RISIS RESEARCH INFRASTRUCTURE FOR RESEARCH AND INNOVATION POLICY STUDIES

Report on the content and technical structure of the

Nano S&T Dynamics

Infrastructure

Université Paris-Est Marne-la-Vallée

RISIS "Research infrastructure for research and innovation policy studies"

FP7, Grant agreement no: 313082

Task 6, Workpackage 6, coordinated by AIT Austrian Institute of Technology Gmbh

SEVENTH FRAMEWORK PROGRAMME

Report on the content and technical structure of the Nano S&T dynamics database (Task 6 of WP6)

Foreword

The report presents the database dealing with nano science and technology dynamics (Nano DB) that will be opened as part of the RISIS infrastructure. The Nano DB is an original database combining data on publications and patents.

The data is retrieved from the Web Of Science (scientific publications) and from Patstat (patents) through an original dynamic query developed by IFRIS. It gathers 1 182 344 publications and 735 834 priority patents between 1991 and 2010. Three enrichments have been produced dealing with affiliations (type of organisation and harmonisation of institution's names); geolocalisation of all addresses of authors and inventors; geographical aggregation (clusters).

In order to provide comprehensive information on the content and technical structure of the Nano DB this report follows the common structure adopted for RISIS reports: basic characteristics, information on substantive content, legal issues, technical structure, and further developments preparing the opening of the Nano DB.

A specific appendix presents the methodology developed for the dynamic query.

Table of contents

Figures and tables

Report on the content and technical structure of the *Nano S&T dynamics* dataset (Task 1 of WP6)

1 Basic characteristics

a) Name and short description of the infrastructure

The dataset focuses on nano science and technology, considered by many analysts of science dynamics as the new leading science of the time (Bonaccorsi, 2008). The dataset, developed by IFRIS, gathers scientific publications (1991-2010) and patents filed between 1991 and 2009. This represents 1182344 publications (out of which 979517 articles) and 2682429 patent applications (out of which 735834 priority patent applications, i.e. 5% of total world production over the period).

The dataset is organised around 3 major dimensions:

- Organisational with the affiliations of authors and patent grantees
- Geographical dealing with authors, grantees and inventors based on addresses (countries, cities and clusters)
- Thematic based on the subject categories of the WoS, and on technological specialisations of patents.

b) Aim of the database (context of data acquisition)

The focus of the dataset is to develop all the approaches, methods and techniques to identify, characterise and analyse the dynamics of emerging and fast growing fields of knowledge. This is applied on nano science and technology as the new "dominant science" (understanding matter and phenomena at the nano scale). It is warranted by numerous authors as part of the 'converging sciences' and as the new 'general purpose technology' of the time. It is in most countries a strong component of research and innovation policies, from the regional to the international level (in particular at European level).

One key methodological stake is that the perimeter of data collected needs to evolve periodically as the "domain" progressively structures itself. This has driven the producers, after having developed a first 'static' query (see Mogoutov and Kahane 2007) to develop a new dynamic query, enabling to capture both the stabilised part of the field and, year after year, to adapt to the evolutions of the scientific and technical vocabulary produced. Its objective is to cover both the central established core and its developments, and the on-going explorations (many of which will be short lived but are essential to understand on-going dynamics).

c) Legal name of operating organization

UPEM - IFRIS

d) Database location and type of access

The Nano database is hosted on a MySql server at IFRIS and is accessible on site at IFRIS at Marne-la-Vallée (France).

2 Information on substantive content of the *Nano S&T dynamics dataset*

The following section describes the two units of analysis – patents and publications – and explains the selection process through which the dataset is built (i.e. the lexical query developed which is fully presented in appendix 1). It will then enter the three harmonisation processes entered into: geographical, institutional and thematic. These harmonisation processes are critical since they enable to link both units (patents and publications) at the geographical, institutional and thematic level to describe overall dynamics.

2.1 Definition and description of observations

a) Units and definition of observations

Patents

The database gives information related to priority patent applications anywhere in the world to protect an invention. The priority date is used to determine the novelty of the invention (it is an important aspect in patent procedures). For statistical purpose, the priority date is the closest date to the date of invention.

Scientific publications

The database gathers publications that correspond to the domain of nano sciences and technologies in journals that are indexed in the web of science and give access to titles, abstracts and author keywords. Their selection is linked to the query developed see below and annex.

b) Number of observations

The dataset contains 1182344 publications (with 2179065 addresses, i.e. on average 1.85 addresses per publication, the focus being on organisations and places not on individual authors). For patents the dataset includes 2682429 patents (with 8832556 inventors). Our analyses focus on priority patents: 735834 over the period with 1834604 inventor addresses (i.e. 2.5 inventor addresses per priority patent).

2.2 Data acquisition and processing (e.g. data cleaning)

a) Where are the data retrieved from

Patents

We used the PATSTAT-IFRIS database, built on PATSTAT version October 2009 provided by EPO (see next section for detailed description on data gathered).

Scientific publications

Data is retrieved from the Web of Science [1] and, within it, the following sources have been interrogated:

- Science Citation Index Expanded™ (SCI™ Expanded): Science Citation Index Expanded is a multidisciplinary index to the journal literature of the sciences. It fully indexes over 6,650 major journals across 150 scientific disciplines and includes all cited references captured from indexed articles.
- Social Sciences Citation Index® (SSCI®): The Social Sciences Citation Index is a multidisciplinary index to the journal literature of the social sciences. It fully indexes over 1,950 journals across 50 social sciences disciplines. It also indexes individually selected, relevant items from over 3,300 of the world's leading scientific and technical journals.
- Arts & Humanities Citation Index® (A&HCI®): Arts & Humanities Citation Citation Index is a multidisciplinary index to the journal literature of the arts and humanities. It fully covers 1,160 of the world's leading arts and humanities journals. It also indexes individually selected, relevant items from over 6,800 major science and social science journals.
- Conference Proceedings Citation Index - Science (CPCI-S) This citation index covers conference literature in all scientific and technical fields.
- Conference Proceedings Citation Index - Social Sciences & Humanities (CPCI-SSH): This citation index covers conference literature in all fields of social sciences, arts, and humanities.
- Index Chemicus® (IC®): Index Chemicus contains the structures and critical supporting data for novel organic compounds reported in leading international journals. Many records show the reaction flow from starting material to final product. Index Chemicus is a vital source of new information on biologically active compounds and natural products.

[1] http://images.webofknowledge.com/WOK46/help/WOK/h_database.html

b) *How are the data processed in terms of data cleaning (e.g. harmonisation of organization names, etc.)*

The query developed

Scientific publications

Most queries start with a seed made of a few keywords that are core to the domain under study – here it is linked like other queries to the prefix 'nano' (with classical exceptions like nanoliter). This seed has generated 517050 publications. Semantic analyses of keywords are made on this seed identifying their internal specificity (on the core dataset) and then checked about their external specificity (on the whole WoS). Articles are then extracted on the basis of the vocabulary selected. This is done on the whole period covered (1991-2010) giving what we call the 'static' extension. And this is done year after year giving annual vocabularies to download the 'dynamic' part of the dataset. Appendix 1 develops in detail this methodology and the different steps to make it operational and robust. As shown in the table the nanostring represents 44% of the dataset while each extension (static and dynamic) represents 28%. Overtime the share of the nanostring increases from 14% in 1991 to 50% in 2006. Since it has remained between 50 and 55%. Overall the growth rate until 2008 is 14% but goes down below 10% over the last two years.

Layer	*Number of publications*
Nanostring	517 050
Static	332 739
Dynamic	332 555
Total	***1 182 344***

Fig & tab 1. *Scientific publications: overview*

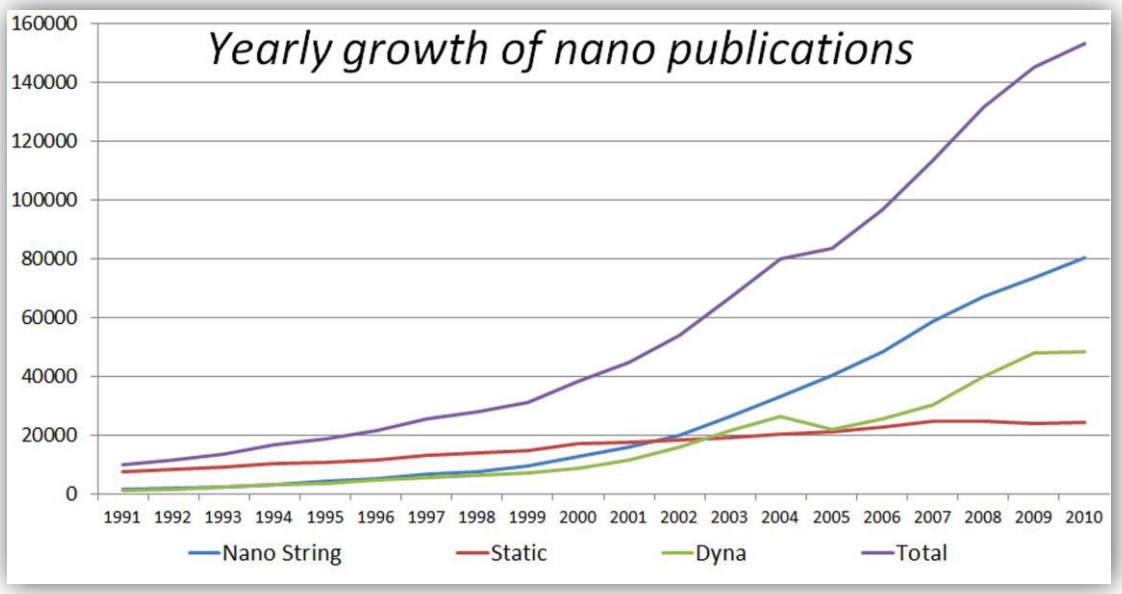


Fig & tab 2. *Growth of scientific publications*

Patents

The present version of the dataset did not develop a new approach but used the vocabulary defined for publication for both static and dynamic extensions. A recent test was made on one year (2005) comparing this approach with the reproduction of the methodology applied for publications on patents. It highlighted a very different core vocabulary but when we compared the 2 sets of patents selected there was over 85% of overlap. Thus, though this is not optimal, the present approach gives a quite fait account of the dynamics in technology.

Note: We plan in the next version of the dataset (end of 2015) to apply the full methodology also for patents.

The nano patent dataset is made only of priority patents. Three aspects need to be noted:

- There is a long delay between the application and the publication by national patent offices of that application (this is linked to the handling of patents by patent offices ; but there are also in some offices like the US patent office, possibilities for firms to ask for a longer delay for publication). This explains why, even if we cover up to the end of 2009, we have only probably half the patents for 2008 and nearly nothing for 2009.
- A specific dimension of patents lie in the existence of interconnected patents through 'families' built upon the connection made by grantees between the new patent and existing patents (for a more in depth definition see the information presented on Patstat in the other report produced by IFRIS on the dataset on large firms, CIB). Families are important since they represent a significant share of the total patents considered: 49% overall, but this percentage goes down regularly: from 71% in 1991 to 36% in 2006-07.
- The pattern of growth is very different compared to the growth of scientific publications. Overall we have a fast growth from 1991 (around 15000 patents) to 2000 (with just over 60000 patents). Then the annual number has rather decreased than increased moving to around 55000 since 2003.

Layer	*Numer of priority applications*	*Total of applications*
Nanostring	44 955	96 025
Static	148 267	467 872
Dynamic	181 635	842 132
Total	***374 857***	***1 406 029***
FamInpadoc	360 977	2 682 429

Fig & tab 3. *Patents: overview*

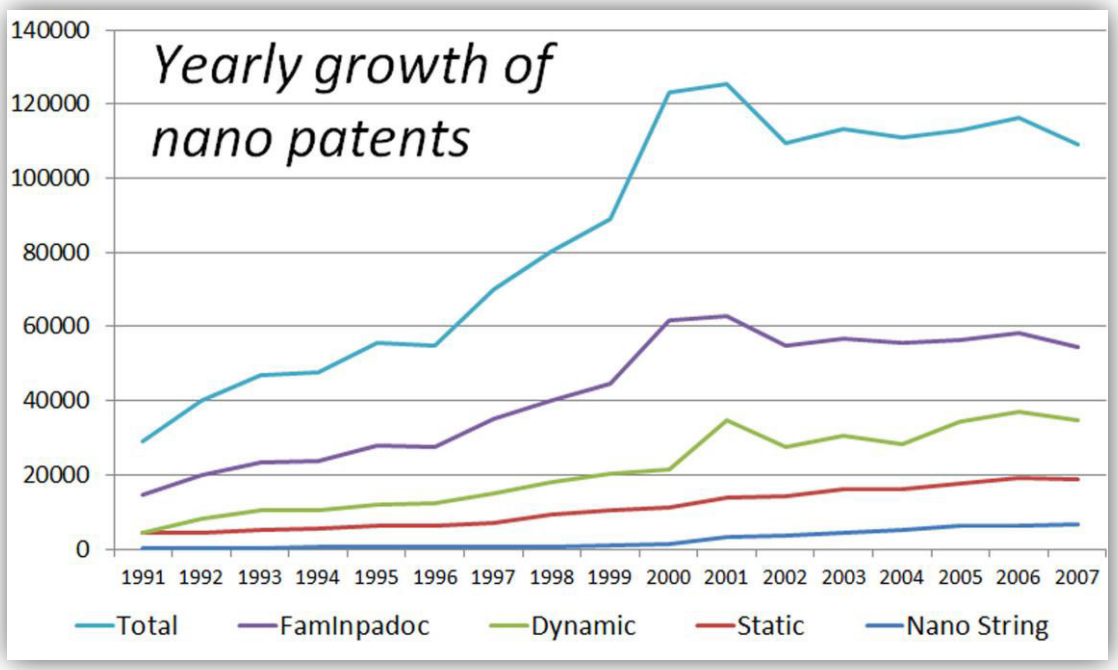


Fig & tab 4. *Growth of patents*

Harmonising and categorising institutions

Here we face an interesting paradox. While the information available has witnessed an explosive growth, we still remain very limited in our ability to follow the involvement of organisations whether public and private. We have shown with the CIB dataset (see its presentation report) how difficult it was to rebuild the large firms from the hundreds of legal names under which they patent, either because writing differs from one patent to another, because firms change names and even more because many of their subsidiaries patent. Similar issues are raised for public institutions often associated with multiple spelling, even after the standardisation efforts made by databases (in particular publications and the differences between legal institutions and their schools, departments, institutes and research groups).

The nano database capitalises years of work on multiple projects. They are grouped under two complementary aspects:

- Automatic treatments: we mobilise recognised algorithms to compare chains of characters and when similarity is considered large enough, to associate them to a unique entity
- Construction of a reference dataset: these automatic treatments have been checked, validated and complemented manually in a number of projects, being progressively integrated in a reference dataset that is enriched project after project.

The way to characterise the institutional dimension differs widely between patents and publications. In publications it is part of the address and builds the first two or three elements of the address (with the lab, the department and/or school, the faculty and the

organisation per se). In patents we have the name of the patent filer as it appears in the patent document and the name as Patstat has standardised it.
We have used a two-step process for cleaning names: first allocation to a type, and then harmonisation of names.

Step one. In order to better understand the roles of the different types of institutions, we have built a categorisation around 5 main types: University, Government organisation, Hospital, Firm and other. Appendix 2 lists the queries developed to allocate addresses to one of these types. One immediate issue is that a given written form may belong to more than one entity. The automatic classification offers as a solution the most frequent one encountered. Manual checking has been used to validate it.
Overall it gives 75% of addresses for universities, 18% for government organisations (mostly but not exclusively public research organisations), 5% to firms, 2% for hospitals and 1% other. There are however two issues pending: one is the split between universities and PRO (especially with joint units, quite common in France but also growing in other countries in Europe), the other is the split between universities and hospitals (when being faced with university hospitals). The pragmatic choice made (until RISIS develops a more robust approach) has been to follow the choice of authors in indicating their primary affiliation.

Step two deals with the harmonisation of names. The 2 million plus addresses for publications gave nearly 100000 different written forms (see table below). We used then two complementary treatments: (a) automatic: we measured the distance between written forms using the Levenshtein distance weighted by a Jaccard distance (see Cohen, Ravikumar & Fienberg 2003 for a comparison between similarity measures). This has enabled to reduce by 27% the number of different forms. (b) We have compared with our reference database complemented by a manual check: the latter has enabled another reduction of 32%, leaving us with some 40000 different 'organisations'.

Nano publications 1991 - 2010		***%***
2 179 065 addresses		
Number of different writing forms	97 194	
1 - Harmonisation based on similarity measures	26 242	27%
2 - Manual harmonisation of written forms	31 187	32%
Total number of harmonised written forms	39 765	41%

Fig & tab 5. *Harmonisation of the name of institutions*

Data geolocalisation

The geolocalisation of data from publications and patents has been done in 3 steps

- Pre-processing: cleaning (getting rid of commas, hyphens and the like) and harmonizing (using approaches described before).
- Patterns extraction: extraction and harmonization of geographical information (cities, postal codes…).
- Geolocalisation: matching with geo-databases or toponymy recognition.

For the last two steps we have used two methods in sequence. First we have compared the geographical data extracted to those of the database GeoName[2]. For bringing addresses together, we have identified recurrent patterns linked to the ways addressed are written in different nations (e.g. the position of names of cities, the number of characters of postal codes) and their translation in regular formats. We have operated a matching per country using three levels: the postal code, the city and the region. We have obtained geographical coordinates (latitude, longitude) and enriched the toponyms linked to our addresses (in particular states in federal countries). All ambiguous situations have been left aside. Still it has geolocalised 91% of total addresses (in publications and patents).

For the 9% left aside, we have proposed the addresses we had to the geolocalisation webservice based (among other) on the Google engine[3]. This search engine offers 9 levels of accuracy (see table below): any level from 4 to 9 has been considered enough to geolocalise addresses. This has enabled to complement the coverage to nearly 99% (1.19% addresses not covered).

Accuracy[4]	*Description*
Value	
0	Unknown accuracy.
1	Country level accuracy.
2	Region (state, province, prefecture, etc.) level accuracy.
3	Sub-region (county, municipality, etc.) level accuracy.
4	Town (city, village) level accuracy.
5	Post code (zip code) level accuracy.
6	Street level accuracy.
7	Intersection level accuracy.
8	Address level accuracy.
9	Premise (building name, property name, shopping center, etc.) level accuracy.

Fig & tab 6. *Geolocalisation accuracy*

The table below gives the percentage of geolocalised addresses for all countries with more than 10000 addresses.

Countries with more than 10 000 author's addresses	*Harnonised country*	*Number of addresses*	*Addresses geolocalised*	*%*
Total for all the 166 countries		***2 176 376***	***2 153 142***	***98,93%***
UNITED STATES	US	471352	471322	99,99%
CHINA	CN	268630	268488	99,95%

[2] Geoname, http://download.geonames.org/export/dump/, 09/2012

[3] Geobatch, www.findlatitudeandlongitude.com/batch-geocode/

[4] Accuracy codes with the web service GeoBatch
https://developers.google.com/maps/documentation/javascript/v2/reference?csw=1#GGeoAddressAccuracy

JAPAN	JP	216934	215834	99,49%
GERMANY	DE	138001	137994	99,99%
FRANCE	FR	109136	109118	99,98%
SOUTH KOREA	KR	101996	85863	84,18%
UNITED KINGDOM	GB	83113	83103	99,99%
ITALY	IT	73211	73203	99,99%
INDIA	IN	57754	57700	99,91%
TAIWAN	TW	56723	56723	100%
RUSSIA	RU	49725	49599	99,75%
SPAIN	ES	49060	49054	99,99%
CANADA	CA	43182	43182	100%
BRAZIL	BR	31320	31319	100%
AUSTRALIA	AU	30501	28327	92,87%
NETHERLANDS	NL	26337	26335	99,99%
POLAND	PL	24480	24479	100%
SWITZERLAND	CH	23739	23737	99,99%
SINGAPORE	SG	21476	21476	100%
SWEDEN	SE	21365	21364	100%
BELGIUM	BE	17875	17870	99,97%
ISRAEL	IL	15025	15025	100%
IRAN	IR	14992	14992	100%
MEXICO	MX	14038	14038	100%
TURKEY	TR	13801	13801	100%
HONG KONG	HK	13099	13099	100%
AUSTRIA	AT	12360	12358	99,98%
CZECH REPUBLIC	CZ	12026	12024	99,98%
FINLAND	FI	11217	11204	99,88%
PORTUGAL	PT	11191	11190	99,99%
GREECE	GR	10574	10536	99,64%
ROMANIA	RO	10474	10452	99,79%
UKRAINE	UA	10005	10001	99,96%

Fig & tab 7. *Geolocalisation of publications*

A similar level of achievement is observed for patents (see table of countries with more than 1000 addresses for inventors), once account is taken of patents without addresses. The retrieval rate is quite weak (overall 31%) even when integrating the different add-ups produced by other organisations than EPO (in particular Regpat by OECD). The holes are concentrated on 3 Asian countries (China 40%, retrieval rate near to nil), South Korea (16%, retrieval rate 2%), Taiwan (2% retrieval rate 36%) and on Russia (5%, retrieval rate 2%). For these countries, there is little to expect but waiting on new versions of Patstat by EPO (one every 6 months). What has surprised us is the high level of holes (no inventor address) for European countries: 40% of total addresses, representing 10% of world holes. There are two ways to address these problems: one is to go back to national databases and see whether a matching by patent number enables

to fill inventor addresses; the other is, in the case patents are part of a family, to develop an algorithm that enables to tap potential addresses. We are developing both and results should be integrated before opening.

Countries with more than 500 inventor's addresses	Inventor's addresses	Filled addresses	%	Geolocalised addresses	%
Total for the 126 countries	***554 855***	***173 521***	***31%***	***172 900***	***98,63%***
CHINA	193074	777	**0%**	761	97,94%
UNITED STATES	120183	115641	**96%**	115582	99,95%
SOUTH KOREA	80869	1306	**2%**	1269	97,17%
GERMANY	40809	4724	**12%**	4714	99,79%
RUSSIA	24470	515	**2%**	504	97,86%
FRANCE	21100	20482	**97%**	20450	99,84%
TAIWAN	18494	6672	**36%**	6452	96,70%
JAPAN	10426	7484	**72%**	7320	97,81%
CANADA	5424	2902	**54%**	2902	100,00%
SPAIN	5068	548	**11%**	547	99,82%
UKRAINE	4903	52	**1%**	52	100,00%
UNITED KINGDOM	4104	1288	**31%**	1284	99,69%
ITALY	2704	1184	**44%**	1179	99,58%
NETHERLANDS	2585	1635	**63%**	1633	99,88%
SWITZERLAND	2173	1541	**71%**	1538	99,81%
BELGIUM	1865	1444	**77%**	1438	99,58%
INDIA	1268	997	**79%**	996	99,90%
POLAND	1207	61	**5%**	61	100,00%
CZECH REPUBLIC	1077	22	**2%**	22	100,00%
ROMANIA	830	13	**2%**	13	100,00%
ARGENTINA	827	13	**2%**	13	100,00%
ISRAEL	818	626	**77%**	626	100,00%
AUSTRIA	773	218	**28%**	215	98,62%
PORTUGAL	693	24	**3%**	24	100,00%
SINGAPORE	686	485	**71%**	483	99,59%
HONG KONG	630	291	**46%**	291	100,00%
MOLDOVA	619	1	**0%**	1	100,00%
SWEDEN	584	442	**76%**	438	99,10%
BULGARIA	576	1	**0%**	1	100,00%
FINLAND	523	404	**77%**	403	99,75%

Fig & tab 8. *Geolocalisation of patents*

Clustering the scientific and technological activities

After the location (identification of toponyms), and geolocalisation (allocation of longitude and latitude coordinates), the phenomena of concentration of activities can be approached through an analysis of geographic clustering. This step aims at analyzing the

distribution of S&T activities and at measuring agglomeration effects by identifying the existing clusters. The ambition is to look at clustering effects as they happen and not by considering administrative borders that widely differ between countries.

The simplest solution is to take agglomerates already constructed for other studies. This is the case in the US of metropolitan areas that are updated by the statistics office periodically. But this does not exist elsewhere. So most studies adopt administrative borders (e.g. NUTS 2 or 3 in Europe which are known not to represent any economic or social reality). This is why when studying phenomena at the world level and wanting to avoid 'political definitions' of space, studies are bound to develop clustering approaches.

A simple solution to clustering is to work only on distance starting with supposed geographical central locations (in this case working at the city level). This approach supposes that we define a threshold about what a relevant central city is (in our case in our first attempt it was 1000 publications in 10 years), and defining a radius (60km reduced to 30km in some Asian countries). A measure of overlaps (say 20% of joint publications) between clusters formed this way drives to agglomerating initial clusters. We used this first technique on the prototype database with quite good results: 203 clusters worldwide agglomerating over 75% of publications.

However this methodology has a number of drawbacks: it misses fully distributed clusters; limits are defined artificially; there is no continuity in term of size of clusters while a central result was their very uneven distribution. This drove to tailor a new method that addresses these issues. We project all addresses in the geographic space and build a first partition using a density algorithm, DBScan (Ester, Kriegel & Sander, 1996), which is parametered using a minimal distance between two points (two couples of coordinates) and a minimal value for building a cluster (sum of addresses in a given point). The borders of the convex hulls are built using these points. These envelopes build the smallest geographical divisions (initial clusters).

Using the documents (publications and/or patents) as units of analysis, we apply the CHAMELEON method (Karypis, Han & Kumar, 1999). It enables to identify collaboration regimes between individuals (authors and inventors) located in given geographical spaces. We then compare nearby initial clusters (i.e. that are less than 100 km distance) two by two, using the internal and external collaboration values of the initial clusters. Two measures are calculated: relative inter-connectivity and relative closeness that enable defining thresholds for agglomerating clusters (see box 1 for their calculation).

Box 1 : defining Relative Interconnectivity and Relative Closeness

A cluster is defined by the number of nodes (with different geographical coordinates, T_i), the links between these nodes (E_i) and the value of the links is the number of collaborations between the 2 nodes connected by this link (C_i).
The relations between 2 clusters are defined by the number of links between these two clusters ($E_{(i,j)}$) and the total number of collaborations supported by these links ($C_{(i,j)}$).

Calculating the Relative Interconnectivity

RI is the ratio between the total number of collaborations between the two clusters ($C_{(i,j)}$) and the average number of internal collaborations of the two clusters.

$$RI_{(i,j)} = \frac{C_{(i,j)}}{\frac{C_i + C_j}{2}}$$

Calculating the Relative Closeness

The internal closeness (Cl_i) of a cluster (or intra-closeness) is the ratio between the total number of collaborations within that cluster (C_i) and the number of links observed within that cluster (E_i).

$$Cl_i = \frac{C_i}{E_i}$$

Similarly the absolute closeness between two clusters (or inter-cluster closeness, $Cl_{(i,j)}$) is the ratio between the total collaborations observed between the two clusters ($C_{(i,j)}$) and the number of links between these two clusters ($E_{(i,j)}$).

$$Cl_{(i,j)} = \frac{C_{(i,j)}}{E_{(i,j)}}$$

The relative closeness between 2 clusters ($RC_{(i,j)}$) is the ratio between the absolute closeness of the two clusters and the average internal closeness of the two clusters (based upon the number of nodes of the 2 clusters, $T_i + T_j$).

$$RC_{(i,j)} = \frac{Cl_{(i,j)}}{\frac{T_i}{T_i+T_j} \times Cl_i + \frac{T_j}{T_i+T_j} \times Cl_j}$$

Below is an illustration at European level of the application of this combined approach.

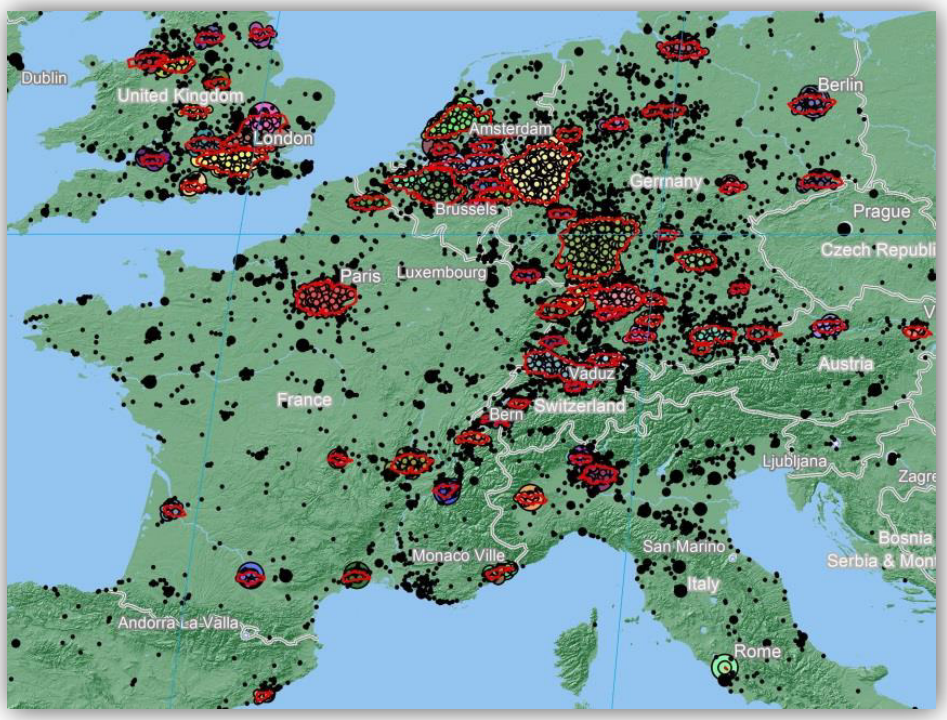


Fig & tab 9. *Geographical aggregation of inventor addresses in Europe (patents)*

Criteria	Publications	Patents
Relative Interconnectivity	1.2	2
Relative Closeness	20	12.50
Minimal weight (addresses)	1 000	1 250
Maximum distance	20 Km	20 Km
Number of addresses analysed	2 160 748	1 523 093
Number of initial clusters (after DBScan)	411	186
Number of final clusters (after Chameleon)	383	155

Fig & tab 10. *Parameters for DBScan and Chameleon*

The tests made show that the method is robust. It addresses the three issues raised by the previous method. We are now in the process of defining a method for labeling them. These developments will be finalised by the end of 2014.

2.3 Information on all variables/indicators

a) Description of all variables and/or indicators that are given for the main units of observation (e.g. number of publications, number of patents, etc.)

Patents

Main units of observation	Examples of variables
Applicant's name	Number of priority patents for applicants
Inventor's name	Number of priority patents for inventors
Subfield, field and domain based on IPC classes	Number of priority patents per subfields
Filing year	Number of priority patents per year
Titles	Vocabulary and frequency of keywords
Abstract	Vocabulary and keyword frequency
INPADOC Family	Number of priority patents per family inpadoc
Geolocalisation of inventor addresses	Number of priority patents per location, Number of priority patents shared by locations
Geolocalisation of applicant addresses	Number of priority patents per location, Number of priority patents shared by locations
Cluster of inventor addresses	Number of priority patents per cluster, Number of priority patents shared by clusters
Cluster of applicant addresses	Number of priority patents per cluster, Number of priority patents shared by clusters

Fig & tab 11. *Main units of observation for nano patents*

Scientific publications

Main units of observation	*Examples of variables*
Author Full Name	Number of collaborations
Document Title	Vocabulary and keyword frequency
Author Keywords	Keyword frequency
Keywords Plus®	Keyword frequency
Abstract	Vocabulary and keyword frequency
Cited References	Cited-citing counting
Web of Science Times Cited Count	Sum of Times Cited for a specific entity
Year Published	Number of publications per year
Web of Science Category	Number of publications per scientific Category
Subject Area	Number of publications per scientific Category
Harmonized institution	Number of publications for an institution, Number of collaborations for an institution
Type of institution	Number of publications for a type of institution, Number of publications for a type of institutions per countries
Geolocalisation	Number of publications for a location, Number of collaborations for a location
Cluster	Number of publications for a cluster, Number of collaborations between clusters

Fig & tab 12. *Main units of observation for nano publications*

2.4 Sectorial, temporal and geographical coverage

For patents tables are calculated using priority patents in table tls 201 (unless other specifications) for the three layers: NanoString, Static and Dynamic. All the calculations are made from 1991 to 2009 for the patents, and 1991 to 2010 for the publications.

a) Information on the temporal coverage used

Patents

Filing Year	*Total*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*
NanoString	*44955*	279	269	347	386	506	497	554	723	972
Static	*148267*	3970	4082	4945	4992	5656	5756	6611	8451	9452
Dynamic	*181635*	16	3939	5087	5039	5973	6145	7926	8875	9780
Total	***374857***	***4265***	***8290***	***10379***	***10417***	***12135***	***12398***	***15091***	***18049***	***20204***

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
NanoString	*1481*	*3245*	*3421*	*4551*	*5130*	*6075*	*6276*	*6672*	*3496*	*75*
Static	9632	10552	10800	11421	10995	11452	12706	12040	4583	171
Dynamic	10468	20985	13372	14445	12223	16654	18204	15957	6369	178
Total	***21581***	***34782***	***27593***	***30417***	***28348***	***34181***	***37186***	***34669***	***14448***	***424***

Fig & tab 13. *Number of priority patents per year and query*

Scientific publications

Year	Total	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
NanoString	*53758*	1336	1711	2312	3074	4199	5207	6528	7619	9302	12470
Static	*116006*	7491	8220	9116	10268	10748	11573	13240	13802	14480	17068
Dynamic	*43669*	980	1450	2193	3132	3520	4578	5515	6275	7191	8835
Total	***213433***	***9807***	***11381***	***13621***	***16474***	***18467***	***21358***	***25283***	***27696***	***30973***	***38373***

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
NanoString	15981	19945	26213	33168	40356	48091	58513	67134	73466	80425
Static	17542	18145	19105	20395	21054	22803	24630	24804	23939	24316
Dynamic	11285	15936	21325	26433	22012	25630	30323	39771	47847	48324
Total	***44808***	***54026***	***66643***	***79996***	***83422***	***96524***	***113466***	***131709***	***145252***	***153065***

Fig & tab 14. *Number of scientific publications per year and query*

b) Geographic coverage

Patents

Continent	Number of inventor's addresses
Total of filled addresses	***554 855***
Africa	142
Asia	307056
Europe	119597
Latin America and the Caribbean	2063
Northern America	125607
Oceania	390

Fig & tab 15. *Number of inventor's addresses per continent*

Region	*Number of inventor's addresses*
Total of filled addresses	***554 855***
Australia and New Zealand	384
Central America	425
Eastern Africa	18
Eastern Asia	303498
Eastern Europe	34780
Middle Africa	6
Northern Africa	55
Northern America	125607
Northern Europe	6447
Polynesia	6
South America	1435
South-central Asia	1459
South-eastern Asia	930
Southern Africa	48
Southern Europe	8903
the Caribbean	203
Western Africa	15
Western Asia	1169
Western Europe	69467

Fig & tab 16. *Number of inventor's addresses per sub-continent*

Country with more than 500 inventor's addresses	*Country harmonized*	*Inventor's addresses*	*With filled addresses*	*Geolocalized addresses*	*%*
Total for the 126 countries		***554 855***	***173 521***	***172 900***	***98,63%***
CHINA	CN	193074	777	761	97,94 %
UNITED STATES	US	120183	115641	115582	99,95 %
SOUTH KOREA	KR	80869	1306	1269	97,17 %
GERMANY	DE	40809	4724	4714	99,79 %
RUSSIA	RU	24470	515	504	97,86 %
FRANCE	FR	21100	20482	20450	99,84 %
TAIWAN	TW	18494	6672	6452	96,7 %
JAPAN	JP	10426	7484	7320	97,81 %
CANADA	CA	5424	2902	2902	100 %
SPAIN	ES	5068	548	547	99,82 %
UKRAINE	UA	4903	52	52	100 %
UNITED KINGDOM	GB	4104	1288	1284	99,69 %
ITALY	IT	2704	1184	1179	99,58 %
NETHERLANDS	NL	2585	1635	1633	99,88 %
SWITZERLAND	CH	2173	1541	1538	99,81 %

BELGIUM	BE	1865	1444	1438	99,58 %
INDIA	IN	1268	997	996	99,9 %
POLAND	PL	1207	61	61	100 %
CZECH REPUBLIC	CZ	1077	22	22	100 %
ROMANIA	RO	830	13	13	100 %
ARGENTINA	AR	827	13	13	100 %
ISRAEL	IL	818	626	626	100 %
AUSTRIA	AT	773	218	215	98,62 %
PORTUGAL	PT	693	24	24	100 %
SINGAPORE	SG	686	485	483	99,59 %
HONG KONG	HK	630	291	291	100 %
MOLDOVA	MD	619	1	1	100 %
SWEDEN	SE	584	442	438	99,1 %
BULGARIA	BG	576	1	1	100 %
FINLAND	FI	523	404	403	99,75 %

Fig & tab 17. *Number of inventor's addresses per country*

ISO 3166 [5] is a standard developed for the current names of countries, dependencies, and other areas of particular geopolitical interest, on the basis of lists of country names obtained from the United Nations and maintained by the ISO 3166 Maintenance Agency established by the ISO Council, the International Organization for Standardization (ISO). The international two letter country code (ISO alpha-2) is used as Harmonized country code.

Scientific publications

Continent	*Number of addresses*
Total	***2 176 376***
Africa	16836
Asia	801142
Europe	748763
Latin America and the Caribbean	60886
Northern America	514534
Oceania	34215

Fig & tab 18. *Number of author's addresses per continent*

Region	*Number of addresses*
Total	***2 176 376***
Australia and New Zealand	34195
Central America	14242

[5] ISO 3166 alpha 2, http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2

Eastern Africa	620
Eastern Asia	657412
Eastern Europe	126873
Melanesia	19
Micronesia	1
Middle Africa	223
Northern Africa	10745
Northern America	514534
Northern Europe	141267
South America	45698
South-central Asia	76208
South-eastern Asia	33520
Southern Africa	4347
Southern Europe	152825
the Caribbean	946
Western Africa	901
Western Asia	34002
Western Europe	327798

Fig & tab 19. *Number of author's addresses per sub-continent*

Main countries with more than 10 000 addresses	*Country harmonized*	*Number of addresses*	*Addresses geolocalized*	*%*
Total for all the 166 countries		***2 176 376***	***2 153 142***	***97%***
AUSTRIA	AT	12360	12358	99,98 %
AUSTRALIA	AU	30501	28327	92,87 %
BELGIUM	BE	17875	17870	99,97 %
BRAZIL	BR	31320	31319	100 %
CANADA	CA	43182	43182	100 %
SWITZERLAND	CH	23739	23737	99,99 %
CHINA	CN	268630	268488	99,95 %
CZECH REPUBLIC	CZ	12026	12024	99,98 %
GERMANY	DE	138001	137994	99,99 %
SPAIN	ES	49060	49054	99,99 %
FINLAND	FI	11217	11204	99,88 %
FRANCE	FR	109136	109118	99,98 %
UNITED KINGDOM	GB	83113	83103	99,99 %
GREECE	GR	10574	10536	99,64 %
HONG KONG	HK	13099	13099	100 %
ISRAEL	IL	15025	15025	100 %
INDIA	IN	57754	57700	99,91 %
IRAN	IR	14992	14992	100 %
ITALY	IT	73211	73203	99,99 %

JAPAN	JP	216934	215834	99,49 %
SOUTH KOREA	KR	101996	85863	84,18 %
MEXICO	MX	14038	14038	100 %
NETHERLANDS	NL	26337	26335	99,99 %
POLAND	PL	24480	24479	100 %
PORTUGAL	PT	11191	11190	99,99 %
ROMANIA	RO	10474	10452	99,79 %
RUSSIA	RU	49725	49599	99,75 %
SWEDEN	SE	21365	21364	100 %
SINGAPORE	SG	21476	21476	100 %
TURKEY	TR	13801	13801	100 %
TAIWAN	TW	56723	56723	100 %
UKRAINE	UA	10005	10001	99,96 %
UNITED STATES	US	471352	471322	99,99 %

Fig & tab 20. *Number of author's addresses per country*

c) Technological fields and scientific domains

Patents

Patents are classified according technology categories. The categories used are those build by WIPO (see PATSTAT-IFRIS report for detailed information): 5 domains, 35 fields and 374 subfields.

The distributions of priority patent domains and fields and subfields are shown below (Note: a patent can be affected to several domains, fields and sub-fields)

Domain code	*Domain name*	*Number of priority applications*
Total		***951 807***
TD01	Electrical engineering	193947
TD02	Instruments	221638
TD03	Chemistry	379890
TD04	Mechanical engineering	115026
TD05	Other fields	41306

Fig & tab 21. *Number of priority patents per domain*

Field code	*Field name*	*Number of priority applications*
Total		***951 807***
TF01	Electrical machinery, apparatus, energy	61499

TF02	Audio-visual technology	54660
TF03	Telecommunications	6365
TF04	Digital communication	1492
TF05	Basic communication processes	5992
TF06	Computer technology	9661
TF07	IT methods for management	635
TF08	Semiconductors	53643
TF09	Optics	160585
TF10	Measurement	31358
TF11	Analysis of biological materials	5076
TF12	Control	4874
TF13	Medical technology	19745
TF14	Organic fine chemistry	40618
TF15	Biotechnology	10858
TF16	Pharmaceuticals	34998
TF17	Macromolecular chemistry, polymers	41455
TF18	Food chemistry	12597
TF19	Basic materials chemistry	55023
TF20	Materials, metallurgy	67019
TF21	Surface technology, coating	48028
TF22	Micro-structural and nano-technology	6503
TF23	Chemical engineering	39840
TF24	Environmental technology	22951
TF25	Handling	13987
TF26	Machine tools	13066
TF27	Engines, pumps, turbines	8932
TF28	Textile and paper machines	19437
TF29	Other special machines	27249
TF30	Thermal processes and apparatus	9338
TF31	Mechanical elements	9735
TF32	Transport	13282
TF33	Furniture, games	13304
TF34	Other consumer goods	14208
TF35	Civil engineering	13794

Fig & tab 22. *Number of priority patents per field*

Main subfields with more than 5 000 patents		***Number of priority applications***
	Total for all the 374 subfields	***951 807***
T01F01	Lighting	8369
T01F02	Displaying Advertising	11932
T01F05	Basic Electronic Circuitry	5992

T01F06	Computing	8204
T01F08	Semiconductor Devices	53643
T01F09	Optical Elements Systems	140685
T01F11	Material Analysis by Chem Phys Properties	5076
T01F14	Cosmetic Preparations	8710
T01F16	Medical Preparations	23824
T01F20	Casting and Powder Metallurgy	6829
T01F32	Vehicles	8095
T01F33	Furniture and Domestic Equipment	8887
T02F02	Arrangements for Control	20076
T02F16	Therapeutic Activity of Chemical Compounds	11174
T02F20	Inorganic Chemistry	24741
T02F22	Nano-Technology	5484
T02F26	Mechanical Metal-Working	6131
T02F31	Engineering Elts or Units	8494
T03F02	Information Storage Based Record Carrier	11622
T03F17	Macromol with C-To-C Unsaturated Bonds	9840
T03F21	Layered Products	7346
T04F14	Acyclic or Carbocyclic Compounds	16323
T04F17	Macromol Withouth C-To-C Unsaturated Bonds	6573
T04F20	Refractories	14233
T04F21	Coating Metallic Material	14681
T04F25	Containers for Storage of Articles	7335
T05F09	Photomechanics of Surfaces	6810
T05F14	Heterocyclic Compounds	6814
T05F20	Metallurgy of Iron	5183
T06F17	Inorg or Non-Macromolr Organ Subst	11089
T06F20	Metallurgy	11963
T06F21	Crystal Growth	18078
T07F01	Discharge Lamps	15385
T07F17	Compositions of Macromolecular Compounds	13207
T07F19	Paints and Inks	9817
T07F28	Threads or Fibers	5520
T08F09	Devices Using Stimulated Emission	5919
T08F13	Methods for Sterilising	5129
T09F01	Batteries and Related	13922
T10F18	Other Foods	5051
T10F24	Solid Waste and Contaminated Soils	6208
T11F10	Chemical Physical Analyses	15075
T12F01	Boards for the Distribution of Electricity	5031
T12F19	Materials for Miscellaneous Applications	13180
T13F24	Absorbing Noise from Roads	10838
T14F23	Chem or Phys Lab Apparatus	24972

T17F29	Working of Plastics	9500
T20F29	Processes of Compounding	5748

Fig & tab 23. *Number of priority patents per sub-field*

Scientific publications

In a first stage we use the ISI-WoS subject classifications to characterise domains and subject areas. Note: a publication has several subject categories

Main subject categories with more than 10 000 publications	*Number of publications*
Total for all the 274 subject categories	***4 032 893***
Physics	389630
Materials Science	351481
Chemistry	314217
Materials Science, Multidisciplinary	280878
Physics, Applied	226973
Chemistry, Physical	166803
Engineering	154761
Physics, Condensed Matter	148371
Optics	121152
Chemistry, Multidisciplinary	117392
Science & Technology - Other Topics	104777
Polymer Science	104146
Nanoscience & Nanotechnology	93476
Metallurgy & Metallurgical Engineering	92796
Electrochemistry	75890
Engineering, Electrical & Electronic	73766
Biochemistry & Molecular Biology	61399
Instruments & Instrumentation	46420
Physics, Multidisciplinary	45248
Materials Science, Coatings & Films	40574
Pharmacology & Pharmacy	37183
Engineering, Chemical	36916
Crystallography	31944
Chemistry, Analytical	31917
Materials Science, Ceramics	31142
Energy & Fuels	31032
Physics, Atomic, Molecular & Chemical	28175
Biophysics	26128
Biotechnology & Applied Microbiology	24628
Nuclear Science & Technology	21592
Cell Biology	21284
Mechanics	20998

Environmental Sciences & Ecology	20033
Spectroscopy	19010
Chemistry, Applied	17983
Environmental Sciences	17757
Chemistry, Inorganic & Nuclear	17090
Engineering, Mechanical	15890
Microscopy	14946
Geochemistry & Geophysics	14548
Engineering, Biomedical	13470
Microbiology	12594
Mineralogy	11726
Chemistry, Organic	11400
Computer Science	11386
Multidisciplinary Sciences	11301
Radiology, Nuclear Medicine & Medical Imaging	11252
Materials Science, Biomaterials	10860
Water Resources	10858
Biochemical Research Methods	10721
Materials Science, Composites	10719
Engineering, Environmental	10514
Astronomy & Astrophysics	10498

Fig & tab 24. *Number of publications per subject categories*

d) Coverage for institutions

Patents

Work in progress

Scientific publications

Table 25 presents the partition of addresses in a country along the type of institution. Overall University addresses represent 75% of the total, Government labs 18%, firms 5% and hospital and others 2% together. The 5 major institutions per country are listed in table 26 (see appendix 3).

Country name	Country code harmonized	Number of addresses for the countries	firm	gvt	hosp	other	univ
Total of addresses		***19 83 667***	***93 971***	***3 59 617***	***37 357***	***13 312***	***1 479 390***
ARGENTINA	AR	7719	43	2720	152	51	4753
AUSTRALIA	AU	30486	751	3794	980	252	24709
AUSTRIA	AT	12342	701	1107	260	420	9854
BELGIUM	BE	17858	904	451	473	191	15839

BRAZIL	BR	31275	309	2755	445	57	27709
CANADA	CA	43152	2179	4594	1661	281	34436
CHINA	CN	281531	3869	57298	1098	66	219199
CZECH REPUBLIC	CZ	12110	313	5726	192	29	5850
DENMARK	DK	9885	809	264	649	96	8067
FINLAND	FI	11217	515	1329	422	24	8927
FRANCE	FR	109105	5299	32251	3358	675	67522
GERMANY	DE	137885	8571	39202	2441	1035	86636
GREECE	GR	10572	137	3035	302	48	7050
HUNGARY	HU	7941	221	3289	41	11	4379
INDIA	IN	57718	882	18384	390	773	37288
IRAN	IR	14975	193	1741	46	62	12933
IRELAND	IE	6073	173	57	192		5651
ISRAEL	IL	15027	421	2635	587	12	11372
ITALY	IT	73134	2326	18070	2239	366	50133
JAPAN	JP	92928	280	57			92591
MEXICO	MX	14079	84	2638	119	9	11229
NETHERLANDS	NL	26323	2418	2509	3068	135	18193
NORWAY	NO	5106	330	920	473	9	3374
POLAND	PL	24455	136	7649	116	44	16510
PORTUGAL	PT	11185	70	764	91	404	9856
ROMANIA	RO	10465	276	4722	86	21	5360
RUSSIA	RU	49884	1108	32568	48	30	16130
SINGAPORE	SG	21484	576	4420	243	8	16237
SOUTH KOREA	KR	101947	6760	12967	354	192	81674
SPAIN	ES	49027	793	13553	1313	555	32813
SWEDEN	SE	21353	1504	589	412	222	18626
SWITZERLAND	CH	23729	2195	1683	809	323	18719
TAIWAN	TW	56811	1402	6258	1202	29	47920
THAILAND	TH	6127	39	635	28	8	5417
TURKEY	TR	13786	88	369	318	11	13000
UKRAINE	UA	10028	92	6838	3	1	3094
UNITED KINGDOM	GB	83045	5375	4398	2704	1122	69446
UNITED STATES	US	471900	41829	57378	10042	5740	356894

Fig & tab 25. *Institutional coverage: classification of the institutions*

2.5 Quality and accuracy of data

a) Information on the number of missing values

Missing data for geographical information

Patents

For patents 273000 inventor addresses do not have any harmonized country. They have thus not been geolocalised at this stage. This represents 33% of total inventor addresses. Furthermore in a number of countries Patstat does not include yet inventor addresses: this is systematic for instance in China the second largest country in patenting for nanotechnology. Overall at this stage, we have a quite weak coverage. Work is being done for improving it before opening.

Note on different cases of missing information:

- The Patstat database gives no information on any inventor (or applicant) of a patent. In that case, the patent application is present (with its appln_id) in table firm_tls201_appln_ifris but absent of the table invt_adr_ifris_epwofr_frac_ctry that includes only patents with at least a few information on inventors (or table applt_adr_ifris_epwofr_frac_ctry which considers patents with at least a few information on applicants).
- The number of priority patents (appln_id in firm_tls201_appln_ifris) without any information on inventors is 273 753.
- The database gives limited information on inventor (or applicant) of a patent but this information may be partial (absence of the inventor or applicant name, address, country). In that case, the patent application is present (with its appln_id) in table invt_addr_ifris (or table applt_addr_ifris) but some fields are still empty after the filling steps described in the PATSTAT-IFRIS section.

The geolocalisation of filled addresses is very high by integrating the information and/or add-ups from REGPAT (OECD)[6], the analysis of the DB of the French Patent Office (INPI) and from the internal analysis of Inpadoc families.

Country Name	*Country Harmonized*	*Addresses for priority patents*	*Addresses for non singleton priority patents*	*Non singleton priority patents with addresses*	*Geolocalised addresses form non singleton priority patents*	*Non singleton addresses geolocalised*

[6] OECD REGPAT, http://www.oecd.org/science/inno/40794372.pdf

Total		*827 618*	*239 053*	*131 533*	*131 168*	*45,95 %*
Vide	Vide	273 753	45 055	59	0	0 %
CHINA	CN	193074	4087	355	345	8,44 %
UNITED STATES	US	120183	90955	88964	88919	97,76 %
SOUTH KOREA	KR	80869	18540	683	667	3,6 %
GERMANY	DE	40809	27068	3941	3933	14,53 %
RUSSIA	RU	24470	1511	337	328	21,71 %
FRANCE	FR	21100	17492	17129	17104	97,78 %
TAIWAN	TW	18494	6219	2399	2336	37,56 %
JAPAN	JP	10426	6561	5376	5289	80,61 %
CANADA	CA	5424	3474	2258	2258	65 %
SPAIN	ES	5068	2695	415	415	15,4 %
UKRAINE	UA	4903	211	29	29	13,74 %
UNITED KINGDOM	GB	4104	2697	1018	1015	37,63 %
ITALY	IT	2704	1310	914	910	69,47 %
NETHERLANDS	NL	2585	1719	1417	1415	82,32 %
SWITZERLAND	CH	2173	1606	1236	1234	76,84 %
BELGIUM	BE	1865	1589	1292	1287	80,99 %
INDIA	IN	1268	723	633	632	87,41 %
POLAND	PL	1207	102	42	42	41,18 %
CZECH REPUBLIC	CZ	1077	360	21	21	5,83 %
ROMANIA	RO	830	16	8	8	50 %
ARGENTINA	AR	827	95	7	7	7,37 %
ISRAEL	IL	818	573	483	483	84,29 %
AUSTRIA	AT	773	517	165	162	31,33 %
PORTUGAL	PT	693	266	17	17	6,39 %
SINGAPORE	SG	686	485	389	387	79,79 %
HONG KONG	HK	630	264	176	176	66,67 %
MOLDOVA	MD	619	15	0	0	0 %
SWEDEN	SE	584	443	365	364	82,17 %
BULGARIA	BG	576	248	1	1	0,4 %
FINLAND	FI	523	338	297	297	87,87 %
HUNGARY	HU	421	134	60	60	44,78 %
MEXICO	MX	398	136	91	90	66,18 %
BELARUS	BY	390	39	8	8	20,51 %
DENMARK	DK	381	214	142	138	64,49 %
AUSTRALIA	AU	339	215	200	200	93,02 %
SLOVENIA	SI	334	169	24	18	10,65 %
SLOVAKIA	SK	287	56	10	9	16,07 %
BRAZIL	BR	261	99	62	62	62,63 %
NORWAY	NO	246	192	101	100	52,08 %

LATVIA	LV	218	22	1	1	4,55 %
CUBA	CU	182	68	8	7	10,29 %
IRELAND	IE	160	101	69	68	67,33 %
LITHUANIA	LT	150	9	7	7	77,78 %
MALAYSIA	MY	145	58	50	50	86,21 %
TURKEY	TR	137	40	22	22	55 %
LUXEMBOURG	LU	126	111	108	105	94,59 %
VENEZUELA	VE	95	79	75	73	92,41 %
PERU	PE	84	1	0	0	0 %
SAUDI ARABIA	SA	75	12	8	8	66,67 %
URUGUAY	UY	74	64	61	61	95,31 %
Other 76 countries		0	0	0	0	33,2 %

Fig & tab 26. *Missing geographical data for patents*

Scientific publications

Only 1.19% of addresses have not been geolocalised.

Missing information for institution and technological domains

Patents

Only 0,7% (6 677) priority patents have no technological domain.

Scientific publications

5.69% (124 064) of addresses do not have a harmonised institutional name and 5.75% (125 389) have not been allocated to one of the 5 types.

b) Estimation of data quality issues with respect to data acquisition, reliability of retrieving system

Various sources of noise have been identified:

- Patents: we have used the PATSTAT standardisation name as a source for harmonising institutions.
- Publications and patents: in quite a few cases toponyms are ambiguous (even within a country) and this leads to a failure in geolocalisation.

3 Legal issues encountered and access conditions

a) Legal issues concerning access of the database

The dataset is only accessible for research purposes (no commercial use is authorised). An important dimension of the database is that information at the individual level (one publication and one patent) remains confidential; only aggregated data can be published in public reports and/or academic journals.

b) Legal necessities for potential opening procedures

Users need to belong to an institution that has both a subscription to the Web of Science and to Patstat.

4 Technical structure of the dataset

4.1 Information on the data base system

a) Current data base system used

The current data base system is My SQL 5.1.63 with MyISAM as the default storage engine. In term of maintainability and backup, the main advantage of this storage engine is to use three different files for each table of a database:

- the data file has a .MYD (MYData) extension;
- the index file has a .MYI (MYIndex) extension;
- the structure file has a .frm extension.

MySQL is optimized for an intensive usage: a high level of accessibility and efficiency, for a low amount of users.

b) Planned future technical changes concerning data base system

No changes planned

4.2 Technical variable definition

Labelling of all variables. Data type of all variables (e.g., float, string, etc.). Current use and definition of unique identifiers (if applicable).

a) Variables for patents

tls209_appln_ipc_ifris_epwofr	
appln_id	INT(10)
ipc_class_symbol	CHAR(15)
ipc_class_level	CHAR(1)
ipc_version	DATE NULL DEFAULT NULL,
ipc_value	CHAR(1)
ipc_position	CHAR(1)
ipc_gener_auth	CHAR(2)
PRIMARY KEY	appln_id, ipc_class_symbol, ipc_class_level

tls210_appln_n_cls	
appln_id	INT(10)
nat_class_symbol	CHAR(15)
PRIMARY KEY	appln_id, nat_class_symbol

tls205_tech_rel	
appln_id	INT(10)
tech_rel_appln_id	INT(10)
PRIMARY KEY	appln_id, tech_rel_appln_id

tls202_appln_title	
appln_id	INT(10)
appln_title	VARCHAR(3500)
PRIMARY KEY	appln_id

tls207_pers_appln	
person_id	INT(10)
appln_id	INT(10)
applt_seq_nr	SMALLINT(4)
invt_seq_nr	SMALLINT(4)
PRIMARY KEY	appln_id, person_id

tls206_person	
person_id	INT(10)
person_ctry_code	VARCHAR(3)
doc_std_name_id	INT(10)
person_name	VARCHAR(300)
person_address	VARCHAR(500)
PRIMARY KEY	person_id

tls208_doc_std_nms

doc_std_name_id	INT(10)
doc_std_name	CHAR(100)
PRIMARY KEY	doc_std_name_id

tls204_appln_prior_ifris	
appln_id	INT(10)
prior_appln_id	INT(10)
prior_appln_seq_nr	SMALLINT(4)
appln_priority_year	INT(4)
PRIMARY KEY	appln_id, prior_appln_id

tls201_appln_ifris	
key_appln	VARCHAR(19)
appln_id	INT(10)
appln_auth	CHAR(2)
appln_nr	CHAR(15)
appln_kind	CHAR(2)
appln_filing_date	DATE NULL DEFAULT NULL,
ipr_type	CHAR(2)
appln_title_lg	CHAR(2)
appln_abstract_lg	CHAR(2)
internat_appln_id	INT(10)
appln_filing_year	INT(4)
appln_first_priority_year	INT(4)
no_appt_invt	INT(1)
no_ipc	INT(1)
artificial	INT(1)
singleton	INT(1)
layer	VARCHAR(4)
PRIMARY KEY	appln_id

tls211_pat_publn_ifris	
key_publn	VARCHAR(19)
pat_publn_id	INT(10)
publn_auth	CHAR(2)
publn_nr	CHAR(15)
publn_kind	CHAR(2)
appln_id	INT(10)
publn_date	DATE NULL DEFAULT NULL,
publn_lg	CHAR(2)
publn_first_grant	SMALLINT(2)
publn_year	INT(4)
PRIMARY KEY	pat_publn_id

tls203_appln_abstr	
appln_id	INT(10)
appln_abstract	VARCHAR(4000)
PRIMARY KEY	appln_id

tls218_docdb_fam	
appln_id	INT(10)
docdb_family_id	INT(10)
PRIMARY KEY	appln_id, docdb_family_id

tls219_inpadoc_fam	
appln_id	INT(10)
inpadoc_family_id	INT(10)
PRIMARY KEY	appln_id

tls214_npl_publn	
npl_publn_id	INT(10)
npl_biblio	VARCHAR(2100)
PRIMARY KEY	npl_publn_id

tls212_citation	
pat_publn_id	INT(10)
citn_id	SMALLINT(4)
cited_pat_publn_id	INT(10)
npl_publn_id	INT(10)
pat_citn_seq_nr	SMALLINT(4)
npl_citn_seq_nr	SMALLINT(4)
citn_origin	CHAR(5)
PRIMARY KEY	pat_publn_id, citn_id

ipc_technology_frac_ifris	
appln_id	INT(10)
ipc_class_symbol	CHAR(15)
ipc_class_level	CHAR(1)
ipc_version	DATE NULL DEFAULT NULL,
ipc_value	CHAR(1)
ipc_position	CHAR(1)
ipc_gener_auth	CHAR(2)
nb_ipc	BIGINT(21)
frac_ipc	DECIMAL(5,4)
domaines	VARCHAR(4)
fields	VARCHAR(4)

sfields	VARCHAR(6)
PRIMARY KEY	appln_id, ipc_class_symbol

applt_adr_ifris_epwofr_frac_ctry	
key_applt	VARBINARY(26)
appln_id	INT(10)
person_id	INT(10)
doc_std_name_id	INT(10)
person_name	VARCHAR(300)
person_address	VARCHAR(500)
person_ctry_code	VARCHAR(3)
SOURCE	VARCHAR(7)
PM_PP_COMP	VARCHAR(1)
QUALIF_COMP	VARCHAR(4)
NAME_COMP	VARCHAR(300)
FIRSTNAME_COMP	VARCHAR(300)
ADR_COMP	VARCHAR(500)
STREET_COMP	VARCHAR(40)
ZIP_CODE_COMP	VARCHAR(5)
CITY_COMP	VARCHAR(50)
CTRY_COMP	VARCHAR(2)
METHODE	VARCHAR(4)
frac_applt	DECIMAL(5,4)
adr_final	VARCHAR(500)
ctry_final	VARCHAR(2)
ctry_lib_ifris_ctry_final	CHAR(2)

ctry_lib_ifris	
ctry_final	CHAR(2)
lib_ctry_harm	VARCHAR(100)
ctry_harm	CHAR(2)
nomen_geo_ifris_lib_ctry_harm	VARCHAR(100)
PRIMARY KEY	ctry_final

nano_corpus	
appln_id	INT(10)
layer	VARCHAR(10)
fields	VARCHAR(5)
PRIMARY KEY	appln_id

invt_adr_ifris_epwofr_frac_ctry	
key_invt	VARBINARY(26)
appln_id	INT(10)

person_id	INT(10)
doc_std_name_id	INT(10)
person_name	VARCHAR(300)
person_address	VARCHAR(500)
person_ctry_code	VARCHAR(3)
SOURCE	VARCHAR(7)
QUALIF_COMP	VARCHAR(4)
NAME_COMP	VARCHAR(300)
FIRSTNAME_COMP	VARCHAR(300)
ADR_COMP	VARCHAR(500)
STREET_COMP	VARCHAR(40)
ZIP_CODE_COMP	VARCHAR(5)
CITY_COMP	VARCHAR(50)
CTRY_COMP	VARCHAR(2)
METHODE	VARCHAR(4)
frac_invt	DECIMAL(5,4)
adr_final	VARCHAR(500)
ctry_final	VARCHAR(2)
ctry_lib_ifris_ctry_final	CHAR(2)

Fig & tab 27. *List of fields, types of variable and primary keys for the Nano Patents*

b) Variables for scientific publications

tab_address	
address_id	int(10)
address	varchar(255)
address_org	varchar(127)
address_typorg	smallint(2)
address_ctycod	char(2)
address_city	varchar(40)
address_state	varchar(45)
address_poscod	varchar(11)
address_lat	decimal(10,7)
address_lng	decimal(10,7)
PRIMARY KEY	(address_id)

tab_author	
author_id	int(10)
document_id	int(10)
author_abbr	varchar(63)
author_full	varchar(127)

author_order	smallint(2)
author_type	smallint(2)
author_email	varchar(320)
PRIMARY KEY	(author_id)

tab_author_has_address	
author_id	int(10)
address_id	int(10)
PRIMARY KEY	(author_id,address_id)

tab_author_identifier	
identifier	varchar(255)
author_id	int(10)
identifier_type	smallint(2)
PRIMARY KEY	(identifier)

tab_citation	
citation_id	int(10)
citation	varchar(500)
cited_author	varchar(63)
cited_year	year(4)
cited_pub	varchar(90)
cited_volu	varchar(10)
cited_bpag	varchar(10)
cited_doi	varchar(255)
cited_pat	varchar(90)
cited_document_id	int(10)
cited_publication_id	int(10)
PRIMARY KEY	(citation_id)

tab_citation_info	
citation_info_id	int(10)
document_id	int(10)
citation_info_source	smallint(2)
citation_info_count	smallint(4)
citation_info_cited	smallint(5)
citation_info_date	date
PRIMARY KEY	(citation_info_id),

tab_corpus_info	
corpus_info_id	int(10)

corpus_info_name	varchar(255)
corpus_info_parent_id	int(10)
PRIMARY KEY	(corpus_info_id)

tab_document	
document_id	int(10)
language_id	smallint(3)
document_title	varchar(767)
document_abs	varchar(10000)
PRIMARY KEY	(document_id)

tab_document_has_address	
document_id	int(10)
address_id	int(10)
PRIMARY KEY	(document_id,address_id)

tab_document_has_citation	
document_id	int(10)
citation_id	int(10)
citation_order	smallint(4)
citation_type	smallint(2)
citation_source	smallint(2)
PRIMARY KEY	(document_id,citation_id,citation_order),

tab_document_has_keyword	
document_id	int(10)
keyword_id	int(10)
keyword_source	smallint(2)
keyword_order	smallint(4)
PRIMARY KEY	(document_id,keyword_id,keyword_source),

tab_document_has_publication	
document_id	int(10)
publication_id	int(10)
doc_pub_date	varchar(15)
doc_pub_year	year(4)
doc_pub_volu	varchar(10)
doc_pub_issue	varchar(7)
doc_pub_bpag	varchar(10)
doc_pub_epag	varchar(10)
doc_pub_npag	smallint(3)

doc_pub_url	varchar(511)
PRIMARY KEY	(document_id,publication_id)

tab_document_has_subject	
document_id	int(10)
subject_id	smallint(4)
subject_source	smallint(2)
subject_order	smallint(4)
PRIMARY KEY	(document_id,subject_id,subject_source)

tab_document_has_type	
document_id	int(10)
type_id	smallint(2)
PRIMARY KEY	(document_id,type_id)

tab_document_identifier	
identifier	varchar(255)
document_id	int(10)
identifier_type	char(3)
PRIMARY KEY	(identifier)

tab_file_info	
file_info_id	int(10)
corpus_info_id	int(10)
file_info_path	varchar(255)
file_info_name	varchar(255)
file_info_type	smallint(2)
file_info_date	timestamp
PRIMARY KEY	(file_info_id)

tab_keyword	
keyword_id	int(10)
keyword	varchar(255)
PRIMARY KEY	(keyword_id)

tab_language	
language_id	smallint(3)
language	varchar(20)
language_iso1	char(2)
PRIMARY KEY	(language_id)

tab_publication	
publication_id	int(10)
publisher_id	int(10)
publication	varchar(255)
publication_subt	varchar(255)
publication_type	char(1)
publication_aiso	varchar(90)
publication_a29	varchar(29)
publication_issn	char(8)
publication_isbn	char(13)
PRIMARY KEY	(publication_id),

tab_publisher	
publisher_id	int(10)
publisher	varchar(127)
publisher_addr	varchar(255)
publisher_city	varchar(40)
publisher_web	varchar(2083)
PRIMARY KEY	(publisher_id)

tab_reference	
reference_id	int(10)
document_id	int(10)
file_info_id	int(10)
reference_proc	tinyint(1)
PRIMARY KEY	(reference_id,file_info_id,document_id)

tab_reference_has_tag	
reference_id	int(10)
tag_id	int(10)
tab_order	smallint(2)
tag_bol	int(10)
tag_eol	int(10)
tag_used	tinyint(1)
PRIMARY KEY	(reference_id,tag_id,tab_order)

tab_subject	
subject_id	smallint(4)
subject	varchar(255)
PRIMARY KEY	(subject_id)

tab_tag	
tag_id	int(10)
tag	varchar(3)
PRIMARY KEY	(tag_id)

tab_type	
type_id	smallint(2)
type	varchar(40)
PRIMARY KEY	(type_id)

Fig & tab 28. *List of fields, types of variable and primary keys for the Nano Publications*

4.3 Description of the Entity Relationship Model of Nano

Below is a description of all tables that are specific to the nano database. This means that tables that are present in the Patstat IFRIS database are described in the Patstat IFRIS appendix of the CIB database.

Whole relational diagram of Nano Patents with PATSTAT IFRIS

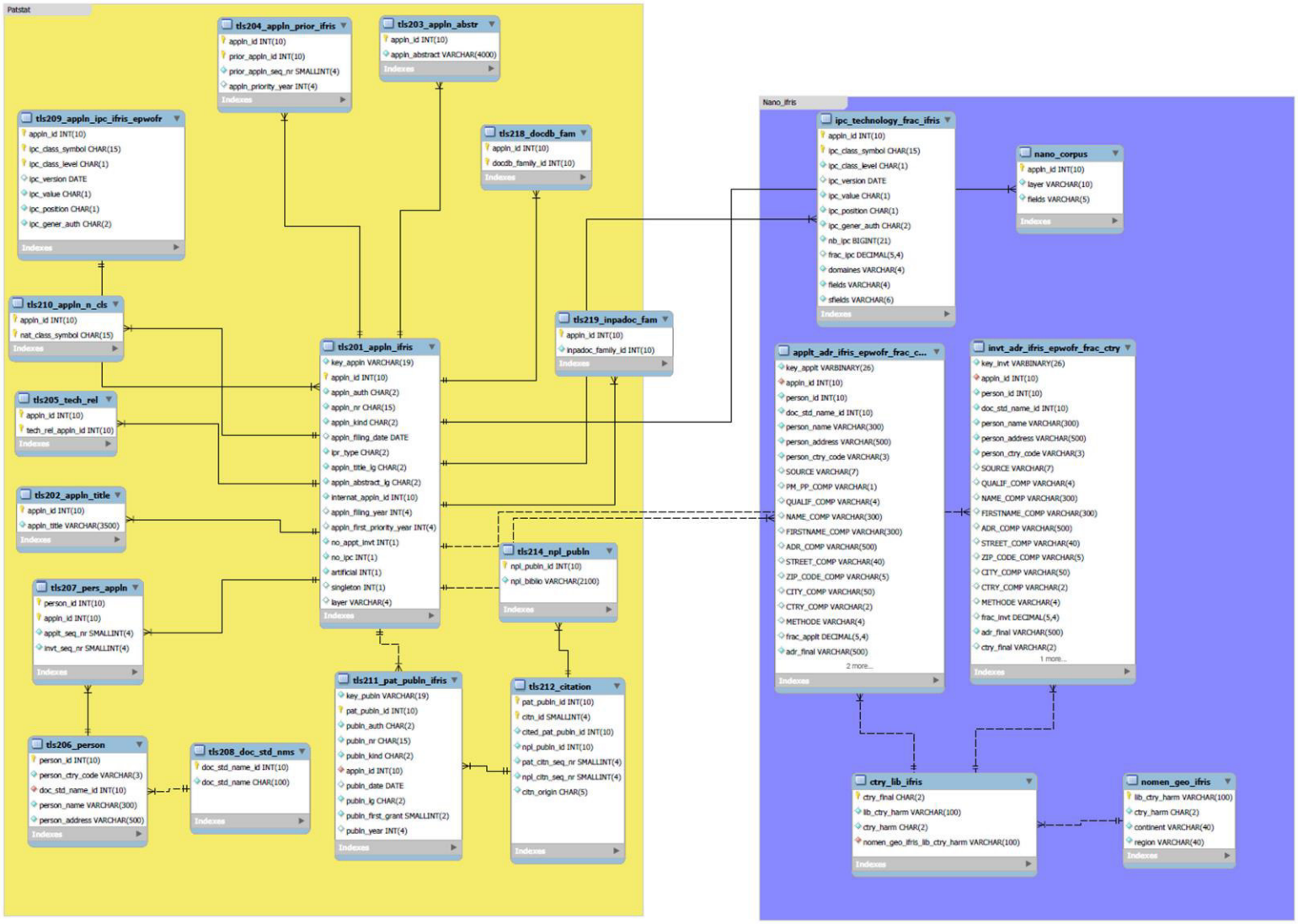


Fig & tab 29. *Relational diagram of Nano Patents*

Scientific publications relational diagram

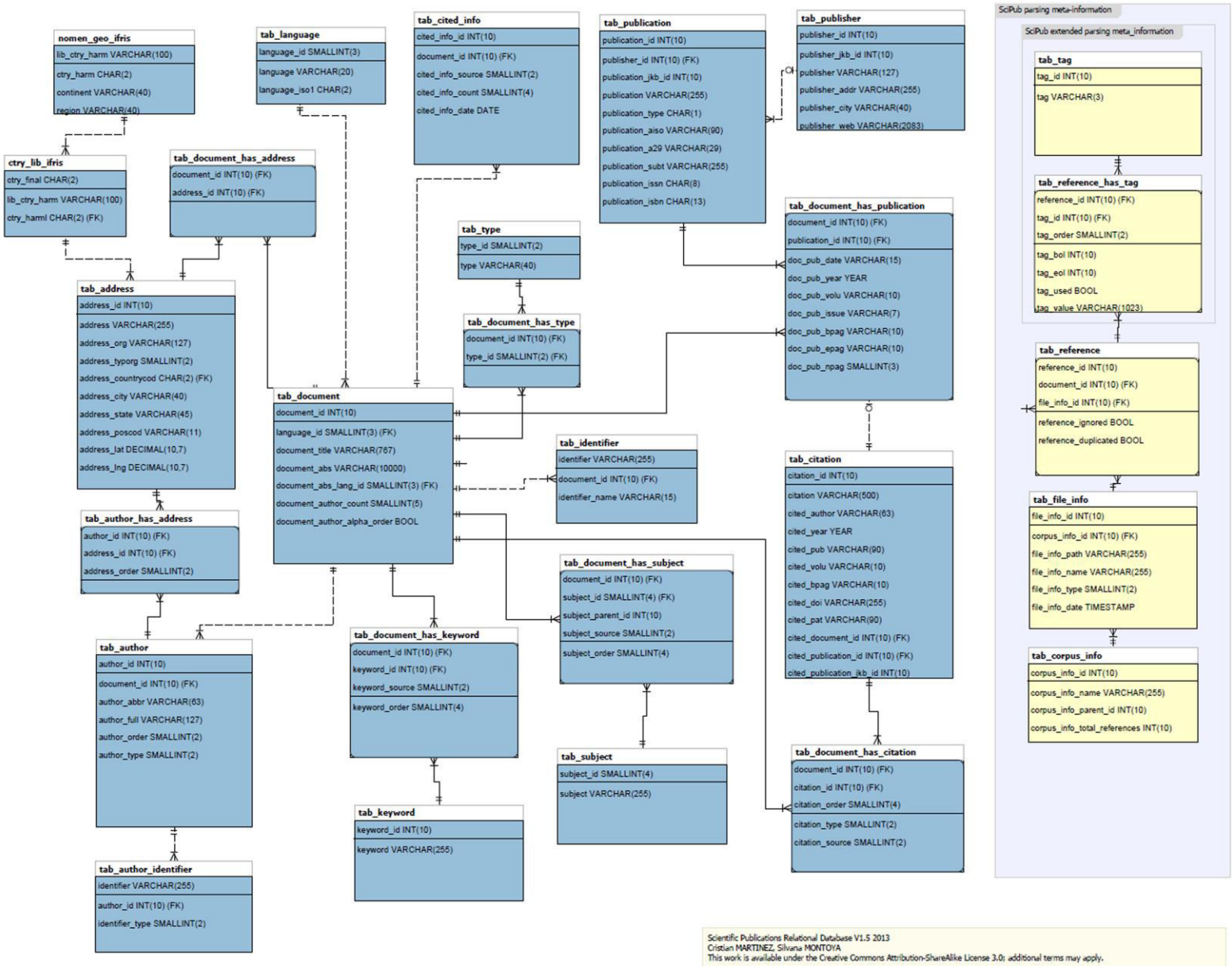


Fig & tab 30. *Relational diagram of Nano Publications*

4.4 Interfaces for access and to other infrastructures

The databases are accessible with a local access through the software MySQL Workbench[7].

[7] MySQL WorkBench, www.mysql.fr/products/workbench/

5 Further planning of the opening of *Nano*

a) Document concrete steps towards opening of the respective dataset

We have 3 on-going developments dealing with: (a) finalizing institutional harmonization (entering changes in the whole dataset); (b) patents (filling missing inventor addresses); and (c) finalizing clustering (naming and qualifying)

This will be done before March 2015.

b) Necessary updates and/or technical changes

We have two major developments starting (which we hope to integrate before the opening of the dataset): (i) move from the 2011 to 2014 Patstat dataset integrating all developments being made on Patstat-IFRIS; (ii) integrate a semantic thematic clustering based on CORTEXT Manager[8].

c) Changing legal conditions for accessing the dataset or parts of the dataset

None

d) Suggestions

The geographical and institutional harmonisations are critical for articulating this dataset to others. If this can be done we could

a) link engagement in nanotechnology of European universities with their characteristics (ETER dataset), their academic standing (Leiden ranking) and their presence at European level (EUPRO)
b) See how large firm engagement in nanotechnology matches with their degree of internationalisation (CIB)
c) See whether the hypothesis developed about the targeted role of small firms in nano (mostly "B to R") translates into their positioning in the world of start-up firms (VICO dataset)

[8] CorText Manager web app, http://manager.cortext.net/

Main references

Bonaccorsi A (2008) Search regimes and the industrial dynamics of science. *Minerva* 46: 285–315.

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. *In Proceedings of the IJCAI* (pp. 73–78). Acapulco, Mexico.

Ester, M., Kriegel, H.-P., & Sander, J. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In S. Evangelos, J. Han, & U. M. Fayyad (Eds.), *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* (pp. 226–231). AAAI Press.

Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8), 68–75.

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. doi:10.1016/j.respol.2007.02.005

Appendix 1: Detailed presentation of the dynamic query

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

Kahane B, Mogoutov A., Cointet J.P., Villard L., Larédo P.

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[9] 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[10]. 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.

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

[10] 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

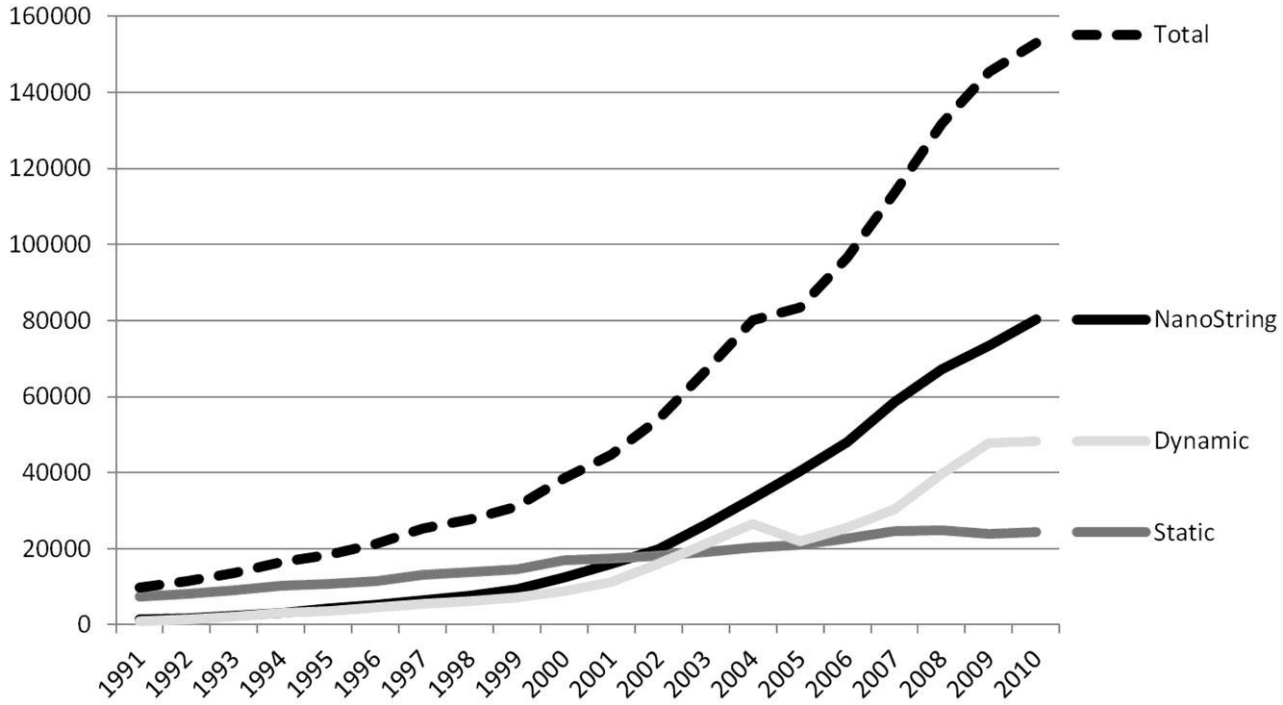
Note: the introduction by the WoS interface of a lemmatisation has simplified life for managing 'main forms' (see below) but it has limited the use of "*" in the construction of the query for abstracts and keywords (TS) driving to a different query as this for titles (TI).

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[11]. 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".[12]
- '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[13]. 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'

---

[11] 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.

[12] 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.

[13] 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.

and no longer require decomposing it using pre-existing fields (we had 8 such sub-fields 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[14]. 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

[14] 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.

that a persistent term translating a successful aggregation of knowledge has to be 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)[15]. 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

[15] 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).

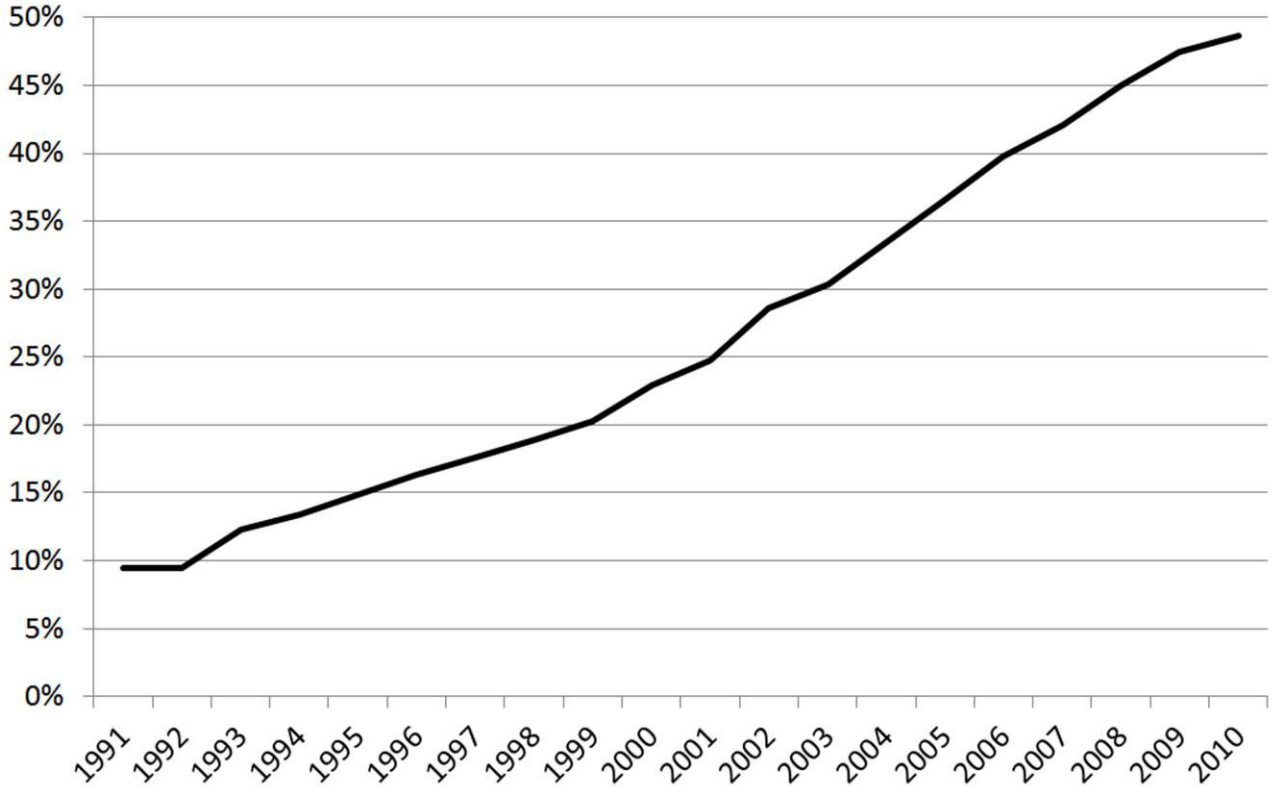
For moving to the next step, we considered another result of Bonaccorsi (2010) on computer science. He shows that over time the exploration does not diminish and that the rate of renewal of keywords does also not diminish overtime; only does the selection by colleagues become harsher, most terms remaining 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[16].

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).

[16] As we stop just below the multi-term that trespasses the annual quantitative threshold, the implication of this repetitition 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
00001110000000000000	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

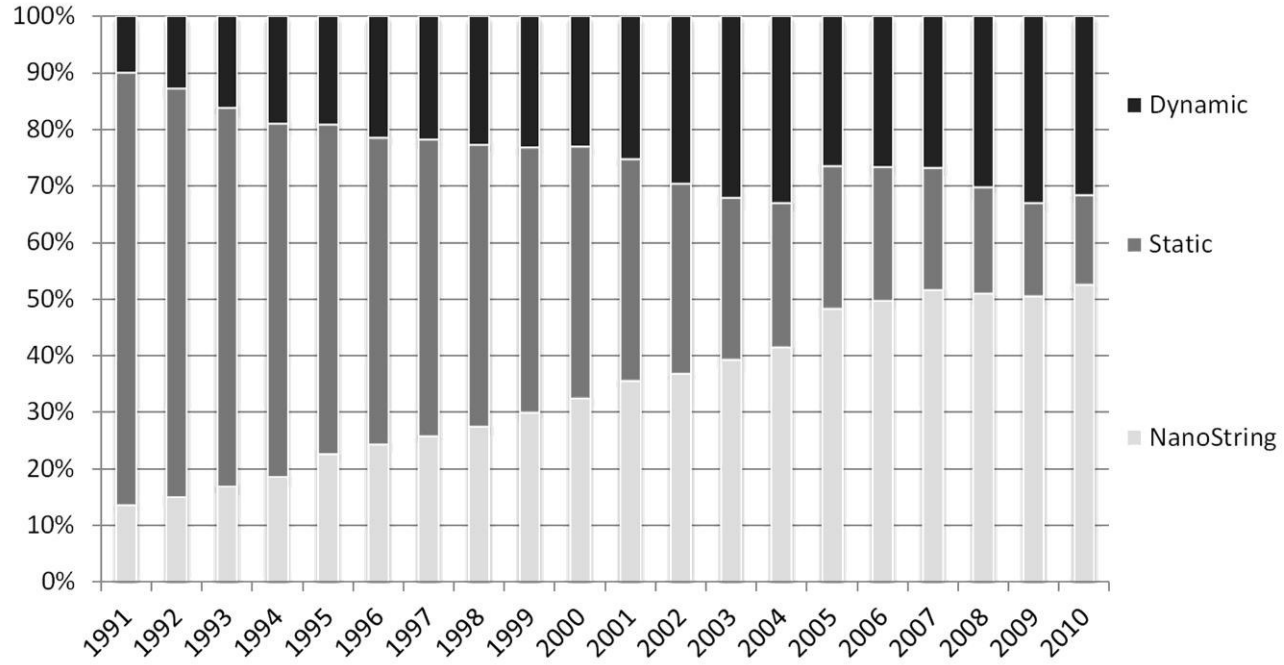
	Terms nber	Terms %	Articles nanostring	Potential articles	%
Manipulation observation	32	20%	93374	86035	19%
Production fabrication	30	19%	108120	123419	28%
Applications	12	7%	11778	13924	3%
Materials	38	23%	125363	122284	28%
Measures	22	14%	38331	34748	8%
Characteristics/ properties	28	17%	63815	62083	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.

Appendix 2: Automatic allocation of addresses to given types of actors

We have 5 types: universities, government organisations, hospitals, firms and other. For each we have established automatic allocations that are presented below

University / Abbreviations detection
"*_Coll_*" or "*_Fac_*" or "*_Fak_*" or "*_sch_*" or "*_Hsch_*" or "*_TH_*" or "*_Univ_*" or "*_Grad_*" or "*_Ecole_*" or "*_Scuola_*" or "*_School_*" or "*_Polytech_*" or "*_Polytecn_*"

University / Main institutions detection
"*_Caltech_*" or "*_CUNY_*" or "*_ETH_*" or "*_IIT_*" or "*_MIT_*" or "*_NYU_*" or "*_NTH_*" or "*_SUNY_*" or "*_UNWIST_*" or "*_georgia_tech_*" or "*_ASU_*" or "*_cornel*" or "*_FSU_*" or "*_FAMU_*" or "*_Harvard_*" or "*_ISU_*" or "*_LSU_*" or "*_ETH_*" or "*_Mem_Sloan_Kettering_Canc_Ctr_*" or "*_Manchester_Mat_Sci_Ctr_*" or "*_Leiden_Amsterdam_Ctr_Drug_Res_*" or "*_ITN_*" or "*_IPN_*" or "*_IPCMS_*" or "*_IHP_*" or "*_Forschungszentrum_Karlsruhe_*" or "*_ENSERG_*" or "*_ENSEEG_*" or "*_EMPA_*" or "*_Virginia_Tech_*" or "*_UTBM_*" or "*_USP_*" or "*_UNINOVA_*" or "*_UNICAMP_*" or "*_UNESP_*" or "*_UNAM_*" or "*_UMIST_*" or "*_UFSCar_*" or "*_UFRJ_*" or "*_UFPR_*" or "*_TU_Chemnitz_*" or "*_Supelec_*"

Government institute / Abbreviations detection (first part of the address)
"*_Akad_*" or "*_Acad_*" or "*_Mil_*" or "*_Minist_*" or "*_Def_*" or "*_Estab_*" or "*_Govt_*" or "*_Agcy_*" or "*_Inst_*" or "*_Ist_*" or "*_Lab_*" or "*_natl_lab*" or "*_ntl_lab*" or "*_natl_Inst_*" or "*_ntl_inst_*" or "*_ntl_ctr_*" or "*_ntl_res_*" or "*_natl_ctr_*" or "*_natl_res_*" or "*_Acad_*" or "*_adm_*" or "*_Fed_*" or "*_bur_*" or "*_office_*" or "*_survey_*" or "*_metrop_*" or "*_Fraunhofer_*"

Government institute / Abbreviations detection (second part of the address)
"*_Akad_*" or "*_Acad_*" or "*_Mil_*" or "*_Minist_*" or "*_Def_*" or "*_Estab_*" or "*_Govt_*" or "*_Agcy_*" or "*_natl_lab*" or "*_ntl_lab*" or "*_natl_Inst_*" or "*_ntl_inst_*" or "*_ntl_ctr_*" or "*_ntl_res_*" or "*_natl_ctr_*" or "*_natl_res_*" or "*_Acad_*" or "*_adm_*" or "*_Fed_*" or "*_bur_*" or "*_office_*" or "*_survey_*" or "*_metrop_*" or "*_Fraunhofer_*"

Government institute / Main institutions detection
"*_AFRC_*" or "*_ANL_*" or "*_AERE_*" or "*_BNL_*" or "*_CDC_*" or "*_CDCP_*" or "*_CENS_*" or "*_CEN_*" or "*_CEA_*" or "*_CERN_*" or "*_CAB_*" or "*_CNEN_*" or "*_CNRS_*" or "*_CSIRO_*" or "*_CSIC_*" or "*_CNR_*" or "*_CSIR_*" or "*_DESY_*" or "*_DFVLR_*" or "*_EURATOM_*" or "*_FAA_*" or "*_FCC_*" or "*_FAO_*" or "*_INRA_*" or "*_INSERM_*" or "*_KFA_JULICH_*" or "*_MRC_*" or "*_MAFF_*" or "*_NASA_*" or "*_NCI_*" or "*_NEI_*" or "*_NHLBI_*" or "*_NIAID_*" or "*_NIAMDD_*" or "*_NICHHD_*" or "*_NIDR_*" or "*_NIMH_*" or "*_NIH_*" or "*_NOAA_*" or "*_NERC_*" or "*_ORNL_*" or "*_ORSTOM_*" or "*_SERC_*" or "*_FOM_*" or "*_TNO_*" or "*_UKAEA_*" or "*_USAF_*" or "*_USDA_*" or "*_EPA_*" or "*_FDA_*" or "*_USN_*" or "*_Euratom_Assoc_*" or "*_AFSOR_*" or "*_AIST_*" or "*_Russia_Sci_Ctr_*" or "*_ANDRA_*" or "*_Angstrom_Technol_Partnership_*" or "*_ASCR_*" or "*_Australian_Nucl_Sci_&_Technol_Org_*" or "*_BESSY_*" or "*_Bhabha_Atom_Res_Ctr_*" or "*_Bundesanstalt_Mat_Forsch_&_Prufung_*" or "*_Bur_Rech_Geol_&_Minieres_*" or "*_CAS_*" or "*_CCAST_*" or "*_CCLRC_*" or "*_CEIT_*" or "*_Chem_&_Chem_Engn_Res_Ctr_Iran_*" or "*_China_Ctr_Adv_Sci_&_Technol_*" or "*_Chinese_Ctr_Adv_Sci_&_Technol_*" or "*_CIEMAT_*" or "*_CLRC_*" or "*_CMRDI_*" or "*_CNR_*" or "*_CNRS_*" or "*_CNRSM_*" or "*_Combinatorial_Mat_Explorat_&_Technol_*" or "*_Comis_Nacl_Energia_Atom_*" or "*_Commiss_European_Communities_*" or "*_Consejo_Nacl_Invest_Cient_&_Tecn_*" or "*_CREST_*" or "*_CSIC_*" or "*_Ctr_Adv_Studies_Sci_&_Technol_Microstruct_*" or "*_Ctr_Adv_Technol_*" or "*_Ctr_Atom_Bariloche_*" or "*_Ctr_Brasileiro_Pesquisas_Fis_*" or "*_Ctr_Dis_Control_&_Prevent_*" or "*_Ctr_Int_Laser_*" or "*_Ctr_Invest_Quim_Aplicada_*" or "*_Ctr_Mat_Elect_Technol_*" or "*_Ctr_Nacl_Aceleradores_*" or "*_Ctr_Nucl_Sci_*" or "*_darpa_*" or "*_DEMOCRITOS_Natl_Simulat_Ctr_*" or "*_Dept_Vet_Affairs_Med_Ctr_*" or "*_DERA_*" or "*_DIPC_*" or "*_DLR_*" or "*_doe_*" or "*_Donostia_Int_Phys_Ctr_*" or "*_ECN_Solar_Energy_*" or "*_Elect_Res_&_Serv_Org_*" or "*_Electrochem_Res_Ctr_*" or "*_ENEA_*" or "*_Energy_Res_Ctr_Netherlands_*" or "*_ESRF_*" or "*_ETRI_*" or "*_Synchrotron_*" or "*_Forschungszentrum_*" or "*_Fraunhoffer_*" or

"*_Fujitsu_Labs_Ltd_*" or "*_Geoforschungszentrum_Potsdam_*" or "*_Geol_Survey_Norway_*" or "*_German_Canc_Res_Ctr_*" or "*_GIST_*" or "*_High_Energy_Accelerator_Res_Org_*" or "*_IFW_*" or "*_IFW_Dresden_*" or "*_ILL_*" or "*_IMEC_*" or "*_IMRE_*" or "*_Indira_Gandhi_Ctr_Atom_Res_*" or "*_INFM_*" or "*_INRS_Energie_*" or "*_INSTM_*" or "*_Int_Adv_Res_Ctr_Powder_Met_&_New_Mat_*" or "*_Int_Supercond*" or "*_Interuniv_Microelect_Ctr_*" or "*_IPICyT_*" or "*_ISTEC_*" or "*_JAERI_*" or "*_Japan_Fine_Ceram_Ctr_*" or "*_Japan_Sci_&_Technol_*" or "*_JASRI_*" or "*_Jawaharlal_Nehru_Ctr_Adv_Sci_Res_*" or "*_Joint_Res_Ctr_Atom_Technol_*" or "*_JRCAT_*" or "*_JST_*" or "*_K_JIST_*" or "*_Kansai_Adv_Res_Ctr_*" or "*_KIST_*" or "*_lawrence_berkeley__*" or "*_LBL_*" or "*_LBNL_*" or "*_livermore_*" or "*_LURE_*" or "*_Max_Delbruck_Ctr_Mol_Med_*" or "*_Max_Planck_*" or "*_MPG_*" or "*_NASU_*" or "*_Nat_Hist_Museum_*" or "*_Natl_Cardiovasc_Ctr_*" or "*_Natl_Met_&_Mat_Technol_Ctr_*" or "*_Natl_Microelect_Res_Ctr_*" or "*_Natl_Nano_Device_Labs_*" or "*_Natl_Sci_Council_*" or "*_NCSR_*" or "*_New_York_State_Dept_Hlth_*" or "*_NIMS_*" or "*_NIPER_*" or "*_NIST_*" or "*_NIST_*" or "*_NMRC_*" or "*_NREL_*" or "*_Nucl_Res_Ctr_Negev_*" or "*_Nucl_Sci_Ctr_*" or "*_Off_Natl_Etud_&_Rech_Aerosp_*" or "*_Off_Naval_Res_*" or "*_Palo_Alto_Res_Ctr_*" or "*_Phys_Tech_Bundesanstalt_*" or "*_PRESTO_*" or "*_RAS_*" or "*_Res_Ctr_Energy_Convers_&_Storage_*" or "*_Res_Ctr_Rossendorf_*" or "*_Riken_*" or "*_RIKEN_*" or "*_Russian_Res_Ctr_*" or "*_sandia_*" or "*_SAS_*" or "*_Sincrotrone_Trieste_*" or "*_SINOPEC_*" or "*_SINTEF_*" or "*_Stanford_Linear_Accelerator_Ctr_*" or "*_TRIUMF_*" or "*_UFRGS_*" or "*_UNESP_*" or "*_UOP_LLC_*" or "*_US_DOE_*" or "*_US_Geol_Survey_*" or "*_USN_*" or "*_Vet_Adm_Med_Ctr_*" or "*_Vet_Affairs_Med_Ctr_*" or "*_VTT_*" or "*_WHO_*" or "*_Zentrum_Sonnenenergie_&_Wasserstof*" or "*_Zentrum_Sonnenergie_&_Wasserstoff_*" or "*_ZSW_*"

Hospital / Abbreviations detection

"*_Hop_*" or "*_Hosp_*" or "*_Osped_*" or "*_Med_cent_*" or "*_CHR_*" or "*_CHU_*" or "*_med_ctr_*" or "*_clin_*" or "*_eye_ctr_*" or "*_vet_*" or "*_ctr_med_*" or "*_Canc_Ctr_*" or "*_canc_res_*" or "*_heart_*" or "*_john_hopkins_*" or "*_Kaiser_Permanente_*" or "*_eye_*" or "*_blood_*"

Firm / Abbreviations detection

"*_Co_*" or "*_Corp_*" or "*_gesell_*" or "*_inc_*" or "*_gmbh_*" or "*_associate*" or "*_bhd_*" or "*_consult_*" or "*_llc_*" or "*_ltd_*" or "*_SA_*" or "*_semicon_*" or "*_venture_*" or "*_ab_*" or "*_Spa_*" or "*_AG_*" or "*_PLC_*" or "*_SARL_*" or "*_EURL_*"

Firm / Main institutions detection

"*_AEG_*" or "*_ALCOA_*" or "*_ABC_*" or "*_BASF_*" or "*_CBS_*" or "*_DUPONT_*" or "*_GEC_*" or "*_ICI_PLC_*" or "*_3M_CO_*" or "*_NBC_*" or "*_SKF_*" or "*_SK&F_*"

Other

All other names that don't contain the pattern used for the other classes, and also with : "*_Assoc_*" or "*_Fdn_*"

Appendix 3 - Institutions standardised, the five main institutions per country

Main (top 5) institutions per country	*Country code harmonized*	*Total of publications for the countries*	*Institutions standardized*	*Number of addresses per institutions*
ARGENTINA	AR	7721	cnea	1329
ARGENTINA	AR	7721	univ buenos aires	1288
ARGENTINA	AR	7721	conicet	1174
ARGENTINA	AR	7721	univ nacl la plata	925
ARGENTINA	AR	7721	univ nacl cordoba	425
AUSTRIA	AT	12357	univ tech vienna	2381
AUSTRIA	AT	12357	univ vienna	1887
AUSTRIA	AT	12357	univ linz	1166
AUSTRIA	AT	12357	univ tech graz	887
AUSTRIA	AT	12357	univ graz	727
AUSTRALIA	AU	30507	univ queensland	2838
AUSTRALIA	AU	30507	univ new s wales	2727
AUSTRALIA	AU	30507	univ sydney	2727
AUSTRALIA	AU	30507	CSIRO	2160
AUSTRALIA	AU	30507	univ melbourne	2122
BELGIUM	BE	17872	kul	3549
BELGIUM	BE	17872	univ cathol louvain	1984
BELGIUM	BE	17872	univ antwerp	1947
BELGIUM	BE	17872	univ ghent	1891
BELGIUM	BE	17872	imec	1595
BRAZIL	BR	31330	univ sao paulo	5346
BRAZIL	BR	31330	univ campinas	2956
BRAZIL	BR	31330	univ estad paulista	1983
BRAZIL	BR	31330	univ fed rio janeiro	1696
BRAZIL	BR	31330	univ fed sao carlos	1654
CANADA	CA	43183	univ toronto	4798
CANADA	CA	43183	univ mcgill	3002
CANADA	CA	43183	nrc	2934
CANADA	CA	43183	univ alberta	2668
CANADA	CA	43183	univ british columbia	2401
SWITZERLAND	CH	23739	ethz	7243
SWITZERLAND	CH	23739	EPFL	5092
SWITZERLAND	CH	23739	univ basel	1890
SWITZERLAND	CH	23739	univ geneva	1141
SWITZERLAND	CH	23739	EMPA	1029
CHINA	CN	281585	chinese acad sci	48763

CHINA	CN	281585	univ tsinghua	9797
CHINA	CN	281585	univ nanjing	8967
CHINA	CN	281585	univ zhejiang	8094
CHINA	CN	281585	univ china sci & technol	7799
CZECH REPUBLIC	CZ	12117	czech acad sci	5387
CZECH REPUBLIC	CZ	12117	univ charles	1946
CZECH REPUBLIC	CZ	12117	prague inst chem technol	668
CZECH REPUBLIC	CZ	12117	Univ Masaryk brno	532
CZECH REPUBLIC	CZ	12117	univ tech czech	425
GERMANY	DE	138005	max planck	14927
GERMANY	DE	138005	helmholtz	9302
GERMANY	DE	138005	leibniz	6760
GERMANY	DE	138005	KIT	5285
GERMANY	DE	138005	univ munchen	3660
DENMARK	DK	9892	univ tech denmark	3126
DENMARK	DK	9892	univ aarhus	1879
DENMARK	DK	9892	univ copenhagen	1678
DENMARK	DK	9892	univ so denmark	649
DENMARK	DK	9892	univ aalborg	450
SPAIN	ES	49044	csic	10229
SPAIN	ES	49044	Univ Barcelona	3201
SPAIN	ES	49044	univ aut madrid	2468
SPAIN	ES	49044	univ complutense madrid	2285
SPAIN	ES	49044	univ pais vasco	2157
FINLAND	FI	11217	univ helsinki	1994
FINLAND	FI	11217	univ tech helsinki	1939
FINLAND	FI	11217	univ turku	922
FINLAND	FI	11217	univ abo akad	817
FINLAND	FI	11217	vtt	692
FRANCE	FR	109156	cnrs	17053
FRANCE	FR	109156	cea	6904
FRANCE	FR	109156	upmc	5545
FRANCE	FR	109156	univ paris sud	4838
FRANCE	FR	109156	univ lyon claude bernard	4194
UNITED KINGDOM	GB	83092	univ cambridge	7498
UNITED KINGDOM	GB	83092	univ oxford	4700
UNITED KINGDOM	GB	83092	imperial college	4139
UNITED KINGDOM	GB	83092	ucl	3825
UNITED KINGDOM	GB	83092	univ manchester	3638
GREECE	GR	10575	univ thessaloniki	1806
GREECE	GR	10575	ncsr demokritos	1434
GREECE	GR	10575	univ patras	1411
GREECE	GR	10575	FORTH	1003

GREECE	GR	10575	univ tech athens	797
HUNGARY	HU	7946	hungarian acad sci	2912
HUNGARY	HU	7946	univ szeged	1248
HUNGARY	HU	7946	univ eotvos lorand	969
HUNGARY	HU	7946	univ tech & eco budapest	920
HUNGARY	HU	7946	univ debrecen	417
IRELAND	IE	6078	trinity coll dublin	1760
IRELAND	IE	6078	univ coll cork	1137
IRELAND	IE	6078	univ dublin city	682
IRELAND	IE	6078	univ limerick	659
IRELAND	IE	6078	univ coll dublin	611
ISRAEL	IL	15031	technion	2798
ISRAEL	IL	15031	univ hebrew jerusalem	2735
ISRAEL	IL	15031	univ tel aviv	2134
ISRAEL	IL	15031	weizmann inst sci	2042
ISRAEL	IL	15031	Univ ben gurion negev	1731
INDIA	IN	57760	csir	7567
INDIA	IN	57760	indian inst sci	2933
INDIA	IN	57760	bhabha atom res ctr	2413
INDIA	IN	57760	iit kharagpur	2253
INDIA	IN	57760	indian assoc cultivat sci	1617
IRAN	IR	14998	univ tehran	1355
IRAN	IR	14998	univ tech sharif	1244
IRAN	IR	14998	univ islam azad	1117
IRAN	IR	14998	univ tarbiat modarres	805
IRAN	IR	14998	univ tech teheran amirkabir	689
ITALY	IT	73200	cnr	12358
ITALY	IT	73200	univ padua	2628
ITALY	IT	73200	univ roma la sapienza	2419
ITALY	IT	73200	univ bologna	2399
ITALY	IT	73200	univ milano	2256
JAPAN	JP	92937	univ tohoku	12226
JAPAN	JP	92937	univ osaka	10637
JAPAN	JP	92937	univ kyoto	5915
JAPAN	JP	92937	univ kyushu	5567
JAPAN	JP	92937	univ nagoya	5315
SOUTH KOREA	KR	102001	univ natl seoul	9718
SOUTH KOREA	KR	102001	kaist	5795
SOUTH KOREA	KR	102001	univ hanyang	5025
SOUTH KOREA	KR	102001	univ yonsei	4593
SOUTH KOREA	KR	102001	postech	4166
MEXICO	MX	14086	univ nacl aut mexico	4047
MEXICO	MX	14086	inst politecn Nacl	2498

MEXICO	MX	14086	conacyt	1423
MEXICO	MX	14086	univ aut metropolitan mexico	969
MEXICO	MX	14086	Inst Mexicano Petr	606
NETHERLANDS	NL	26336	univ tech delft	3706
NETHERLANDS	NL	26336	univ twente	2752
NETHERLANDS	NL	26336	univ tech eindhoven	2722
NETHERLANDS	NL	26336	Univ utrecht	2176
NETHERLANDS	NL	26336	univ groningen	1949
NORWAY	NO	5113	ntnu	1244
NORWAY	NO	5113	univ oslo	1185
NORWAY	NO	5113	univ bergen	444
NORWAY	NO	5113	sintef	372
NORWAY	NO	5113	univ tromso	190
POLAND	PL	24479	polish acad sci	5628
POLAND	PL	24479	univ tech warsaw	1508
POLAND	PL	24479	univ warsaw	1310
POLAND	PL	24479	univ tech wroclaw	1241
POLAND	PL	24479	univ jagiellonian	986
PORTUGAL	PT	11191	univ aveiro	2250
PORTUGAL	PT	11191	univ porto	1358
PORTUGAL	PT	11191	univ tech lisbon	1358
PORTUGAL	PT	11191	univ minho	1122
PORTUGAL	PT	11191	univ coimbra	909
ROMANIA	RO	10476	romanian acad sci	1279
ROMANIA	RO	10476	natl Inst mat phys	1003
ROMANIA	RO	10476	univ polytech bucharest	898
ROMANIA	RO	10476	univ babes bolyai	779
ROMANIA	RO	10476	univ bucharest	753
RUSSIA	RU	49926	russian acad sci	27843
RUSSIA	RU	49926	univ state moscow lomonosov	4614
RUSSIA	RU	49926	univ state st petersburg	1255
RUSSIA	RU	49926	kurchatov inst	605
RUSSIA	RU	49926	univ state novosibirsk	543
SWEDEN	SE	21366	univ uppsala	3217
SWEDEN	SE	21366	Univ lund	2940
SWEDEN	SE	21366	KTH	2670
SWEDEN	SE	21366	univ tech chalmers	2620
SWEDEN	SE	21366	univ linkoping	1873
SINGAPORE	SG	21484	univ natl singapore	9530
SINGAPORE	SG	21484	univ tech nanyang	6372
SINGAPORE	SG	21484	a.star	3723
SINGAPORE	SG	21484	a star	600

SINGAPORE	SG	21484	chartered semicond mfg ltd	219
THAILAND	TH	6129	univ chulalongkorn	1624
THAILAND	TH	6129	univ mahidol	756
THAILAND	TH	6129	univ chiang mai	715
THAILAND	TH	6129	natl sci & technol dev agcy	449
THAILAND	TH	6129	king mongkuts inst technol	372
TURKEY	TR	13799	univ tech middle east	988
TURKEY	TR	13799	univ hacettepe	910
TURKEY	TR	13799	univ tech istanbul	839
TURKEY	TR	13799	univ bilkent	717
TURKEY	TR	13799	univ gazi	529
TAIWAN	TW	56816	univ natl taiwan	7057
TAIWAN	TW	56816	univ natl cheng kung	5872
TAIWAN	TW	56816	univ natl tsing hua	4861
TAIWAN	TW	56816	univ natl chiao tung	4749
TAIWAN	TW	56816	acad sinica	2341
UKRAINE	UA	10036	ukrainian acad sci	6464
UKRAINE	UA	10036	univ natl kiev	765
UKRAINE	UA	10036	Univ natl lvov	267
UKRAINE	UA	10036	univ natl tech kiev	244
UKRAINE	UA	10036	univ natl kharkov	228
UNITED STATES	US	472303	MIT	9212
UNITED STATES	US	472303	Univ Illinois Urbana	8377
UNITED STATES	US	472303	univ penn state	7073
UNITED STATES	US	472303	univ michigan	7021
UNITED STATES	US	472303	univ northwestern	6955

Fig & tab 31. *Main (top 5) institutions per country*