RISIS

Research infrastructure for research and innovation policy studies

SEVENTH FRAMEWORK PROGRAMME

The nano S&T dataset

Philippe Larédo, Lionel Villard, Michel Revollo

Marne la Vallée, 8/10/2015

UNIVERSITÉ PARIS-EST

École des Ponts ParisTech

IFRIS

MANCHESTER 1824

The University of Manchester

MIoIR

The presentation

- Why this dataset
- 3 major developments for building the DB
- The DB structure
- Some perspectives

Why Nano S&T (1)

- 1990s – a fast emerging new science capturing the attention of multiple audiences (including science fiction, see Drexler & the engines of creation)
- 2000 – the US National Nanotechnology initiative (NNI) is created to organise federal level priority (turned into law in 2002). Since then a national priority in most countries (BRIC included).
- A growing 'hype' nurtured by strong market anticipations (e.g. Lux Research) …
- And a growing social debate: what future society (e.g. human enhancement) and what risks (ELSA)? The multiplication of public debates

Why nano S&T DB (2)

- But multiple questions about what do we speak about
- Some articles (even recent ones) speak of some thousands patents while others speak of some hundreds thousands
- Are we speaking of a new science per se (emerging from interdisciplinary research) or of simply a new reduced scale of analysis for existing disciplines
- Some analysts consider that a new industry will emerge (like for computer science) or that an industry will be radically transformed (like biotechnology), others have different views (based on a renewed view of general purpose technologies)
- Where does it take place? In the same countries and the same places as usual? Or do we witness transformations?

Ref: Larédo et al, 2010

Why nano S&T DB (3)

- The nano S&T DB tries to address these questions
- By collecting publications and patents dealing with nano S&T
- By identifying actors (focusing on organisations & not individuals)
- By geolocalizing activities and delineating spaces where activities take place
- By enabling thematic analyses (both using existing categories, or mobilising semantic tools (and in particular CorText manager, another development of IFRIS and RISIS)

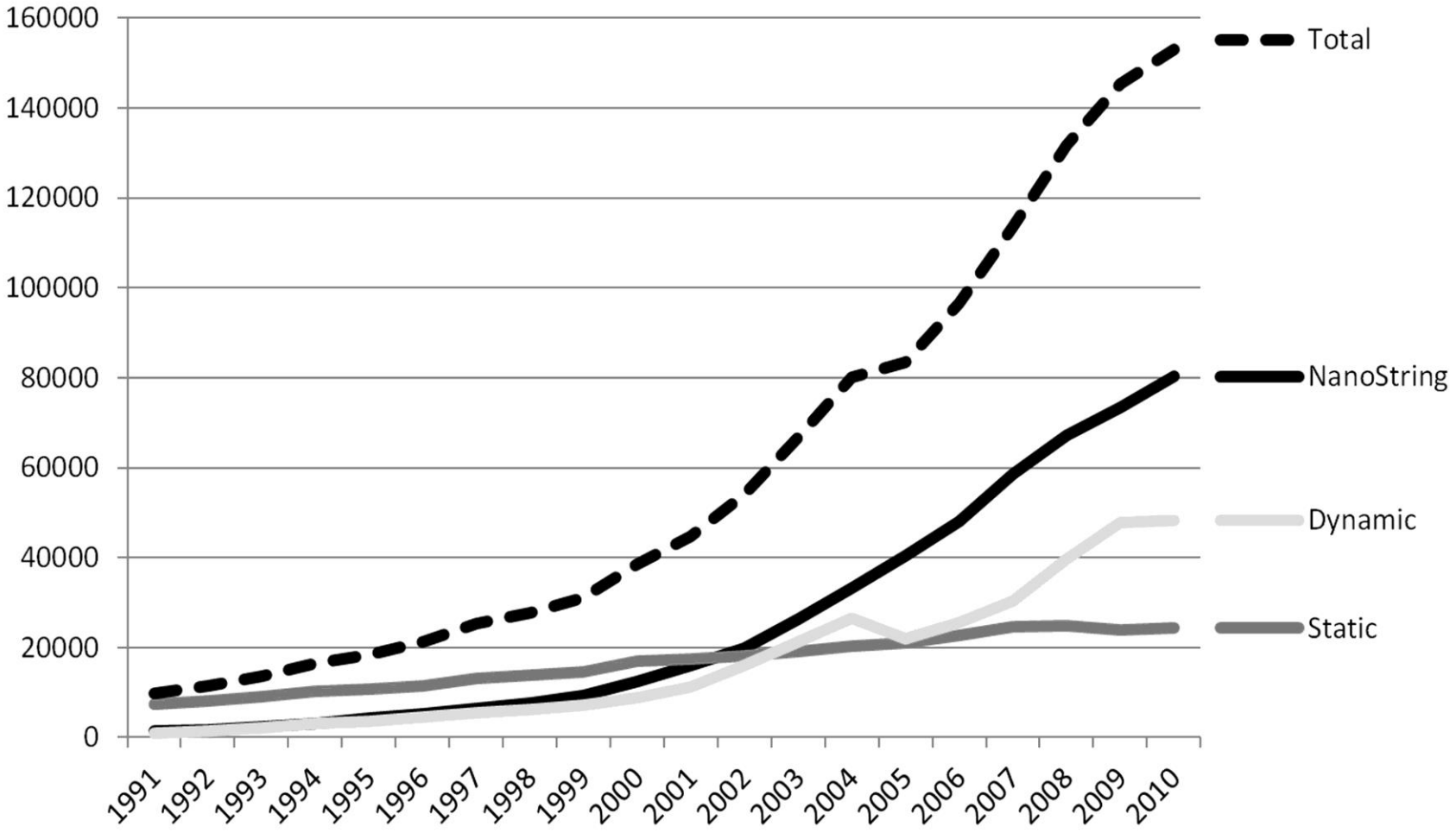
3 major developments for building the nano DB

- Develop a delineation process of nano papers & patents from the WoS and Patstat
- Develop new computer-based approaches for geolocalisation and clustering
- Use a manual treatment of organisation names based upon internal reference databases built for other projects

D1 Delineating ‘nano’ papers - 1

- A fully lexical method developed in 2 steps
  - a static approach – Mogoutov & Kahane 2007
  - a dynamic extension – Kahane et al. 2014
- A 5 steps approach
  (i) retrieval of a core ‘nano’ dataset
  (ii) lexical analysis of this core for identifying keywords to consider
  (iii) definition of principles for extension
  (iv) static extension
  (v) dynamic extension

An overview of results



Delineating ‘nano’ papers - 2

1- Building the core dataset

- The choice for a simple query for initial downloading
  - far faster to download
  - using then the different papers produced by colleagues for further targeted exclusions (e.g. plankton, flagel, or nanomolar)

The query for the nanostring

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*))

Delineating ‘nano’ papers - 3

- Important results:
  - 517 000 articles from 1991 to 2010
  - a yearly growth of 20% during 15 years (40000 articles in 2005), doubling in 2010 (80000 articles)
  - overall the nanostring represents 14% of papers in 1991, 30% in 1999, 48% in 2005, stable since around 50%

Delineating 'nano' papers - 4

2- Preparing the lexical extension

- The use of CorText manager for this step
- (a) selection of candidate 'noun phrases'
- (b) selection of most relevant multi-terms based on

- high 'unithood' (using C value, see Frantzi & Ananiadou, 2000). Central assumptions: 'pertinent terms tend to appear more frequently and longer phrases are more likely to be relevant'
- sorting by descending order, selection of 4N terms (N being the objective looked for, here 2000 per year)
- high 'termhood' aims at discarding terms with low meaning (ie with unbiased distribution, Van Eck et al, 2011), builds '*internal specificity*'

- Results:

- there are less than 2000 terms until 1997 and an overall theoretical number of over 34000 multi-terms
- but de facto only **3900 different multi-terms** (after cleaning non nano terms, see exclusions before)
- overall 17 million occurrences i.e. **33 multi-terms per article**!

Delineating ‘nano’ papers - 5

3- Principles for the extension

•One important result of early works: most authors did not use the ‘nano’ prefix, so need for tapping relevant papers
→ need to tap core terms over the period: static extension

•Bonaccorsi (2010): the importance in emerging sciences of explorations many of which fail but explain the effective trajectory → need to tap year-after-year relevant terms

•No process to define the level of extension to be done, work by experience looking at results arrived at through other approaches → drives to multiply on average by 3 the initial seed
→ our choice: each increase (static and dynamic) will theoretically double the seed.

Delineating ‘nano’ papers - 6

RISIS

4- the static extension

- Building the external specificity of multi-terms

- ratio presence in nanostring/total presence in WoS
- done year by year, provides yearly sequence, individual year & average ratios

- Selecting multi-terms

- select terms present over half of the period and part of the 250 highest termhoods (1105 terms identified),
- rank them in descending order of external specificity
- stop at the theoretical addition of articles nearest to the core set (517000 articles).

- Results:

- we find the same **external specificity threshold** as in 1st dataset: 26%
- static extension based on 114 multi-terms present 18 years out of 20 (with a skewed distribution: 13 terms bring 66%)

Analysis of multi-terms of the static extension

Type of multi-term	Nber of multi-terms	% of theoretical addition
Observation, manipulation & control (TEM, AFM, STM…)	30	57%
Materials (TIO2, graphene,..) & CNT, nanowires	37	23%
Characteristics & properties of materials, molecules, genes	36	12%
Fabrication / expression techniques	11	8%
	114	100%

Delineating 'nano' papers - 7

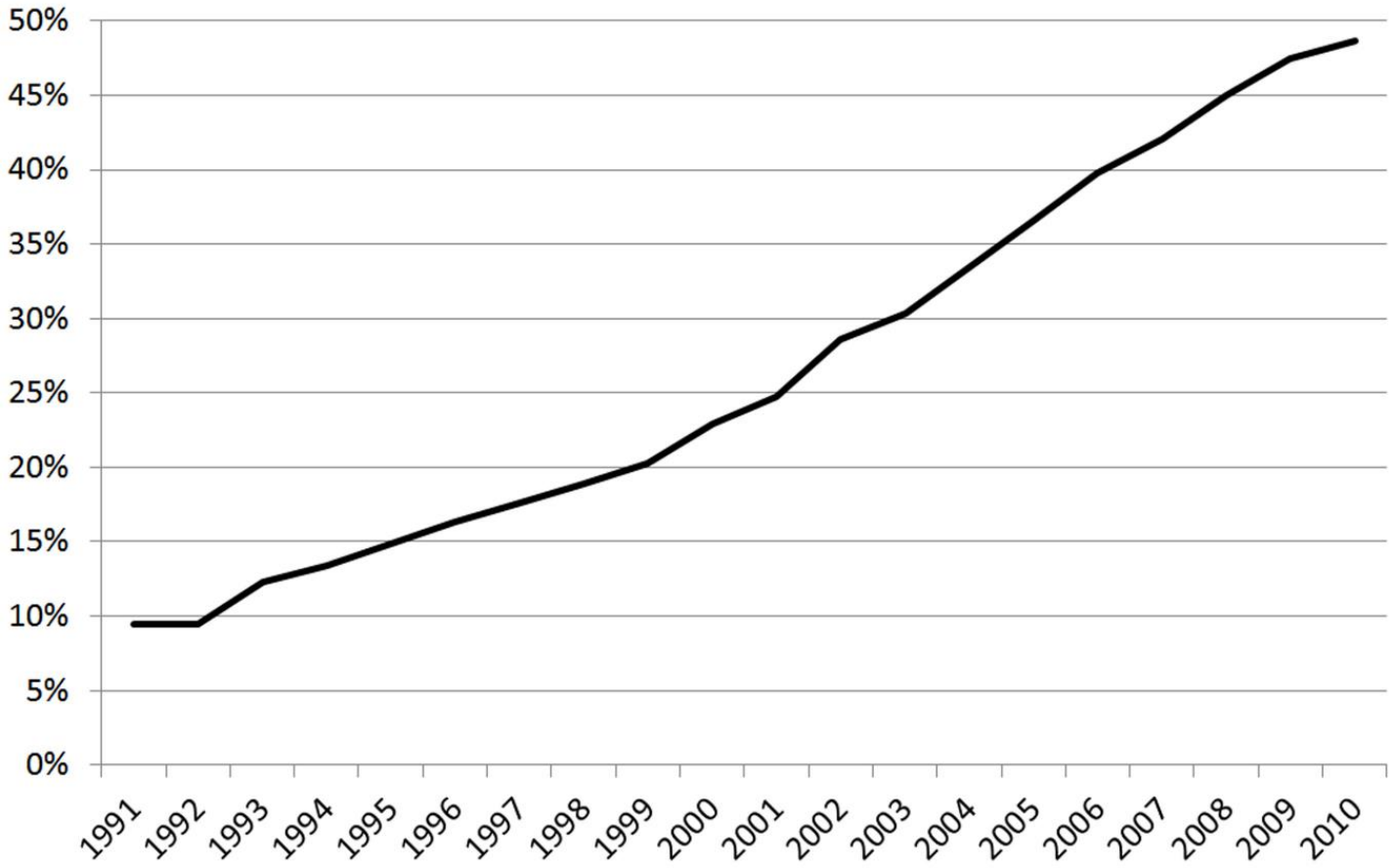
5- The dynamic extension

- 2 objectives: (i) extend the corpus thanks to terms present in certain years, translating explorations; (ii) take hold of on-going transformations
- Calculating external specificity: same approach as for the static extension using all terms but static ones & ranking them for each year by descending external specificity
- Defining the theoretical number of articles per year as equal to the number that year in the nanostring.
- This gives us a list of relevant terms per year: 558 in total with an **average presence of 5 years**, 30% bringing 90% of the theoretical addition
  - moves from 1% of total terms in 1991-92 to 25% in 1996-99, and 41% in 2008-2010
  - 27% appeared & died in 1st decade; 52% emerged in the 2nd decade and 22% moved from one decade to the other
  - the external specificity threshold increases from 10 to 50% over the years

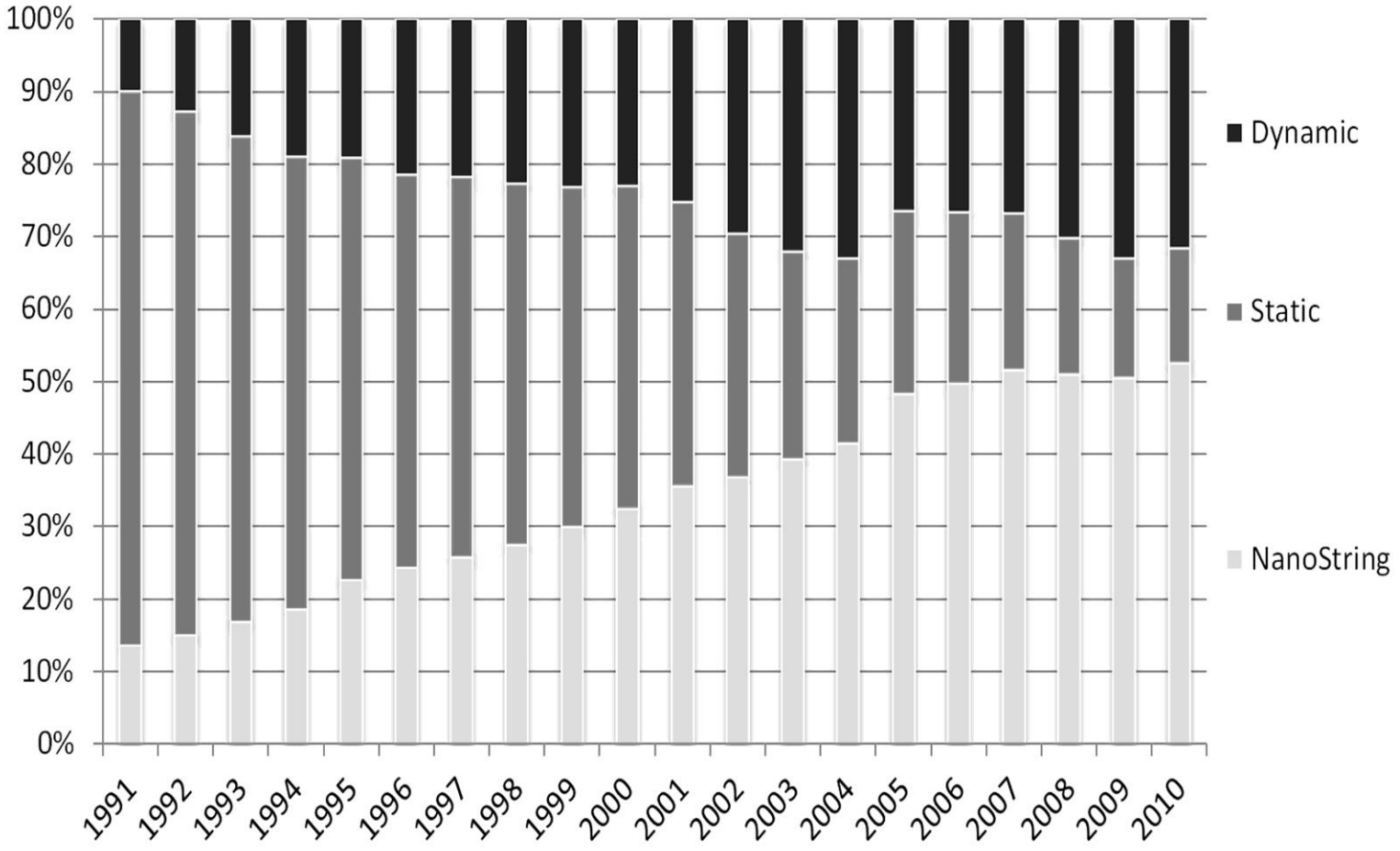
Analysis of multi-terms of the dynamic extension

Type of multi-term	Nber of multi-terms	% of theoretical addition
Observation, manipulation & control (TEM, AFM, STM…)	32	**19%**
Materials (TIO2, graphene,..) & CNT, nanowires	38	28%
Characteristics & properties of materials, molecules, genes	28	14%
Measures	22	8%
Fabrication / Production	30	**28%**
Applications	12	3%
	162	100%

Evolution of the external specificity over time for the dynamic extension



The overall publication dataset



Moving from publications to patents 1

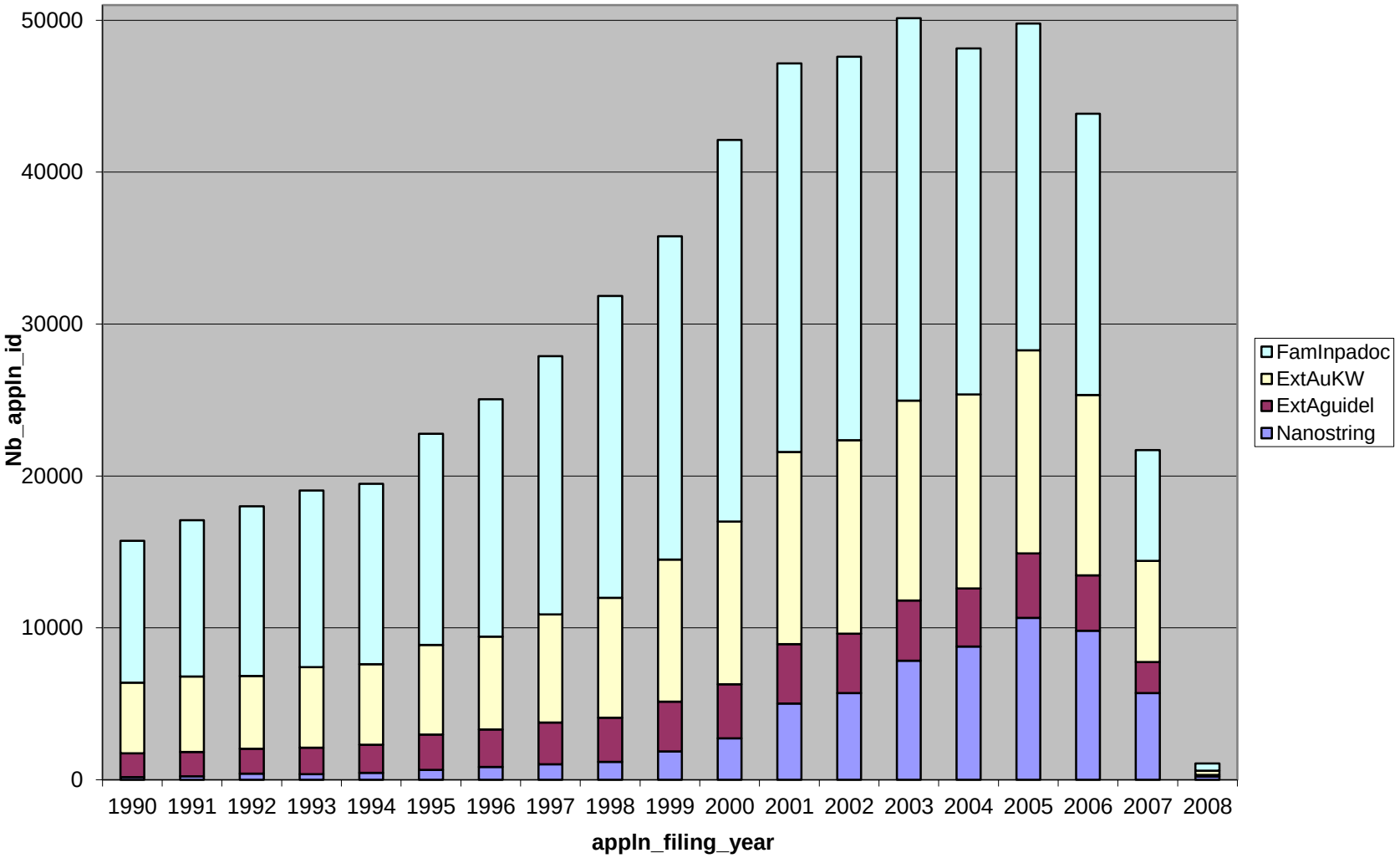
- Looking for technology dynamics, thus the choice of applications and not granted patents, and the choice of the first filing date (in order to have the nearest date to technology development)
- Focusing on organisations and on location (to follow dynamics) → requires to focus on inventor addresses to avoid the bias of headquarters strategies in patenting
- The limits of US PTO for non US inventor addresses → thus the choice of Patstat as the database of all databases of patent offices
- A focus on priority patents, but not loosing track of extensions (as proof of 'technology value')

Moving from publications to patents 2

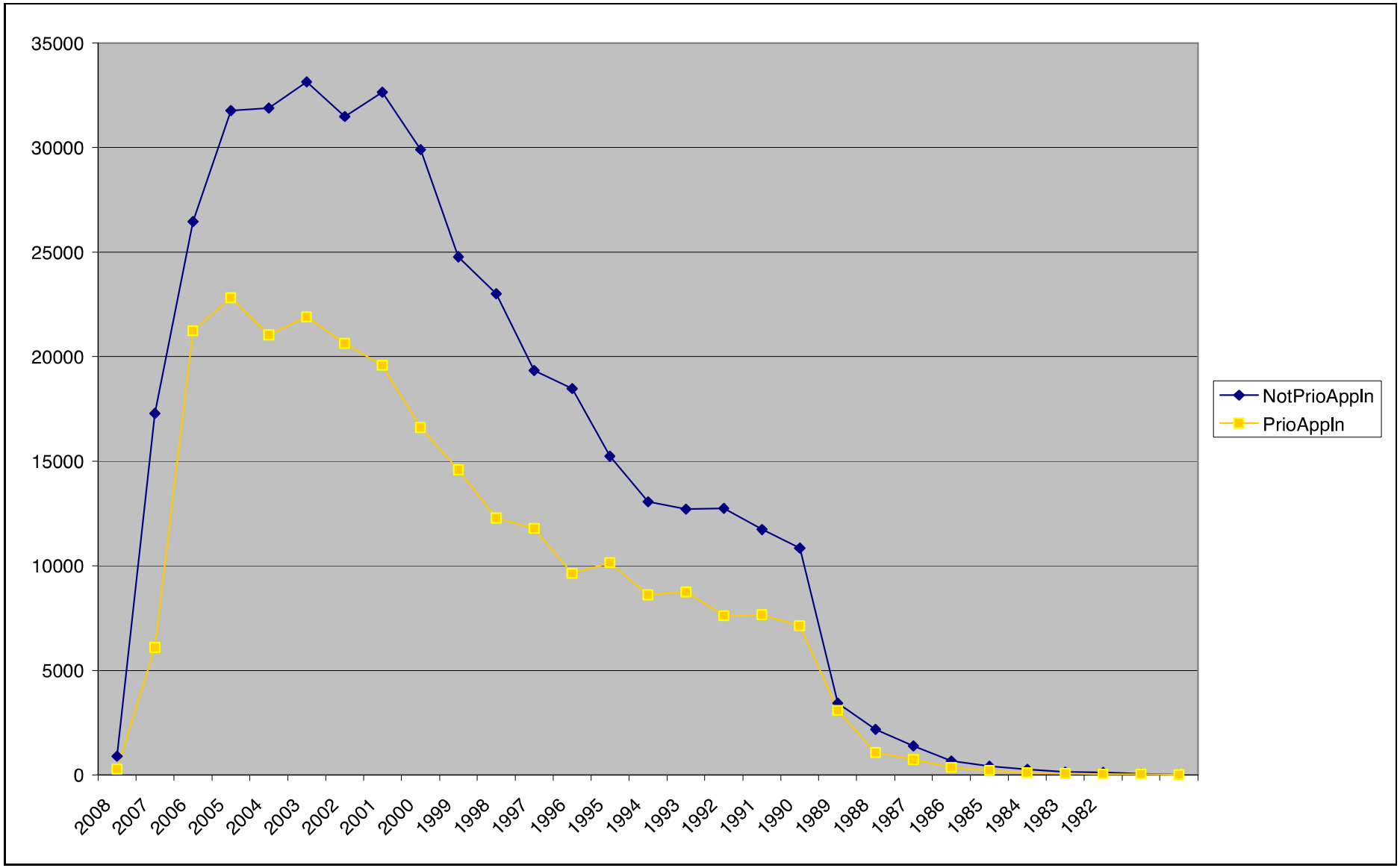
- NanoDB V1
  - we use the same vocabulary as for publications for selecting applications (on top of nanostring)
  - argument: Bonaccorsi (2007): 70% of nanoinventors are also academic authors.
- NanoDB V2 (under construction)
  - testing the method on 1 year: 80% of the vocabulary is different but 80% of the patents are the same…
  - testing the full method: works with the static extension (5 times more multi-terms), but not with the dynamic one (words too specific driving to very low external specificity thresholds)
  - developing a combined approach: using publication multi-terms not present in the patent static extension, selecting those above the average presence (5 years) and the average external specificity (25%).

Nano patstat V1

- Construction in 3 steps
  - using patstat 2009 (ie until 2006)
  - step 1: extracting the patents using the nanostring → gives 63580 patents
  - step 2: extracting the extension following the 2007 approach (Mogoutov & Kahane) → gives 207000 patents
  - step 3: incorporating patents from INPADOC families → gives 327000 patents
- Results
  - overall 660000 patents but only **254000 priority patents, 190000 INPADOC families**


50000
40000
30000
20000
10000
0
Nb_appln_id
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
appln_filing_year
FamInpadoc
ExtAuKW
ExtAguidel
Nanostring

A fast growth followed by a plateau since 2000



The importance of ‘one-off’ patents

Nbr applications per iNPADOC family	Nber of inpadoc families	Nber of priority patents	% priority patents
1 only (singletons)	97800	97800	38%
2 to 5	61216	74862	29%
6 to 10	21610	35800	14%
11 to 50	8540	29205	11%
51 to 100	314	4577	2%
101 to 500	164	6631	3%
501 to 1000	10	2476	1%
More than 1000	4	2733	1%
total	189658	254084	100%

Few countries concentrate priority patents …

office	nber
US	93759
CA	798
AU	1929
RU	2827
IN	330
ZA	103
BR	Below 100

office	nber
JP	68937
KR	27060
CN	24798
TW	5076
SG	165

office	nber
DE	12568
GB	3863
EPO	3088
FR	2893
SE	697
IT	661
CH	414
NL	378
IL	337

…BUT

- Are all patents equal
- Can we trust offices as a source for identifying where knowledge is produced?

Strong country differences
if we consider the 'value' of patents
(i.e. those valorised are those 'extended' (not singletons)

office	total	% single-ton	Net total
US	93759	11,5%	82966
JP	68937	47,5%	36189
KR	27060	65,5%	9335
CN	24798	95,1%	1218
EU	25240	21,4%	19829

region	% crude total	% net total
US	38,3%	54,7%
JP	28,2%	23,9%
Other Asia	23,3%	8,3%
EU	10,3%	13,1%

D2-Localising inventive activities

- Principle: Moving from offices to the addresses of actors
- Choice: not stay at organisation's address and consider the location of inventors
- Limitation: many addresses not filled (wrongly located by Patstat, artificial patents, simply missing) → drove to important developments (OECD, Munich, Leuven, and Paris); see their integration in Patstat IFRIS.
- Hypotheses:
  - very strong agglomeration process in metropolitan areas
  - important collaborations mostly between metropolitan areas
  → countries thus not a good entry point to understand dynamics; need for developing clustering methods

Localising inventive activities - 2

- Three steps:
  (i) reshaping addresses → see presentation on Patstat IFRIS
  (ii) geolocalising addresses
  (iii) clustering addresses
  → see overall chart next slide
- Here only key elements presented
- Reference for full presentation: Villard-Revollo, 2015 'geographical concentration of S&T activities', RISIS website

Methodological steps

Pre-processing
Preparing addresses

M1.1 Data expansion
- External sources (e.g. RegPat, INPI)
- Addresses propagation (e.g. INPADOC families)
- Statistical propagation (e.g. using standardized organisation names)

M1.2 Cleaning addresses
- Punctuation suppression (except comma)
- Special characters suppression (e.g. "Š" "Æ")
- Country codes standardization (ISO 3166-1 alpha-2)

M1.3 Parsing addresses
- Parsing using comma
- Selection of the three last values
- Grouping distinct values

Geocoding
Identifying toponyms and attribution of the geographic coordinates

Postal codes phase — 45.5%

M2.1 Geocoding with postal codes

Constraint by country codes
AND
Regular expression"/[0-9]{4,}/"
- Finding Geonames "postal codes"

M2.2 Geocoding with postal codes

Constraint by country codes
- Finding Geonames "place name"

Toponym's names phase — 46.5%

M3.1 Geocoding common toponyme's names

Constraint by country codes
AND
specific type of toponymes (e.g. ADM2, ADM3)
- Matching on the toponyme names
- Finding "place names"

M3.2 Geocoding with toponyme's vernacular names

Constraint by country codes
- Matching on the toponyme vernacular names
- Finding the corresponding "place names"

Missing cities and coordinates — 6.3%

M4.1 Missing citiy names identification
- Genral address structure (Postal codes, Regional codes, City) for 25 countries (e.g. FR, CH, BE, AT...)
- Use of pivot toponymes with rules (e.g. "-shi" for japan's cities, state names in USA) for GB, DE, AU, CA, KR, MX, US, JP

M4.2 BatchGeo engine

Use Google API
- Cities, (in some cases, state), countries

Clustering
Identification of the areas of aggregation of the S&T activities

M5.1 Initial clusters

Density based cluster
- Customized DBScan algorithm
- Minimal amount of addresses inside a perimeter

M5.2 Merging initial clusters

Relational and structural approach
- CHAMELEON algorithm
- Collaborations between clusters and similarities on the density of collaborations inside clusters

M5.3 Building geograpical objects
- Boundaries of clusters using a convex hull algorihtm
- Calculating barycentres for merged clusters

Geolocalisation: 3 phases

- use of postal codes by comparing extracted postal code with GeoNames postal codes (over 900000 codes proposed with their place-name, latitude & longitude) → 45.5% of addresses geocoded
- use of toponymes by comparing extracted city names with corresponding toponymes of GeoNames (using country code & selected administrative & PPL features of GeoNames (for disambiguation) → 46.5% of addresses geocoded
- Dealing with non identified addresses → specific address structure identified for 35 countries (including state/regional division for 8) & submitted to BatchGeocode (using the 9 level accuracy filtering) → 6.3% of addresses geocode
- Result: Only 1.7% of total addresses not geocoded

Geolocalisation: an important comment

- As the geolocalisation process works well, the coverage of the DB depends on the existence of addresses → we decided recently to **extend coverage**
- Even with Patstat IFRIS solutions for additions (addresses in other Patstat fields, Regpat & other additions, propagation for artificial patents), we have only **50% of addresses filled**.
- We thus decided of a **logical extension** (called standard name extension) after clusterisation and organisation identification**, in 2 steps**
  (i) allocating the institution address to authors of singleton applications: brings 7% more addresses
  (ii) extending the geocoding within 1 patent for non covered authors when all addresses for that patent are in one cluster and that there is a number of such patents by the same organisation in that cluster: adds 62% addresses to the initial set!
- Overall this drives **to 88% total coverage of addresses** without changing the **coverage of patent applications: 85%**
- **Note: all data below do not take into account this extension**

Building clusters 1

- 2 successive methods applied for Nano DB V1 and newcoming nano DB V2
- Both are based on 2 sequential processes
  (i) initial clustering:
  - in V1 we use core cities (with more than 1000 publications) and aggregate all points in a radius of 50 km (depending on countries & their population density)
  - in V2 we use a density-based algorithm (DBScan that requires feeding 2 parameters: the minimal number of points surrounding each point; and the maximum distance between 2 points)

Building clusters -2

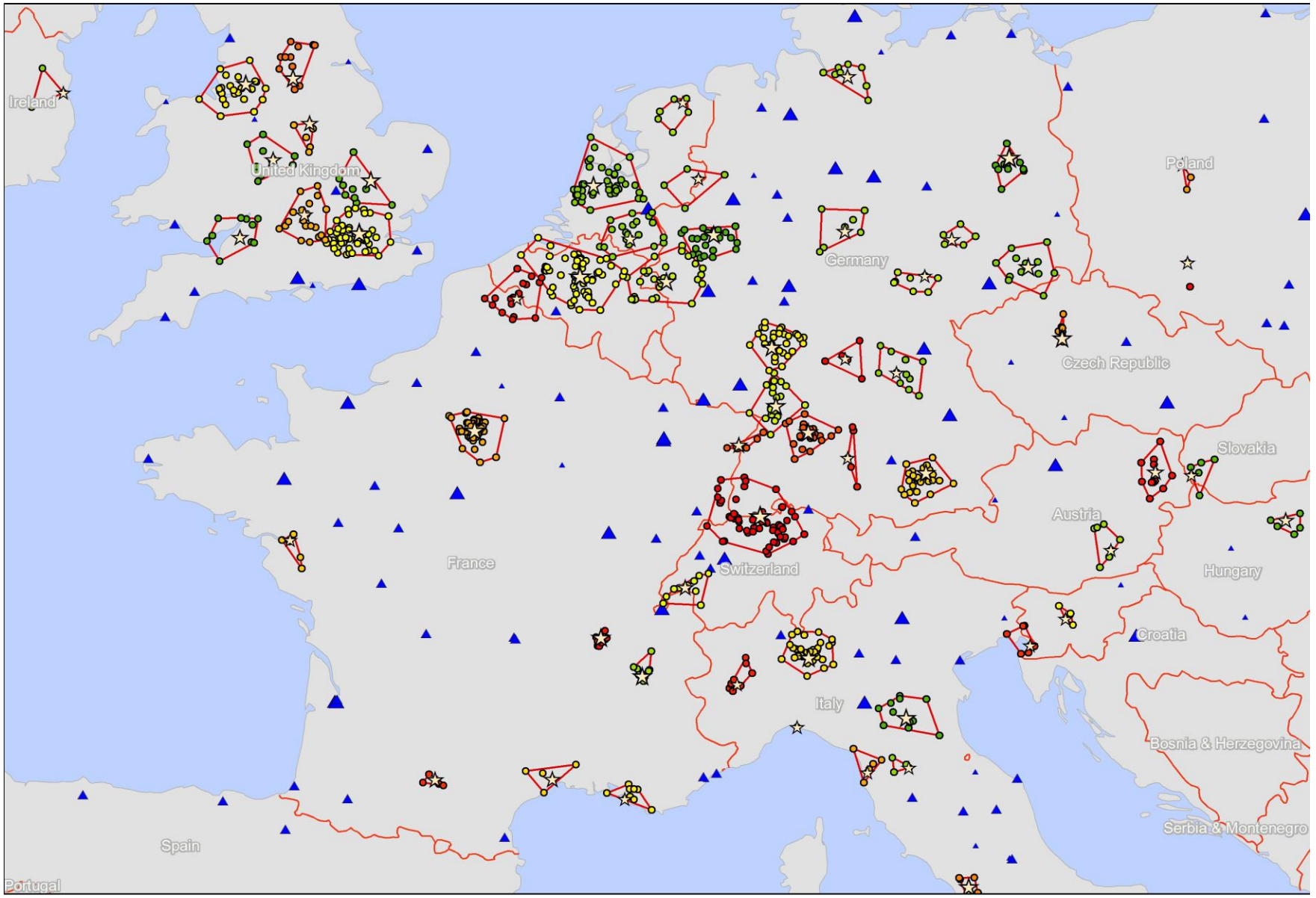
- (ii) agglomerating interconnected clusters
  - in V1 we use the overlap between clusters in term of addresses (clusters are merged when overlap > 20%)
  - in V2 we use the 'chameleon approach' (Karypis, Han & Kumar, 1999) which considers 2 parameters
  * relative interconnectivity (measured as a ratio between bilateral links and internal links) under a maximum geographical constraint
  * relative closeness: measures the similarity of collaboration profiles of the 2 clusters

Building clusters -3

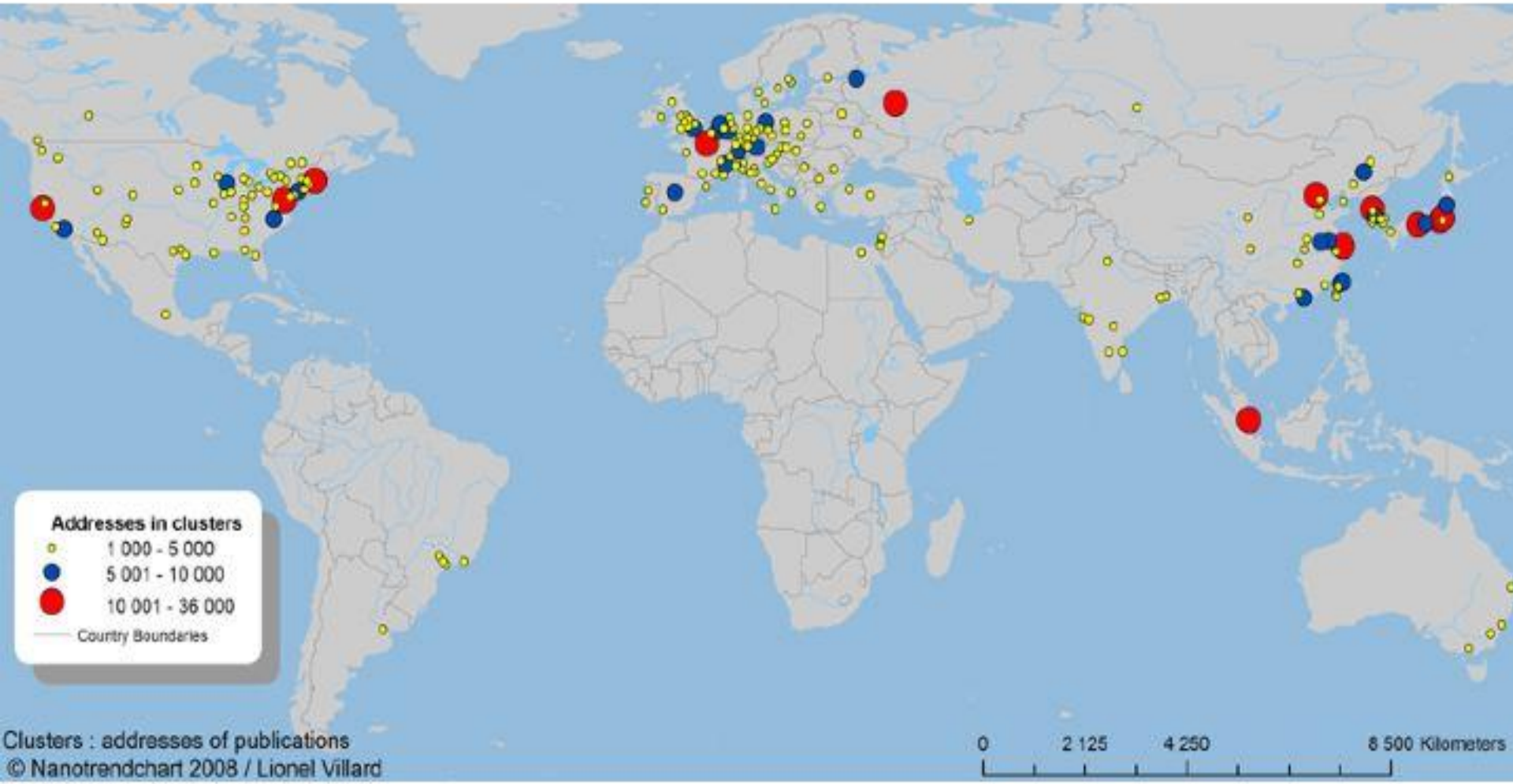
- We chose to build clusters on nano publications (1998-2006)
  → 203 clusters identified gathering 77% of addresses
- We applied the clusters to nano patents (1998-2006)*
  → the 203 clusters cover 75% of inventor addresses
- Some results:
  - A very high concentration (not considering singleton applis)
  - 10 clusters represent 53% of addresses (1998-2006) while the last 2/3rds represent only 8%
  - of the 52 top clusters (more than 1000 applis), 48% are from the 25 US clusters (San Francisco representing nearly 1/3rd alone), 40% from the 12 Asian clusters (Tokyo representing 40%) and 11% from the 14 EU clusters (8 being German)
- On-going developments: look at complementary clusters (patent specific, or by adding publications & patents)

* Not taking into account the standard name extension

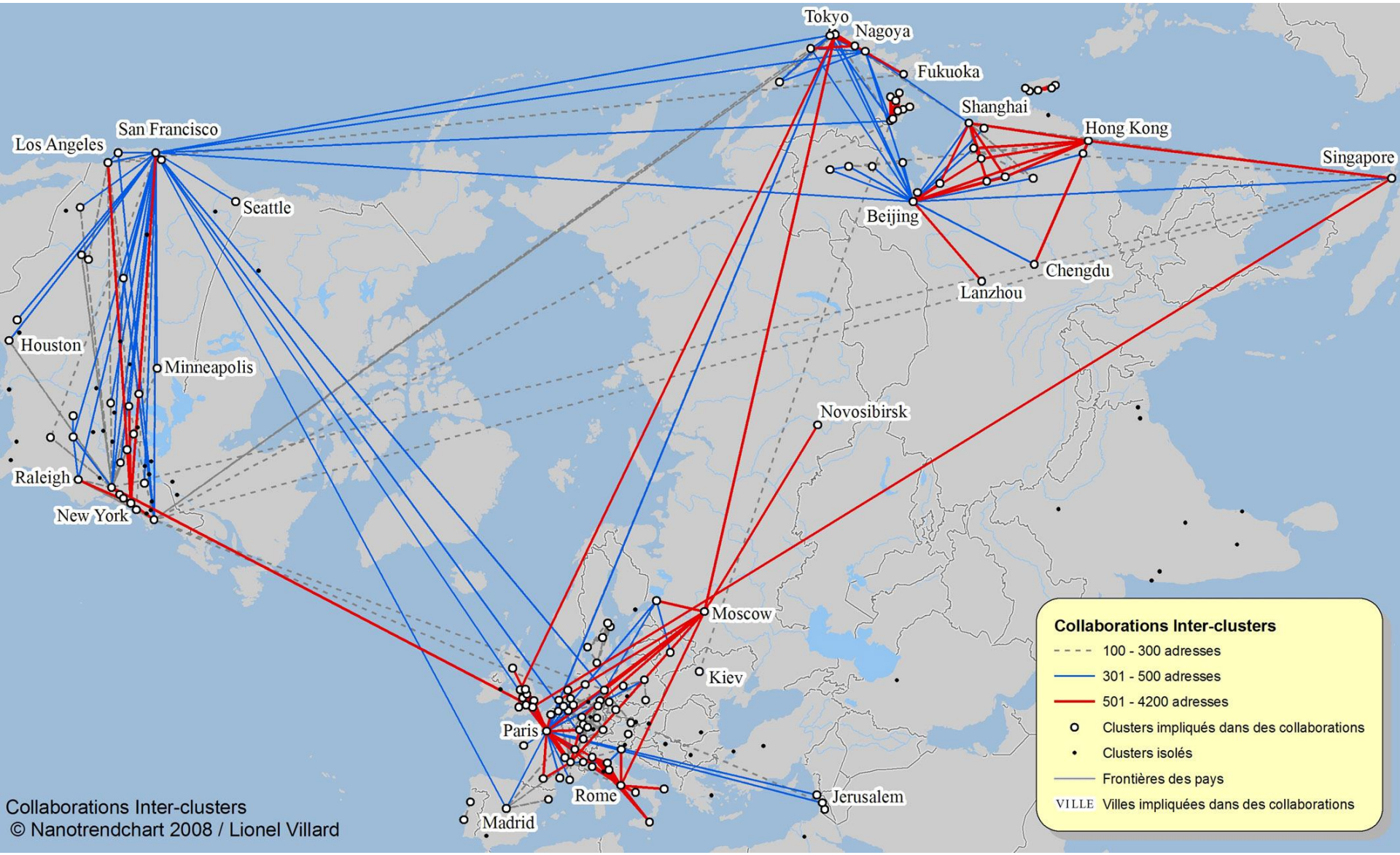
The construction of clusters: Europe as an example



The 203 clusters identified based on 1998-2006 outputs



Intercontinental collaborations concentrated in a few clusters



Characterising institutions- 1

- Based on Patstat standard names
- A lengthy manual process after using classical automated tools/
  - initial markers enabling to tag type (firm, univ, govt including PRO, other incl. hospitals)
  - addressing name similarities (manual check of proposals by automated treatment)
- An important support over time: the reference databases built in house
  - in particular CIB for large firms & their subsidiaries,
  - using other projects (on specific countries or sectors) in particular for PRO (e.g. handling Helmholtz institutes)
  - remaining problems: handling university hospitals, level of agglomeration (e.g. US DOE or individual labs from DOE)

Characterising institutions- 2

- Time consuming approach to both identifying organisations & finalizing tagging:
  (i) work within clusters, and for organisations not in clusters, work at country level.
  (ii) work on aggregated data per organisation for checking (insures robustness of large actors)
- Some results
  - 13300 different organisations, 61 firms with 1000+applis (37% total)
  - Patents with more than 1 assignee represent around 10% of total patents, 60% are other firms, 20% public assignees and 20% individuals
  - 87% of patents filed are by firms 7% by universities, 5% by PRO (and governments), 1% by others (including hospitals)

Top applicants (1998-2006)*

Samsung	13021	CEA	1636	Univ california	2243
Mitsubishi	7590	JST (Japan)	1571	Univ Tsinghua	825
LG	7565	KIST (Korea)	1391	Univ Texas	695
Sumitomo	5923	ITRI (Taiwan)	1341	Univ shanghai jiaotong	547
IBM	4884	Fraunhofer	1209	Caltech	512
Seiko Epson	4828	AIST (Japan)	1095		
Sony	4531	CNRS	943		
		CSIRO (Australia)	636	Harvard	332
BASF	3927			Univ oxford	131
3M	3525			Univ Cambridge	169
Gen Electric	3375				

* 2010 treatment

Some EU examples of Cluster composition*

	global	Paris	Zurich/Basle**	Leuven	Berlin	Atlanta	
Firm	87%	74%	67%	75%	63%	68%	
Univ	7%	3%	23%	24%	7%	31%	
PRO/gov	5%	22%	9%		29%	1%	
other	1%						
		77 firms			80 firms	110 firms	

* Only taking clusters around 1000 priority patent applications (1998-2006)
* Transborder cluster (PRO= Fraunhofer!)

Is nano linked with specialised industries?

firmes du DTI scoreboard	total	nano	%
Electronic & electrical equipment	103	70	68%
Technology hardware & equipment	226	150	66%
Chemicals	96	84	88%
Pharmaceuticals & biotechnology	153	73	48%
Health care equipment & services	53	39	74%
Automobiles & transport	86	59	69%
Aerospace & defence	35	24	69%
Materials & construction	55	42	76%
Oil, Gas & Electricity	53	39	74%
Food producers inc. Beverages)	32	16	50%
General industrials	38	24	63%
Household & personal goods	40	21	53%
Industrial engineering	70	35	50%
Telecom & media	32	14	44%
Software & computer services	110	14	13%
banks, insurance, retail, leisure	49	6	12%
total	1231	710	58%

Thematic dynamics

- The dataset includes the Patstat IFRIS technical classification
- We have used simplified classifications to understand overall dynamics (see next slide, the case of publications)
- Next step: As shown by firms nano crosses categories, we expect existing categories to offer limited understanding of how technologies cluster, thus we intend to mobilise CorText for providing complementary thematic clustering (2016)

Publications: 3 main domains with a central role for Chemistry and materials

Biotechnologies
Life Science
12% total
+112% in 8 years

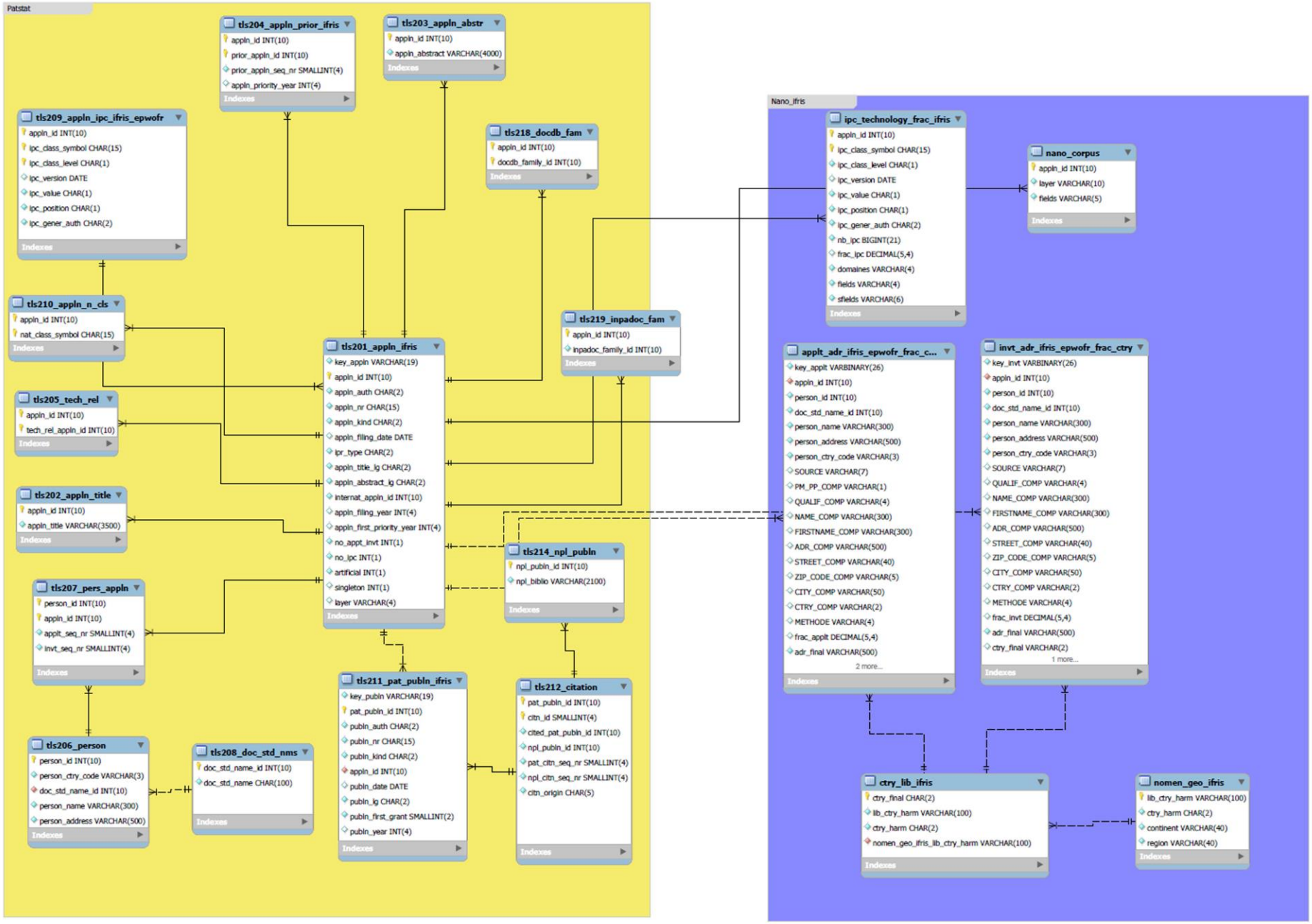
Electronics Physics
34% total
+104% in 8 years

Chemistry & Materials
52% total
+170% in 8 years

Database composition

- 11 main units of analysis
- The architecture is similar to Patstat IFRIS

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


Patstat
tls204_appln_prior_ifris
appln_id INT(10)
prior_appln_id INT(10)
prior_appln_seq_nr SMALLINT(4)
appln_priority_year INT(4)
Indexes
tls203_appln_abstr
appln_id INT(10)
appln_abstract VARCHAR(4000)
Indexes
tls209_appln_ipc_ifris_epwofr
appln_id INT(10)
ipc_class_symbol CHAR(15)
ipc_class_level CHAR(1)
ipc_version DATE
ipc_value CHAR(1)
ipc_position CHAR(1)
ipc_gener_auth CHAR(2)
Indexes
tls218_docdb_fam
appln_id INT(10)
docdb_family_id INT(10)
Indexes
tls210_appln_n_cls
appln_id INT(10)
nat_class_symbol CHAR(15)
Indexes
tls219_inpadoc_fam
appln_id INT(10)
inpadoc_family_id INT(10)
Indexes
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
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)
Indexes
tls205_tech_rel
appln_id INT(10)
tech_rel_appln_id INT(10)
Indexes
tls202_appln_title
appln_id INT(10)
appln_title VARCHAR(3500)
Indexes
tls214_npl_publn
npl_publn_id INT(10)
npl_biblio VARCHAR(2100)
Indexes
tls207_pers_appln
person_id INT(10)
appln_id INT(10)
applt_seq_nr SMALLINT(4)
invt_seq_nr SMALLINT(4)
Indexes
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
publn_lg CHAR(2)
publn_first_grant SMALLINT(2)
publn_year INT(4)
Indexes
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)
Indexes
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)
Indexes
tls208_doc_std_nms
doc_std_name_id INT(10)
doc_std_name CHAR(100)
Indexes
Nano_ifris
ipc_technology_frac_ifris
appln_id INT(10)
ipc_class_symbol CHAR(15)
ipc_class_level CHAR(1)
ipc_version DATE
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)
Indexes
nano_corpus
appln_id INT(10)
layer VARCHAR(10)
fields VARCHAR(5)
Indexes
applt_adr_ifris_epwofr_frac_c...
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)
2 more...
Indexes
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)
1 more...
Indexes
ctry_lib_ifris
ctry_final CHAR(2)
lib_ctry_harm VARCHAR(100)
ctry_harm CHAR(2)
nomen_geo_ifris_lib_ctry_harm VARCHAR(100)
Indexes
nomen_geo_ifris
lib_ctry_harm VARCHAR(100)
ctry_harm CHAR(2)
continent VARCHAR(40)
region VARCHAR(40)
Indexes

Next steps

- Nano Patstat V1 is available with all its improved characteristics (covering until 2006)
- In 6 months it will be complemented by Nano Patstat V2
  - based on Patstat IFRIS 2014 (until 2012)
  - with full geolocalisation process
  - completely harmonised with CIB for large firms
  - with new approach for clustering (a standard choice, but can be refined for ad-hoc questions)
- We shall progressively integrate RISIS features: identifying midsize and venture capital supported start-up firms; providing OECD standardised metropolitan areas (FUA)

Some possibilities

- One critical aspect is that we follow publications and patents, insuring harmonisation of organisations and geographical dimensions
- We have only worked on global dynamics, little has been done on specific technologies or industries (critical if it is a general purpose technology)
- Very little has been done on collaborative patterns, at actor, geographical and even more at cognitive level

Some references

- Larédo P., Delemarle A., Kahane B., 2010, Dynamics of nanoS&T: policy implications, STI policy review, 1, 43-62
- Larédo P. & Villard L., 2013/2015, La construction d'une Europe de la recherche et ses effets sur les relations transatlantiques, cahiers rouges 215, fondation Jean Monnet
- Villard L., Revollo M., 2015, Geographical concentration of S&T activities, RISIS website
- Kahane B, Mogoutov A., Cointet J.P., Villard L., Larédo P., 2014, A dynamic query to delineate emergent science and technology: the case of nano science and technology, RISIS website
- IFRIS team, 2014 (November) Report on the content and technical structure of the nano S&T dynamics database, RISIS website