. 2007), a broad perspective on

[2] See: http://cns.asu.edu/.

[3] The CREA project involved researchers from the Fraunhofer Institute for Systems and Innovations Research (Fhg-ISI), Germany, the Technology Policy and Assessment Center at Georgia Institute of Technology (USA), and Science and Technology Policy Research (SPRU) at Sussex University, UK, with sponsorship from the European Union's program in New and Emerging Science and Technologies (NEST), see Heinze et al. (2007).

[4] See: http://www.thevantagepoint.com/.

nanoscale science and engineering (NSE) (ETC 2003), and an infometrics treatment of nano (Zitt and Bassecoulard 2006; Bassecoulard et al. 2007). The Huang et al. (2003, 2004) articles also provided insights on nanotechnology trends. An examination of the definitions in this literature suggests that the proposed search approaches varied considerably in how they treated the interface between biotechnology and nanotechnology, and the extent to which they captured research in other nanotechnology sub-fields.

These approaches provided the basis upon which to develop an alternative search strategy. We conducted a number of exploratory search comparisons based on the definitions in the aforementioned prior work. To begin, the behavior of research publications in the 2005 CREA search was compared with that of Kostoff et al. (2006a, b). The results, presented in Table 1, provide a sense of how diffuse the nanotechnology domain is, and how challenging it becomes to generate a refined search strategy. The table shows that these two search algorithms yield comparable numbers of nanotechnology publication "hits" (45,000 for 1 year of WOS). However, almost 30% of the publications retrieved using each search strategy yielded uniquely differing outcomes. Detailed comparisons suggested that particular term phrasings were responsible for these differences. Additionally, given that some definitions were developed several years ago, we wondered whether emerging sub-topics were adequately captured. In early 2006, we proceeded to extend this prior work by developing an alternative nanotechnology research publication and patent data search definition.

Table 1 Comparing web of science coverage of two nano search algorithms (for 2005)

Search result summary table

Search	Records	Description
CREA	45,168	CREA total
Kostoff	45,845	Kostoff Total
CREA OR Kostoff	58,559	Union
CREA AND Kostoff	32,454	Intersection
CREA NOT Kostoff	12,714	Records Unique to CREA
Kostoff NOT CREA	13,391	Records Unique to Kostoff

Source: Georgia Tech TPAC analysis of publications for 2005 from WoS using nanotechnology search terms employed by Project on Creative Capabilities and the Promotion of Highly Innovative Research in Europe and the United States (CREA) (using definition of Fraunhofer Institute for Systems and Innovations Research 2002) and Kostoff et al. (2006a, b)

Differing approaches to the development of search terminologies exist. Our use of the Boolean search term approach can be counterposed with an alternative that can be termed "bootstrapping." The Georgia Tech team considered the iterative and expansive (or "bootstrap") search methods carried out by the colleagues at the Nanobank at the University of California Los Angeles (UCLA), Duke University, the European PRIME network, and others.[5] While the specific methods of these methods differ, in general they take a core set of nanotechnology papers as their starting point for further elaboration. Elaboration can involve examination of other papers by authors of nanotechnology focused papers that may not use "nano-terms" per se in titles, keywords, or abstracts. Another mode of extension is to consider papers referenced by, or referencing, the core nanotechnology set. The rationale of this method is that these subsequent works reflect research knowledge transfer with the core nanotechnology publication set, hence, are apt to be highly salient. In some cases, review by nanotechnology authors or experts is used to fine-tune the expanded sets, discard less related work, and add further relevant pieces. Expert review has the advantage of not being limited to use of classification codes (indexes), keywords, or prominent terms (in titles, abstracts, patent claims, etc.). However, the ongoing and extensive use of expert-based approaches is costly, very time-consuming, and challenging to replicate such that the same outcomes result.

We were motivated by the CNS-ASU project's theme of "real-time technology assessment" (Guston and Sarewitz 2002) and the NCSU project's need for available data to inform particular initiatives. To

[5] Prior to determining our search strategy, we consulted with others in the nanotechnology research community. In December 2005, we participated in a conference call involving members of the UCLA Nanobank team, CNS-ASU, CNS-UCSB, and other nano projects to discuss nano search strategies and information sharing. We also initiated contacts with Duke University (Giannela) and the European Union PRIME network (Mangematin) to share ideas and, potentially, to share nano information. We also interact on an ongoing basis with Georgia Tech colleague Stuart Graham, who is working on a UCLA-Harvard nano project, primarily focusing on nanopatenting.

support those, we sought to have a viable nanotechnology information resource at hand for further subdataset probing within a reasonably short period of time. We thus decided to use a Boolean search method which would produce usable results more quickly (and less expensively), but which could also be modified and tuned in subsequent rounds.

We chose a modular Boolean term search approach, augmented by class code enhanced patent searching. We established three key criteria for the inclusion of search terms, namely that they should be:

(1) encompassing–the term should be associated with a sizeable quantity of articles while at the same time being relevant to the domain
(2) transparent—researchers should be able to determine how well a topic of interest is covered by the search; and
(3) elastic—it should be easy to add/remove/modify terms from a search to adjust the record set to meet differing research interests as the field of nanotechnology evolves.

Regarding the latter point, we can track the emergence of new terms over time and adjust the search algorithms dynamically in updating nanotechnology datasets. Having the data available in text mining software (VantagePoint) enables the extraction of subsets of records to profile activity relating to a particular theme, or by a particular organization. The details of this approach are described in the following section.

Overview of nano search definition method and process

Our approach to developing a nanotechnology bibliometric search definition involves three major steps. First, we created a pilot "field scope," drawing upon and combining search terms and insights from prior efforts to define nanotechnology search terms. Second, we asked multiple nanotechnology experts to review our pilot field scope and, in so doing, received recommendations to delete, modify, add, or retain terms. Third, we further evaluated candidate terms by testing and assessing results against the publication and patent data. Over time, we can adjust the terms and class codes to address observed weaknesses in the dataset and follow emerging research trails.

From our comparisons of the CREA and early Kostoff search results (summarized in Table 1; more recent results from Kostoff et al. (2006) work is described in Sect. "Initial base analyses") supplemented by insights gained from the other search definitions, we developed a pilot "field scope" definition. This included a schematic Venn diagram representing four overlapping fields: (1) metrology and nanoprocesses, (2) nanostructure chemistry and materials, (3) nanodevices and nanoelectronics, and (4) nano-medicine and nano-biotechnology (Fig. 1). Within each field, examples of key terms were included.

We then developed a detailed search algorithm comprised of eight major sections and a series of

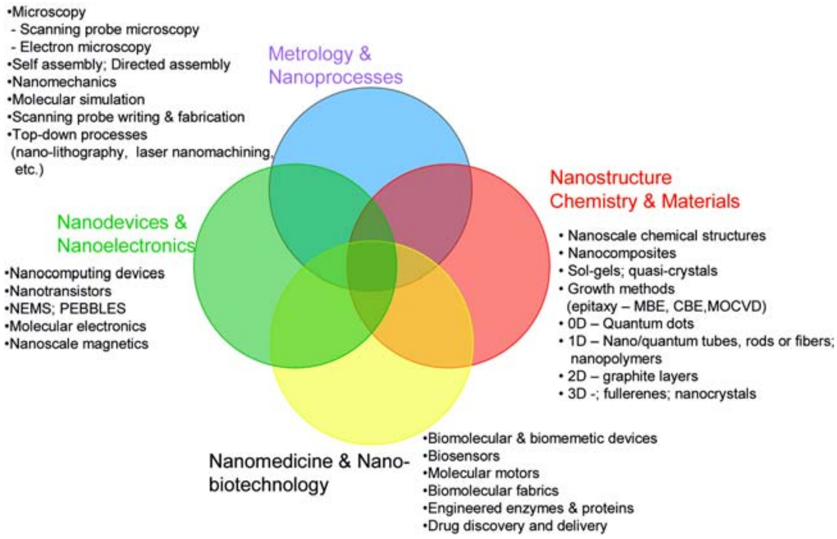
•Microscopy
- Scanning probe microscopy
- Electron microscopy
•Self assembly; Directed assembly
•Nanomechanics
•Molecular simulation
•Scanning probe writing & fabrication
•Top-down processes
(nano-lithography, laser nanomachining, etc.)
Metrology & Nanoprocesses
Nanostructure Chemistry & Materials
• Nanoscale chemical structures
• Nanocomposites
• Sol-gels; quasi-crystals
• Growth methods
(epitaxy – MBE, CBE,MOCVD)
• 0D – Quantum dots
• 1D – Nano/quantum tubes, rods or fibers; nanopolymers
• 2D – graphite layers
• 3D -; fullerenes; nanocrystals
Nanodevices & Nanoelectronics
•Nanocomputing devices
•Nanotransistors
•NEMS; PEBBLES
•Molecular electronics
•Nanoscale magnetics
Nanomedicine & Nano-biotechnology
•Biomolecular & biomemetic devices
•Biosensors
•Molecular motors
•Biomolecular fabrics
•Engineered enzymes & proteins
•Drug discovery and delivery

Fig. 1 Venn diagram of intersecting nano emphases. *Source*: "Field Scope" of Nanotechnology, developed by the Georgia Tech Technology and Assessment Center (GT CNS-ASU Group)

search terms. We conducted a survey from February to April 2006 to share our preliminary model and search algorithm with some 45 nanoscientists with various backgrounds. The 19 who provided substantive responses included 13 academics and 6 non-academics, including industry and government experts. The Venn diagram proved especially helpful for eliciting feedback from these nanotechnology scientists and engineers (who were less inclined to wade through search algorithm details). While these respondents largely endorsed our model, they nominated several terms to add and to remove. We evaluated candidate terms by testing and assessing results against the publication and patent data.

Crafting the candidate pilot search entailed many "gray area" choices. The candidate term set started with terms incorporated by other searches—especially Kostoff et al. (2006) and the CREA search. The set was enriched from Alencar et al. (2007); ETC (2003) and Zitt and Bassecoulard (2006) and Bassecoulard et al. (2007). The list was further extended by suggestions of the 19 nanoscientist and engineering respondents. One of the most daunting challenges concerned how to capture bio-nano research without casting too broad a net with respect to basic biology research. Another challenge concerned the extent to which the multitude of microscopy terms (e.g., transmission electron microscopy or TEM) should be included.

Additional questions involved whether to include particular terms. For example, should the term "quantum" be deemed sufficient in and of itself to characterize a publication as nanotechnology or must it be combined with other terms to fulfill this characterization? For many specialized terms we searched in WOS and/or EI Village, checking quick analysis summaries (e.g., on INSPEC keywords in the case of EI Village) to determine the extent to which a given search resulted in high convergence with other nanotechnology oriented terms. This is not a foolproof approach. For instance, terms co-occurring frequently with nano* (that is, nano as prefix to various extensions) include the relevant ("atomic force microscopy") and the very general ("silicon"). We spot-checked small samples of records (e.g., 10 at a time) to assess whether we deemed a high share (at least 70%) to be nano-related. This term assessment provided its own form of bootstrapping, as it surfaced related terms that also required checking for relevance to the nanotechnology domain. For example, this approach uncovered the term "NEXAFS" (near edge X-ray absorption fine structure spectroscopy), but additional research deemed it not overtly nano by itself, that is, we judged that too low a percentage of sample records produced from that term were clearly nano-related.

Two complementary search criteria commonly used for analysis of bibliometric search terms are recall and precision. Recall seeks to minimize the number of truly relevant records missed. Precision seeks to minimize the number of irrelevant records retrieved. As per our declared primary search criterion of giving greater emphasis to being relatively more encompassing, we gave greater weight to recall than precision. For a huge, diffuse domain like "nanotechnology" there is no absolute standard to gauge recall and precision. Simply put, opinions about what should be included vary with the range being quite broad. Were one to stick to a Drexlerian "bottom-up" emphasis (Drexler and Peterson 1991), the amount of nanotechnology research would be reduced by orders of magnitude; the number of nanotechnology articles for 2005 might be closer to 500 than to 50,000 (Table 1). Conversely, were one to decide that "novel properties at nanoscale" (refer to Footnote 1) was the focus, the tallies of publications relevant to that definition might increase by a factor of 10 or 100 in the other direction. Moreover, relying on scientists to distinguish "nano" from "non-nano" is subject to judgmental bias as well, for example, given the current favorability in research funding awards having a nanotechnology orientation. (Khushf 2004, p. 22). That said, our key operational criterion for determining if a particular search phrase should be included was our judgment that at least 70% of the items retrieved belong (guided by the NNI definition, Footnote 1), with selective confirmatory review by nanoscientists.

The limitations of such search term processes—whether endogenous (bootstrap) or exogenous (Boolean term based)—were driven home in discussions with colleagues. Rafols and Meyer (2007) did in-depth case analyses to understand the nature of collaboration in a particular bionano research group. Rafols and Meyer compare one research endeavor that truly integrates two previously independent research streams with another endeavor that draws upon discipline-bounded research knowledge already highly familiar to the scientists. The referencing patterns of both show similar blends of journals based on Institute for Scientific Information (ISI) subject categories. This implies that use of class codes (including the new nanotechnology ISI subject category and nanotechnology patent classes) are inherently imperfect. This same "bionano" research area evidences weaknesses of our Boolean term search as well. None of our "bionano" search terms are apt to capture the full set of relevant molecular structures research, as much of that work will not use those terms, but rather will emphasize more detailed terminology specific to particular narrow study areas, for example, tools and results in manipulating kinesin and myosin molecules. The bootstrap approach has better prospects of capturing such research, but it also suffers from the aforementioned weaknesses. Results from boostrapping gain by engaging a multitude of busy researchers to help discern truly relevant areas from the less relevant. Without that, if one took "all" the research of a given scientist, or "all" the research on myosin, precision would be poor.

The lesson from these examples is that one needs to be clear about the intended uses of the nanotechnology datasets being prepared. This knowledge will inform tradeoffs between broader and tighter focus and sensitivities to errors of precision versus recall. Given that our main uses entail extracting records from our nanotechnology datasets relating to particular themes or organizations, high levels of recall are more important than high degrees of precision. Whatever approach one takes, one should remain vigilant as to the limitations of the approach and its implications.

We explicitly evaluated a substantial list of candidate terms. Some that ultimately were not incorporated into the final search algorithm included spintronic, molecular beam epitaxy, extreme ultraviolet lithography, molecular beacon, molecular sensor, molecular modeling, quantum computing, quantum model, and biochip. These terms generated a mix of seemingly nano-relevant and not so relevant results. It was therefore determined to require these terms to co-occur with other terms for inclusion into the nanotechnology publication database. We applied a relatively inclusive "molecular environment" (MolEnv-I) term-set (second row in Table 2) in conjunction with certain words or phrases (e.g., the self-assembly terms). For other terms, we further constricted the search, requiring co-occurrence with a more restrictive "molecular environment" (MolEnv-R) terms (third row, Table 2); this is the case for one of the "nano-pertinent" term sets (#6 and #7 sets in Table 2). Other terms were searched without such qualifiers (e.g., certain "quantum" phrases; #2 set in Table 2).

In the end, any given term was incorporated into the search based on a comparison of search results phrased in different ways and an assessment of whether the results largely fit within the sense of scope of nanotechnology. A selectivity ratio was constructed to calculate the percentage of publications resulting from searches that intersect the set of phases known as MolEnv (either I or R), and terms beginning with the nano-prefix. The following is an illustration which focuses on the results of publication searches based on key words involving the term microscopy. The selectivity ratio for the full microscopy term set (which includes expressions such as photoelecton*, spectroscop*, X-ray photoelecton*, spectroscop*, auger electon*, spectroscop*, AES, electron energy loss spectroscop*, and tunnel* microcsop*) was 38% when pairing microscopy terms with MolEnv-I, but 42% when pairing microscopy terms with the MolEnv-R. Although these percentages appear close, given that the search represents a large number of publications, this difference suggests that a higher percentage of microscopy-oriented publications would be predicted to be found in the nanotechnology domain—and thus can be deemed "on-target"—when they are paired with MolEnv-R delimiters than when paired with MolEnv-I delimiters.

The resulting modular search algorithm appears as Table 2. The root search is nano*, augmented by seven additional modules ("Quantum" through "Additional items in nano journals"). The Molecular Environment—Inclusive and Molecular Environment—Restrictive term sets (referenced as MolEnv-I and MolEnv-R) are used as modifiers, limiting certain of the modular searches as indicated. Note the critical role of exclusions (Table 3) applied to the data after downloading. We used this two-step approach to make the modular search algorithm more usable in alternative search engines, some of which restrict the length or number of terms in a given search phrase.

Table 2 illustrates results in WOS, based on a search of the ISI Web of Knowledge site with a restriction to the Science Citation Index (SCI) on a

Table 2 Georgia tech modular nano search algorithm: phase 1 database download*

Search	Terms	RESULT:SCI 2005 as of 4/22/06
MolEnv-I (inclusive)	(monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array* or molecul* or polymer* or (co-polymer*) or copolymer* or mater* or biolog* or supramolecul*)	>100,000
Or		
MolEnv-R (more restrictive	(monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array*)	78,390
And		
1. Nano*	nano*	39,101
2. Quantum	(quantum dot* OR quantum well* OR quantum wire*) NOT nano*	3,633
3. Self-Assembly	(((SELF ASSEMBL*) or (SELF ORGANIZ*) or (DIRECTED ASSEMBL*)) AND MolEnv-I) NOT nano*	3,532
4. Terms to include as Nano without other delimiters	((molecul* motor*) or (molecul* ruler*) or (molecul* wir*) or (molecul* devic*) or (molecular engineering) or (molecular electronic*) or (single molecul*) or (fullerene*) or (coulomb blockad*) or (bionano*) or (langmuir-blodgett) or (Coulomb-staircase*) or (PDMS stamp*)) NOT nano*	3,550
5. Microscopy - terms to include but limit to the molecular environment	((TEM or STM or EDX or AFM or HRTEM or SEM or EELS) or (atom* force microscop*) or (tunnel* microscop*) or (scanning probe microscop*) or (transmission electron microscop*) or (scanning electron microscop*) or (energy dispersive X-ray) or (X-ray photoelectron*) or (electron energy loss spectroscop*)) AND MolEnv-I) NOT nano*	11,665
6. Nano-pertinent; Limit to the Molecular Environment - More Inclusively	(pebbles OR NEMS OR Quasicrystal* OR (quasi-crystal*)) AND MolEnv-I) NOT nano*	128
7. Nano-pertinent; limit to the Molecular Environment - More Restrictive	(biosensor* or (sol gel* or solgel*) or dendrimer* or soft lithograph* or molecular simul* or quantum effect* or molecular sieve* or mesoporous material*) AND (MolEnv-R)) NOT nano*	2,104
	1 or 2 or 3 or 4 or 5 or 6 or 7	61,173
8. Additional Items in Nano Journals	fullerene* or ieee transactions on nano* or journal of nano* or nano* or materials science & engineering C - biomimetic and supramolecular systems (in JOURNAL title field) NOT nano*	506
Total	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8	61,479

* Efforts at replication using multiple databases should employ hyphenation, wildcards, categories, and the like with care

particular date.[6] This relatively comprehensive nanosearch facilitates the extraction of more specialized subsets, for example, the subset of records associated with "nanorods."

The next question concerned to which databases these search terms should be applied to best measure the nanotechnology research domain. We evaluated nano research article coverage by 18 databases using DialIndex. The results were not unambiguous. Obvious search terms like "nano*" hit truncation limits, requiring that comparisons of multiple terms be made for multiple time periods. The resulting publication hit rates determined that four databases stood out as most relevant: SCI, INSPEC, EI Compendex, and Chemical Abstracts. SCI drew the highest number of publications initiated by the nano* prefix (nearly 24,000 in 2005). EI Compendex drew the highest number of publications involving microscopy terms (more than 21,000) followed by INSPEC with more than 16,000 in 2005. All four databases drew roughly the same number of publications in 2005 with respect to self-assembly related search terms (ranging from 2,000 to 3,800) and molecular terms (1,600–2,300).

[6] Modifications of this search string for the EI Village databases (INSPEC and Compendex) are available on request.

Table 3 Georgia tech modular nano search algorithm: phase 2 exclusions*

Exclusion terms	
Records containing these terms are removed from "Nano*" dataset	Exclude any nano* records containing only one of these terms and no other nano terms
Plankton*	Nanometer*
n*Plankton	Nanosecond*
m*Plankton	Nanomolar*
b*Plankton	Nanogram*
p*Plankton	Nanoliter*
z*Plankton	Nano-second
NanoFlagel*	Nano-meter
NanoAlga*	Nano-molar
NanoProtist*	Nano-gram
Nanofauna*	Nano-liter
Nano*aryote*	
Nanoheterotroph*	
Nanophtalm*	
Nanomeli*	
Nanophyto*	
Nanobacteri*	
nano2*, nano3*, nanos_, nanog_, nanor_, nanor_, nanoa_, nanoa_, nano-, nanog-, nanoa-, nanor-	

* Terms excluded from Search #1 (Nano*) are deleted from the dataset

Source: Search terms and exclusion terms for nanotechnology, Georgia Tech Technology and Assessment Center (GT CNS-ASU Group), May 2006

Quantum-related search terms attracted the largest number of publications in the Chemical Abstracts database (more than 5,000), followed by INSPEC (nearly 2,500). Because Chemical Abstracts restricts analyses of their records to use of their proprietary software, and with the exception of quantum-related research it does not appear to add a significant number of records (at least in aggregate), our approach was to rely upon SCI, Compendex, and INSPEC to compile large swaths of the global research literature in nanotechnology. Coverage of nanotechnology research in these three databases is not complete for other reasons. The databases decide which sources to include. They favor English language publications, although they do reach well beyond to abstract articles appearing in other language-based journals. As of 2006, SCI covers more than 6,600 journals that deal with physical and life sciences, plus medical and engineering sciences); SCI does not generally cover conferences. INSPEC emphasizes electrical and production engineering, computer and information sciences, and physics via coverage of some 3,500 journals and 1,500 conference proceedings. EI Compendex covers engineering broadly through some 5,000 journals, conferences, and technical reports. We thus have good coverage of NSE, but certainly not every article published. Database coverage overlaps, so consolidation of results from SCI, INSPEC, and EI Compendex is important.

Searching and downloading of nano-related abstract records began in May of 2006. We applied necessary variations of the search algorithm in each of the three databases for the 1990–2006 time period. Downloading was finished in August, 2006. Hence, publications obtained in 2006 represent partial year results. Although space limits discussion of all the nuances, we recognized that different search engines used different parameters and rules to enable access. We determined not to incorporate proximity in our search algorithm to facilitate generalization across search engines. Translating our search algorithm crafted for SCI/WOS into EI Village (and subsequently patent searching) was not straightforward. Among the issues were how to handle hyphenation variations, wildcards, exact phrases, classifications, and the like. To illustrate the sensitivities for readers contemplating performing their own nanotechnology searches, a comparison of search variations of an important nano term—self-assembly—in EI Village (for INSPEC on July 18, 2006) yielded the following results:

- Self-assembly: 11,289 records
- Self-assembl*: 0 records
- Self assembly: 13,093 records
- Self assembl*: 17,376 records

These results are reported based on the selection of "autostemming OFF"; results are the same with "autostemming ON" except that the third term count increases to 17,315. The message is to check parameters and rule alternatives to assess the sensitivities of the search engine.

Concurrently, we explored patent database access. The major patent authorities, especially US Patent and Trade Office (USPTO), European Patent Office

(EPO), and Japanese Patent Office (JPO) provide free web-based access, in English. However the format of access offered by these offices is oriented toward people who are searching for a relatively few patents to view in an indepth manner. We were seeking convenient access to huge numbers of patents to download for further "mining" with software assistance. We tried out Cassis, EI Village patents, FreePatentsOnline, and Community of Science patents, but found inadequacies in each relative to the needs of the project. Eventually we determined that the MicroPatent database was the optimal source for this analysis. However, it was necessary to adapt the nanotechnology search algorithm described above to patent searching, not just because of software front-end limitations but also to take into consideration specific nanotechnology patent classes. The search for international nanopatents covered the USPTO, EPO, JPO, World Intellectual Property Office (WIPO), and patent offices of Germany, Great Britain, and France. To augment these records, the search also included INPADOC records to cover about 70 countries. The INPADOC search excluded the aforementioned patent offices covered in the MicroPatent search. INPADOC does not allow searching of claims and many documents are not translated into English. However, this combined approach allowed for the development of a more globally-indicative patent data set. This represented a significant advance on prior studies that tend to focus on only one PTO (usually USPTO or EPO) or the triad of USPTO, EPO, and JPO. In addition, a patent citation database was developed using the patent numbers from the prior searches to identify US patents cited by those US nanopatents, and US patents citing those nanopatents.

The keyword strategy was adapted for patent searches. The base searches covered titles, abstracts, and claims (where available). These searches were done using the expressions nano*, bionano*, or bio-nano* and several other of our nano search terms, modified as necessary for the MicroPatent search engine. In addition, searches in the nanotechnology patent classification (e.g., IPC-B82 and US Class 977) were conducted. A MicroPatent function was applied to the results of these combined searches that reduces results to just one record per patent family (i.e., a patent family includes variations of the same invention being filed with multiple patent authorities). By early August 2006, these patent data were made available to us for further work and analysis. We subsequently undertook a cleaning process to identify and remove any further duplicates and apply exclusion terms (as discussed in the next section).

Processing the patent data also presented challenges. To get location information on inventors and assignees required a separate search and download of the INPADOC files. These came as full text XML individual records. Due to their size they had to be downloaded in many packets and then re-consolidated based on the extraction of essential information. Basic patent information was available for all patents from PTOs. However, geographic information was not available from all PTOs (for inventors and assignees). The USPTO had relatively complete coverage of geographic information whereas INPADOC did not have this information.

Data exclusions

This section discusses the processes to apply the exclusion terms. Prior to the application of exclusion terms, researchers removed duplicate records from publication and patent databases. The identification number associated with the publications in SCI, Compendex, and INSPEC was used to remove duplicates. In the case of patents, the MicroPatent facility was used to reduce the number of patents to one member per family of patents (i.e., the same patent awarded by multiple patent and trade offices).

Phase 2 of the search term process (Table 3) is very important. In this phase, we *exclude* certain retrieved publication and patent abstract records from each dataset based on the presence or absence of particular terms. There are two types of exclusion terms. The first are terms excluded without condition because they clearly do not involve nanotechnology. Some of these terms refer to water- or land-based organisms (example, nanoplankton, nanofauna) that do not involve the manipulation or engineering of matter. Others refer to chemical formulas (NaNO2, NaNO3) rather than nanotechnology matter. The second type of exclusion terms are designated "conditional" because they are excluded unless they are paired with another nanotechnology search term. One example is nanometer, which is excluded (when it refers to size alone) unless that publication's raw record also contains another nanotechnology search term, such as film. This condition was determined to be important after reviewing a sample of records that were initially excluded based on the appearance of a size-oriented search term, but subsequently found to refer to technology-related research or patenting.

This exclusion phase prompted the removal of several thousand records from our databases. Results were as follows: 18,139 records were excluded from SCI; 6,240 records were excluded from Compendex; and 2,661 records were excluded from INSPEC. Very few patent records were affected by the Phase 2 exclusion process, however.

Initial base analyses

The nano research publication activity trend, based on WOS-SCI, is displayed in Fig. 2. Not counting a few hundred records prior to 1990 that were picked up, we have 406,967 nano R&D publication abstracts from WOS-SCI. For normalization purposes, we obtained total record counts for SCI for full-year activity 2006, which we could use to calculate an approximate ratio to adjust the nano tally for the part-year 2006 data, although this normalization is not presented in Fig. 2. Our nanotechnology publication set comprises 2.7% of the total WOS-SCI hits over the 1990–2006 time period and 4.1% for the 2005–2006 time period. Our international nanopatent file currently contains 53,720 patent abstracts**.** Figure 3 shows a trend chart for nanopatents. Initial inspection suggests three acceleration points for nano patents: 1998, 2001–2002 and 2005.

We have compared results from the search strategy described above and two alternative search strategies. The search strategy described by Zitt and Bassecoulard (Zitt and Bassecoulard 2006; Bassecoulard et al. 2007) combines lexical queries, as used in this research, with citation coupling. Based on communications with the authors, the total counts of publication records in SCI resulting from the Zitt-Bassecoulard and our search strategies for the period of 1999–2003 were observed to be within range of one another: 168,200 for the Zitt-Bassecoulard database compared with nearly 158,000 for our searches. Top keywords in the Zitt-Bassecoulard search were compared with our database using the percentage distribution of publications by keyword. Somewhat higher percentages of publications were present in our database across many keywords in the comparison. Keywords involving film, microscopy, and semiconductors were particularly associated with higher publication shares in our nanotechnology publication database than in the Zitt-Bassecoulard search. In contrast, there were slightly more publications involving terms such as atomic force and scanning tunneling in the Zitt-Bassecoulard search. However, most of the keywords had relatively

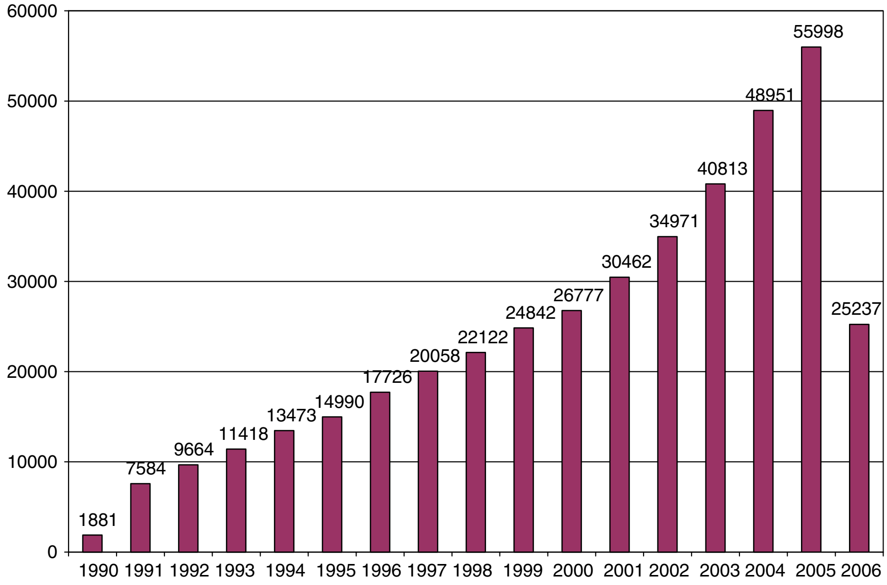

Fig. 2 Nano research publications, 1990–2006 (August), from web of science—science citation index. *Source*: Web of Science, using Georgia Tech nano search definition (publications). Data for 2006 is for part year

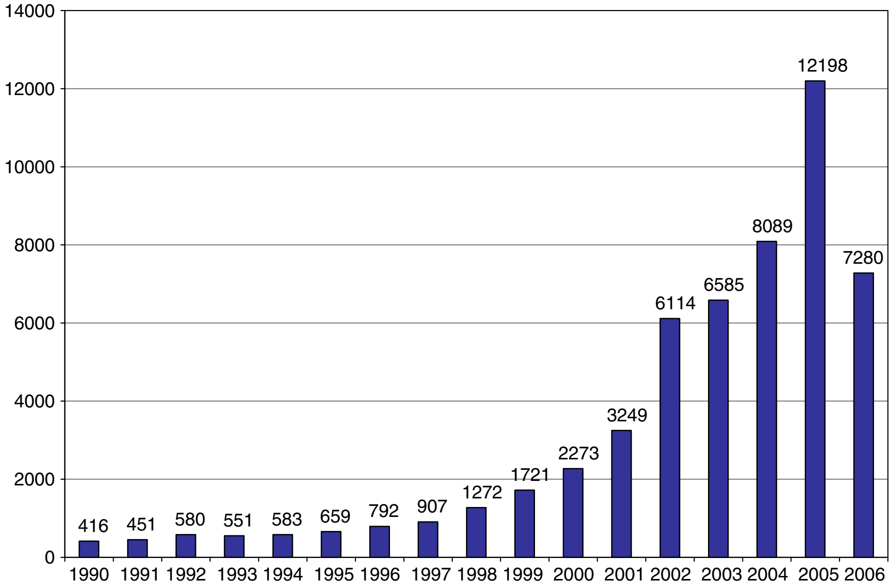

Fig. 3 Nanopatents, 1990–2006 (August), from MicroPatent and INPADOC. *Source*: Micropatent and INPADOC, using Georgia Tech nano search definition (patents). Data for 2006 is for part year

comparable percentages of publications across the two databases.

Assessments were also made of the similarities and differences with our database and that of the CREA project. Several of the 19 nanoscientists who commented on our search strategy expressed concern about definition of bio-nano research. They suggested that some of the bio-oriented keywords used in the CREA definition of nanotechnology largely fell outside the boundaries of nanotechnology because they included too much basic biology or non-nano biotech. In response, our search strategy made a particular effort to exclude terms such as DNA and RNA unless they appeared with a core nanotechnology keyword, such as nanoarray or self-assembly. Table 4 compares the numbers and percentages of publications resulting from the CREA and our search strategy with respect to the most common bio-oriented keywords in these databases over the 1999–2003 time period. This time period was chosen because it comprised the most recent set of complete years in the CREA search of SCI. The results show that the percentages of publications in the CREA project are higher throughout all of the most common bio-oriented keywords in these databases. This finding suggests that the search strategy described in this paper achieved the goal of narrowing the degree to which basic biological research is manifested in the nanotechnology publication profile.

Following the completion of our search and database retrieval, Kostoff et al. (2007) have updated their nanotechnology searching and analyses. They profile many interesting dimensions, breaking out highly cited papers and clustering topical emphases to generate a multi-tier tree. It is interesting to compare search results since Kostoff's earlier search formulation served as a basis for our own. The comparison suggests that the overall nano publication trend shows a very similar trajectory to our Fig. 2 in that the numbers of articles double from 2000 to 2005. Country trends are quite aligned as well; for example, publications from the US lead both databases with approximately 15,000 articles in 2005, followed by sharply rising publications from China at approximately 12,000 articles, followed by Japan and Germany. Our search of WOS was restricted to SCI, while Kostoff et al. searched the Social Science Citation Index as well. Thus, our approach retrieves approximately 56,000 articles for 2005 while their approach yields 65,000 for the same year.

Figure 4 compares the two sources based on selected topical areas, authors and source journals. The main profile elements correspond well, but there also are notable second-tier differences. For instance, in coverage of the first topic—nanocomposite*—Japan and South Korea are in reverse order. Likewise, while both approaches produce the same listing of top

Table 4 Comparison of bio-oriented search terms in CREA and georgia tech (CNS-ASU) definitions

Most Common Bio-oriented Keywords in FhG ISI Nanotechnology Database*	CREA/FhG ISI*		GT (CNS-ASU)**	
	Number of Nanotechnology Publications	% of Total Nanotechnology Publications	Number of Nanotechnology Publications	% of Total Nanotechnology Publications
DNA	5,853	7.8%	4,103	2.6%
Protein	9,928	13.3%	8,232	5.2%
Oligonucleotide	1,208	1.6%	694	0.4%
Biosensor	4,470	6.0%	2,093	1.3%
Encapsulation	1,036	1.4%	1,066	0.7%
Gene delivery	742	1.0%	256	0.2%
Tissue engineering	289	0.4%	319	0.2%
Gene therapy	1,036	1.4%	200	0.1%
Drug targeting	103	0.1%	35	0.0%
Drug delivery	1,234	1.6%	705	0.4%
Immobilized	689	0.9%	365	0.2%
Biocompatibility	340	0.5%	504	0.3%
Bloodcompatibility	95	0.1%	110	0.1%
Cell seeding	3	0.0%	4	0.0%
Tissue repair	43	0.1%	19	0.0%
Cell therapy	20	0.0%	13	0.0%
Cell adhesion	677	0.9%	375	0.2%
Biochip	120	0.2%	152	0.1%
Extracellular matrix	281	0.4%	489	0.3%
Immunosensor	347	0.5%	220	0.1%
Total number of publications, 1999–2003	74,806		157,865	

*CREA project analysis of nanotechnology publications from Science Citation Index, 1999–2003, based on nano definition of Fraunhofer Institute for Systems and Innovations Research (FhG-ISI) (2002)

**GT (CNS-ASU) = Georgia Tech Program in Science, Technology and Innovation Policy and the Center for Nanotechnology and Society (CNS-ASU). Analysis of Nanotechnology Publications from Science Citation Index, 1999–2003 based on Georgia Tech nanotechnology definition

journals, the ordering of these journals switches places.

Reflections and analytical directions

Defining research and commercialization domains using publication and patent databases for analysis of emergent technology has been undertaken for over a decade (Porter and Cunningham 2005). Still nanotechnology presents special challenges. First the search nuances are multidimensional. Nanotechnology is extremely cross-disciplinary and its boundaries are ill-defined. Emerging science and technology fields typically take time to consolidate their identity and terminology. This is particularly true of nanotechnology given its breadth and degree of flux. Tracking the stabilization of terms could prove an interesting indicator in its own right.

Second, the scale of research related to nanotechnology is vast. As mentioned, our nanotechnology search collects 4.1% of all research in the Science Citation Index for 2005 and partial year 2006. This poses special challenges in data downloading and processing. Abstract research publication and patent records lend themselves to these processes because of their field-structuring and metadata characteristics, but the very size of these files stress available desktop computing capabilities.

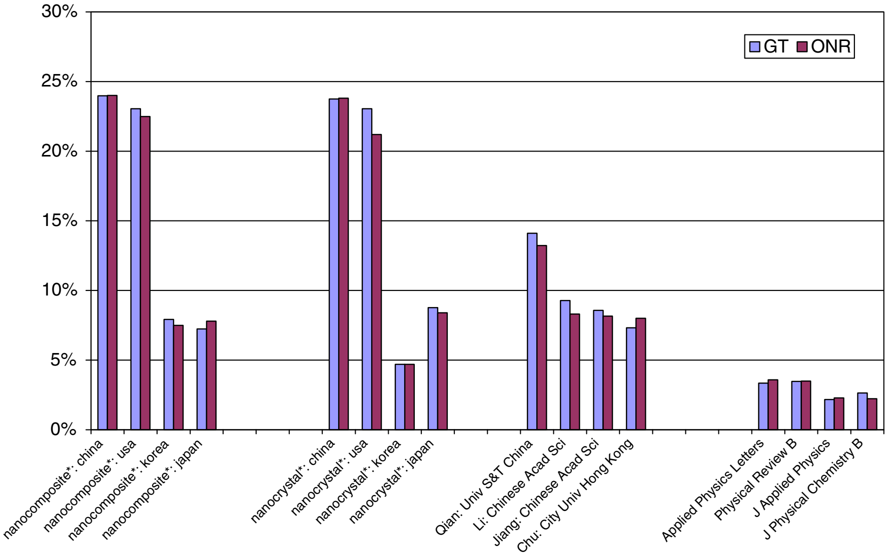

Fig. 4 Comparison of Selected Search Results with Kostoff et al. (2007)*. *GT = Georgia Tech or our search strategy; ONR = Office of Naval Research or the search strategy described in Kostoff et al. (2007). For the first two sets of topical comparisons, the results represent the percentage of articles on that topic with authors in one of these leading countries. For the third set—leading authors associated with a particular organization—percentage of nano articles in the set in 2005 is actually multiplied by 100 for scaling purposes. Our topical tallies are based on searching the key terms (i.e., title NLP phrases, author keywords and keywords plus). We also compared searches based on the full raw record in our database. Some differences arise. For example, for the term nanocomposites, a search of the entire record results in the US edging ahead of China (740–733 articles). We employ our comparison using the tallies in Kostoff et al. (2007) for the terms "nanocomposite*" and "nanocrystal*" taken from Figs. 3B and 3C which provide sufficient accuracy for these purposes

The challenge of developing a database to help understand NSE research is to balance the need for rigor with the need for research findings. Emergent technologies such as nanotechnology can benefit from timely information for decision making. At the same time, one does not wish to rely on knowledge bases that are so quickly drawn that they do not take into consideration important boundary elements that will eventually delineate the field. Development of a base of knowledge about research and commercialization in emerging fields such as nanotechnology requires a measure of experimentation and craft to address both the need for rigor and timely information.

We are currently analyzing our nanodata to examine and model nano research and innovation trajectories, the emergence of nano as a general purpose technology, regional nanodistricts and clusters, and mid-term nano applications. Additionally, we are collaborating with colleagues at CNS-ASU, NCSU, and elsewhere on special topics and studies. The more we draw on these data, the more we will become effective and efficient in extracting intelligence from them, in cleaning the data, and in linking them with other data sources, quantitative and qualitative. We believe that analyses of this information will raise important issues about nanotechnology development. We will be probing further to extract intelligence on the leading research centers and emerging topical thrusts. Building on those results, we intend to explore future nanotechnology developmental pathway prospects. We will assess those findings to gauge potential socio-economic impacts. Taken together, these analyses can offer a unique, evidence-driven vantage point to illuminate useful R&D interventions and to probe emerging policy questions.

Acknowledgments Significant research assistance in search definition and database development was provided by Li Tang, Sharyn Finney, Pratik Mehta and Luke McCloud. Rich Kolar developed and generated the patent searches and databases. This research was undertaken at Georgia Tech with support by the Center for Nanotechnology in Society at Arizona State University (CNS-ASU), funded by the National Science Foundation (Award No. 0531194), and support by the

National Partnership for Managing Upstream Innovation: The Case of Nanoscience and Technology (North Carolina State University; Award No. EEC-0438684). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alencar MSM, Porter AL, Antunes AMS (2007) Nanopatenting patterns in relation to product life cycle. Technol Forecast Soc Change, forthcoming

Bassecoulard E, Lelu A, Zitt M (2007) Mapping nanosciences by citation flows: a preliminary analysis. Scientometrics 70(3):859–880

Drexler E, Peterson C (1991) Unbounding the future: the nanotechnology revolution. William Morrow and Company, New York

ETC Group (2003) From genomes to atoms: the big down. The etc Group, Winnipeg, Canada

Fraunhofer Institute for Systems, Innovations Research (2002) Search methodology for mapping nanotechnology patents. Karlsruhe, Germany

Heinze T, Shapira P, Senker J, Kuhlmann S (2007) Identifying creative research accomplishments: Methodology and results for nanotechnology and human genetics. Scientometrics 70(1):125–152

Huang Z, Chen H, Yip A, Ng G, Guo F, Chen ZK, Roco MC (2003) Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field. J Nanoparticle Res 5(3-4):333–363

Huang Z, Chen C, Chen A-K, Roco MC (2004) International nanotechnology development in 2003: country, institution, and technology field analysis based on USPTO patent database. J Nanoparticle Res 6(4):325–354

Guston DH, Sarewitz D (2002) Real-time technology assessment. Technol Soc 24:93–109

Khushf G (2004) A hierarchical architecture for nano-scale science and technology: taking stock of the claims about science made by advocates of NBIC convergence. In: Baird D, Nordmann A, Schummer J (eds) Discovering the nanoscale. IOS Press, Amsterdam

Kostoff RN, Koytcheff R, Lau CGY (2007) Structure of the global nanoscience and nanotechnology research literature. Available at http://www.onr.navy.mil/sci_tech/33/332/techno_watch_publications_textmine.asp. Cited 7 June 2007

Kostoff RN, Murday JS, Lau CGY, Tolles WM (2006) The seminal literature of nanotechnology research. J Nanoparticle Res 8(2):193–213

Kostoff RN, Murday JS, Lau CGY, Tolles WM (2006a) The seminal literature of nanotechnology research. J Nanoparticle Res 8(2):193–213

Kostoff RN, Stump JA, Johnson D, Murday JS, Lau CGY, Tolles WM (2006b) The structure and infrastructure of the global nanotechnology literature. J Nanoparticle Res 8(3–4):301–321

PCAST (2005) The National Nanotechnology Initiative at 5 years. Washington, DC: President's Council of Advisors on Science and Technology, Executive Office of the President

Porter AL, Cunningham SW (2005) Tech mining: exploiting new technologies for competitive advantage. Wiley, New York

Rafols I, Meyer M (2007) Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. Proceedings of the 11th international conference of the international society for scientometrics and informetrics, Madrid, June, 2007. Available at http://www.sussex.ac.uk/spru/irafols. Cited 7 June 2007

Zitt M, Bassecoulard E (2006) Delineating complex scientific fields by a hybrid lexical-citation method: an application to nanosciences. Inform Processing Management 42(6):1513–1531