A dynamic query to delineate emergent science and technology: the case of nano science and technology.

Working paper / Kahane Bernard, Mogoutov Andrei, Cointet Jean-Philippe, Villard Lionel, Larédo Philippe, "A dynamic query to delineate emergent science and technology: the case of nano science and technology", *In* Villard Lionel, Revollo Michel, Laredo Philippe, *Content and technical structure of the Nano S&T Dynamics Infrastructure*, RISIS, 2015, pp. 47-70
http://risis.eu/wp-content/uploads/2015/03/Report-Task1-Nano.pdf

1- Introduction

Building a larger and relevant database out of an initial seed without relying, because of potential bias, on experts is a common challenge for those who wish to study or track a scientific or technological field. Publications and patents are not the only, but definitely an important component of knowledge generation and dissemination and one of the potential sources for innovation. Scientists communicate their findings through publications. Similarly, patents are legal documents to claim ownership of an invention but they also build a public paper trail of technology advancement. Thus publications and patents are an important, relevant and useful tool to follow and represent results of scientific and technological endeavours (Huang, 2010). Data mining is the extraction of relevant and useful information from large volume of data. Publication and Patent data systematically collected in worldwide databases such as the WoS and Patstat are used to track science and technology dynamic. Data mining faces an important challenge in a context of emergence when new technologies experience explosive growth, evolve rapidly and often cross and subvert existing scientific and technology fields. Emerging science and technology (biotechnology in the 1980s, nanotechnology today, other science and technology fields tomorrow), which often carry strong implications and potentialities for science, business and society, add to the challenge. Their content and dynamic are difficult to track at a time when they are struggling to define who they are, what they include and exclude and how they organize themselves internally.

Such is the case for nanotechnology, where the quest for a relevant reliable and replicable way to extract relevant publications and patents, is an on-going process involving several teams worldwide (Glanzel 2003, Noyons 2003, Mogoutov and Kahane, 2007, Porter et al., 2008, Kostoff 2007, Leydesdorff and Zhou, 2007). Nanotechnology is a rapidly evolving emerging and dynamic field. Analysts argue that it is likely to be a "general purpose technology" (Youtie 2008, Laredo et al. 2010) with a potential impact across an entire range of industries and great implications on human health, the environment, sustainability and national security. The perceived potential value of nanotechnologies has led to the increased will of governments, academic institutions,

firms and other societal actors to better understand what is happening in the field, who is active and where. There is thus an important challenge to develop robust methods to track the nanotechnology field while it rapidly develops and evolves. As a matter of fact, good quality and comprehensive extraction of data is a prerequisite for meaningful understanding and analysis. Huang 2010 as well as L'huillery et al. 2010 have compared the different methodologies developed, and reported on their robustness as well as on the similarities and discrepancies of results obtained. They confirmed the robustness and interest of the evolutionary lexical methodology we have developed (Mogoutov and Kahane, 2007). At that time, three requirements were central to the approach developed. First, it should not depend upon experts. Indeed, the on-going and extensive use of expert-based approaches is costly, time-consuming, and challenging to replicate such that the same outcomes result. This is an important restriction when facing a highly dynamic field where borders are constantly evolving requiring terminology requalification at different times. Second, it should allow updates in order to replicate and compare results while the nanotechnology field (and its lexicon) develop and expand. And third, it should be able to track the relative evolution of subfields inside nanotechnologies: in 2007 we translated this into a third requirement of being "modular".

While the initial development of our methodology was performed in order to extract data from 1998 to 2006, we later engaged in producing an update that could expand the database backward and forward in order to cover years 1991-2011. In our initial methodology, the selection of relevant terms was performed with knowledge built and keywords selected on one single year (2003). A simple solution was to reproduce the selection of terms for 2011, driving us to two semantic universes of nanotechnology, respectively built in 2003 and 2011. However Bonaccorsi (2010) has demonstrated that in a dynamic field such as nanotechnology, keywords often display short life and experience a type of Darwinian selection process. Using this approach, the characterisation of the evolution of the field over 20 years would have only relied on two years for the identification of relevant keywords. There would thus be a risk that we miss the richness of the exploration that shapes the dynamics of knowledge production. Not considering transient keywords that might have emerged and then disappeared, would be a serious drawback in such a dynamic field. There are multiple reasons for this. Two are of particular importance. One is about the learning that a stream of research, even if it goes on with a life of its own, has been experimented but proved not to be useful for colleagues at the time. The other lies in the fact that streams of research which for a while turn to be a dead end, can nevertheless reappear later and become a key resource as demonstrated in many instances. Such a limitation becomes even more visible when taking the whole period under review for identifying relevant keywords. This drove us to add a fourth requirement for such an approach: What is needed is a methodology, which allows us to incorporate and discard in real time relevant terms as they appear and disappear in the nanotechnology story. We need a methodology that

allows us to track keywords as characters appear and disappear along the storyline in a movie.

Thus, using nanotechnology as a showcase, we here report a data search strategy made of three consecutive steps. As in all the data search strategies for nanotechnology, we start with an initial seed built through the nanostring. We then use the same principle that we applied in our previous approach, that is expanding the initial seed through a dual process where additional keywords observed during a given period are sorted according to their internal specificity (e.g. the extent to which they provide value added meaning to a publication) and then tested in the overall database for 'external specificity' (e.g. the ratio of articles in the seed vs. articles in the overall database of publications). This selection of keywords is first applied on the whole dataset covering the 20 years, enabling a "static extension". The third step builds the "dynamic extension" where additional keywords are identified through a yearly analysis of internal specificity within the nanostring, and selected depending upon their 'external specificity'.
Besides being applied in a specific way for nanotechnology, we claim that such a three steps strategy has universal value to describe the dynamics of emergent and fast evolving fields, transcending pre-existing classifications.

The article is built as follows. First, it provides a literature review of different search strategies, pointing to their limitations and explaining how our choices were made. Second, it looks at specific requirements needed when studying nanotechnology and explains how and why we decided to address them. Third, it provides the rationale and the description for the successive steps of our methodology. Fourth, some lessons derived from the nanotechnology example are derived for other emerging fields.

2- Evolutionary query requirements and methodology

As reported by Huang (2010), four different methodologies are used to search nanotechnology articles in the publication databases. They are lexical query, evolutionary lexical query, citation analysis and harvesting publications in core journals. We review them with our four requirements in mind: easiness (enabling wide access by research teams), portability (enabling reproducing results from one place to another), updating (to accommodate for the need for periodic characterisation of evolutions) and capturing dynamics of search (a critical issue in fluid fields facing wide exploration).

Lexical query

Most works and methodologies dealing with emerging fields rely on slight variations of an initial query, often built on a few terms that help define the field with some exclusion of obvious non-relevant terms. In the case of nanotechnology, it defines a nano-string built with the word "nano" plus a joker ("nano*"). For nanotechnology, such an initial

query was developed by Fraunhofer-ISI in 2002[1] and is still at the core of most publications analysing the content and evolution of the field, whether in publications or in patents (Glanzer et al., 2003; Noyons et al., 2003; Porter et al., 2008). Two limitations exist with this approach. On the one hand, some words like NaNo2 or nanosecond need to be excluded. On the other, in emerging technologies with fast expansion, authors become increasingly attracted and introduce alternative keywords for labelling the field, which need to be incorporated in the search[2]. Indeed, we have shown that the core of related keywords experience an even more rapid growth than the entire database of nanotechnology publications (Mogoutov and Kahane, 2007). In both cases, the more precise the exclusion or the inclusion, the greater will be the need for complementary keywords. One possible solution is the use of experts, but Huang, reviewing the existing approaches, underlines the possible bias associated with their subjectivity (Huang, 2010). Thus, automatic methods are needed while manual exclusion or inclusion have to be kept to the minimum. This applies as well for defining the initial seed: in our initial nanostring seed, only nanoliter, nanosecond and chemical formula of NaNO2, NaNO3, NaNO and NaNO5 are excluded.

Automatic evolutionary extension of keywords

In a similar vein (avoid experts subjectivity and bias), automatic and iterative ways of obtaining search keywords have been developed as an alternative to manual extension (Zucker et al, 2007; Mogoutov and Kahane 2007). Out of a first dataset built through the nanostring, a set of keywords is harvested. Keywords are then ranked by their level of relevance to the field, based upon their frequency of appearance (alone or in combination). A mathematical threshold is built on keywords profile and/or an iterative process is mobilized in order to assess the relevance of keywords. As this relevance is assessed within the initial seed only, we speak of internal relevance and later internal specificity of the keywords. The iterative process looks at publications convergence on a relatively consistent set of keywords that change only slightly between iterations (Zucker et al., 2007; Kostoff et al., 2006) or at data distribution (Mogoutov and Kahane 2007). This selection of keywords is dependent on the initial seed collected. This is the drawback of minimizing expert intervention, and the limitations associated with their subjectivity. Most approaches have witnessed successive improvements of the method they use to measure the internal relevance of keywords. Compared to our previous publication, we propose here a new alternative method, which we claim to be of better quality.

[1] Note that at that time the bulk of present nano publications and relevant keywords did not exist.

[2] Early bibliometric analysis by, for instance Braun and al 1997 have shown that extraction through the use of the simple term "nano*" suffered from the omission of biotechnology-related publications whose keywords were less likely to contain the prefix "nano".

Automatic Citation analysis

Zitt and Bassecoulard (2006) demonstrated an alternative hybrid lexical-citation approach to extend publications beyond the nanostring. There the second step is done by identification of a "core" literature cited by the seed literature. To extend the seed, they extract other publications citing this core literature while controlling by use of a parameter that strikes a balance between the specificity and the coverage of the publications in order to get a good "noise to silence" ratio (Huang 2010). As for the previous evolutionary extension method, subjectivity of expert intervention is limited while the way the inclusion/exclusion parameter is defined becomes the key factor. The trade-off is between too much "noise" vs. "silence". Nevertheless, this approach adds another difficulty since its implementation requires setting up a citation linkage between all the papers in the WoS database. This limits this approach to no more than a dozen institutions worldwide with such capacity to access the full web of science database to use the pre-built citation linkages (Mogoutov and Kahane 2007). Thus, as in our previous publication, we discarded this approach in order to keep the portability and feasibility by other teams that we wished in order to achieve dissemination and comparative analysis.

Publications in the core nanotechnology journals

Leydesdorff and Zhou (2007) use journals as the unit of analysis and extract articles from a set of core journals. Using "betweeness centrality" as an indicator for measuring the interdisciplinarity of scientific journals, they distinguish a set of three core nanotechnology journals and a group of 85 journals related to them from which they identify ten core journals on nanotechnology. One of the drawbacks of this approach is that it only covers a small share of the literature. Thus, as demonstrated by Huang (2010), the total number of publications harvested by this approach is 5 to 10 times smaller to what is obtained through other strategies. Moreover, as the technology is emerging and evolving, the set of journals, which publish nanotechnology related articles, is also changing. The analysis based on a very limited number of the core journals chosen at a certain time would thus impair results.

This last argument points to the specific issue of an emerging field and its evolving nature. This result emphasizes the need and requirement for an approach, which will display a strong capacity to reflect and track the intense dynamic of the field. It is in line with the work of Bonaccorsi on search regimes (Bonaccorsi 2008) and its results about the rapidly evolving nature of emerging fields and about the need for approaches and queries that take into account keywords life. This requirement challenged our previous methodology which was built on a modular basis allowing specific subfield analysis but which did not offer any tool to follow on going evolutions. Studying computer science, Bonaccorsi (2010) points to two central phenomena, which happen in an emergent field with rapid expansion and intense dynamics. Firstly, very few research lines and associated keywords succeed in establishing themselves on a long-term basis. In order

to capture these, we developed a first "static" extension that looks at keywords, which have established a significant presence in the field when the whole period of analysis is considered. Besides these success, Bonaccorsi shows that many other tentative lines of research and their associated keywords struggle but do not succeed in maintaining a presence in the field on a long term basis. Thus, without taking on board these exploratory lines of research we would miss a large share of the dynamics, which characterizes the evolution of nanotechnology. Further, we would not be able to catch researches and keywords at the end of the period studied since there are great chances that their presence is still too limited to overcome the limitation of a few years of presence in the database. Thus, in order to capture these tentative lines of research, we had to develop another kind of extension that we call "dynamic extension". We now report below the different steps through which the initial nanostring is built and then expanded.

3- Methodology

Our approach is based on a multiple step procedure of query building and tests. The methodology is made of the following steps:

- Extraction of publications through the nano string giving the nanostring database
- Selection and cleaning of "main forms" from the nanostring database, giving the universe of keywords to consider
- Extraction of the main forms selected from the entire period in order to build the "static extension" database
- Extraction of main forms selected year by year in order to build the "dynamic extension database"

Step 1: Retrieval of a core 'nano' dataset: Extraction of publications through the nanostring

In line with the previous method, we applied the same formal nominalist simple search with the 'nano' substring as used in most other methods. In order to limit and reduce bias to minimal we excluded as before only a few terms containing this string but not related to the nanotechnology field (nanosecond, NaNO2, NaNO3, NaNO4, NaNO5). It is presented in the box below, which takes into account evolutions of the interface proposed by the WoS at the time of downloading.

Box 1 - The query for the nanostring

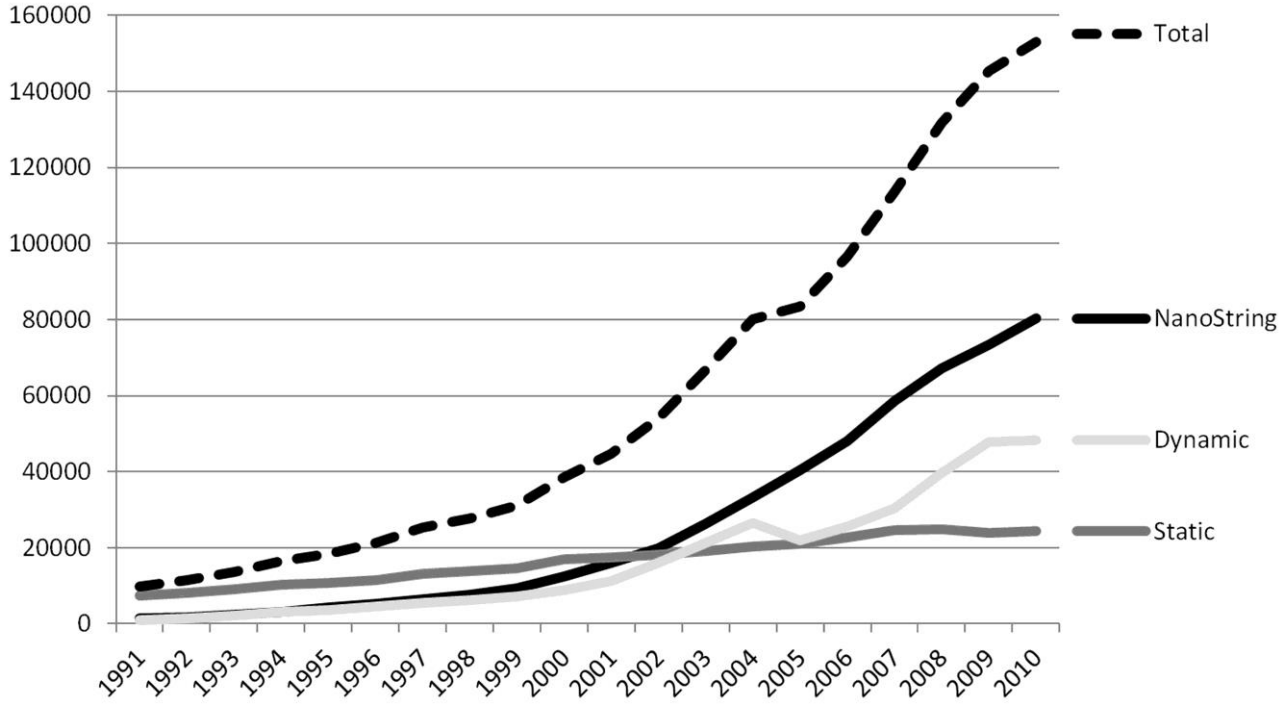
Note: the introduction by the WoS interface of a lemmatisation has simplified life for managing 'main forms' (see below) but it has limited the use of "*" in the construction of the query for abstracts and keywords (TS) driving to a different query as this for titles (TI).

TI=((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*)) OR TS=((NANO*) NOT (NANO2 OR NANO3 OR NANO4 OR NANO5 OR NANOSECOND* OR NANOLITER*))

From 1991 to 2010, this extraction gave 517050 articles, with an impressive growth of 20% for 15 years rising to 40000 articles in 2005, and then a doubling in 5 years to 80425 articles. We shall see later that the share of this coreset, the nanostring, will regularly increase in relative importance during the first period from 14% in 1991 to 30% in 1999 and 48% in 2005. Since then it has fluctuated around 50% and been on average for the last five years 51%.

Graph 1: Nano science: an overview



Box 2 – Technical notes on the nanostring

The note addresses two issues: coverage and exclusions.
Coverage: the query has been simplified for technical reasons about downloading from the WoS. After tests we decided not to keep for abstracts the same rule as for titles to insure the full presence of words that do not start with the prefix nano. The tests showed that it would reduce the overall volume by 0,6%. As a consequence we decided that extensions would verify that we had not missed too many articles (where the specific term such as subnano* would be in the abstracts only). This approach that reduced downloading time significantly proved to be relevant: for instance, the 'raw' static extension (see below) contains 73 multi-terms including nano that theoretically represent more than the total nanostring. Still we only retrieved 777 potential articles showing that this time optimisation was very efficient.

Exclusions: We decided to concentrate 'targeted' exclusions afterwards, that is on the effective dataset built. The argument is dual: technical (one of simplicity and efficiency in downloading) and substantive (we do not master the multi-terms built around the classical exclusion terms – e.g. subnanosecond - and we do not know their potential articulation to other 'relevant' multi-terms of the vocabulary). Another interesting aspect lies in the role that the extensions made provide in term of testing the specificity and relevance of the multi-terms identified in the nanostring: this is very efficient in identifying problematic areas (such as those around the measure of the amount of substance concentration).
This has also enabled to take advantage of the progressive work done in particular by Grieneisen and Zhang (2011) and Arora et al. (2013).
We put here the main exclusions operated from the final dataset:
- The 270 taxonomic organisms and species identified by Grieneisen and Zhang (2011).
- The classical terms around plankton (nano & pico), satellites (nanosatellites) and flagel (e.g. nanoflagellates)
- The classical exclusions around grams and moles (all the variations around nanogram, including nanog, and nanomolar).
The latter represents by far the largest set of articles excluded while all the others have only a marginal effect on the dataset.

Step 2: Data set preparation and lexical expansion and extraction

At this stage, we adopt a lexical extension methodology different from the one used in our previous publication. Step 1 provides us with a core dataset of publications related to nanotechnology that needs to be expanded to better cover relevant publications. Expansion requires first extracting terms pertaining to a given corpus. Similarly to what was done in the previously published methodology, titles from articles are extracted and pre-processed from the dataset obtained on step 1: a complete indexation of words present in these titles is performed as well as a lemmatization in order to reduce the number of words with similar meaning in further analysis. Then methodologies diverge. We have made two central changes compared to our previous approach.
The first one deals with the selection of candidate terms for the selection of articles, and the other deals with the approach to the way of defining sub-datasets for computing. First, in our previous method, "word combinations" in titles and abstracts were classified according to their frequency in order to select candidates for further automatic relevant selection. Now the Natural Language Processing (NLP) tools we apply, allow us to identify not only simple terms (e.g. nanotube) but also multi-terms (e.g. carbon nanotube or tubular carbon nanotube) (also called n-grams). While automatic multi-terms extraction is a classical task in NLP, the existing tools are not always well suited when one wishes to extract only the most salient terms. We thus mobilised methods for measuring their specificity. However specificity computing drives to an exponential growth of computer time and resource as datasets grow larger. We have thus developed an automatic method that helps reducing computing requirements. This lexical extraction strategy is not directly applied to the entire corpus. We first split the corpus into 20 sub-corpuses, one per year (each sub-corpus gathers all publications published a given year). Lexical extraction is then applied on each sub-corpus and the

2000 most relevant multi-terms are extracted for each year. Hence we can be confident that we do not miss important terms that only occur in the early times or which are only important during a limited time period.

The selection of the relevant multi-terms is made in two stages. First classical linguistic processes end up defining sets of candidate noun phrases. Second, the most relevant multi-term stems are selected.

a) Defining candidate noun phrases
- We use a Part-of-Speech Tagging tool to classify each word of the text according to its grammatical type: noun, adjective, verb, adverb, etc. This allows focusing on potentially meaningful terms for analysis (nouns and possibly adjectives), leaving aside less interesting terms (such as verbs or adverbs).
- 'Chunking' associates to each word of the text a tag describing its type. As shown in the example below, a noun phrase is then defined as a pattern of successive nouns and adjectives. This step builds the universe of multi-terms. It helps define and extract the minimal meaningful units on which to build further analysis.

Box 3 - Example of chunking process

Therefore<CC> a<DET> finite-volume<ADJ> discretization<N> of<CC> the<DET> 3d<ADJ> self-consistent<ADJ> model<N> was<V> implemented<V>...

Results: two different noun phrases are obtained
- finite-volume<ADJ> discretization<N>
- 3d<ADJ> self-consistent<ADJ> model<N>

- 'Normalizing' corrects small orthographical differences between multi-terms regarding the presence or absence of hyphens. For example, we consider that the multi-terms "single-strand polymer" and "single strand polymer" belong to the same class.
- 'Stemming' drives to gather multi-terms together into a single class if they share the same stem. For example, singular and plurals are automatically grouped into the same class (e.g. "fullerene" and "fullerenes" are two possible forms of the stem "fullerene").

b) Selection of most relevant multi-terms stems
This first processing based on grammatical constraints provides an exhaustive list of possible multi-terms grouped into stemmed classes. The second stage aims at selecting the *N* most relevant terms.
Following an approach defined by Kageura and Umino (1996), we are looking for groups of relevant terms which convey the most interesting semantic unit (high ***unithood)*** using as a proxy those multi-terms appearing more frequently and being in the longer

phrases[3]. Meanwhile, we wish these terms to convey strong meaning *(*high ***termhood***) and thus to discard those which may be very frequent in the corpus but do not help characterizing the content of the text. These are for example terms like "review of literature" or "past articles". For this purpose, we proceed in four stages:

- 'Counting': we count each stem according to corresponding multi-terms found in the whole corpus to obtain their total number of occurrences (frequency). In this step, if two candidate multi-terms are nested, we only increment the frequency of the larger chain. For example if "spherical fullerenes" is found, we only increment the multi-stem "spherical fullerene" but not the smaller stem "fullerene".[4]
- 'C-value unithood calculation': for each multi-term stem, we associate the C-value as proposed by Frantzi & Ananiadou (2000). This provides each stem with a unithood value defined as $u(i)= \log(l_i) f_i$ where l_i is the number of terms involved in the multi-term *i* and f_i designates its frequency.
- 'Sorting': Items are then sorted according to their unithood value (Van Eck et al., 2011) and the list is pruned to 4 times the number of multi-terms looked for (see above) starting from the highest C-value. This step removes less frequent multi-term stems.
- 'Selecting': A second-order analysis is performed on the 4N list obtained of the terms with highest unithood value in order to exclude those who do not carry special meaning. We adopt the approach proposed by Van Eck et al. (2011) to identify multi-term stems with low termhood. The rationale that we follow is that irrelevant terms should have an unbiased distribution compared to other terms in the list. These terms may appear in any documents in the corpus whatever the precise thematic they address. We first compute the co-occurrence matrix *M* between each item in the list. We then define the termhood θ of a multi-stem as the sum of the chi-square values it takes with every other class in the list[5]. We rank the list according to θ and only the *N* most specific multi-stems are conserved.

Thus, through this yearly double process of identifying sets of candidate noun phrases and then of sorting multi-term stems according to their relevance through their unithood and termhood, the final output of our analysis comes to a list of multi-term stems (from now on we shall speak of multi-terms to qualify them) which display both high unithood value and termhood and which can now be ranked according to what we call their **internal specificity**. The power of NLP and the approach developed entailed one important implication: we can work directly at the level of the whole 'nanostring' and no longer require decomposing it using pre-existing fields (we had 8 such sub-fields

[3] This unithood qualification builds on two pragmatic assumptions classically made in multi-word automatic term recognition tasks: pertinent terms tend to appear more frequently and longer phrases are more likely to be relevant.

[4] Nested terms need to be treated carefully because they may induce false positive - for example when the multi-term "self organizing map" is found in a text, we should not count the multi-term "organizing map", otherwise we would overestimate its unithood even though it does not convey any unit of meaning.

[5] The endogenous specificity of term i is $\theta(i)= \sum_{j \neq i} (M_{ij} - M_i M_j)^2 / (M_i M_j)$ where $M(i)= \sum_j M_{ij}$. This measure accommodates both the possible bias of item *i* toward certain other items and still takes into account terms frequency.

in the 2007 query). This drove us to abandon the 'modular' approach designed (in part for pragmatic reasons) in our previous approach. This has one important consequence: before we had to consider specifically all potential 'long-distance' interdisciplinary papers (i.e. between the selected fields identified) while they are now de facto taken into consideration.

Box 4- Main results of Step 2 on the nanostring
When performing the identification of multi-terms we arrive only at 2000 different multi-terms in 1997, giving a theoretical total number of 34191 multi-terms over the whole period (1991-2010). Redundancy is very high as the total vocabulary is only 4189 different multi-terms with in total more than 17 million occurrences. This means that on average one article is defined by 33 multi-terms, which builds a very rich characterisation.

Introducing the two step extension

The two next steps aim at identifying within the relevant multi-terms selected in the initial seed, those that can be considered as specific to nanotechnology and which we shall use to retrieve complementary articles to those already included in the nanostring. In our previous query we only had a 'static' extension, selecting the most relevant multi-terms over the whole period. It aims at enriching the core knowledge that has demonstrated over the period its ability to aggregate scholars and their publications. We propose in this new query to add a dynamic extension. The purpose of such an extension is not to loose track of the explorations made year after year even if they have not succeeded to become 'core'. This is also important since otherwise by only having a static extension we would not take into account on-going developments. Doing so requires making choices about the overall size of the dataset and caring about the noise-silence ratio. The literature is not very rich about these issues that most of the times remain unaddressed by developers. Looking at our previous query, which covered 9 years only, the lexical extension multiplied the core by 2.6 times. We found similar multipliers in other queries. We thus considered that keeping in line would be a reasonable solution and that we should aim at a theoretical tripling of the nanostring balanced between the static and dynamic extensions. As extensions drive to select more than once articles (if only between the two extensions), and knowing empirically that overtime papers refer more and more explicitly to nanotechnology (Arora et al. 2013, see also the growing share of the nanostring over time), this should drive to a far lower net increase (de facto 2.28 times with each extension representing 28% of the expanded dataset).

In our previous study, we highlighted a very rapid rate of growth (14% per year between 1998 and 2006). We thus took into account that, even if with size the rate of growth might slightly reduce, it would continue to grow arriving to very large yearly levels (de facto the number of publications in 2010 is equal to the total of the first 9

years of the dataset -1991-1999). This drove us to look carefully at the results of Bonaccorsi (2010) in computer science (even though its rate of growth was slower).
First, many new lines of research constantly emerge with new associated keywords and only a few of these new lines of research and associated keywords establish themselves to become persistent. An extension must thus give credit to research directions that have succeeded in becoming persistent. This was already at the core of our previous approach and we kept it: this builds the "static extension".
Second, this also means that most new lines of research that emerged had only temporary existence. They translate the fact that many researchers at some point explore a new direction (associated with new keywords), and that the evaluation made by colleagues (here measured through their take-up of keywords) was that it was not relevant at this stage. The previous approach did not consider them at all (which can be acceptable over a short period of time), but for a 20 year coverage associated with a 14% yearly growth rate, it becomes difficult to forget all the explorations made that did not prove fruitful (at least at the time of analysis): this would drastically reduce our understanding of de facto dynamics. It would simply forget, in a fast growing emerging field, all the attempts that are made to progressively structure its dynamics. And if we follow Bonaccorsi that large exploration pattern is characteristics of all 'new' fields of science. This is why we have added a "dynamic extension". We now present the two extensions in turn.

Step 3: Static extension query

Step 3 aims at enlarging the dataset around the central dynamics observed in the corset produced, the nanostring.

- Defining the external specificity of multi-terms

We define external specificity as a ratio representing the occurrence of a given multi-term in the nanostring compared to its occurrence in the whole science. This is done by calculating, multi-term by multi-term and year by year, the number of articles that appear in the whole WoS[6]. The external specificity ratio of a multi-term is thus calculated yearly. We use their mean over the 20 years of the database for the static extension. All candidate multi-terms are then ranked by their mean external specificity ratio.

- Selecting relevant multi-terms

Our next challenge is then to decide where to cut on the level of external specificity, thus deciding on a threshold above which multi-terms are considered as relevant and selected for downloading new articles. Looking at the literature does not give any robust indication on how to proceed. We decided on a two-step procedure. First, we considered that a persistent term translating a successful aggregation of knowledge has to be

[6] For this we use only the main form (that is the most frequent form) that appears in the N candidate multi-term stems. This is all the more feasible that the WoS, through its interface, operates a lemmatisation that de facto enables to retrieve the majority of the other forms identified, keeping the order of terms in multi-terms.

central for a minimum number of years. We translated this in one central criterion: it must be within the 250 terms with the highest termhood in the different years of presence. This drove to a first selection of 1105 different terms (from the 3930 overall vocabulary and out of a theoretical possibility of some 23500 multi-terms). The second step was to decide upon a threshold. First tests were made on the Web of Science to have an idea of what different thresholds mean: they showed that a threshold of 20% would in theory bring 1.5 million articles (nanostring included), a threshold of 25%, 990000 articles and a threshold of 30% 745000 articles. This reinforced us in our approach to match in size the theoretical addition to the nanostring. For doing so we used our list of terms ranked by declining levels of specificity and measured what each multi-term could theoretically bring (i.e. the expected increment is the total occurrences in the WoS less those of the nanostring). We stop when the theoretical level matches this of the nanostring (that is 517000 theoretical additions)[7]. This drove to an effective external specificity threshold of 26% brought by 114 multi-terms that represent the static extension (see box for the characterisation of the static extension). The effective number of new articles was of course far lower: when taking into account duplicate articles (similar articles attracted by two different multi-terms), it de facto increased the seed by a factor of 1.65, adding 330000 articles to the 517000 articles of the nanostring.

Box 5- Positioning the static extension

A Preliminary note: arriving to the effective static extension

When operating the extension, we decided not to exclude any 'nano' term, and thus not to consider potential exclusions of not-relevant nano terms (such as nanomolar) (Only the 'nano' standing alone was excluded). It gave 210 'raw' multi terms.

The second step is to consider the check that is conducted on all 'nano' terms (see box 2). This concerns 73 multi terms (that theoretically overall bring more than the effective nanostring, 590000 potential articles vs. 517000 effective ones). This brings only, as mentioned in box 2, 777 potential new articles, showing that the choice made for simplifying the nanostring was quite efficient.

The third step done (afterwards to characterise the effective extension) is to check for the excluded vocabulary: we in fact find in the raw static extension 24 terms, 19 being fully specific and 4 (linked to subnano* in abstracts only) adding 2790 potential articles. This also provides a measure of their presence in the nanostring – a theoretical total of 26000 articles out of which 71% are linked to multi-terms associated with nanomolar and 19% to multi-terms associated with nanogram.

The effective extension is then built on 114 multi-terms that could theoretically add some 497000 articles.

B Characterising the effective static extension

[7] The technial choice made for extracting articles was to use the possibilities offred by the WoS for mumti term words, that is using NEAR/0 that activates lemmatisation; we also have been careful not to accept transitivity in multi-terms; each multi-term has thus a query that rejects the reversed format, as shown in the following example for chemical deposition and for year 2007: TS=((chemical NEAR/0 deposition) NOT ("deposition chemical")) AND PY=2007

The distribution is very skewed showing that only a few multi-terms bring the core of the theoretical expansion: 6 terms bring 51%, 13 terms 66%, 21 terms 75% and 44 terms 90%. It means at the other extreme that 29 multi-terms together bring less than 1% of the theoretical extension,

The thematic orientation of multi-terms is revealing:
- 30 multi-terms deal with observation, manipulation and control techniques (TEM, AFM, STM, NSOM) and make the majority of the theoretical extension (57%).
- The second group concerns materials: TIO2, CDS, graphene, (nano)porous AAO, carbon based nanotubes & quantum-based (dots, wire…): it gathers 37 multi-terms and altogether 23% of the theoretical extension.
- The third group is linked with the characteristics/properties and characterisation of materials, molecules or genes at the nanoscale: it gathers 36 multi-terms and 12% of the theoretical extension. Finally, and contrary to the dynamic extension (see below) there are few multi-terms dealing with fabrication / expression techniques (11 multi-terms bringing 8% of the theoretical extension).

A third characteristic is linked with their presence over time. Tables 2 and 3 below show that on average nano-based terms (our 73) have been present for nearly 18 years and non nano-based ones (our 114) for one year more, whatever level of presence. There is a progressive appearance of terms during the first decade (starting at 47% in 1991, standing at 84% in 1995 and being all but one present in 2000. For instance, we already speak of nanofabrication in 1991 and carbon nanotubes appear in 1992, as does graphene (20 years before the Nobel price).

Moving from the theoretical to the effective extension drives to a severe reduction in new articles, due to a high level of articles containing more than one multi-term: the static extension is only made of 332000 different articles and represents 28% of the total dataset, multiplying the core set by only 1.65.
The effect of the static extension varies strongly with time: it increases the nanostring by a factor of 5 at the beginning (1991) and this multiplier strongly decreases over time, being below 1 in 2002, below 50% in 2006 to end at 30% on 2009-2010.

Tables 2 and 3: time composition of the static extension

Years of presence	Total	20	19	18	17	16	15	14	13	12	Average
Nano-based extension	73	31	6	6	11	5	6	3	2	3	17,8
Non nano-based extension	114	67	9	12	8	7	2	2	4	3	18,6
Total	187	98	15	18	19	12	8	5	6	6	18,3

Date of presence of multi terms	Total	1991	1995	2000	1991	1995	2000
Nano-based extension	73	21	58	72	29%	79%	99%
Non nano-based extension	114	67	100	114	59%	88%	100%
Total	187	88	158	186	47%	84%	99%

Step 4: Dynamic query

The characteristics of the static extension show the interest of having a more refined extension looking at explorations made year by year. Though many of the selected terms do not display a significant presence over the whole period (measured both through presence and internal specificity), they nevertheless have been strong in some specific

years. The principle of the dynamic extension is to mobilise them for expanding the corpus only for those years where they have had a strong presence and provided they show also a relevant external specificity.

The starting point of the approach is similar to this adopted for the static extension but based on all terms (less those already selected for the static extension), i.e. 4189 terms minus the 210 terms of the raw static extension. We then calculate their external specificity, but here to avoid too brutal variations we use three-year moving averages. This also gives us year by year their expected theoretical increment to the dataset (the overall number of articles in the WoS minus the articles in the nanostring).

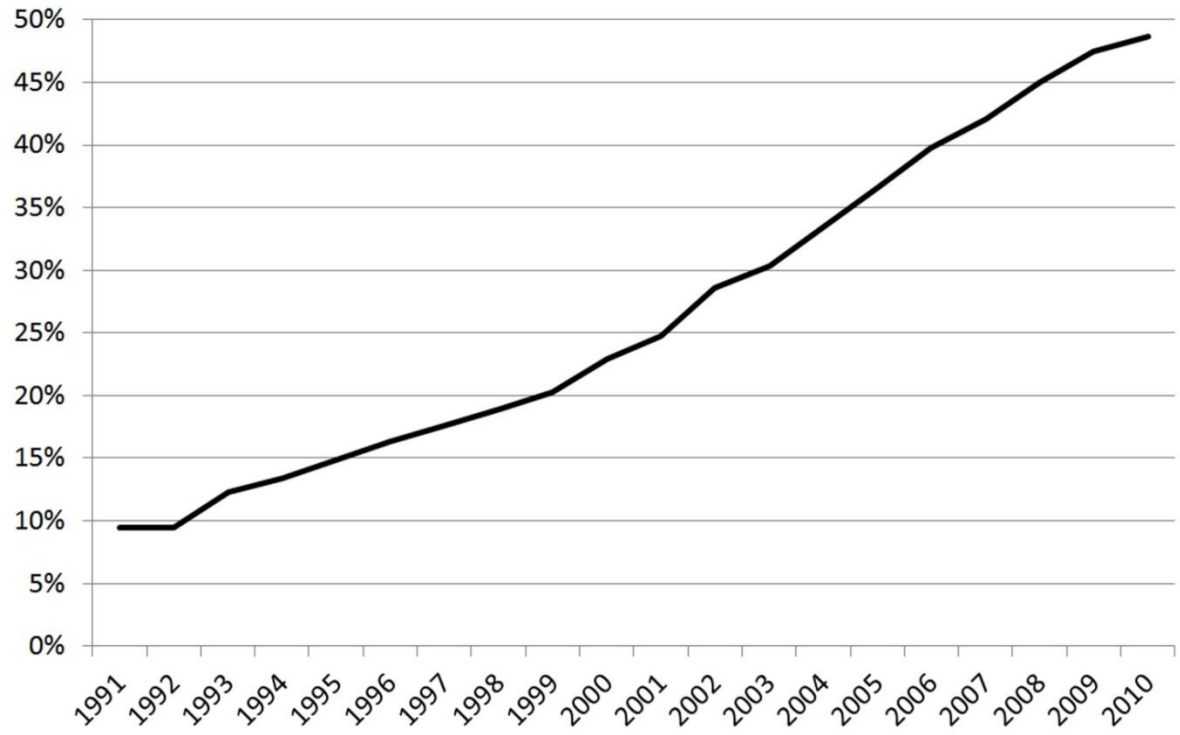
For moving to the next step, we considered another result of Bonaccorsi (2010) on computer science. He shows that over time the exploration does not diminish and that the rate of renewal of keywords does also not diminish overtime; only does the selection by colleagues become harsher, most terms remaining orphan (i.e. with very low uptakes). This means that we should be careful not to reduce the level of exploration over the years. This has driven us to adopt a yearly approach to our principle of a theoretical tripling of the nanostring balanced between the static and the dynamic query. As for the static query we thus look for a theoretical doubling of the nanostring (adding 517000 potential new articles). However, contrary to the static extension, we do not do it over the whole period, but year by year[8].

This drives to calculate the nanostring for each year, defining for each given year the theoretical number of publications that need to be extracted through the dynamic extension. We then go back to the yearly list of multi terms ranked in descending order of external specificity. Adding the potential additions term by term, we define the last term to be included to match the nanostring that year. This enables to identify the external specificity threshold that needs to be applied for the corresponding year. We retain, for this year, only the multi terms above this threshold to download articles.

A key feature of the dynamics is that with time (and with the fast rise of publications), the threshold will increase year after year: it moves from 9% in 1991 to 23% in 2000 and 49% in 2010 (see graph below).

[8] As we stop just below the multi-term that trespasses the annual quantitative threshold, the implication of this repetition over 20 years is that the de facto total is just under 500000 potential new articles, and not near to 517000 articles.

Graph 2: yearly external specificity threshold for the dynamic extension



This process drives to a selection table that crosses multi-terms and years. We arrive to 742 different multi-terms. Box 6 provides a detailed analysis of the composition. We had few multi-terms to exclude that are furthermore only concentrated on the first years of the extension. Like in the static extension, the check done on nano-based terms (171) shows that the simplification adopted in the query for the nanostring is relevant. And we end with 558 different multi-terms appearing on average just over 5 years. In itself this is an interesting validation of Bonaccorsi's hypothesis about wide ranging exploration. A second important finding is that the number of terms appearing in one year increases regularly, representing at the end of the period (2010) 40% of the selected vocabulary: this reinforces the discussion engaged by Arora et al. (2014) about the progressive enrichment of a 'common nano-technology lexicon'. One interesting feature is to consider the typical sequences observed over the 20 years of analysis (table 4). Box 6 also shows interesting differences between on one side the overall vocabulary (gathered in 7 major themes) and the 'core' vocabulary that gathers 90% of the potential extension, and on the other side between the static and the dynamic extension with one clear central difference, the former privileging observation/manipulation techniques and the latter fabrication/production ones.

Table 4: typical patterns of yearly occurrences of multi-terms in the dynamic extension; the 20 most frequent patterns

Hash	NbMainForm	NbConcecutYear	FristYear	LastYear
00000000000000000011	38	2	2009	2010
00000000000000000111	24	3	2008	2010
00001110000000000000	24	3	1995	1997
00000000000000001111	23	4	2007	2010
00000000000011111111	22	8	2003	2010
00001111000000000000	21	3	1996	1998
00000000000000011111	19	5	2006	2010
00000000000000111111	19	6	2005	2010
00011100000000000000	17	3	1994	1996
00000000000001111111	17	7	2004	2010
00000000000111111111	16	9	2002	2010
00000000000011100000	15	3	2003	2005
00000000001111111111	15	10	2001	2010
00000000011111111111	11	11	2000	2010
00000011100000000000	11	3	1997	1999
00000000000111000000	10	3	2002	2004
00000000000000001110	10	3	2007	2009
00000000011100000000	9	3	2000	2002
00000000111111111111	9	12	1999	2010
00000000001110000000	9	3	2001	2003

Box 6 – A detailed analysis of the dynamic extension

The year-by-year selection process of relevant couples (multi term x year) drives to 742 different multi terms selected.

a) Excluded terms only appear at very low levels of specificity thresholds, between 1991 and 1995
There are 8 different terms building 36 couples term-year selected and representing 440 occurrences in the nanostring. Only 12 add 1298 potential new articles (specificity below 1) with only 4 couples linked to “micromolar” bringing 75% of the total.

b) The testing of multi-terms containing ‘nano’ gathers 171 terms representing 1308 couples term-year, an average just under 8 years of appearance.

The test shows once more the relevance of the simplification made for the 'nanostring' since these terms appear nearly 164000 times in the nanostring, while they only generate 243 potential new articles (thus linked to terms only present in the abstracts).
Looking more in detail on the dynamics of terms, we see a fast increase from an average of 7 terms only in 1991-92 to 70 in 1995 and a peak of 90 terms annually between 2002 and 2005, before going down to 70 terms on average between 2006 and 2010.
We have organised words by main themes in order to measure their respective importance and follow their dynamics (Table 5). This shows three interesting results.
First in term of composition: materials mobilised come first (31% of presence) with measure (21%) and characterisation dimensions (18%). Both these terms share an important feature: their importance reduces in relative terms between the two decades observed, in favour of terms dealing with application (still limited in importance 9%) and even more vis-à-vis terms dealing with three dominant types (tubes, wires and films, 20% of total presence and nearly 80% in the second decade).

Table 5 – the nano-based vocabulary of the dynamic extension

Themes	Terms Nber	Terms %	Presence Nber	%	Share of 2nd decade
Nanomaterials (gold…)	49	29%	406	31%	65%
Nano tubes/wires/films	39	23%	268	20%	79%
Nano applications (fibers, powders…)	20	12%	121	9%	67%
Characterisation	31	18%	239	18%	53%
Measure	32	19%	274	21%	40%
total	171	100%	1308	100%	61%

Box 6 continued

c) The dynamic extension per se is made of 558 multi-terms representing 2856 couples term-year, just over 5 years of presence per term on average.
We witness an interesting evolution over time: the number of multi-terms per year compared to the total population (558) increases at the same time the external specificity threshold does (see graph 3): it starts with 1% of the total vocabulary in 1990-91, is around 20 to 25% between 1996-2000, then moves to an average of 31% between 2001-2005 and to 41% in 2008-2010.
An interesting feature is linked to the life cycle of multi-terms depending upon the fact they emerged and died in the first decade (27%), they emerged after 2000 (52%) or they emerged during the first decade and went on in the second decade (21%). Their respective life cycle was 2.9 years for the first, 3.9 years for the second and 8.4 years for the third.
As for the static query there is a clear concentration effect: the first 100 couples bring 42% of the potential extension, the following 100 13%, the following 300 19%. The last 2000 couples only bring 14% of the total potential extension.

The composition shows interesting features compared to the static extension (table 6)
- Observation/manipulation techniques play an important role (12% of terms, 14% of total presence) as in the static extension but to a lesser degree (26% in the static extension). This is

the exact reverse for production/fabrication (16% of terms and of presence in the dynamic extension, vs 8% in the static extension).
- Materials (21%) complemented by nano tubes/wires and films (6%) are less important than in the static extension (32%)
- A clear difference between the static and the dynamic extensions lies in the richness of the measurement and characterisation vocabulary, respectively 33+8% and 32%; while applications only appear in the dynamic query but at a marginal level (4%, and 10% if we include nano tubes, wires and films).
The difference is even wider with the nano-based dynamic extension (see above) that has nearly no term dealing with observation, manipulation and production / fabrication techniques, a very different balance between measures and characterisation, and nearly 50% of terms associated with materials and nanotubes/wires and films.

To better grasp the role of the multi-terms in the dynamic extension, we have selected all the terms that potentially bring more than 500 articles, i.e. 162 terms out of the overall 558. Together they potentially bring 442000 articles, compared to an overall total of 492000 potential articles (90%), once excluded multi-terms have been excluded and once account is taken of the selection process implemented.
This is illustrative of the difference between the overall vocabulary and the core vocabulary that generates significant numbers of new articles (table 7): most terms related to nanotubes (without the term nano) do not generate any significant number of articles (they are all in the nanostring). Observation, manipulation, production and fabrication techniques represent overall 28% of the vocabulary; their role in generating articles in far more important (39% of the key vocabulary and 47% of total articles). On the contrary characteristics and properties represent only half of their share of the vocabulary (17% vs 33%) bringing only 14% of total potential articles.

Finally, the static and the dynamic extensions share in common the importance of observation and manipulation techniques, but levels differ widely: 57% of the total potential static extension against only 19% for the dynamic extension. This is counterbalanced by the contrasted importance given to fabrication techniques (respectively 8% and 28% of the respective potential extensions). Both extensions share a near to similar importance given to materials (respectively 23 and 28%) and to characterisation (respectively 12 and 14%)..

Table 6 – A thematic analysis of the vocabulary of the dynamic extension

Themes	Terms nber	Terms %	Presence nber	%	Years pres
Observation/manipulation techniques	69	12%	395	14%	5,11
Production / fabrication processes	88	16%	451	16%	5,13
Materials	116	21%	573	20%	4,94
Nano tubes, wires, films, ribbons	28	5%	176	6%	6,29
Applications	29	5%	119	4%	4,1
Measures	44	8%	337	12%	7,66
Characterisation	184	33%	805	28%	4,38
Total	558	100%	2856	100%	5,11827957

Table 7- Core vocabulary generating 90% of the expected dynamic extension

	Terms nber	Terms %	Articles nanostring	Potential articles	%
Manipulation observation	32	20%	93374	86035	19%

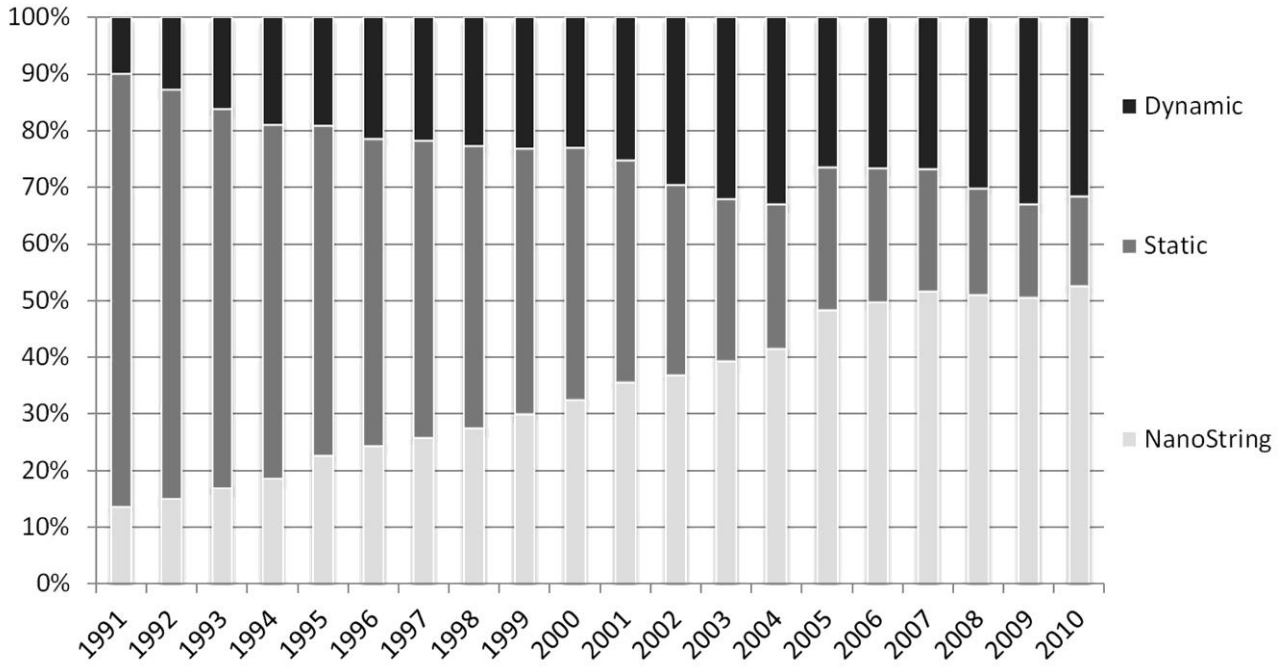
Production fabrication	30	19%	108120	123419	28%
Applications	12	7%	11778	13924	3%
Materials	38	23%	125363	122284	28%
Measures	22	14%	38331	34748	8%
Characteristics/ properties	28	17%	63815	62083	14%
	162	100%	440781	442493	100%

Box 7 gives the main characteristics of the effective extension and its role in the overall dataset. One interesting feature is the opposite relative roles of the static and the dynamic extensions over time in an overall dataset where the nanostring has regularly increased in importance to over half of the total since 2006: while the role of the static extension moves from 75% to 15% in 20 years, the dynamic one starts with 10% to finish with 30%. Graph 3 also shows that since 2002-3 the role of the dynamic extension remains stable while the relative growth of the nanostring is linked with a regular decrease of the static extension, as if the common vocabulary over the period is less and less relevant to define the new dynamics at work.

Box 7 – The effective dynamic extension – Main figures

This dynamic extension has been conceived to double the nanostring as was the static extension. Redundancy in the characterisation of articles is far less than expected, bringing the overall total of new articles included to 332000, nearly the same amount brought by the static extension, representing 28% of the whole dataset.
However this 64% increase is obtained very differently than for the static extension: it plays a minimal role in the overall dataset at the beginning of the period moving from 10% in 1991 to an average of 22% in 1996-2000 and then oscillating around 30% since.

Graph 3- the nano DB: evolution of the role of respective layers 1991-2010



4-Conclusion

The ambition of this article is to propose a new automatic evolutionary lexical query to address emerging fields. This query is made of a core component based upon the central keywords associated with the emerging field (in our example "nano"), and of two extensions that tap on one side the progressive 'stabilisation' of the field, and on the other the continuous exploration that characterises 'new dominant sciences' to follow Bonaccorsi.

This new approach follows our previous one (Mogoutov and Kahane 2007) taking advantage of three developments in primary datasets (in particular the new lemmatisation capacity offered by the WoS), in new approaches and software to analyse contents and extract relevant multi-terms, and in power computing that enabled to move from tens of thousands to millions of units of analysis. To circumvent limitations in our previous query we had to develop a modular approach to extension, while here we propose one, which does not require any ex-ante content choice.

We have made key choices that require further discussions within the community. We think that extension beyond a core set is critical since in an emerging field, both established categories poorly address the emerging field, and since also the vocabulary being not stabilised there is enormous variation in central keywords used for positioning the emerging field. But other studies have reduced their coverage to the core set alone or limited expert-based extensions. Arora et al. (2014) show that after 20 years of development, the scope and variety of the nano-based vocabulary is such that we might have a good image of the present dynamics only using it. We share their results but not the conclusions: we think that this drives to lose all the explorations made in the way, and thus gives a limited image of the effective 'search regime' and we think, always

following Bonaccorsi and his conclusions on computer science, that the 'nano' vocabulary has all chances to miss most of the on-going exploration at the present and still instable frontier of the field. This is why we consider critical to keep extensions until a field is fully institutionalised. (Remember that following existing categorisations, for instance in comparing public research organisations - cf. science metrix 2013 on European PRO - drives to measure the relative performance of different organisations in a disciplinary framework that ignores all new fast growing fields). However the nature and the level of the extension to be made, remain to be discussed. Here we have proposed to differentiate between two types of extensions: a 'static' and a 'dynamic' extension. The former takes hold of those aspects that are 'core' to the emerging field over the whole period of observation, while the latter reflects the variety and multiplicity of explorations made about the potential content and directions of the new field. We think that the results exposed above clearly demonstrate the utility of this dual approach. What remains important to discuss is the extent of the extension. We have read widely and have found no satisfying answer, and often no discussion at all, about this level. Taking work done by the main teams in nano science and technology, we have arrived at an empirical estimate of tripling the initial seed. And we have proposed two complementary methods that we consider relevant for both the static and the dynamic extensions. There is thus further research to be done to better address this question. Meanwhile, if our pragmatic solution is considered satisfactory, we offer a fully reproducible method for any new emerging field.

References

Arora S.K, Porter A.L., Youtie J., Shapira P, 2013, capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs, *Scientometrics*, 95,1, 351-370.

Arora S.K, Youtie J., Carley S., Porter A.L., Shapira P., 2014, Measuring the development of a common scientific lexicon in nanotechnology, *Journal of Nanoparticle Research* 16, 2194, DOI 10.1007/s11051-013-2194-0

Bonaccorsi, A., 2008, Search regimes and the industrial dynamics of science, *Minerva*, 46, 285-315

Bonaccorsi, A., & Vargas, J., 2010, Proliferation dynamics in new sciences, *Research Policy*, 39, 8, 1034-1050.

Braun T., Schubert A., Zsindely S., 1997, Nanoscience and nanotechnology on the balance, *Scientometrics*, 38, 2, 321-325.

Frantzi, K., & Ananiadou, S., 2000, Automatic recognition of multi-word terms: the C-value/NC-value method, *International Journal on Digital Libraries* , 3.2, 115-130

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

Glanzel W. et al., 2003, *Nanotechnology, analysis of an emerging domain of scientific and technological endeavour*, KU Leuven, Leuven, July, 73 pages.

Grieneisen M.L., Zhang M., 2011, Nanoscience and nanotechnology : evolving definitions and growing footprint on the scientific landscape, *Small*, 7, n°20, 2836-2839.

Huang C, Notten A, Rasters N, 2010, Nanoscience and technology publications and patents: a review of social science studies and search strategies, *Journal of Technology Transfer* 36, 2, 145–172.

Kageura, K., & Umino, B. (1996), Methods of automatic term recognition: a review. *Terminology*, 3(2), 259–289.

Kostoff R.N., Murday J.S., Lau C.G.Y., Tolles W.M., 2006, The seminal literature of nanotechnology research, *Journal of Nanoparticle Research* 8, 2, 193–213.

Kostoff R.N., Koytcheff R., Lau C.G.Y., 2007, Global nanotechnology research metrics, *Scientometrics*, 70, 3, 565-601

Larédo P., Delemarle A., Kahane B., 2010, Dynamics of nanosciences and technologies: policy implications, *STI Policy Review* 1, 43-62.

Leydesdorff L. and Zhou P., 2007, Nanotechnology as a Field of Science: Its Ddelineation in Tterms of journals and patents, *Scientometrics,* 70, 3, 693-713

L'huillery S., Raffo J., Foray D., 2010, le positionnement et les perspectives stratégiques des nanotechnologies en France, *rapport pour le Ministère de la recherche et de l'enseignement supérieur*, EPFL, Février, 188 pages.

Mogoutov A., Kahane B., 2007, Data Search Strategy for Science and Technology Emergence: A Scalable and Evolutionary Query for Nanotechnology Tracking, *Research Policy*, 36, 6, 893-903.

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*, Leiden University & Fraunhofer ISI, October, 114 pages.

Porter A.L., Youtie J., Shapira P., Schoeneck D.J., 2008, Refining search terms for nanotechnology, *Journal of Nanoparticle Research* 10,5, 715–728.

Van Eck, N. J., & Waltman, L.,2011, Text mining and visualization using VOSviewer. *Arxiv preprint arXiv: 1109.2058*.

Youtie J., Iacopetta M., Graham S., 2008, Assessing the nature of nanotechnology: can we uncover an emerging general purpose technology, *Journal of Technology Transfer*, 33, 315-329.

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

Zucker L., Darby M., Funer J., Liu R., Ma H., 2007, Minerva unbound: Knowledge stocks, knowledge flows and new knowledge production, *Research Policy*, 36, 6, 850-863.