



Refining Search Terms for Nanotechnology

Briefing Paper:
Refining Search Terms for
Nanotechnology

Alan Porter
Georgia Institute of Technology

Jan Youtie
Georgia Institute of Technology

and

Philip Shapira
Georgia Institute of Technology

Center for Nanotechnology in Society
Arizona State University

2006

This research was conducted as part of the Center for Nanotechnology in Society, Arizona State University (CNS-ASU). CNS-ASU research, education, and outreach activities are supported by the National Science Foundation under cooperative agreement #0531194.



Refining Search Terms for Nanotechnology

Briefing Paper

Alan Porter^a
Jan Youtie^b
Philip Shapira^c
David J. Schoeneck^d

Center for Nanotechnology in Society (CNS-ASU)
Program in Nanotechnology Research and Innovation Systems Analysis

Technology Policy and Assessment Center
School of Public Policy, Georgia Institute of Technology
Atlanta, GA 30332-0345, USA

August 2006

Prepared for Presentation at National Science Foundation, First Year Review of the Center for Nanotechnology in Society (CNS-ASU), Arizona State University, Tempe, AZ, August 24, 2006.

Author Information

- a. Professor Emeritus of Public Policy, Georgia Institute of Technology, Atlanta, USA
 - b. Senior Researcher, Georgia Tech Enterprise Innovation Institute
 - c. Professor of Public Policy, Georgia Institute of Technology, Atlanta, USA; Program Leader, Center for Nanotechnology in Society (CNS-ASU), Program in Nanotechnology Research and Innovation Systems Analysis.
 - d. Research Associate, Search Technology, Inc., Norcross, GA, USA
- * Email address for correspondence: pshapira@gatech.edu (Philip Shapira)

Acknowledgments

Significant research assistance in search definition and database development was provided by David J. Schoeneck, Li Tang, Sharyn Finney, and Luke McCloud. This research was undertaken at Georgia Tech through the Center for Nanotechnology in Society (CNS-ASU), supported 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; NSF award # 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.

Refining Search Terms for Nanotechnology

1. Introduction

The Center for Nanotechnology in Society (CNS-ASU) team at Georgia Tech is undertaking assessments of the evolving nanotechnology research and innovation system in the US and internationally. A fundamental building block of this work involves refining *in specific bibliometric terms* our definition of nanotechnology.

We know that nanotechnology is the manipulation of molecular-sized materials to create new products and processes.¹ It encompasses contributions from fields 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, we need to define in greater detail the multiple sub-fields within the nanotechnology domain. This will enable us 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 make available some comparisons with other nanotechnology search definitions, discuss our approach to downloading and cleaning data, and report initial results.² The paper concludes with reflections on our search process and planned and potential applications of the databases we are developing.

2. Overview of Nano Search Definition Method and Process

Our approach to developing a refined nanotechnology bibliometric search definition involves several steps. First, we developed 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 confirm terms; third, we further evaluated candidate terms by testing and assessing results against the publication and patent data.

One of the prior efforts that we built upon was work undertaken by members of Georgia Tech’s Technology Policy and Assessment Center (TPAC) as part of the project on Creative Capabilities

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 to 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).

² As of August 2006, we are still finalizing the cleaning and validation of our large-scale data sets on research publications and patents. Results reported in this paper are thus subject to revision.

and the Promotion of Highly Innovative Research in Europe and the United States (CREA).³ This project is analyzing creative research in the domains of nanotechnology and human genetics. As part of the CREA project, TPAC downloaded in 2005 over 100,000 Web of Science (WOS) records and over 10,000 US patents relating to nanoscience and engineering (NSE). The bibliometric definition use 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 1994 and 2005.⁵

We have also leveraged National Science Foundation support of a North Carolina State University (NCSU) Partnership for Innovation project on nanotechnology on which we are a sub-contractor.⁶ That project seeks to foster knowledge transfer from non-industrial research to promote industry innovation. With guidance especially from Angus Kingon (a nanoscientist at NCSU), we explored various nano search algorithms.

We also gained perspective from a variety of nano profiling efforts, including: Kostoff et al. (2005), a Brazilian nanopatent search (Alencar et al., under submission), a broad nano perspective (ETC, 2003), and an infometrics treatment of nano (Zitt and Bassecoulard, in press). The Huang et al. (2003, 2004) papers also provided insights on nano-trends.

As we considered these analyses and the insights they offered, we concluded that they provided a base upon which to build but also that improvements could be made. For example, the definitions varied considerably in how they treated the interface between biotechnology and nanotechnology and the extent to which they captured research in other nano sub-fields. We conducted a number of search comparisons. To begin, we compared the behavior of our 2005 CREA search and that of Kostoff et al. (2005). Table 1 gives a sense of how diffuse the nano domain is, and how challenging it becomes to generate the “best” search strategy. Here are two search algorithms that yield comparable numbers of nano hits (45,000 for one year of WOS), yet almost 30% of each search is unique from the other. Detailed comparisons suggested strengths and weaknesses of particular term phrasing. Additionally, given that some definitions were developed several years ago, we wondered if emerging current sub-topics were adequately captured. In early 2006, we proceeded to extend this prior work by developing a refined nano research publication and patent data search definition. We established three key criteria for the search terms, namely that they

³ The team leader of the CREA project is the Fraunhofer Institute for Systems and Innovations Research (ISI), Germany; the team partners are the Georgia Tech Technology Policy and Assessment Center and SPRU at Sussex University, UK. CREA is sponsored by the European Union’s NEST Programme.

⁴ *VantagePoint* development was initiated at TPAC in 1993 to help exploit science & technology information resources. Search Technology, Inc. and Georgia Tech continued development supported mainly by DARPA (Defense Advanced Research Project Agency), the U.S. Army, and ONR (Office of Naval Research). *TechOASIS* is the version of the software available for U.S. Government use and *Thomson Data Analyzer* is another commercial version available internationally. The software is especially effective at mining field-structured data (e.g., abstract records), for which it facilitates data cleaning and consolidation, text and numerical analyses, and generation of information products. Many auxiliary enhancements (especially scripts) key on analyzing S&T and patent database search results.

⁵ The TPAC team involved in the CREA nano search was led by Philip Shapira, with research assistance provided by Ajay Bhaskarabhatla and Li Tang.

⁶ The TPAC team involved in the NCSU project was led by Alan Porter, with research assistance provided by David Schoeneck.

should be: (1) encompassing – as inclusive as sensible; (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 this search to adjust the record set to meet differing research interests, and as the field of nanotechnology evolves.

There are quite different search strategies possible. We considered an iterative, expansive, “bootstrap” search method (being tried by the UCLA Nanobank, Duke, PRIME, and others).⁷ While the specific methods of these researchers differ, broadly they take a core set of nano papers, then extend. Extensions can involve examination of other papers by nano-authors that may not use “nano-terms” per se in titles, keywords, or abstracts. Another mode of outreach is to consider papers referenced by, or referencing, the core nano set. The rationale is that these reflect research knowledge transfer with nano-research, hence, are apt to be highly salient. In some cases, review by nano authors or experts is used to fine-tune the expanded sets, i.e., to discard less related work and add additional relevant pieces. This has the advantage of not being limited to use of classification codes (indexes), keywords, or prominent terms (in titles, abstracts, patent claims, etc.). But, these approaches are costly and very time-consuming.

Consistent with CNS-ASU’s focal theme of “real-time technology assessment”, we decided to use instead a 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 can track emergence of new terms over time and adjust the search algorithms dynamically in updating our nano datasets.

3. The Georgia Tech Nano Search

From our comparisons of the CREA and Kostoff search results (summarized in Table 1), supplemented by insights gained from the other search definitions, we developed a pilot “field scope” definition. This included a schematic Venn diagram with four overlapping fields - metrology and nanoproducts; nanostructure chemistry and materials; nanodevices and nanoelectronics; and nano-medicine and nano-biotechnology (current version appears as Figure 1). Within each field, examples of key terms were included. We also developed a more detailed search algorithm comprised of 8 major sections and a series of search terms. Between February and April 2006, we shared our preliminary model and search algorithm with some 75 nanoscientists, with various backgrounds. The 19 who responded substantively included 13 academics and 6 non-academics. While they largely endorsed our model, they nominated many additional terms and also terms to remove. We evaluated candidate terms by testing and assessing results against the publication and patent data. Various comparisons support our assessment that our final nano search algorithm is quite robust.

Crafting our candidate pilot search entailed many “gray area” choices. Our candidate term set started with terms incorporated by other searches -- especially Kostoff et al. (2005) and our CREA

⁷ Prior to determining our search strategy, we consulted with others in the nanotechnology research community. On December 9, 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 contact with Duke University (Giannella) 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.

search; enriched from Alencar et al., under submission; ETC, 2003; and Zitt and Bassecoulard, in press. The list was further extended by suggestions of our 19 nanoscientist reviewers. Tough choices concerned how to capture bio-nano research without casting too broad a bio-net and whether to include the multitude of microscopy terms (e.g., TEM – transmission electron microscopy).

We had to decide whether to include particular terms (e.g., should “quantum” mark a record as nano?). For many specialized terms we searched in Web of Science and/or EI Village. We checked quick analysis summaries (e.g., on INSPEC keywords) to see the extent to which a given search resulted in high convergence with other nano-terms. This is not foolproof – for instance, terms co-occurring frequently with nano* (i.e., 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 a high fraction (70% or so) seemed to be nano-related. Our term assessment provided its own form of bootstrapping, as we came upon related terms to check for nano-relevance (e.g., we assessed NEXAFS – near edge x-ray absorption fine structure spectroscopy, but deemed it not overtly nano).

We explicitly evaluated a substantial list of candidate terms. Others that did not get included in our search algorithm are: spintronic, molecular beam epitaxy, extreme ultraviolet lithography, molecular beacon, molecular sensor, molecular modeling, quantum computing, quantum model, and biochip. Many of these terms appeared to generate a mix of nano-relevant and not so relevant results. We therefore determined to require those terms to co-occur with other terms for inclusion. We applied a relatively inclusive “molecular environment” term-set (top row in Table 2) in conjunction with certain terms (e.g., the “self-assembly” terms). For other terms, we further constricted the search, requiring co-occurrence with more restrictive “molecular environment” terms (second row in Table 2); this is the case for one of the “nano-pertinent” term sets (later in Table 2). Other terms were searched without such qualifiers (e.g., certain “quantum” phrases; see Table 2). We decided how to treat a given term by comparing search results alternative ways and assessing whether they largely fit within our sense of scope for nano.⁸

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 are used as modifiers, limiting certain of the modular searches as indicated. Note the critical role of exclusions (Table 2 – Phase 2) applied to the data after downloading. We did that to make this modular search algorithm more usable in alternative search engines, some of which restrict the length or number of terms in a given search phrase.

⁸ We constructed a “**selectivity ratio**” – what % of the term intersections with MolEnv (either I or R) also are in nano*? An illustration -- surprisingly the MolEnv-I or -R do not improve the hit rate for our microscopy terms much; but MolEnv-R does better (at the expense of losing half of the MolEnv-I 10514 hits); so if nano* co-occurrence is a good index, we would predict that microscopy search using MolEnvI would get us 4000 on-target records along with 6500 off-target. [Note that we are doing these searches because we believe many of those “off-target” are nano-related without using a nano* term.] Using MolEnv-R would get us 2230 on-target with 3100 off-target. Raw microscopy would get us 9100 on-target and 18500 off-target. [We chose to use the MolEnv-I as delimiter with the microscopy terms. This means that some users might well ask us to additionally exclude records that just have these tools mentioned without additional nano terminology.]

Table 2 illustrates results in Web of Science (“WOS” – here we actually search the ISI Web of Knowledge site, restricting to the Science Citation Index – “SCI”) on a particular date. Available on request are the search modifications for EI Village (for INSPEC and Compendex databases). This relatively comprehensive nano-search allows us to extract more specialized subsets whenever so desired within *VantagePoint* – e.g., to pull out records associated with “nanorods.”

We evaluated nano research article coverage by 18 databases using DialIndex.⁹ We determined that four databases stood out as richest: WOS, INSPEC, EI Compendex, and Chem Abstracts. Figure 2 compares their hit rates for selected terms for 2005. We had full search, download, and analytical access to the first three, so determined to use these. (ChemAbstracts restricts analyses of their records to use of their software or MS Excel, and we do not further use in our analysis.) Our approach is to rely upon these three leading science and engineering databases to compile large swaths of the global research literature, from which we then search and download nano-related abstract records. Coverage of nano research is not “100%.” The databases decide which sources to include. They favor English language publications, but reach well beyond as well. Web of Knowledge covers articles published in some 22,000 journals (we are presently focusing on the SCI subset that covers 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.¹⁰

In May, 2006, search and download began. We investigated the possibility of using the CREA WOS records, but these lacked cited reference information. Likewise, we determined not to try to salvage the patents from Community of Science, but rather to do a complete, new search. We applied necessary variations of the search algorithm in each of the three databases for 1990-2006 (searching from May into August).¹¹ Download was finished in August, 2006. Note that 2006 results are partial.

⁹ Comparison is not unambiguous. Obvious search terms like “nano*” hit truncation limits. We compared multiple terms for multiple time periods.

¹⁰ An additional approach is to consider internet nano coverage. This is a fascinating resource that we believe complements and enriches the solid base built from database searching. Our close colleague, Scott Cunningham and his associate, Telli van der Lei, describe an intriguing initial analysis of nano web sites and their linkages (2006).

¹¹ Not to attempt to treat all the nuances, but different search engines (i.e., the way we access different databases) use different parameters and rules. We determined not to incorporate proximity in our search algorithm to facilitate generalization across search engines. Translating our search algorithm crafted for WOS to EI Village and the patent searching (Micropatent) was not straightforward. Among the issues are how to handle hyphenation variations, wildcards, exact phrases, classifications, etc. Just to illustrate the sensitivities for readers contemplating performing their own nano searches, here’s a comparison of search variations of an important nano term – self-assembly – in EI Village (for INSPEC on July 18, 2006):

| | |
|-------|---------------|
| 11289 | Self-assembly |
| 0 | Self-assembl* |
| 13093 | Self assembly |
| 17376 | Self assembl* |

These are with “autostemming OFF”; results are the same with “autostemming ON” except that the third term count increases to 17315. The message is to check exactly what you are getting.

Concurrently, during Spring 2006, we explored **patent** database access.¹² We tried out Cassis, EI Village patents, FreePatentsOnline, and Community of Science patents, but found each wanting to meet our needs. Rich Kolar at EKMS, a veteran patent analyst, worked with us to adapt our nano search algorithm to patent searching (e.g., to also search on the nano patent classes). We worked with Kolar to complete a search with the following features:

1. The search for international nanopatents covered the USPTO (U.S.), EPO (European Patent Office), JPO (Japan), WIPO (World Intellectual Property Office), and patent offices of Germany, Great Britain, and France (using Micropatent). To augment these records, the searched also included INPADOC to cover about 70 countries. The INPADOC search excluded the patent authorities covered in the Micropatent search. The INPADOC document does not allow searching of claims and many documents are not translated into English. However, this combined approach does allow us to develop a globally-indicative patent data set – this is a significant advance on prior studies which tend to focus on only one PTO (usually USPTO or EPO) or the triad of USPTO, EPO, and JPO.
2. The core searches were downloaded in tab-delimited format. To retrieve address and location information, as available for inventors and assignees, separate full-text or nearly full-text records, as available, were downloaded in XML format.
3. Additionally, a search of the EKMS patent citation database was completed using the patent numbers from the prior searches to identify a) US patents cited by those US nanopatents, and b) US patents citing those nanopatents.¹³

The keyword strategy was adapted for patent searches. The base searches covered titles, abstracts, and claims (where available). These were done on nano*, bionano*, or bio-nano* and other of our nano search terms, modified as necessary for the MicroPatent search engine.¹⁴ In addition, we searched in the nano codes (IPC-B82 and US Class 977). On the results of these combined searches, a MicroPatent function was applied that reduces results to just one record per patent family (i.e., essentially the same invention being filed with multiple patent authorities). By early August 2006, these patent data was 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).

¹² The major patent authorities (especially USPTO, EPO, JPO) provide free web-based access, in English. However this access is oriented for people who are searching for a relatively few patents, probably to read. We are seeking convenient access to huge numbers of patents to download for further “mining” with software assistance. Hence, we need patent database access to obtain these.

¹³ The data was split into two files. In these files GEN-0 are the US patents from the nano search and GEN-1 are the US patents cited by the nano patents. Only for “Gen-0” (our nanopatent set) patents that have multiple assignees, the Excel output is such that if this patent has multiple citations to, or citations by, the additional rows showing those will only note the Gen-0 first assignee. So, in analyzing networking, we need to tag those Gen-0 patents with multiple assignees, and for them, add Gen-0 co-assignee information to each citation.

¹⁴ The modified patent search term = (“quantum dot” or “quantum dots” or “quantum wire” or “quantum wires” or “self assembly” or “self-assembly” or “self assembled” or “self-assembled” or fullerene* or “PDMS stamp” or quasicrystal* or quasi-crystal* or “soft lithography” or “soft lithographic” or “mesoporous material” or “mesoporous materials” or dendrimer* or “coulomb blockade” or “langmuir-blodgett” or “molecular motor” or “molecular motors” or “molecular ruler” or “molecular rulers” or “molecular wire” or “molecular wires” or “molecular wired” or “molecular wiring” or “molecular device” or “molecular devices” or NEMS)

4. Data Cleaning

This section discusses the steps taken in cleaning the results from publication and patent data searches.

Nano-publications

We initiated data cleaning in July, 2006. This is a big task, entailing:

- Removal of duplicate records from the several searches
- Applying the exclusion terms (see Table 2 – Phase 2).
- Cleaning the author information (e.g., consolidating variations on the same names from multiple sources)
- Cleaning the affiliation and organizational address information
- Consolidating records across databases

We have had to innovate our processing. For starters, the SCI search yielded some 500,000 records before duplicate removal. These are big abstract records because we gathered cited reference information along with many base abstract record fields. We increased RAM on several desktop computers and purchased a new faster PC to facilitate data manipulations.¹⁵ One effective innovation entailed:

- Writing new “lite” import filters for *VantagePoint* to just bring in selected key identification fields for each record
- Process these lite files to identify duplicate records and perform exclusion operations and to devise thesauri for author/inventor and affiliation cleaning.
- Import the full records and use the thesauri and removal lists to cull and clean the “real” file.

At this time, the WOS-SCI records have all been imported into *VantagePoint*. Actually, three formats were applied:

1. “Lite” records – minimal raw record fields brought in
2. “Enhanced Lite” – many of the small fields in these raw records
3. Full records – the entire abstract records as downloaded from WOS; in particular, these include Cited References

The Lite records have been consolidated into a single file (all years – 1990-2006). Duplicates have been removed and the exclusions applied.

We completed searching EI Village in early August for nano R&D publications. We also separately searched and downloaded from the two databases: INSPEC and EI Compendex.

Remaining on our agenda are to:

- Further clean the authors (consolidate name variations), first within database, then across;

¹⁵ Even with a new desktop PC, we still find that data processing takes many hours per run. Even more powerful computing capabilities on which we can run *Vantage Point* would be desirable.

- Further clean author affiliations (consolidate organization name variations; develop “type thesauri to distinguish academic, industrial, and government & non-governmental organization labs).¹⁶
- Associate geography to records, in particular metropolitan scale agglomerations.
- Combine the WOS-SCI, INSPEC, and EI Compendex records (while retaining the separate files as well for certain analyses)

Nanopatents

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 zipped. These had to be consolidated, with special “lite” filters devised to extract the location and other essential information. This then had to be fused into the 60,000 or so regular MicroPatent search records. These files also presented processing challenges as these records are large.

All of the tab-delimited MicroPatent “enhanced lite” records were imported into VantagePoint. We needed several fields to perform the exclusions (e.g., to remove NaNO₂ and plankton related records). The INPADOC records do not contain Claims information, which the other MicroPatent records have (for now, helpful in performing exclusions). The 9,000 or so INPADOC records were imported as full records and merged with the MicroPatent “enhanced lite” records.

Basic patent info was available for all patents from PTOs. However, geographic information was not available from all PTOs (for inventors and assignees), so we have built this in another XML file.

The resulting main nanopatents record file is in VantagePoint. Duplicates have been removed (Including those common to INPADOC and MicroPatent). The exclusion terms (Table 2 – Phase 2) have been applied. The exclusion terms were applied not only to patent titles, but also to the full raw records.

Remaining on our agenda are to:

- Consolidate in the location information from the XML records
- Further cleaning of author and inventor information (e.g., consolidating variations on the same names from multiple sources)
- Further cleaning of affiliation and assignee organization information, including geographical identifiers.
- Process the patent citation information

¹⁶ Developing effective organizational thesauri is time consuming, in that we are cleaning on a global scale with thousands of institutional and corporate names around the world. We have significant “local” knowledge in our team, including researchers knowledgeable about the US, Europe, China, India, and Japan; nonetheless significant checking is required, generally using web-based sources. The advantage of the thesauri approach is that once the variations on an organizational affiliation are verified, the thesaurus can be reapplied to multiple analyses, subject to the need for updating from period to period.

5. Initial Base Analyses

The nano research publication activity trend, based on WOS-SCI, is displayed in Figure 3. Not counting a few hundred records prior to 1990 that were picked up, we have **420,775 nano R&D publication** abstracts from WOS-SCI. For normalization purposes, we obtained total record counts for SCI. Based on recent full-year activity, we can use an approximate ratio of 1,300,000 to 600,000 to adjust the nano tally for the part-year 2006 (not yet done in Figure 3). Interestingly, our nano set is 2.7% of the total WOS-SCI hits over this time period; 4.1%, for 2005-06 to date. The database will grow larger when we incorporate non-duplicated INSPEC and EI Compendex publication records.

Our international nanopatent file currently contains **61,174 patent abstracts**. Figure 4 shows a trend chart for nanopatents. We have to do further analysis before any results can be substantiated. However, initial inspection suggests three “acceleration” points for nano patents: 1998, 2001 and 2005.

6. Reflections and Analytical Directions

TPAC researchers have undertaken “tech mining” (Porter and Cunningham, 2005) for over a decade, but “nano” presents truly special challenges. First the search nuances are multi-dimensional. Nano is extremely cross-disciplinary and its boundaries are ill-defined. Emerging science and technology fields typically take time to consolidate their identity and terminology. Nano is extreme due to 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 nano is astounding. As mentioned, our nano search collects 4.1% of all research in the Science Citation Index recently. This poses special challenges in data downloading and processing. Abstract research publication and patent records lend themselves to this (field-structured, “meta” data), but file size stresses available desktop computing. We learned to create “lite” record versions to facilitate processing.

After much hard work, we are poised to initiate productive analyses! There are three major and interrelated directions in which we anticipate our nanotechnology research and innovation systems analyses will proceed over the next period of time (Years 2-4):

1. Drivers and characteristics of nano development

Once the publication and patent data sets are fully processed, we will begin nano profiling – examining the drivers and characteristics of nanotechnology research and commercialization.

This will include a research mapping aspect, focusing on trends, leading research organizations, notable topical thrusts (“hot areas”), and so forth. For instance, we will apply our organizational thesauri to group activity into industry, academic, and other – to examine

relative emphases of each research sector. We will provide CNS-ASU team leaders with initial results and seek guidance on further probes and best ways to represent.¹⁷

A second aspect of this line of research is to probe the characteristics of nanotechnology research and commercialization. We have already started a line of research on whether nanotechnology meets the claims of being a general purpose technology and the related issue of the extent of convergence in nanotechnology.¹⁸ As we know, nanotechnology is frequently argued to be a transformative general technology with fundamental technological, economic and societal consequences. We aim to use our data sets to probe this issue more finely than hitherto, in particular to move beyond a yes/no or hope/hype polarization and to develop evidence on the kinds of combinations, convergences and (perhaps) divergences that are emerging in the nanotechnology field in research, industrial, and regional contexts.

A third line of work (to start in Year 2) will be to extend our datasets using additional sources. To support work in mapping nanotechnology commercialization, we anticipate building a dataset of new nanotechnology-based firms in the US – comprised of data on spin-offs and startups in the US, many of which have venture capital or SBIR support.¹⁹ We will also explore the availability of additional data sources on the nanotechnology activities of incumbent (larger) firms.

2. Emergence of nano districts

One of the emerging debates in nanotechnology analysis is about the rise of “nano districts” – regional agglomerations where nanotechnology research and commercialization activities are closely combined with supporting public and private infrastructures and networks. In previous rounds of technological development, the US has established strong leadership in developing model technological agglomerations, particularly in electronics (in locations such as Silicon Valley and Boston’s Route 128) and biotechnology (in the San Francisco Bay Area, San Diego, Boston, and elsewhere). Will the same regional locations lead in developing nanotechnology? This is an issue of intense policy interest and activity in the US, particularly in regions of the country that seek to emerge as new nanotechnology complexes. But it is also a global question. In Europe, some analysts argue that nanotechnology provides unique opportunities for leadership, which locations such as Grenoble (France) and Mesa+ (Netherlands) seek to exploit. However, it is also apparent

¹⁷ We are aiming toward the capability to generate special nano interest sub-datasets. One potentially useful strategy is to devise a basic special topic template. We can then quickly generate a basic set of information on specific nano sub-topics. This might include: *What* – topical breakouts; clusters; *Who* – leading research organizations and patent assignees, leading authors and inventors; *Where* – activity by country or region; *When* – trends over time and trajectories. All of these are conducive to combined breakouts, such as: *Profiles* – Contrast the R&D emphases of leading organizations in a given sub-area; *Maps* – Geo-mapping of concentrations of R&D activity; *Network Maps* – Depicting collaboration patterns of various sorts.

¹⁸ Jan Youtie and Maurizio Iacopetta (Georgia Tech CNS-ASU team) are already working with Stuart Graham (Georgia Tech UCLA-Harvard team) on the general purpose technology issue. An initial paper (drawing on another patent dataset) will be presented at the 2006 Technology Transfer Society Conference in Atlanta (September 27-29, 2006). Philip Shapira, Alan Porter, and Jan Youtie are beginning preliminary work on a paper that is planned to address the nanotechnology convergence debate (this will be a 2007 product).

¹⁹ Jue Wang will lead this data-building effort, as part of her doctoral thesis research on nanotechnology small firms and university research linkages.

that – in addition to Japan – other locations in Asia, particularly in China, are contenders to emerge as leading nano districts.

Our globalized publication and patent databases provide a unique resource to enter into this debate, not only to explore where nano agglomerations are emerging, but also to probe the specializations and characteristics of emerging nano districts. We have completed some initial work on nano districts in the US South.²⁰ We anticipate significantly extending this work in Year 2. In addition to analyses that we can conduct within the GT CNS-ASU team, there is also a need to collaborate with others in CNS-ASU (including CNS-ASU colleagues working on employment and nano development policy issues) and elsewhere, to pursue comparative cases and to explore the broader societal and policy issues associated with the rise of nano districts.²¹

3. Potential nano applications

Building on the work already discussed in mapping the drivers and characteristics of nanotechnology development, we anticipate drawing on our data sources to probe leading issues and trends related to nanotechnology applications. Again, this is an important area of debate, fraught with poles of hype, hope and concern. Current applications of nanotechnology generally seem focused on incremental product and process improvements, and have yet to engender the radical consequences that some project and others fear. But this may well change in the medium term (in the next 5-10 years). In Year 2, we will begin work that begins to probe potential and likely trajectories of nanotechnology, focusing on the two themes central to CNS-ASU: Freedom, Privacy, & Security (FPS); and Human Identity, Enhancement, & Biology (HIEB).

This research direction will need close coordination with other groups in CNS-ASU (and maybe also CNS-UCSB and other nano centers). For example, CNS-ASU colleagues have already shared with us initial ideas of projected nanotechnology applications as part of the planned scenario analyses. By using our data on research and innovation activities, we anticipate that we may be able to offer some triangulation on such scenarios.²² We will also consider the feasibility and utility of extending broader analyses in the two CNS-ASU

²⁰ This early work is reported in: Southern Growth Policies Board and the Georgia Tech Program in Science, Technology and Innovation Policy, Connecting the Dots: Creating a Southern Nanotechnology Network. Southern Growth Policies Board, Research Triangle Park, NC: April 2006. A journal paper, to be submitted to a Journal of Technology Transfer special issue on innovation metrics, is in preparation: Jan Youtie and Philip Shapira, “Mapping the Nanotechnology Enterprise: A Multi-indicator Analysis of Emerging Nanodistricts in the US South.”

²¹ We have been involved in discussing a potential collaboration on nano districts with colleagues in the European Union PRIME Network, although specific roles and activities are not agreed as of the time of writing.

²² We have completed a model of what might be done here using some pilot indicators on nanorods (drawing on preliminary publication and patent data). We offered initial responses to: Who are the leading US research groups? We generated several tables based on publications, breaking out the most prolific research organizations, their prominent researchers, and topical foci. We also offered complementary tabulations of leading patenting organizations in this domain and which non-industrial organizations their patents cite [available on request]. We also plotted SCI publication activity against citations. This plot suggested four US universities as stand-outs. This illustrates the notion of massaging raw data to derive valuable research intelligence on emerging nano sub-topics. A next step would be to expose such representations to nanoscientists, industry experts, and other project researchers to promote dialogue and learning about the future directions and applications of this research and patenting.

themes, picking up on some early work on identifying major topical areas of application within FPS and HIEB.²³

The three lines of research outlined above represent an ambitious agenda for the next three years (CNS-ASU Years 2-4). The research questions do overlap, and all of them will draw on the common large-scale data sources on research and innovation that we are completing. The more we draw on these data, the more we will become effective and efficient in extracting intelligence from it, in improving it, and in linking it with other data sources (quantitative and qualitative) and research groups. The three research lines we have targeted all embody and raise important and highly-debated issues about the societal implications of nanotechnology development; we anticipate not only being able to offer a unique evidence-driven vantage point, but also useful interventions into current and emerging policy questions – thus contributing to CNS-ASU’s real-time technology assessment missions.²⁴

²³ This was initiated by Dirk Liebars at Georgia Tech (Memo on Nanotech Profiling, November 30, 2005); we could not proceed at the time, pending development of data. However, now that data sources are in place, this represents a theme that we are now able to explore.

²⁴ In addition to advancing methods and pursuing policy-relevant nano research and innovation systems questions, we are also training a group of young researchers in the techniques and art of tech mining, itself a highly valuable outcome.

References

- Alencar, M.S.M., Porter, A.L., and Antunes, A.M.S., (2006) Nanopatenting Patterns in Relation to Product Life Cycle, Technological Forecasting & Social Change, submitted
- ETC Group (2003). From Genomes to Atoms: The Big Down
- Fraunhofer Institute for Systems and Innovations Research (2002). Search Methodology for Mapping Nanotechnology Patents. Karlsruhe, Germany.
- Huang, Z., Chen, H., Yip, A., Ng., G., Guo, F., Chen, Z-K, and Roco, M.C. (2003). “Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field,” Journal of Nanoparticle Research, 5, 333-363.
- Huang, Z., Chen, C., Chen, A-K., and Roco, M.C. (2004). “International Nanotechnology Development in 2003: Country, Institution, and Technology Field Analysis based on USPTO Patent Database,” Journal of Nanoparticle Research, Vol. 6, No. 4, 325-354.
- Kostoff, R.N., Murday, J.S., Lau, C.G.Y., and Tolles, W.M. (2005). The Seminal Literature of Nanotechnology Research, Office of Naval Research.
- PCAST (2005). The National Nanotechnology Initiative at Five Year. President’s Council of Advisors on Science and Technology, Executive Office of the President, Washington, DC.
- Porter, A.L., and Cunningham, S.W. (2005). Tech Mining: Exploiting New Technologies for Competitive Advantage, Wiley, New York.
- Van der Lei, T.E., and Cunningham, S.W. (2006). Use of the Internet for Future-oriented Technology Analysis, Second International Seville Seminar on Future-oriented Technology Analysis, September, Seville.
- Zitt, M., and Bassecouard, E. (in press). “Delineating Complex Scientific Fields by a Hybrid Lexical-Citation Method: An Application to Nanosciences,” *Information Processing and Management*.

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) (Fraunhofer ISI 2002 definition) and Kostoff et al. (2005).

Table 2. Georgia Tech Modular Nano Search Algorithm**PHASE 1: Database Download** (In using multiple databases, use hyphenation, wildcards, categories, etc., with care!)

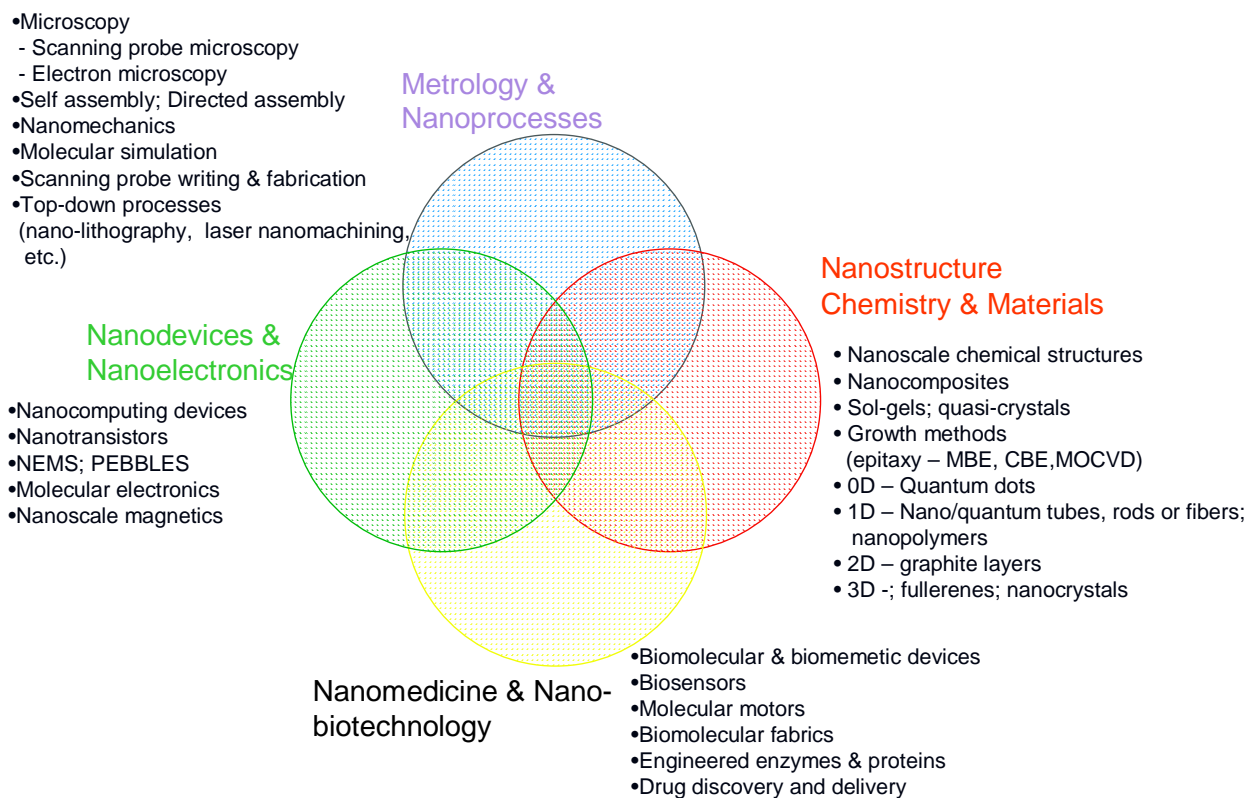
| Search | Terms | RESULT: SCI 2005 as of 4/22/06 |
|---|---|---|
| MolEnv-I (inclusive) | (monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array* or molecu* or polymer* or (co-polymer*) or copolymer* or mater* or biolog* or supramolecul*) | >100000 |
| <i>Or</i> | | |
| MolEnv-R (more restrictive) | (monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array*) | 78390 |
| <i>And</i> | | |
| 1. Nano* | nano* | 39101 |
| 2. Quantum | (quantum dot* OR quantum well* OR quantum wire*) NOT nano* | 3633 |
| 3. Self-Assembly | ((((SELF ASSEMBL*) or (SELF ORGANIZ*) or (DIRECTED ASSEMBL*)) AND MolEnv-I) NOT nano* | 3532 |
| 4. Terms to include as Nano without other delimiters | ((molecu* motor*) or (molecu* ruler*) or (molecu* wir*) or (molecu* devic*) or (molecular engineering) or (molecular electronic*) or (single molecu*) or (fullerene*) or (coulomb blockad*) or (bionano*) or (langmuir-blodgett) or (Coulomb-staircase*) or (PDMS stamp*)) NOT nano* | 3550 |
| 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* | 11665 |
| 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* | 2104 |
| | 1 or 2 or 3 or 4 or 5 or 6 or 7 | 61173 |
| 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 | 61479 |

PHASE 2. Exclusions from Nano*

Terms excluded from Search #1 (Nano*) – these records are deleted from dataset.

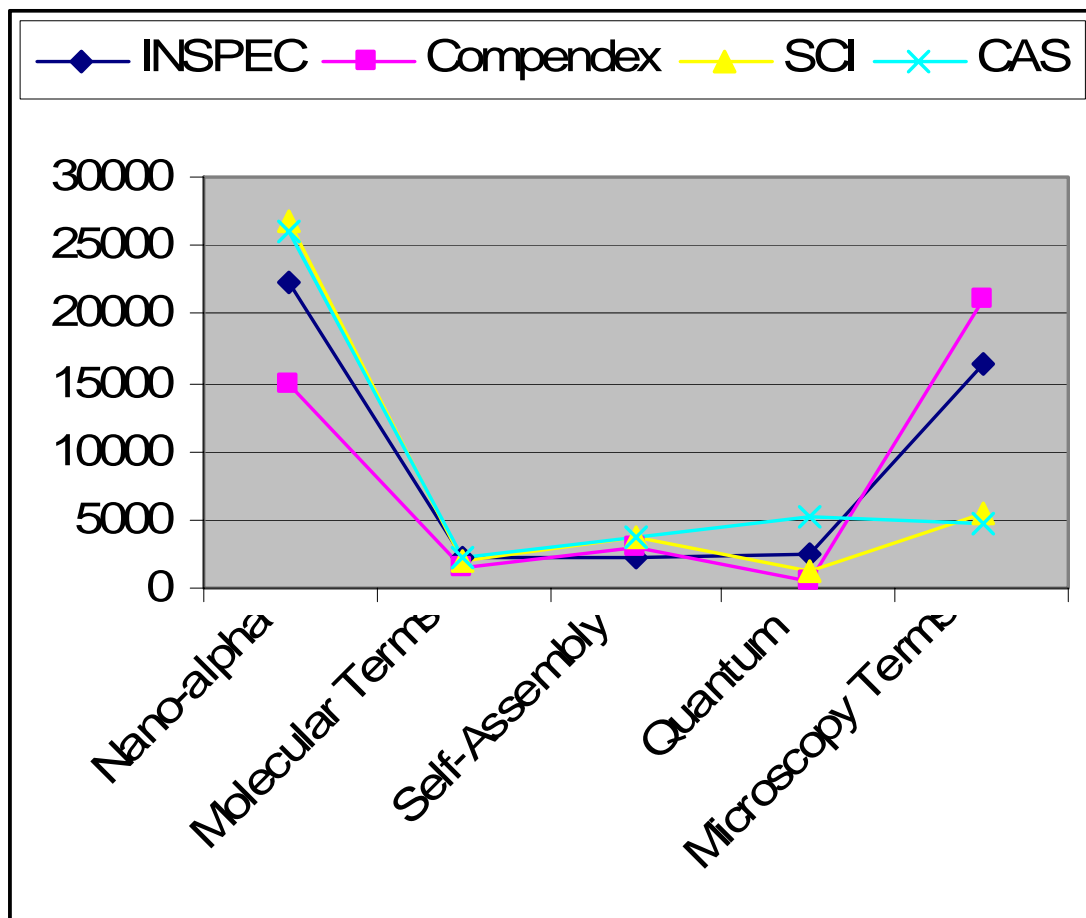
| Exclusion Terms | |
|--|---|
| Records containing these terms should be removed from "Nano*" dataset | Exclude any nano* records containing only one of these terms and no other nano terms |
| plankton* | nanometer* |
| n*plankton | nanosecond* |
| m*plankton | nanomolar* |
| b*plankton | nanogram* |
| 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_, nanao_, nanao_, nano-, nanog-, nanao-, nanor- | |

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

Figure 1. Venn Diagram of Intersecting Nano Emphases

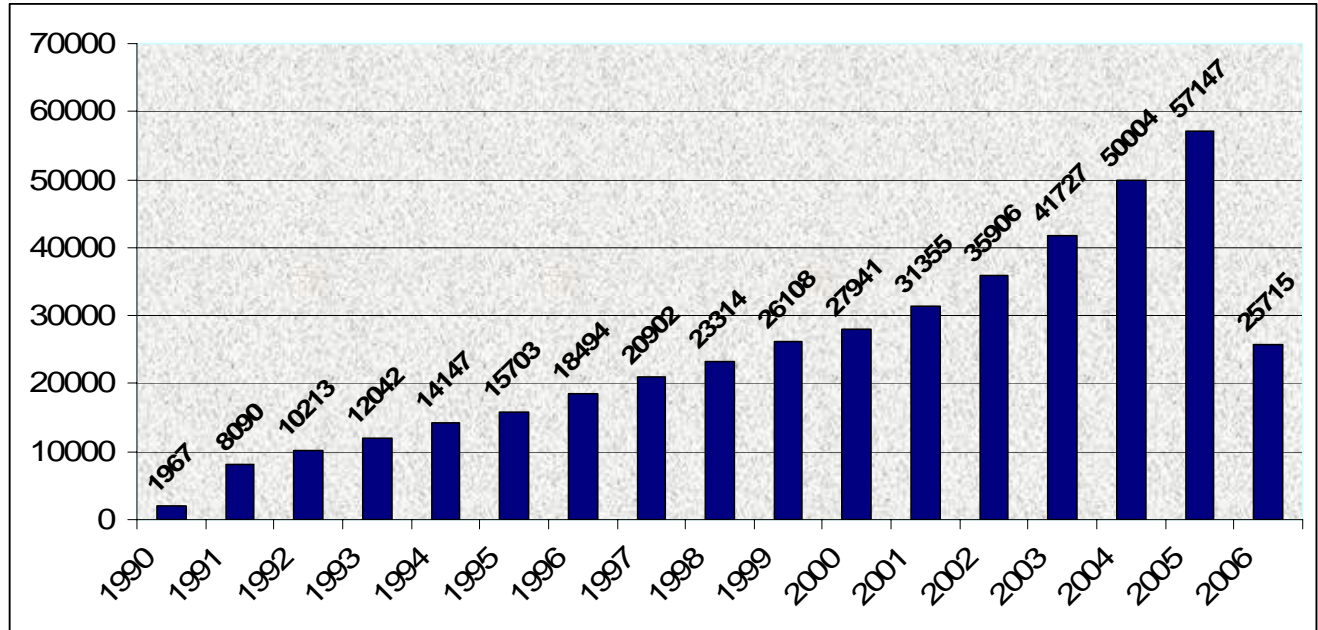
Source: “Field Scope” of Nanotechnology, developed by the Georgia Tech Technology and Assessment Center (GT CNS-ASU Group).

Figure 2. Richness of Nano Coverage by Four Science & Technology Databases



Source: Application of selected nanotechnology terms from Georgia Tech nanotechnology definition to publication records for 2005 in INSPEC, EI Compendex, WOS (SCI) and Chem Abstracts (CAS).

Figure 3. 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.

