

# Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking<sup>☆</sup>

Andrei Mogoutov<sup>a,\*</sup>, Bernard Kahane<sup>b,c,1</sup>

<sup>a</sup> AGUIDEL, 68 Bld de Port Royal, 75005 Paris, France

<sup>b</sup> LATTS (Laboratoire Territoires, Techniques et Sociétés), CNRS/UMLV/ENPC, École Nationale des Ponts et Chaussées, 6-8 avenue Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne La Vallée Cedex 2, France

<sup>c</sup> ISTM (Institut Supérieur de Technologie et Management), Cité Descartes, 93162 Noisy le Grand Cedex, France

Available online 23 April 2007

## Abstract

Nanotechnology, like other emerging technologies that increasingly characterize the dynamic of our era, makes specific demands on datamining to track and interpret efficiently what is happening, through publications and other scientific output. We here propose and describe a strategy based on an automated lexical modular methodology to overcome rapidly evolving content and classification problems, which may otherwise accommodate poor quality of data and expert bias, with potential dire consequences for interpretation, decision and strategy. The proposed methodology is based on an initial nanostring enriched and screened by eight subfields, automatically identified and defined through the journal inter-citation network density displayed in the initial core nanodataset. Relevant keywords linked to each subfield are then tested for their specificity and relevance before being sequentially incorporated to build a modular query. We then, as a first test, compare the database constructed using this methodology for years 2003 and 2005 with those obtained by other approaches previously used to cover and explore the nanotechnology dynamic. Finally, using the inherent transparency, portability and replicability of our methodology, we offer, in order to help our initial query evolve and develop, a set of evaluation processes for tests by researchers in the nano field, other scientometric teams and intelligence experts involved in decision-making processes.

© 2007 Elsevier B.V. All rights reserved.

**Keywords:** Datamining; Nanotechnology; Emergent technologies

## 1. Introduction

Emerging sciences and technologies – biotechnology in the 1980s, nanotechnology today – often have major

implications and potential applications for science, business and society. Their content and dynamic are difficult to track at a time when they are struggling to define what they are, what they include and exclude, and how they organize and classify themselves internally. Data mining is the extraction of relevant and useful information from large volumes of data. Publication data systematically collected in worldwide databases such as those of the ISI/WoS are used to track the dynamic of science and technology. Data mining faces an important challenge in the context of emergent technologies when the latter experience explosive growth, evolve rapidly and cross and subvert existing scientific and technological fields. In the meantime, establishing ad hoc databases or relevant

<sup>☆</sup> The paper is based on a research study carried out within the European Nanodistrict project, which is part of the Network of Excellence PRIME. We gratefully acknowledge EC funding. The paper benefits from stimulating discussions with our colleagues Ph. Laredo & V. Mangematin. We gratefully acknowledge the comments from XX and YY anonymous referees. Usual caveats apply.

\* Corresponding author.

E-mail addresses: [andrei@aguidel.com](mailto:andrei@aguidel.com) (A. Mogoutov), [kahaneb@istm.fr](mailto:kahaneb@istm.fr) (B. Kahane).

<sup>1</sup> Tel.: +33 1 45 92 60 09, fax: +33 1 45 92 60 99.

re-labelling often face delay and may only occur subsequently, once the new technology has settled in. Most of the papers received for the present special issue provide testimony to this situation as they rely on slight variation of an initial query (developed by Fraunhofer-Fhg), using a simple method built on a nano-string (nano\*) with some exclusion and a set of key words, defined by experts in 2001, at a time when the bulk of present nano publications did not exist. Thus, while tracking emergent sciences and technologies may be of great importance for researchers, social scientists and decision makers, it often relies on poorly defined data which may be both too large on the one hand (including false positives, publications that are included but should not be), incomplete on the other (including false negatives, publications that are missing but should have been there), and that further, may reflect an incorrect relative balance between the various disciplines that shape the emerging field. Due to the unreliable quality of data and the fragmentation of knowledge, false assumptions may develop with dire consequences.

Using nanotechnology as a specific case, we here report a data search strategy for emergent sciences and technologies based on the development of an evolutionary query, which minimizes the bias of experts, and allows portability at a limited cost as well as easy updating. We describe the quality of data obtained through this query, first in relation to an initial query that until now has acted as the de facto standard in nanotechnology data mining, then to an alternative methodology whose performance is nevertheless difficult to diffuse and use since it requires full access to the whole WoS as a relational database. We also provide and put online the specific query obtained at this stage in order for other research teams to carry out tests, offer comments and suggest improvements, thus allowing further evolution of this query to help researchers compare their work from one place to another.

## 2. Evolutionary query requirements and methodology

Our aim was to apply to nanoscience and technology a strategy of exhaustive search (to obtain the full extent of relevant only data) that could help to establish a new standard of querying for tracking emergent sciences and technologies at large. All data search strategies start from a similar core of keywords, which in this given field of nanoscience and technology is a nano string ('nano\*') that excludes obvious non relevant terms (NaN<sub>2</sub>, nanosecond, etc.). This string can be applied to titles and/or abstracts. In emerging fields, the core of

keywords related to the string usually expands quickly, since authors are increasingly attracted by the subject, may have different and widening interpretations of what the field 'nano' and its label mean, or may even re-label existing research to get the benefit from the 'serendipity' effect. One of our hypotheses is that as the nano field expands, and structures itself as well as its publication environment, the core of related keywords will, before flattening out, experience an even more rapid growth than the whole nano database. Indeed, authors who had not previously done so will increasingly feel the need, when relevant, to refer to the nanofield, as it becomes too big or too obvious for them to ignore, realising that their subject and publication are indeed concerned and relevant. Thus, the 'nanoisation rate' of relevant publications will increase. It may nevertheless still vary, depending of the nano subfields. For example, electronics would have entered earlier and more intensely than materials or biology, which become concerned by nano later on and possibly, to a lesser extent.

In defining the relevant core of nano-keywords, the data-search strategy aims at complementing the initial nano-string by complementary keywords that better define the field, and extend its borders while keeping outside non-relevant publications. In this regard, no single 'best way' yet exists, as any method has its advantages and drawbacks. Thus, in such a context, it seems best and most efficient to build a strategy that allows progressive incorporation of complementary keywords without discarding those that were previously used. Such a data-search strategy would be evolutionary, building on what already exists, instead of using replacement, discarding what existed in order to start anew. In relation to nano, three types of approaches are mostly used, whose relative performances are characterized by:

- Their core methodology (lexical, citationist or mixed).
- The intervention of experts in the nano field (at what stage and how).
- Their portability (to what degree resources and knowledge are mandatory in order to use them).
- Their transparency (on the choice of key elements that may influence the outcome).
- Their replicability (the extent to which other research teams have access to and understanding of enough information and tools to be able to test what results are obtained using the same methodology and set of initial data).
- Their adaptability (how many previous results remain true when the nano field evolves in its definition, content and concepts).

- Their updating capacity (the question of reprocessing results on a time basis).
- And of course the extent and relevance of the data obtained through them.

The first approach initially applied to nanoscience and technology builds on a lexical query and was developed in 2001 by the Fraunhofer Institute (ISI) and Leyden (Noyons et al., 2003).<sup>2</sup> This approach has strong portability since scientometricians around the world can easily apply the query on the Web of Science, which most researchers have access to. Nevertheless, in this approach, experts are mobilized from the beginning in order to choose which keywords should be added to or excluded from the nano string. Here, the specialized and fragmented knowledge of experts in an emerging field, which often combines many disciplines, may considerably increase bias, not taking into account the vested interests that individuals may have. Thus, the transparency of such an approach is intrinsically limited since the reasons for experts' choices are difficult to decipher. Furthermore, whatever the quality of experts and their choices, they express them in the context of data that are available to them. Applying the initial 2001 query without any evolution to the data of 2003 adds 15,000 publications to the initial number of 28,000 obtained in 2001. Lastly, the choices made in 2002 on the data and dynamic of 2001 may need to evolve, as nanoscience and technology develop and are transformed, incorporating new components and recombining with them.

The second approach developed and tested in the nanodistrict project is both lexical and citationist. Here, the previous (lexical) query, which was built on the initial nanostring, is enriched by (citationist) publication co-citation analysis (Zitt and Bassecouard, *in press*). Only new publications that are cited by at least two authors belonging to the initial database are taken into account. This approach is robust for direct citations, but will also tend to automatically include generic articles of any scientific fields that interact with nano or are part of it. Thus, in order to remain specific and avoid false positives, a statistical limit has to be incorporated known as the 'relevance limit', which builds on the specificity rate, comparing the percentage of citations of a given article in the nano-database to what is observed in the general database. Using this approach for the year 2003, Zitt adds 2000 references: 3000 fewer than what is obtained when the initial lexical query is applied to the same

year. Thus, this approach seems to be more specific, but comparative analysis of the data also shows that publications differ in their distribution and respective weight (Table 3). Biology is less present while physics is more present. Our hypothesis is that the composition of the core of keywords may strongly reflect a context where physics seems to have incorporated its 'nano relevance' more rapidly and more intensely than other fields such as chemistry or biology. There are intrinsic limitations to the usefulness of this approach to data search. However, its interest lies in its power in retrieving the full extent of relevant nano publications. The first limitation concerns its inter-citation component, which needs citations of articles to function, a process that introduces a time lag of 2–3 years, something that is problematic in an emerging field which tends to grow and change rapidly. There is thus a risk of obtaining a correct picture of what nano was 2 or 3 years ago, but an incorrect picture of what nano is at the time the search retrieval is performed. The second limitation is institutional, since this search strategy requires having wide and unlimited access to the full Web of Science database in order to perform the numerous iterations, which are mandatory in an inter-citation approach. To our knowledge, no more than a dozen places worldwide have this kind of access, which strongly limits its portability. Thus, we are in a situation where this approach, in combination with the initial query on which it was built, provides an important validation source for lexical and classification work (which we used, as will be shown below), while we still have to come back to a pure improved lexical query if we want to allow better portability, dissemination and testing of the data search strategy. This drove us to design a third approach, as shown below.

The third approach we propose takes into account the drawbacks and results of previous ones, and combines a set of criteria that differentiate it from existing methods, including those presented above. First, the approach is purely lexical, thus excluding any additional retrieval procedures such as co/inter-citation networks, which diminish portability and introduce a time lag into the results obtained. Second, the approach displays intrinsic simplicity, which allows users' understanding of how and why data qualified, since, unlike most lexical search strategies (including the Fhg/Leyden query), relevant keywords are automatically generated. Bias is thus decreased since experts are not mobilized at the beginning to decide what the relevant terms are and what should be excluded, but at the end, to react to those that have been automatically generated. Third, since the dynamic of sciences and technologies and related knowledge at an emerging stage experience strong growth and

<sup>2</sup> A similar approach is used for patents delineation (Huang et al., 2003; Sampat, 2005).

changes in their boundaries and contents, the approach supports easy updates and traceable incorporation of new terms automatically generated by later querying or extended through expert proposals or testing. Fourth, as emergent technologies and outputs express themselves in several forms such as publications, patents, and the web, the query is designed to allow adaptation for querying of other sets of data needed to provide multi-perspective (science, technology, business, society) emergence tracking. Thus, this approach can be labelled as lexical, automated, cumulative and evolutionary, allowing portability (testing by use with different teams), updating (in the comparison of data for a given year with those of previous or later years through relevant and evolving keywords), and transfer (from one set of data to another in order to provide multi-perspective analysis). This approach is presented below.

### 3. Method

Our approach is based on a multiple step procedure of query building and tests whose full description can be accessed through Internet. ([www.nanodistrict.org](http://www.nanodistrict.org)).

#### 3.1. Step 1 Retrieval of a core ‘nano’ dataset

We applied a formal ‘nominalist’ simple search with the ‘nano’ substring, excluding terms containing this string but not related to the nanotechnology field (NaN<sub>o</sub>2, nanosecond, etc.).

```
TS=((NANO* OR A*NANO* OR B*NANO* OR
C*NANO* OR D*NANO* OR E*NANO* OR
F*NANO* OR G*NANO* OR H*NANO* OR
I*NANO* OR J*NANO* OR K*NANO* OR
L*NANO* OR M*NANO* OR N*NANO* OR
O*NANO* OR P*NANO* OR Q*NANO* OR
R*NANO* OR S*NANO* OR T*NANO* OR
U*NANO* OR V*NANO* OR W*NANO* OR
X*NANO* OR Y*NANO* OR Z*NANO*) NOT
(NANO2 OR NANO3 OR NANO4 OR NANO5 OR
NANOSECOND* OR NANOLITER*))
```

This provides a core ‘nano’ dataset of publications that, on the one hand, does not cover the full extent of nano and, on the other, contains irrelevant data. In particular, the exclusion set was limited to only a few terms that were shown through testing to provide a significant amount of non relevant data and would thus, to a large extent, have polluted the nano-publication data set obtained through the nano-query. Other terms that may not be relevant to our search were tested individu-

ally but were not excluded, since the amount of data they commanded was not significant enough to jeopardize the nature of the data obtained through the nano-query. Furthermore, they would only be present for this initial step, since they would be discarded automatically during the progress of the process described below.

#### 3.2. Step 2 Data set preparation and construction of ‘word combinations’

Step 1 provides us with a nano-data set of publications related to nano with some publications still not relevant. Titles from articles are then extracted and pre-processed: a complete indexation of words present in these titles is performed, as well as a lemmatization in order to reduce the number of title words for further analysis. The ‘word combination’ occurrences of titles are then tested and classified according to their frequency and will be used as candidates for further automatic relevant selection.

#### 3.3. Step 3 Delineation of nano-related subfields

In parallel, the core nano-data set of publications is classified according to a categorization constructed independently on a 2005 nano publications database, by journal co/inter-citation network density, using ReseauLu (Cambrosio et al., in press) proprietary software. Our preliminary unpublished study defines eight distinct subfields labelled according to the network of journals that define them (Table 1).

#### 3.4. Step 4 Co-occurrence analysis of words using statistically relevant word combinations in subfields

Classification of subfields applied to the core nano dataset of publications allows splitting of the initial database in order to extract word combinations of spe-

Table 1  
Subfield classification of nano publications in 2005 according to journals co/inter-citation network density obtained through ReseauLu software

Subfields	Nb publications
Biological chemistry	2879
Analytical chemistry	3499
Physical chemistry	4224
Physics	4741
Material science	4803
Chemistry	4895
Macromolecules	5057
Applied physics	6261

cific titles for each subfield. These word combinations are used to select in the general title word combination list previously obtained from the initial nano core dataset. Similarly, a list of journals may be specifically linked to each subfield.

Word combination retrieval is performed and an index of specificity ( $I$ ) is applied to assess relevance.<sup>3</sup> The index is defined as follows: a discrepancy between the observed frequency ( $O$ ) of word co-occurrence and expected value of this co-occurrence ( $X$ ). The expected values correspond to a '0' hypothesis of the complete statistical independence of word co-occurrences. The expected value is a result of multiplication of corresponding total frequencies of each word divided by the general total. The index of subfield specificity ( $I$ ) calculates the normalized difference between observed and expected values using the following formula  $I = (O - X) / \text{Sqr}(X)$  where  $O$  is the observed value and  $X$ —the expected one. High or low frequency word combination can both display high specificity.

The index of specificity for the nano field is a percentage of publications corresponding to word combination overlapping with the general nano string.

- This selection of the word combinations formally corresponds to the highest values of their frequency and specificity.
- The search strings containing more than two terms have been selected and checked using relation mapping with ReseauLu software, providing multiple combinations of search terms.
- Some of the word combinations obtained are, *a priori*, too broad. Statistical distributions for the specificity index for word combination analysis helps to define at 30% (Fig. A1) the nano statistical relevance used as a cut-off value of the specificity to determine selection in each subfield.

Combinations obtained for each subfield are shown below (Table 2), showing their specificity, both in their respective subfield and in the nano field.

Relevant word combinations extracted from publication titles and selected from subfield specificity will then be incorporated into the query on a subfield step-by-step basis.

Table 2  
Subfield 'word combinations', subfield specificity, and specificity in the core nano dataset

Subfield	Word combinations	Subfield specificity	Specificity nano field
Physics	Walled carbon	8.0	1.00
	Metallic carbon	9.1	0.97
	Semiconducting carbon	7.6	0.97
	Carbon tube*	12.7	0.86
	Mechanical resonator*	7.2	0.64
	Quantum dot*	11.8	0.45
	Single carbon	8.9	0.43
	Surface plasmon	8.9	0.35
Physical chemistry	Walled carbon	8.1	1.00
	Carbon tube*	7.0	0.86
	Tio2 solar	6.9	0.82
	Sensitized tio2	7.3	0.73
	Sensitized solar	8.5	0.69
	Tio2 films	7.4	0.55
	Dye tio2	7.6	0.50
	Li batter*	8.6	0.48
	Dye solar	8.4	0.43
	Single carbon	8.1	0.43
Applied physics	Induced deposition	7.1	0.74
	Field emitter*	8.0	0.69
	Field emission	13.2	0.67
	Crystal* memory	6.9	0.67
	Crystalline diamond	7.0	0.59
	Emission propert*	8.6	0.53
	Vapor deposition	13.3	0.38
	Chemical vapor	12.8	0.38
	Plasma chemical	8.0	0.37
	Carbon film*	8.5	0.35
	Chemistry	Solid lipid	10.5
Gold particle*		7.5	0.64
plga particle*		8.2	0.50
Gold catalyst*		6.1	0.50
Mesoporous silica		8.5	0.45
Co oxidation		8.1	0.40
Drug carrier		6.8	0.39
Analytical chemistry	Enhanced raman	13.8	0.68
	Gold particle*	8.8	0.64
	Direct electrochemistry	8.7	0.57
	Tube* modified	13.4	0.50
	Electrode modified	11.9	0.43
	Resonance light	9.1	0.43
	Immunosensor based	9.3	0.41
	Glucose biosensor	12.0	0.40
	Modified glassy	8.4	0.40
	Raman scattering	8.9	0.38
	Modified electrode	14.5	0.34
	Biosensor based	16.1	0.34
	Electrochemical biosensor	10.1	0.33
	Material science	Ball milling	9.0
Composite powder*		12.2	0.56
Severe plastic		10.2	0.46
Gel method		10.5	0.38
Tribological propert*		9.9	0.38

<sup>3</sup> Zitt (Zitt and Bassecouard, in press), in order to expand nanoscience vocabulary for data mining, has also developed a similar specificity test procedure. Nevertheless, his approach fits his capacity to access and work directly with the full Web of Science, a situation that is beyond the reach of most scientometric teams worldwide.

Table 2 (Continued)

Subfield	Word combinations	Subfield specificity	Specificity nano field
	Amorphous alloy	8.3	0.35
	Plasma sintering	9.4	0.33
	Mechanical alloy	11.3	0.33
	Spark plasma	10.0	0.33
	Composite* coating*	7.9	0.33
	Composite coating*	9.9	0.33
	Metallic glass	16.0	0.31
Macromolecules	Silicate composite*	15.4	0.80
	Clay composite*	14.8	0.69
	/clay composite*	9.4	0.69
	Oligomeric silsesquioxane	10.2	0.68
	Situ polymerization	9.2	0.67
	Poly methacrylate	12.0	0.40
	Block copolymer	16.6	0.38
	Polymer composite*	11.7	0.35
	Composite* prepared	11.8	0.34

### 3.5. Step 5 Query enrichment through specific subfield author keywords

Subfield classifications are used in another round, this time to assess publications for relevant author subfield specific keywords. Specificity is assessed using the same process as described above (Fig. 2). The list of words obtained here is not a word combination extracted from titles but keywords used by authors to qualify their article in the publication's database. They add another layer of words to the query, once again on a subfield basis (Table 3).

### 3.6. Query ( $\alpha$ )

Going from step 1 to step 5 provides us with a query, built according to the nano subfields, thus allowing modularity according to the subfields targeted by users. The query is displayed below (Table 4).

This query allows us to construct a third database, labelled as the Aguidel database, which will then be compared to the databases obtained through other approaches in order to test the quality of our work.

### 3.7. Towards a "Beta" query through evolutionary tests

The request of the unique stable query has a hidden contradiction. Some objects, approaches and artefacts like fullerenes, membranes, surface chemistry, and polymers, which are actually related to the nano sphere, were known and developed before the nano era; meth-

Table 3  
Subfield specific author keywords

Field	Keywords	Subfield specificity	Specificity nano field
Physics	Electrostatic force microscopy	6.0	0.7
	Surface plasmons	6.5	0.4
	Quantum rings	6.0	0.4
Physical chemistry	Dye-sensitized solar cell	8.5	0.8
	Graphitic carbon	8.7	0.6
	Supercapacitor	7.0	0.5
	Transmission electron microscopy	9.3	0.5
	Porous carbon	8.0	0.4
	Chemical vapor deposition	11.6	0.4
Applied physics	Soft magnetic materials	5.8	0.5
	Semiconducting materials	7.6	0.5
	Magnetization reversal	6.9	0.4
	Growth from solutions	8.8	0.4
	Zinc compounds	6.8	0.4
	Diamond-like carbon	6.1	0.3
	Diamond film	6.7	0.3
Biochemistry	Primordial proteins	9.3	1.0
Chemistry	Self-assembly	9.2	0.4
	Mesoporous	6.9	0.4
	Mesoporous materials	10.0	0.3
Analytical chemistry	Surface-enhanced Raman	5.7	0.7
	Ball milling	5.5	0.6
Material science	Mechanical alloying	7.5	0.5
	Spark plasma sintering	5.8	0.3
	Organoclay	13.2	0.9
Macromolecules	Electrospinning	11.3	0.9
	Montmorillonite	15.8	0.5
	Block copolymers	9.1	0.3

ods like transmission electron microscopy could also be applied to different fields, phenomena like self-assembly or self-organization are not necessarily nano-specific. One of the possible solutions should be to obtain distinct datasets for the core "nano\*" field and a collection of partially overlapping sets for related fields and subfields. A measurement and characterization of the overlaps and distances between them should also be provided.

Despite this limitation, our aim was not only to provide an automated query which would meet the relevance, expert-independence, portability and ease of use criteria we had set out at the beginning of our work, but also one which could be updated in order to co-evolve with the nano scientific field it seeks to cover. To reach

Table 4  
Modular query defined through specific subfield word construction

Module	Query
General string	TS=(NANO* OR A*NANO* OR B*NANO* OR C*NANO* OR D*NANO* OR E*NANO* OR F*NANO* OR G*NANO* OR H*NANO* OR I*NANO* OR J*NANO* OR K*NANO* OR L*NANO* OR M*NANO* OR N*NANO* OR O*NANO* OR P*NANO* OR Q*NANO* OR R*NANO* OR S*NANO* OR T*NANO* OR U*NANO* OR V*NANO* OR W*NANO* OR X*NANO* OR Y*NANO* OR Z*NANO*) NOT (NANO2 OR NANO3 OR NANO4 OR NANO5 OR NANOSECOND* OR NANOLITER*)
Physics	(TS=(“walled carbon”) OR TS=(“metallic carbon”) OR TS=(“semiconducting carbon”) OR TS=(“carbon tube”) OR TS=(“mechanical resonator”) OR TS=(“quantum dot”) OR TS=(“single carbon”) OR TS=(“surface plasmon”) OR TS=(“low dimensional system”) OR TS=(“semiconductor structure”) OR TS=(“atomistic simulation”) OR TS=(“finite-difference time-domain method”) OR TS=(“chemisorption”
Physical chemistry	OR TS=(“walled carbon”) OR TS=(“carbon tube”) OR TS=(“tio2 solar”) OR TS=(“sensitized tio2”) OR TS=(“sensitized solar”) OR TS=(“tio2 films”) OR TS=(“dye tio2”) OR TS=(“li batter”) OR TS=(“dye solar”) OR TS=(“single carbon”) OR TS=(“solar cell”) OR TS=(“electrochemical performance”) OR TS=(“carbon composite”) OR TS=(“carbon fiber”
Applied physics	OR TS=(“induced deposition”) OR TS=(“field emitter”) OR TS=(“field emission”) OR TS=(“crystal* memory”) OR TS=(“crystalline diamond”) OR TS=(“emission propert”) OR TS=(“vapor deposition”) OR TS=(“chemical vapor”) OR TS=(“plasma chemical”) OR TS=(“carbon film”) OR TS=(“magnetic fluid”) OR TS=(“ion implantation”) OR TS=(“thin film”) OR TS=(“laser ablation”) OR TS=(“crystalline silicon”) OR TS=(“film* deposit”) OR TS=(“laser deposition”) OR TS=(“beam epitaxy”) OR TS=(“crystal morphology”) OR TS=(“sputtering”) OR TS=(“molecular beam epitaxy”
Biochemistry	OR TS=(“solid lipid”) OR TS=(“gold particle”) OR TS=(“plga particle”) OR TS=(“gold catalyst”) OR TS=(“mesoporous silica”) OR TS=(“co oxidation”) OR TS=(“drug carrier”)
Chemistry	OR TS=(“enhanced raman”) OR TS=(“gold particle”) OR TS=(“direct electrochemistry”) OR TS=(“tube* modified”) OR TS=(“electrode modified”) OR TS=(“resonance light”) OR TS=(“immunosensor based”) OR TS=(“glucose biosensor”) OR TS=(“modified glassy”) OR TS=(“raman scattering”) OR TS=(“modified electrode”) OR TS=(“biosensor based”) OR TS=(“electrochemical biosensor”) TS=(“drug delivery”) OR TS=(“heterogeneous catalyst”) OR TS=(“drug release”) OR TS=(“lipid particle”) OR TS=(“delivery system”) OR TS=(“surface chemistry”) OR TS=(“drug delivery”) OR TS=(“heterogeneous catalysis”) OR TS=(“supramolecular chemistry”) OR TS=(“gene delivery”
Analytical chemistry	OR TS=(“ball milling”) OR TS=(“composite powder”) OR TS=(“severe plastic”) OR TS=(“gel method”) OR TS=(“tribological propert”) OR TS=(“amorphous alloy”) OR TS=(“plasma sintering”) OR TS=(“mechanical alloy”) OR TS=(“spark plasma”) OR TS=(“composite* coating”) OR TS=(“composite coating”) OR TS=(“metallic glass”) OR TS=(“gold electrode”) OR TS=(“carbon electrode”) OR TS=(“biosensor”) OR TS=(“single-molecule”
Material science	OR TS=(“silicate composite”) OR TS=(“clay composite”) OR TS=(“/clay composite”) OR TS=(“oligomeric silsesquioxane”) OR TS=(“situ polymerization”) OR TS=(“poly methacrylate”) OR TS=(“block copolymer”) OR TS=(“polymer composite”) OR TS=(“composite* prepared”) OR TS=(“coating* deposited”) OR TS=(“al2o3 composite”) OR TS=(“coating* produced”) OR TS=(“grain growth”) OR TS=(“plastic deformation”) OR TS=(“microstructural evolution”) OR TS=(“sol* method”) OR TS=(“hydrogen storage material”) OR TS=(“sintering” OR TS=(“microstructure”) OR TS=(“superplasticity”) OR
Macromolecules	(TS=(“surface plasmons”) OR TS=(“electrostatic force microscopy”) OR TS=(“quantum rings”) OR TS=(“chemical vapor deposition”) OR TS=(“transmission electron microscopy”) OR TS=(“graphitic carbon”) OR TS=(“dye-sensitized solar cell”) OR TS=(“porous carbon”) OR TS=(“supercapacitor”) OR TS=(“growth from solutions”) OR TS=(“semiconducting material”) OR TS=(“magnetization reversal”) OR TS=(“zinc compound”) OR TS=(“diamond film”) OR TS=(“diamond-like carbon”) OR TS=(“soft magnetic material”) OR TS=(“primordial protein”) OR TS=(“mesoporous material”) OR TS=(“self-assembly”) OR TS=(“mesoporous”) OR TS=(“surface-enhanced Raman”) OR TS=(“mechanical alloying”) OR TS=(“spark plasma sintering”) OR TS=(“ball milling”) OR TS=(“montmorillonite”) OR TS=(“organoclay”) OR TS=(“electrospinning”) OR TS=(“block copolymer”)

this goal, we here describe the process through which we plan to test and enrich the Alpha version of our query to obtain an updated Beta version. Three sets of tests are defined.

(a) The first set of tests relies on three independent groups of nano experts who will analyze and discuss the data obtained for year 2005, using the Alpha query and classifying the data obtained through the

eight different subfields used above, and through a set of 24 subfields built on inter-citation journal density networks. This may allow testing of the combination of words and author keywords that qualify them for each subfield, and which were selected for incorporation into the query.

(b) The second set of tests is based on a web 2.0 approach, in order to have the query tested and discussed by nano researchers, by scientometrics

experts and by decision makers. This updatable query initiative will publish a list of new ‘emerging’ search terms for validation and completion on a website ([www.nanodistrict.org](http://www.nanodistrict.org)). This is a proposition to develop a Nano-pedia initiative: a Wikipedia-like system for updates of the query and a nano-thesaurus. Using online forms the system should support submission of a new term, and validation of search terms and their combinations.

- (c) The third set of tests will be performed, as has already been done between the Alpha and Beta query, through word filtering for new ‘emerging terms’ from one period to the other, in order to track the evolution in the nano field. Splitting the dataset into two subsets corresponding to two distinct periods of time – last year (months) compared with previous years (months) – allows for a selection of terms using a specificity index for the most recent period of time. This will help intelligence experts to identify and track specific dynamics in the nanofield.

Thus, out of the initial Alpha query, which provides the initial core nano data set and subfield word combination and author keywords, the query will be tested and updated in an evolving process. The query is not only automated and independent of expert intervention in its definition, it is also one, which evolves through the different process described above. The same procedure as step 2 – adding new search terms for the Beta version of the query – thus can and will be repeated. Co-occurrence analysis of words, searching for statistically relevant word combinations, will be performed when needed and will add new search terms to the query. Labelling of the subsets obtained with subfield-specific terms also provides us with the possibility to describe the overlaps and distances between them, just as with the core ‘nano’ set. Partial queries for retrieval of articles zooming into related fields and subfields and/or different time periods become possible. Testing and enrichment by experts and scientometric researchers is easier to achieve: The query as well as the initial nano data set keywords will be both tested by experts in the field (three independent groups have shown their interest in this task) and posted on the web for discussion and testing by other scientometric researchers dealing with nano and/or emerging fields. Delineation and content relevance thus become available for test and improvement while definition of search terms for related fields and subfields with measured proximity to the core nano field are performed. In addition, the query provides search terms not only for the WoS and

patents but for web search too: the ‘related query’ analysis strategy offers relevant terms for webmining and early warning analysis. The system will also provide and report basic statistics, rankings, and trends, as well as describe the nano field and related subfield dynamics and trends in terms of key laboratories, authors, articles, technologies and applications that may be used by experts in the nano field, intelligence experts and decision makers.

### 3.8. Comparative analysis of respective methodology outputs on ISI databases

To begin with, in order to allow testing of the query we had developed, we had at hand two sets of relevant nano publications data already obtained, plus a third one resulting from our own approach. We refer to the Fhg/Leyden database to name the set of publications obtained through the first approach described above through a lexical query strategy (Noyons et al., 2003). This approach, which has often been used for data analysis, has the following disadvantages: the formulation complexity of its experts does not allow easy comprehension and extension, and it has been shown to provide only partial coverage of the nano field (Zitt and Bassecouard, *in press*). We refer to the Zitt database to name the set of publications obtained through the second approach described above which uses a combination of citation networks and lexical network analysis. The incorporation of these two logics allows for more robust outcomes (Zitt and Bassecouard, *in press*) than those obtained through the previous initial ISI purely lexical analysis. Meanwhile, its portability, transparency and replicability are impaired by its requirement for web of science full access, which is out of reach for most scientometric teams. We here name the Aguidel Alpha database, the set of publications obtained through the alpha query we have developed for the year 2003 using a more limited number of relevant word combinations (Table C1: Alpha version of the query for 2003) relevant to the data existing until that year.

Year 2003 was used as a first test case to construct the three databases (Fhg/Leyden database, Zitt database, alpha-Aguidel database), employing each one of the three methodologies (initial labelled lexical query, combination of lexical and citation analysis, enriched automated evolutionary query). Year 2005 was used as a second test case, which, for the Aguidel database, relies on the updated Beta query. Nevertheless, in this 2005 context, the Zitt database can only be approximated through its lexical part since its intercitationist part requires complete local WoS access, something beyond

Table 5  
Size and overlaps between the three sets of database for year 2003<sup>1</sup>

Overlaps (overlaps between databases)	Nb articles
Leyden	47067
Aguidel	43113
Zitt	35766
Zitt and Aguidel	29102
Aguidel and Leyden	41737

<sup>1</sup> Year 2003 test uses the whole Zitt's methodology combining lexical query and citationist extensions, Leyden query and alpha version of the Aguidel query for articles WoS.

Table 6  
Size and overlaps between the three sets of database for year 2005<sup>2</sup>

Overlaps (overlaps between databases)	Nb references
Leyden	44953
Aguidel	87292
Zitt (lexical part)	63168
Aguidel and Leyden	40820
Aguidel and Zitt	43623

<sup>2</sup> Year 2005 test uses lexical part of the Zitt's query (without citationist extensions), Leyden query and beta version of the Aguidel query for any kind of publications WoS.

the reach of most scientometrician teams including our own. Intersections between the contents of the databases give rise to seven subsets as shown below for year 2003 and year 2005 (Tables 5 and 6).

At first glance, these databases, although showing discrepancies that we explore later on, are in the same size range, both for year 2003 and 2005. In order to further explore overlap and differences, each database can be characterized and defined by:

- Its relative distribution using ISI classification.
- A list of key journals and the number of relevant publications it holds.
- A list of keywords with their specificity and frequency.

We here show for year 2003, results using only the first criterion, which shows, beside variation in publication numbers, similar field distribution with slight differences for biomedicine, chemistry and physics (Table 7).

Table 7  
(ISI) field distribution for year 2003 databases

Data set	Bio-medicine (%)	Chemistry (%)	Engineering (%)	Materials Science (%)	Multi-disciplinary (%)	Physics (%)
Zitt	12	34	5	17	2	29
Aguidel	18	30	6	17	3	26
Leyden	20	30	6	16	3	25

Further exploration of differences in publication number and ISI classification distribution between the Aguidel database is performed in year 2005 and displayed in Table D1 of the appendices, showing the subfield profiles of the data set intersections.

Three comments can be made concerning similarities and differences between the various databases obtained.

First, a core component of nano publications is identified, since the number of publications obtained respectively for year 2003 and year 2005 is similar for the different sets of databases within each year. For year 2005, they are also in line with the number of publications obtained both by the Georgiatech team (Porter, 2006) and by the US Office of Naval Research team (Kostoff, 2005) whose cumulative number amounts to 58,559 publications (Porter, 2006). Furthermore, the level of overlaps and differences between the various databases is also in line with the level of intersections observed between GeorgiaTech and ONR data (Porter, 2006). These results help provide a first degree of confidence in the quality and relevance of data obtained through the method of query building that we tested.

Second, both existing similarities and differences between results obtained for year 2003 and 2005 show the need for a methodology that is both stable in its construction and scalable in its content. Scalability and update needs provide a second strong element to our methodology, which allows both easy and traceable updating. In comparison, easy updates are problematic and are delayed with the Zitt methodology because it requires ISI full flat-access, which to our knowledge exists in less than 10 places worldwide. It also has a citationist component, which has an inherent time lag of 2–3 years. Traceable updates are also impaired with the Leyden methodology, which requires expert intervention from the beginning of the query building process. Experts are difficult to stabilize from one expertise session to the next, and deciphering their choices adds another layer of problems. Our automated methodology does not have these drawbacks.

Third, emerging fields not only grow in size, but also evolve in content, some parts of the field appear-

ing or expanding while others disappear or become smaller. Thus publication inter-citation patterns evolve, and would need to be reconstructed accordingly from one period to another. This problem does not exist when tracking stabilized science, but it is a serious drawback for emerging fields. As journal inter-citation patterns are more stable, they can be duplicated with more confidence from one period to another. This adds a third level of confidence to our methodology.

**4. Conclusion**

The automated scalable query that we propose here allows the obtaining of a nano database whose quality content will be further assessed by independent ‘nanoexperts’ and scientometricians. In addition to its easily updatable, traceable, and reproducible quality, it answers two different kinds of needs. On the one hand, scientometricians using a similar query in a different time frame are able to perform a ‘cylindrical’ comparative analysis to show how on a given field structure, publications evolve in their content and number. This can be done forwards, using an old query to assess the comparative publications obtained subsequently. It can be done backwards as well, using a new query to track the comparative seed that already existed years ago. On the other hand, intelligence experts and decision makers are able to compare query, subfield specific words and databases at different time periods in order to track emergence that will help shape their strategy, as shown in the example given in appendices-Table D1.

**Appendix A**

Figs. A1 and A2.

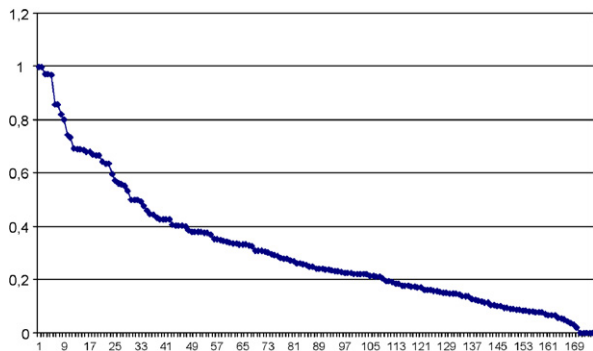


Fig. A1. Distribution of specificity by rank.

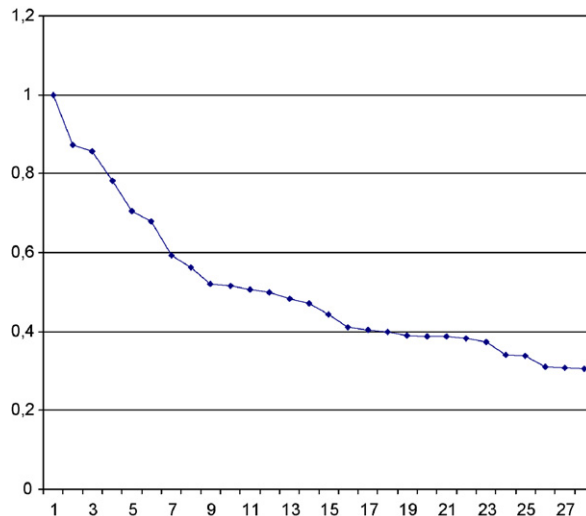


Fig. A2. Distribution of nano field specificity of keywords by rank.

**Appendix B**

Table B1.

Table B1

Alpha version of the query for 2003

```
(((nano*) NOT (nanosec* OR nano2 OR nano3)) OR (Quantum
Dot*) OR (quantum well*) OR (quantum wire*) OR (quantum
structure*) OR (structur* form*) OR (Struct* grow*) OR
(Self-assembl*) OR (Quantum Mirage*) OR
(Microoptoelectromechanic*) OR (Microelectromechanic*) OR
Buckminsterfulleren* OR fulleren* OR SUBMICRO* OR
(SUB-MICRO*) OR (Surface Micromachin*) OR
(LANGMUIR-BLODGETT) OR (Langmuir-Schaefer) OR
(Atomic force microscop*) OR (Molecular electronic*) OR
(Molecular machine*) OR (Molecular manipul*) OR
(Scanning tunneling microscop*) OR (Molecular recognition)
OR (Positional synthesis) OR ((assembl* monolayer*) OR
(assembl* film*) OR (assembl* multilayer*) OR (assembl*
membrane*) OR (atomic imag*) OR (BEAM EPITAX*) OR
(Transmission electron microsc*) OR (Magnetic Resonance
Force Microsc*) OR (Buckyball*) OR (Carbon tube*) OR
(silicon AND ((light AND emit*) OR (purcell AND effect) OR
microcavity OR microdisk OR microtore OR photonic* OR
(laser AND detect*) OR Nanophoton* OR (laser AND
modulat*))) OR ((Single-electron*) OR (Single electron*)) OR
(Lab-on-a-chip*) OR (microarra* OR (DNA chip*)) OR (drug
deliver*))
```

**Appendix C**

Table C1.

**Table C1**  
Profiling of the subsets on year 2005 database using ISI/WoS analysis by subject categories

Aguidel		
Subject category	Record count	% of 87292
Materials science, multidisciplinary	19328	22.1418
Physics, applied	15507	17.7645
Chemistry, physical	11303	12.9485
Physics, condensed matter	10250	11.7422
Chemistry, multidisciplinary	8389	9.6103
Polymer science	4825	5.5274
Nanoscience and nanotechnology	4710	5.3957
Metallurgy and metallurgical engineering	4307	4.9340
Materials science, coatings and films	3519	4.0313
Engineering, electrical and electronic	3507	4.0176
Chemistry, analytical	2949	3.3783
Electrochemistry	2789	3.1950
Physics, multidisciplinary	2695	3.0873
Materials science, ceramics	2611	2.9911
Engineering, chemical	2289	2.6222
Physics, atomic, molecular and chemical	2258	2.5867
Optics	2162	2.4767

#### Leyden

Subject category	Record count	% of 44964
Materials science, multidisciplinary	9183	20.4230
Physics, applied	7404	16.4665
Chemistry, physical	6841	15.2144
Chemistry, multidisciplinary	6215	13.8222
Physics, condensed matter	5152	11.4581
Nanoscience and nanotechnology	3224	7.1702
Polymer science	2780	6.1827
Chemistry, analytical	1905	4.2367
Biochemistry and molecular biology	1740	3.8698
Physics, multidisciplinary	1465	3.2582
Electrochemistry	1435	3.1914
Physics, atomic, molecular and chemical	1387	3.0847
Materials science, coatings and films	1371	3.0491
Engineering, electrical and electronic	1345	2.9913
Metallurgy and metallurgical engineering	1207	2.6844
Biophysics	1165	2.5910
Engineering, chemical	1138	2.5309

#### Zitt

Subject category	Record count	% of 63181
Materials science, multidisciplinary	10171	16.0982
Physics, applied	9019	14.2749
Chemistry, physical	7492	11.8580
Chemistry, multidisciplinary	6873	10.8783
Physics, condensed matter	6392	10.1170
Biochemistry and molecular biology	4218	6.6761
Nanoscience and nanotechnology	4170	6.6001
Polymer science	3101	4.9081
Engineering, electrical and electronic	1958	3.0990
Physics, multidisciplinary	1868	2.9566
Chemistry, analytical	1835	2.9044

**Table C1 (Continued)**

Subject category	Record count	% of 63181
Cell biology	1660	2.6274
Pharmacology and pharmacy	1549	2.4517
Physics, atomic, molecular and chemical	1548	2.4501
Biophysics	1502	2.3773
Materials science, coatings and films	1495	2.3662
Biotechnology and applied microbiology	1479	2.3409

## Appendix D

**Table D1.**

Table D1  
An example of a list of “new” “emerging” terms applicable as a search terms

Aqueous interface
Solar comparison
Diffusion impedance
Sensitized electrochemical
Charge capacitance
Solid devices
Dye electrochemical
State devices
Nanoporous sensitized
Nanoporous dye
Supported palladium
Gold nanoparticles
Pt nanoparticles
Stober silica
Electrochemical solar
Density functional
Nanopillar arrays
Pore system
Nanotube forest

## References

- Cambrosio A, Keating P, Lewison G, Mercier S, Mogoutov A., in press, Mapping the emergence and development of translational cancer research; *European Journal of Cancer*.
- Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.K., Roco, M.C., 2003. Longitudinal patent analysis for nanoscale science and engineering: country, institution and technology field. *Journal of Nanoparticle Research* 5, 333–363.
- Noyons E.C.M., Buter B.K., Van Raan A.F.J., Schmoch U., Heinze T., Hinze S., Rangnow R., 2003, Mapping Excellence in Science and Technology across Europe, Nanoscience and Nanotechnology, Draft report of project EC-PPN CT-2002-0001 to the European Commission.
- Sampat, B.N., 2005, Examining patent examination: An analysis of examiner and applicant generated prior art., Working Paper, Columbia University.
- 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”.