Documentation of RISIS datasets: RISIS CIB2 DATABASE

Patricia Laurens

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

DOCUMENTATION OF RISIS DATASETS

RISIS CIB2 DATABASE

Written by P. Laurens (UGE) 03/2020

Risis Patent was constructed by J.P. Ospina Delgado, P. Laurens, A. Schoen (UGE).
P. Laredo, and L. Villard (UGE) have also contributed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Outline

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

1 Basic Characteristics

The Corporate Invention Board (CIB2) database is designed for the analysis of technological knowledge creation of the worldwide top corporate R&D performers, using patents as a proxy. It includes 3992 companies. It focuses on the priority patents[1] applied by the parent companies and the subsidiaries they control. For patents, it relies on the patent data included in the RISIS Patent Database (RPD)[2]. The list of companies was elaborated using several editions of the EU Industrial R&D Investment Scoreboard and the lists of WIPO top applicants. CIB2 also gives basic data on the companies (locations, sectors, size, financial data).

Overall 6,158,064 priority patents of inventions were applied for from 2000 to 2015[3]. They include 15,712,409 inventors[4].

Besides the date of the filing and the office where the patent was first filed, we specifically consider in CIB2 5 core attributes of priority patents:

- **Their content, using textual sources of information**: the patent titles, the patent abstracts and a text aggregating the definition of the patent IPC (International Patent Classification) classes.

- **Their technological content using the standard technology classification:** (IPC subclasses, aggregation of IPC codes) by technological domains, fields or subfields.

- **The geographical location of inventive activities**. As we are interested in the geography of knowledge creation, we focus on inventor addresses. These addresses are geocoded and associated to functional areas (urban and rural) worldwide.

- **The legal organisations that apply for patents (the applicants)**. We use available information proposed by PATSTAT for the harmonisation of applicant names and allocation of assignees. Among all the legal organisations, we identify worldwide large companies (CIB2 companies).

- **Characteristics linked to the value of the patent**: In this context, the database provides information on whether the patent was granted or not, as well as a set of variables describing the patent families.

The coverage of the database is 2000-2015 (using the PATSTAT 2017 April release as a source). The first version of CIB2 has been released in February 2020. A second one will be made available in spring 2022. The database is developed and maintained by UGE (ex UPEM). The database is located on the servers of UGE at Marne La Vallee, France.

2 Database content

2.1 Definition and description of observations

As a proxy for knowledge creation, the unit of observation is a priority patent application, i.e. the very first patent application, anywhere in the world to protect an invention. The priority

[1] that represent the creation of new knowledge

[2] For an extensive description of the RPD, see the Risis Patent Database documentation https://zenodo.org/record/3342454/files/Documentation_RISIS%20Patstat_Final.pdf?download=1

[3] i.e. ipr_type = PI in PATSTAT

[4] Corresponding to 3,632,109 distinct person_id

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

date is used to determine the novelty of the invention, which implies that it is an important concept in patent procedures. The priority date is the closest date to the date of invention.

We consider only applications of priority patents of invention (i.e. ipr_type = PI in PATSTAT)[5]. Accordingly, the database covers priority patents applied since 2000, whether or not they turn later into granted patents.

When a co-application of patent includes both legal and natural persons, the patent is maintained in the database with its legal applicant only.

A fractional counting was employed to calculate the number of legal applicants in patent applications: all applicants with a fract_applt below 1 have legal co-applicants. There are 494,411 co-applied priority patents.

2.2 Data acquisition and processing (including data cleaning)

2.2.1 Data sources

The CIB2 database uses several sources of information to identify the worldwide top corporate R&D performers and define their consolidated perimeters (3992 parent companies and their subsidiaries owned by the parent company with shares higher than 50.01%). The sources of company names are: the European R&D Industrial Scoreboard (editions from 2008 to 2014), the lists of the PCT top applicants published by the World Intellectual Property Organisation[6] (editions from 2008 to 2014) and the ORBIS database (Bureau van Dijk). The CIB2 patent source is the RISIS Patent Database (RPD), an enriched and simplified Patstat DB (Patstat April 2017). CIB2 benefits from all the enrichments and adjustments made in RPD (see the RPD documentation for details).

2.2.2 Data processing

The main data processing includes:

1 - **Defining the list** of the worldwide top corporate R&D performers

We have first collected:

- The lists of the companies of the European Industrial R&D Investment Scoreboard for seven editions from 2008 to 2014. Each Scoreboard edition ranks a number of companies (European firms and non-European firms) which have invested the largest sums in R&D during the previous year[7]. Using the World and European lists, we have ended with approximatively 17,000 company names with redundancy over the editions.
- The lists of the top PCT patent applicants provided yearly by WIPO (i.e. applicants applying for a minimum of 10 PCT patents in a given year). About 13,500 company names were retrieved when adding names given from 2008 to 2014.

[5] Other types of patent are: Utility Models (ipr_type = UM) and Design patents (ipr_type = 'DP').

[6] https://wipo.int

[7] The number of companies in a given year varies from 2000 to 3500 depending on the Scoreboard edition.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

We ended with approximatively 30,000 company names with a lot of duplicates (several companies were present in both sources for several years with the same name or different names). We then cleaned and harmonised the company names. Using the Orbis database we have identified for all the company names, the parent company. In most cases we have selected the Global Ultimate Owner (GUO) when the GUO was an industrial company or the higher intermediate industrial company owned by the GUO when the GUO was not an industrial company but a country, a family or a holding company. In that latter case, the parent company was tagged as an IGUO (industrial GUO). The final list contains 3992 parent companies.

2- Defining the consolidated perimeter of the worldwide top corporate R&D performers

The consolidated perimeters of the 3992 parent companies (GUO or IGUO) were obtained using the Orbis database[8]. It includes all subsidiaries with a majority share capital owned by the parent companies (share >50.01%). The consolidated perimeter of the parent companies was determined based on information available in Orbis in autumn 2017 (the perimeter for few firms were built in a later phase). The names of the parent companies as well as their subsidiaries were downloaded from Orbis. Moreover, a set of available variables was also retrieved (company name, location, sectors, employees, financial data). We have ended with approximately 320,000 different company names (GUOs, IGUOs and subsidiaries).

3- Matching company names and patent applicants

In order to retrieve from the RISIS Patent Database, the priority patents applied for by the 3992 parent companies, the list of the 320,000 Orbis company names had to be matched with the names of the legal applicants (doc_std_name) in more than 13 millions of priority patents from the RISIS Patent Database.

We used the PAM system (Patent Approximation Matches system) that is a textual analysis tool that has been designed at LISIS (UGE) for matching legal type entities with patent applications, for pairing entities (Figure 1). The PAM system relies on the company name and combines full text search techniques using Elasticsearch with some of the most famous approximate string-matching algorithms such as Jaro-Winkler, Levensthein and Ratcliff (Box 1). Moreover, it uses each of these scores for calculating its own PAM score in order to select the best candidates and dismiss the wrong ones

Box 1 String-Matching Algorithms used

Levenshtein distance: This algorithm is based on the minimum number of single-character changes required in order to have one string converted into the other one that we are comparing.
Jaro-Winkler distance: its similarity score gives more relevance to the fact that two strings are closer when they have in common a longer set of symbols at their beginning than those which contain a mistake in the first few symbols.
The Ratcliff/Obershelp: Its use as the main characteristic the longest substring that S1 and S2 have in common. Then it analyzes the remaining part of the string as if they were new strings.

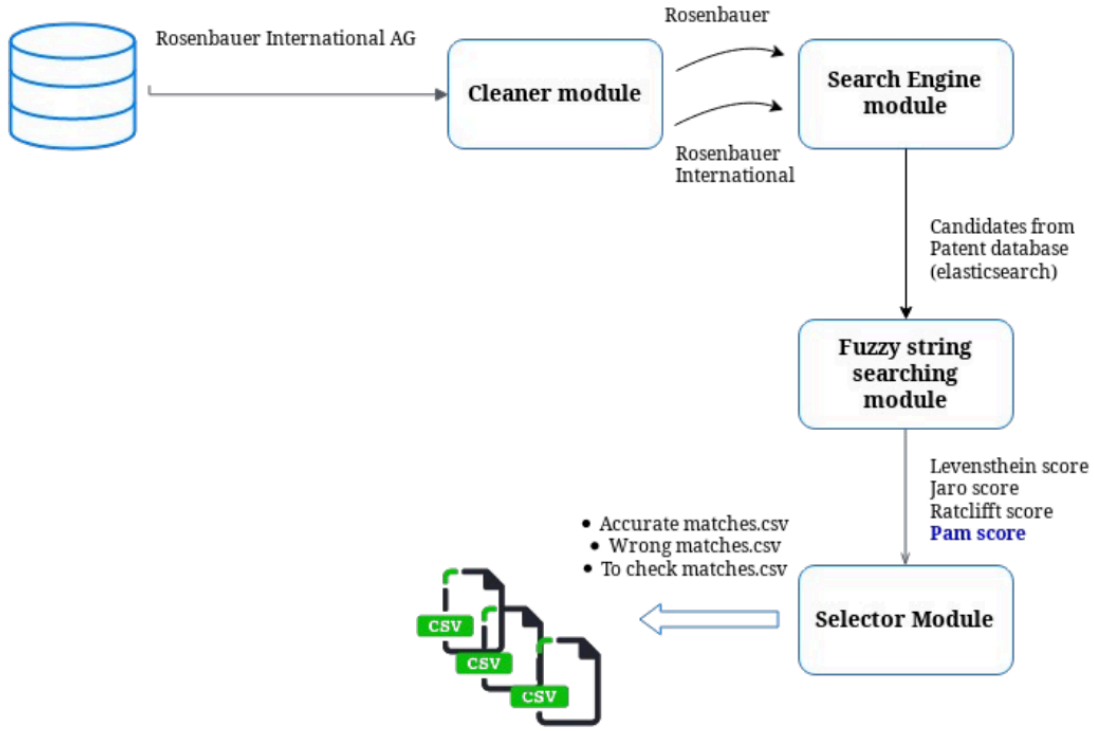
Figure 1: Schematic view of the PAM system

8 A special treatment was applied to identify the subsidiaries to include in the consolidated perimeter of the iGUOs. Their consolidated subsidiaries cannot be identified directly in Orbis. Thereforethe selection has been operated in three steps:
- For a given iGUO, all subsidiaries (up to level 10) have been identified.
- For all these subsidiaries, information has been retrieved for identifying the (ascending) chain of controlling shareholders.
- The consolidated subsidiaries of a company have been then identified and selected as those where this company appears in the chain of controlling shareholders.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Pam System

Patstat Approximate Matches System



In a nutshell, the PAM system is composed of four different modules (Figure 2):

Module 1: A Cleaner module cleans and harmonizes the company name using Magermans methodology and PATSTAT standardization.

- It also enriches the company name with alternative names, and removes stop words like the, on, of, etc.
- It drops every remaining organizational name.
- At last, the module removes names that contain only the name of a country or a city.

Module 2: A Search Engine module that uses Elasticsearch[9]. Elasticsearch splits the names into tokens (words) then it indexes these words in order to create a "reverse index" based on the number of occurrences of the words in the database; the higher is the occurrence of a word, the lower is the score when matching a name with it. For instance, Corporation or University are very frequent words in PATSTAT. Matching a name with one of these words is not very specific and the matching score will not be very high. Conversely, the matching score with Google will be much higher because Google is not very frequent in the patent applicant names. Let us consider a hypothetical case where "Google university corporation" is the name of a patent applicant. Trying to match this name with "University of Gantz", will give a low matching score because it only matches with University; trying to match it with "Google" will give a much higher score because of the uniqueness of the token (word) Google. The Search Engine module uses two types of queries, one using jurisdictions filter (i.e. country) and the other one without this jurisdiction filter.

[9] https://en.wikipedia.org/wiki/Elasticsearch

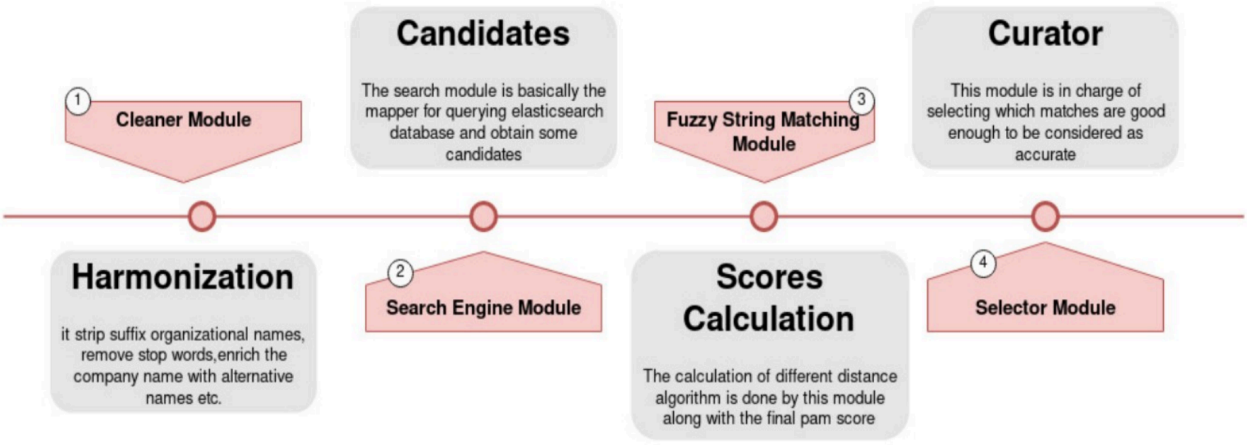
This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

The module is parametrized to retrieve only results where the elastic search score is higher than a defined threshold. It stores the elasticsearch score and all the matches found in the RISIS Patent Database.

Module 3: A Fuzzy string searching module. For every matching obtained by the previous steps the system runs three different algorithms based on edit distance and sequences. Those algorithms are Levensthein distance, Jaro-Winkler, RatcliffObershelp (see box 1). From the resulting scores it calculates a new score called the PAM score, which combines the Elasticsearch score, the algorithm score and the query used as parameters.

Lodule 4: A Selector module. After running tests on relevant samples of the company names, several metrics were elaborated in order to select accurate matches, matches to be checked and wrong matches.

Figure 2: Tasks of the different PAM modules



4- Checking of the matching

Both *accurate* and *to check* matchings were further manually examined when the number of patents was above 30 (about 10,000 manual checks out of 50,000 pairs). For lower number of patents, rules were defined and automatically adopted.

5- Extraction of CIB2 from RISIS Patent Database

The CIB2 database is a subset of the RISIS Patent Database. It thus includes all the tables and variables present in RPD (see the lists and descriptions in the documentation of the RISIS Patent Database)

6- Addition of the tables and variables related to the CIB2

Variables related to the companies are added in five additional different tables (see below for further details of the variables). They include data on the geography, industrial sector and a set of financial information.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.3 Information on all variables/indicators

For each priority patent application, the database gives:

- patent ID number
- date of first filing
- country of first filing
- date of first publication
- date of first granting
- title
- abstract
- IPC categories mentioned: their number and their language description: as many variables as IPC categories, only one linguistic description
- whether the patent is a singleton (only one single application) or not
- whether the patent is a transnational patent, i.e. was the priority application extended in at least one foreign country) or not
- size of the DOCDB family
- size of the INPADOC family
- presence (directly or through extension) in the 5 core patent offices (so called IP 5 families) (5 variables Y/N per office)
- For each applicant: the applicant natural name and its ID; the applicant standardised name and its ID; the applicant standardised name by Leuven University and its ID; the applicant RISIS standardised name and its ID when available
- For each inventor: same variables as for applicants
- For each applicant: presence or not in the CIB database, name and ID of the parent company (GUO or IGUO), presence in FIRMREG database and FIRMREG ID
- For each applicant: The applicant address, the applicant geo coordinates, the applicant urban or rural area it belongs to
- For each inventor: the inventor address, the inventor geo coordinates, the inventor urban or rural area he/she belongs to.

For each company, the CIB2 database gives:

- The company names (GUO or iGUO) including a current name but also previous and alternative company names
- A set of legal information of the company like its date of incorporation; Is it an active firm? Is it a quoted firm? Is it a GUO or an iGUO?
- The geographical location of the headquarter of the company including the address, city, country, latitude, longitude, NUTS3 regions and the functional area name (also named URA for Urban or Rural Area) and ID
- The industrial sectors of the company using NACE2 codes and labels (primary and secondary NACE codes)
- Some economic and financial data of the company including number of employees, turnover, assets, R&D investments, Profit/loss, ROE, ROA. Financial data are given for the most recent available years (up to ten years depending on the data availability.

2.4 Sectorial, temporal and geographical coverage

Sectorial coverage of CIB2 companies

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 1: Distribution of the CIB2 companies by industrial NACE2 sectors (ordered by number of companies per sector)

Sector of the parent companies NACE 2 (4 digits)	Number of parent companies	Share of parent companies	Cumulated share of parent companies
Manufacture of pharmaceutical preparations	309	7,7%	7,7%
Manufacture of electronic components	284	7,1%	14,9%
Other software publishing	136	3,4%	18,3%
Activities of head offices	111	2,8%	21,0%
Computer programming activities	111	2,8%	23,8%
Manufacture of communication equipment	98	2,5%	26,3%
Activities of holding companies	96	2,4%	28,7%
Manufacture of other special-purpose machinery nec	93	2,3%	31,0%
Manufacture of instruments and appliances for measuring, testing and navigation	89	2,2%	33,2%
Manufacture of medical and dental instruments and supplies	86	2,2%	35,4%
Manufacture of other parts and accessories for motor vehicles	84	2,1%	37,5%
Manufacture of computers and peripheral equipment	81	2,0%	39,5%
Manufacture of other chemical products nec	77	1,9%	41,5%
Research and experimental development on biotechnology	68	1,7%	43,2%
Other information technology and computer service activities	67	1,7%	44,8%
Other telecommunications activities	66	1,7%	46,5%
Manufacture of motor vehicles	51	1,3%	47,8%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Unknown	48	1,2%	49,0%
Production of electricity	40	1,0%	50,0%
Total	3992		100,0%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Geographical coverage of CIB2 companies

Table 2: Distribution of the CIB2 companies by country of the company headquarter (ordered by number of companies per country)

Country of headquarter	Number of parent companies	Share of parent companies	Cumulated share of parent companies
US	1067	26,8%	26,8%
JP	558	14,0%	40,7%
DE	330	8,2%	49,0%
GB	316	7,9%	56,9%
CN	268	6,7%	63,6%
FR	156	3,9%	67,5%
KR	143	3,6%	71,1%
TW	106	2,7%	73,7%
SE	107	2,7%	76,4%
IT	83	2,1%	78,5%
NL	76	1,9%	80,4%
CH	71	1,8%	82,2%
KY	69	1,8%	83,9%
FI	62	1,6%	85,5%
IN	59	1,5%	87,0%
AT	54	1,4%	88,3%
DK	53	1,3%	89,7%
CA	56	1,4%	91,1%
BE	47	1,2%	92,2%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

ES	31	0,8%	93,0%
IL	31	0,8%	93,8%
IE	32	0,8%	94,6%
AU	29	0,7%	95,3%
BM	17	0,4%	95,7%
LU	19	0,5%	96,2%
Total	**3992**		**100%**

Sectors of the legal applicants

In the current CIB2 database, we rely on the sectorial information (legal status and/or type of institutions) provided in the raw PATSTAT database[10]. Data are shown in the table below.

Table 3: Legal status of priority patent applicants

Sector of legal applicants	Number of applicants	Share of applicants
COMPANY	121,482	92,7%
COMPANY GOV NON-PROFIT	171	0,13%
COMPANY GOV NON-PROFIT UNIVERSITY	2	0,001,53%
COMPANY HOSPITAL	0	0,0%
COMPANY UNIVERSITY	3	0,0023%
GOV NON-PROFIT	401	0,31%
GOV NON-PROFIT HOSPITAL	0	0,0%
GOV NON-PROFIT UNIVERSITY	2	0,00153%
HOSPITAL	54	0,04%

[10] Harmonizing names and allocation of assignee sectors in Patstat raw data was done by ECOOM (K.U. LEUVEN; http://www.ecoom.be/en/EEE-PPAT)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

INDIVIDUAL	0	0,0%
UNIVERSITY	256	0,20%
UNIVERSITY HOSPITAL	2	0,0%
UNKNOWN	8,658	0,07%
TOTAL	**131,027**	**100,0%**

Technological coverage of patents

There is a thematic coverage of the technology that the patents protect linked to the IPC classification of patents (one patent can belong to multiple classes). Patents are allocated to fields of technology on a fractional count basis according to their IPC. Building on the correspondence table developed by ISI FhG for WIPO - which defines domains (5) and fields (35) of technology, an additional level of sub fields (401) of technology was developed (See Appendix 3 on technological fields in the RISIS Patent Database documentation).

The distributions of patents according to domains and technological fields are shown below.

Table 4: Domains of technology of priority patent applications

Domain code	Domain name	Number of patents (fractional counting)	Share of patents
TD01	Electrical engineering	2,624,101	43,1%
TD02	Instruments	919,172	15,1%
TD03	Chemistry	854,292	14,0%
TD04	Mechanical engineering	1,389,346	22,8%
TD05	Other fields	302,866	5,0%
Total		**6,089,777**	**100,0%**

Table 5: Technology fields of priority patent applications

Field code	Field name	Number of patents (fractional counting)	Share of patents

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

TF01	Electrical machinery, apparatus, energy	511,728	8,4%
TF02	Audio-visual technology	378,506	6,2%
TF03	Telecommunications	265,917	4,4%
TF04	Digital communication	333,704	5,5%
TF05	Basic communication processes	72,000	1,2%
TF06	Computer technology	595,803	9,8%
TF07	IT methods for management	92,151	1,5%
TF08	Semiconductors	374,293	6,1%
TF09	Optics	386,684	6,3%
TF10	Measurement	269,839	4,4%
TF11	Analysis of biological materials	16,549	0,3%
TF12	Control	104,467	1,7%
TF13	Medical technology	141,633	2,3%
TF14	Organic fine chemistry	94,633	1,6%
TF15	Biotechnology	42,930	0,7%
TF16	Pharmaceuticals	100,019	1,6%
TF17	Macromolecular chemistry, polymers	109,810	1,8%
TF18	Food chemistry	29,746	0,5%
TF19	Basic materials chemistry	113,818	1,9%
TF20	Materials, metallurgy	105,292	1,7%
TF21	Surface technology, coating	100,303	1,6%
TF22	Micro-structural and nanotechnology	7,044	0,1%
TF23	Chemical engineering	81,742	1,3%
TF24	Environmental technology	68,956	1,1%
TF25	Handling	144,720	2,4%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

TF26	Machine tools	120,399	2,0%
TF27	Engines, pumps, turbines	208,734	3,4%
TF28	Textile and paper machines	154,285	2,5%
TF29	Other special machines	111,465	1,8%
TF30	Thermal processes and apparatus	110,129	1,8%
TF31	Mechanical elements	188,861	3,1%
TF32	Transport	350,751	5,8%
TF33	Furniture, games	106,771	1,8%
TF34	Other consumer goods	80,561	1,3%
TF35	Civil engineering	115,535	1,9%
Total		**6,089,777**	**100,0%**

Temporal coverage

The database of CIB2 covers patent applications from 2000 to 2015 (based upon PATSTAT 2017 Version April). It is most likely that years 2016 and 2017 are still only partially filled. It includes 6,158,064 applications of priority patents

The number of PI priority patents applied every year is shown below. It has increased by 25% during these 15 years. Year 2015 is probably not fully completed

Table 6: Number of priority patent applications over time

Filing year	Number of PI priority patent applications
2000	339,013
2001	348,252
2002	352,655
2003	365,323
2004	392,924
2005	407,196

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2006	407,316
2007	411,924
2008	419,169
2009	381,487
2010	392,552
2011	409,043
2012	430,997
2013	435,554
2014	424,459
2015	(240,200)[11]
Total	**6,158,064**

Geographical coverage - geography of the PI protection

The data cover all priority patent applications worldwide, i.e. at all regional and national offices in the world (see table below). Approximately 50% of the priority applications are applied for at the Japanese and Chinese offices. The IP5 patent offices (US, EP, JP, CN, KR) cumulate together 87,6% of the priority applications.

Table 7: Number of priority patent according to the patent offices

Patent office	Number of PI patent applications	Distribution of PI patent applications
Total	**6,158,064**	**100,0%**
JP	3,034,010	49,3%
US	1,112,974	18,1%
KR	609,217	9,9%
CN	465,925	7,6%

[11] This number of patents in 2015 should be considered with caution. It should not be interpreted as a decreasing trend of patent application but rather as a lack of completedness of the database for this very year

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

DE	376,079	6,1%
EP	174,887	2,8%
TW	98,358	1,6%
FR	90,078	1,5%
GB	70,644	1,1%
SE	19,042	0,3%
IN	14,089	0,2%
IT	12,184	0,2%
FI	10,477	0,2%
AU	7,940	0,1%
DK	6,960	0,1%
IB	5,557	0,1%
CH	4,448	0,1%
CA	4,351	0,1%
BR	4,344	0,1%
AT	3,998	0,1%
Other	32,502	0,5%

Following a first priority patent application for a new invention, the IP protection can be further extended in several geographical countries considered as potential future markets. The following tables give information on the geography of the patent protection of a given priority patent using data available in its INPADOC family. Most of the patents, for a given invention are applied for in a single patent office and only 35,3% of the priority patents are transnational, i.e. further extended in another (foreign) patent office (28,5% from 2 to 5 patent offices)

Table 8: Number of patent offices where patents are applied for a given invention

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Number of distinct patent offices in the INPADOC family	Number of priority patents	Share of priority patents
1	3,984,687	64,7%
2	708,311	11,5%
3	441,773	7,2%
4	357,745	5,8%
5	244,346	4,0%
More than 5	421,202	6,8%
Total	**6,158,064**	**100,0%**

57% of the inventions are protected in Japan (either in the first or subsequent filings), 24% in China, 39% in US, 20% at EPO, 17% in Korea. 22% of the families include a PCT patent[12]

Table 9: Number and share of patents (priority or secondary applications) applied in the five largest patent offices (IP5 patent offices)

Content of INPADOC family[13]	Number of priority patents	Share of priority patents
Total	6,158,064	100%
US application	2,400,325	39%
EP application	1,234,123	20%
JP application	3,543,388	57,5%
KR application	1,061,296	17,2%
CN application	1,497,215	24,3%
W application	1,361,758	22,1%
Transnational patent	2,173,377	35,3%

[12] ThePatent Cooperation Treaty(PCT) provides a unified procedure for filingpatent applicationsto protectinventionsin each of its contracting states. A patent application filed under the PCT is called aninternational application, orPCT application.
[13] Only ipr_type: PI is considered in INPADOC families.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Geographical coverage - Geography of the inventions

In 93% of the patents, the location of the inventions (inventors geographical location) is known and in 97.7% of them, the location of the applicant is indicated. The shares of patents according to the inventors and applicants addresses are shown in the tables below.

Table 10: Number of priority patent applications according to the country of inventors

Country of inventors	Number of patent applications with an inventor from the country	Share of patent applications with an inventor from the country
Total number of priority patents	6,158,064	
JP	3,050,764	49,5%
US	932,996	15,2%
KR	607,495	9,9%
DE	464,670	7,5%
Unknown	430,703	7,0%
CN	421,493	6,8%
TW	139,289	2,3%
FR	133,642	2,2%
GB	66,770	1,1%
CA	45,642	0,7%
NL	41,424	0,7%
SE	40,270	0,7%
IN	38,507	0,6%
CH	37,882	0,6%
IT	29,310	0,5%
FI	24,527	0,4%
AT	19,826	0,3%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

BE	18,680	0,3%
DK	18,015	0,3%
IL	17,582	0,3%
ES	10,690	0,2%

Table 11: Number of priority patent applications according to the country of applicants

Country of applicant	Number of patents with applicant from the country	Share of patents with applicant from the country
Total number of priority patents	6,158,064	
JP	3,064,415	49,8%
US	946,490	15,4%
KR	622,967	10,1%
DE	419,248	6,8%
CN	333,244	5,4%
TW	144,741	2,4%
Unknown	138,904	2,3%
FR	122,298	2,0%
GB	60,423	1,0%
NL	51,827	0,8%
SE	48,719	0,8%
CH	48,637	0,8%
FI	28,879	0,5%
CA	21,869	0,4%
IL	21,612	0,4%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

DK	15,891	0,3%
IT	13,506	0,2%
BE	12,181	0,2%
AT	10,356	0,2%
IN	9,224	0,1%
SG	4,481	0,1%

2.5 Basic statistics on the patenting activities of CIB2 companies

Basic statistics describing the patenting activities of CIB2 companies are shown in Table 10. It shows a huge heterogeneity of the patenting activities of CIB2 companies.

Table 12: Basic statistics on the number of priority patent applications among CIB2 companies

Statistics (3992 companies)	**Number of priority patents**
Minimum	0
Maximum	154,976
1st quartile	30
2nd quartile	125
3rd quartile	524
Mean	1548

The distribution of the geography of **companies** (**company** HQ) in the four quartiles are shown in the tables below. The shares of the US and European **companies** dominate in the lower quartiles. In the 4th quartile, JP **companies** dominate.

Table 13: Distribution of the company country in the four quartiles depending on their number of priority patents

Table 13-a: companies with 0 to 30 priority patents (1010 companies)

Company HQ Country	**Share of companies**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

US	23,4%
GB	15,9%
DE	8,2%
FR	5,2%
SE	4,9%
CN	4,6%
JP	4,2%
IT	2,9%
NL	2,8%
FI	2,4%
Others	25,8%
Total	**100,0%**

Table 13-b: companies with 31 to 124 priority patents (988 companies)

Company HQ Country	Share of companies
US	32,1%
DE	8,9%
GB	8,5%
CN	6,6%
JP	4,7%
FR	4,0%
SE	3,4%
KR	3,3%
IT	3,2%
CH	2,4%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Others	22,8%
Total	**100,0%**

Table 13-c: companies with 125 to 524 priority patents (995 companies)

Company HQ Country	Share of companies
US	30,3%
JP	11,0%
CN	9,2%
DE	8,3%
GB	4,9%
KR	4,6%
FR	3,5%
TW	2,9%
CH	2,4%
IN	2,2%
Others	20,6%
Total	**100,0%**

Table 13-d: companies with more than 525 priority patents (998 companies)

Company HQ Country	Share of companies
JP	21,3%
US	7,5%
DE	6,5%
CN	5,3%
TW	5,0%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

KR	2,9%
FR	2,3%
GB	1,8%
CH	1,4%
Others	45,9%
Total	**100,0%**

2.6 Quality and accuracy of data

CIB2 database is an enriched subset of the RISIS Patent Database (see https://rcf.risis2.eu/dataset/5/metadata) . Data quality tests already carried out on RISIS Patent Database (see Documentation for RISIS Patent Database: zenodo.org/record/3342454/files/Documentation_RISIS%20Patstat_Final.pdf?download=1) are not reproduced for CIB2. Building the CIB2 database requires matching the names of firms from the Orbis database with the names of applicants in RISIS Patent Database. This was carried out using the PAM algorithm (see section 2.2.3). The algorithm was designed to calculate a PAM score to assess the proximity of the names to match. Based on the PAM score, the quality of the matching is classified and labelled as: *accurate*, *to check*, and to *discard*. Both *accurate* and *to check* matchings were further manually examined when the number of patents was above 30 (about 10,000 manual checks out of 50,000 pairs). For lower number of patents, rules were defined and automatically adopted.

Moreover, we have defined some tests for the structure itself, in which we evaluate the concordance of different components of the database, such as indexes, columns, data types and table size. For the latter, we di not precisely measure the integrity of data, mainly, (due to variation of the company perimeters in CIB2 compare to CIB1), but just evaluate the data consistency, evaluating the size of the CIB2 tables compared to their original sources (the Risis Patent Database) and firm list (see https://github.com/cortext/cib-database/test).

3 Technical Specifications

3.1 Information on the data base system

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.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

For some tables we have also used the storage engine, InnoDB, principally to keep the referential integrity of the firm_reg_id within the firm tables. We did not change the storage engine for the tables related to patent variables (MyISAM) because of the size. As we are above one million of records, it is wiser to use MyISAM for query optimization and speed purposes.

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

Planned future technical changes concerning data base system

None

3.2 Technical variable definition

CIB2 is a database of patents applied by top R&D corporate performers. It is a subset of RISIS Patent Database (RPD) (that contains all priority patent applications). CIB2 displays the same variables as RPD. The name and definition of the variables are detailed in the documentation of RISIS Patent Database (see section 3.2).

CIB2 includes a list of variables related to the companies. They are described in Table 14.

Table 14: Name, definition and tables of the variables of the CIB2 firms

Variable Name	Variable Definition	Name of the Table where is the Variable
FIRMREG_ID	Firm unique identifier coming from RISIS FirmReg (see Section 3.4)	All Tables including firm data
ADDRESS	Addresses of the company HQs	cib_firm_address
LONGITUDE	Addresses of the company HQs	cib_firm_address
LATITUDE	Addresses of the company HQs	cib_firm_address
CITY	City of the company HQs	cib_firm_address
COUNTRY	Country of the company HQs	cib_firm_address
ISO3	ISO3 code of the country of the company HQs	cib_firm_address
NUTS_ID	NUTS3 code of the country	cib_firm_address

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

	of the company HQs	
RURBAN_AREA_ID	Code of the rurban area of the company HQs	cib_firm_address
RURBAN_AREA_NAME	Name of the rurban area of the company HQs	cib_firm_address
YEAR	Year of the financial data	cib_firm_financial_data
OPERATING_REVENUE	Company operating revenue (thousand $)	cib_firm_financial_data
TOTAL_ASSETS	Company total assets (th. $)	cib_firm_financial_data
NUMBER OF EMPLOYEES	Number of employees ithe company	cib_firm_financial_data
PL_BEFORE_TAXES	Profit or loss before taxes (th. $)	cib_firm_financial_data
ROE_USING_PL_BEFORE_TAXES	Return on equity before taxes (th. $)	cib_firm_financial_data
ROA_USING_PL_BEFORE_TAXES	Return on assets before taxes (th. $)	cib_firm_financial_data
R&D_INVEST	R&D investments of the company (m€)	cib_firm_financial_data
PREVIOUS_NAME	Previous names of the company	cib_firm_names
PREVIOUS_NAME_DATE	Date of the change of name	cib_firm_names
ORBIS_NAME	Name of the company given in Orbis database (02/2020)	cib_firm_names
FIRM_REG_NAME	Name of the company given in the FirmReg register database	cib_firm_names

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

AKA_NAME	Alternative name of the company	cib_firm_names
NAME	Current name of the company	cib_firms
ISO_CTRY	ISO2 country code of the HQs	cib_firms
NACE2_MAIN_SECTION	NACE2 main section of the company	cib_firms
CATEGORY	Size category of the company	cib_firms
STATUS	Status of the company activity (Active/Inactive)	cib_firms
QUOTED	Status of the company in terms of quotation (Y/N)	cib_firms
I_GUO	Industrial Global Ultimate Owner (if 1) (when the GUO is not a industrial firm. Ex GUO is a state, a family, an individual)	cib_firms
INCORPORATION_DATE	Date of company incorporation	cib_firms
LAST_YEAR_ACCOUNT_AVAILABLE	Most recent year for company account data	cib_firms
NACE2_PRIMARY_CODE	Primary NACE2 code (4 digits)	cib_firm_sector
NACE2_PRIMARY_LABEL	Label of the primary NACE2 code	cib_firm_sector
NACE2_SECONDARY_CODE	Secondary NACE2 code (4 digits)	cib_firm_sector

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

NACE2_SECONDARY_LABEL	Label of the secondary NACE2 code	cib_firm_sector

3.3 Description of the Entity Relationship Model

The data model of the RISIS Patent database is shown below. It includes on the one hand the 6 tables related to the variables describing the patents (as they are already present in RISIS Patent Database) and on the other hand 5 tables describing the company variables (see Figure 3).

The tables related to patent variables (Figure 4) are:

- Table CIB_Priority_Patent Attributes gives a set of information related to the patent application and publication places and dates,
- Table CIB_Textual_Content includes the pieces of textual information useful for text mining. It aims at defining the core of the protected inventions beside the predefined technological classifications,
- Table CIB_Technological_Classification of patents relates patent to technology on a fractional counting based using existing classification built on IPC codes,
- Table CIB_Patent_Value proposes several indicators to estimate the patent application value
- Table CIB_Geography _Location of inventions informs on the locations of the inventions based on the inventors addresses,
- Table CIB_Actors focuses on the institutions that patent and give information on their type, name and location.

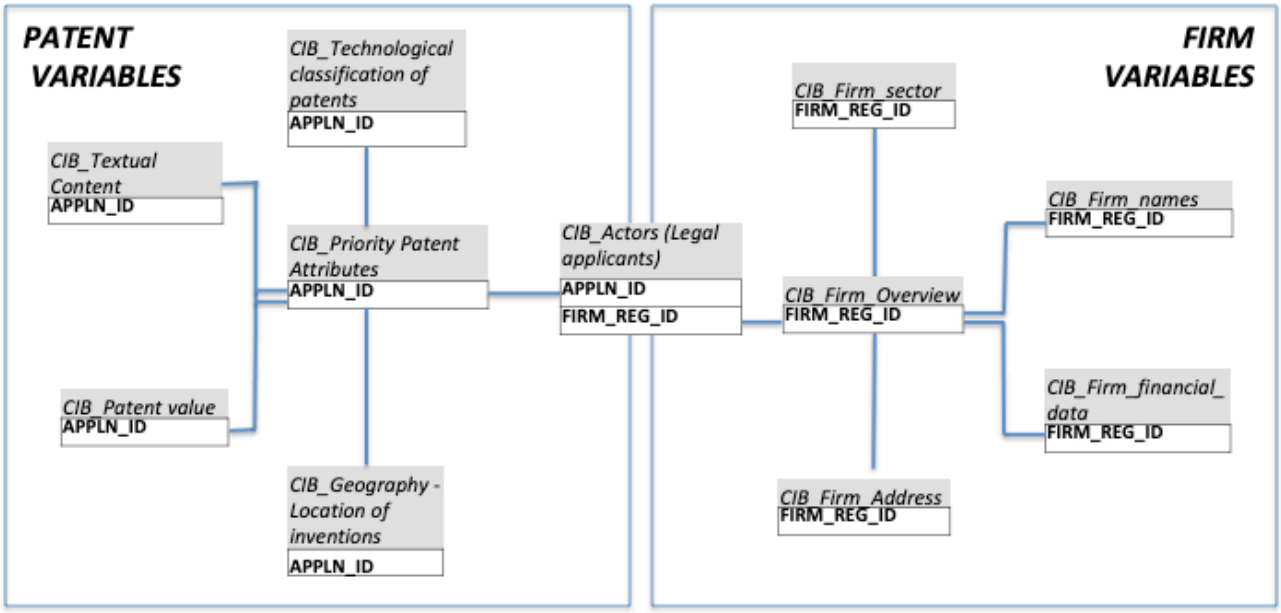
The tables related to the company variables (Figure 5) are:

- Table CIB_firms gives a set of basic information related to the company: Its current simplified name, HQ country, industrial sector, size category, …
- Table CIB_firm address details the location of the company HQ (address, geo coordinates, regions, rurban area),
- Table CIB_sector lists the NACE codes and labels of the companies,
- Table CIB_firm_name proposes the various names (past or in use) for the companies,
- Table CIB_financial data gives information on the financial situation of the companies. It includes revenue, assets, number of employees, profit/loss, return on equity, return on assets. It includes also R&D investments of a set of companies [14].

Figure 3: Data model of RISIS Patent - Global view

[14] R&D investments are available only for companies that were included in EU Industrial R&D scoreboards.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091



This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 4: Data model of RISIS Patent -View of patent variables

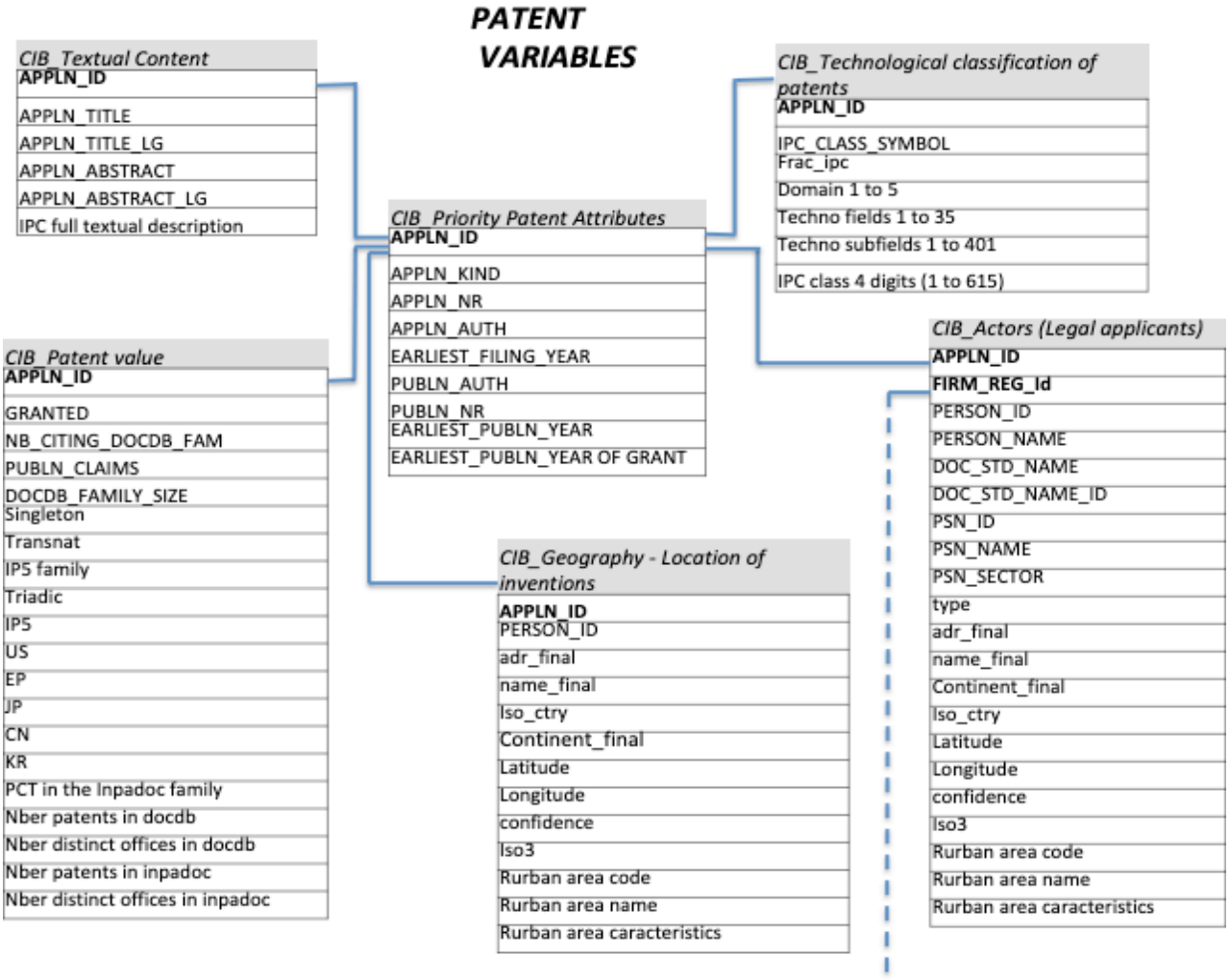
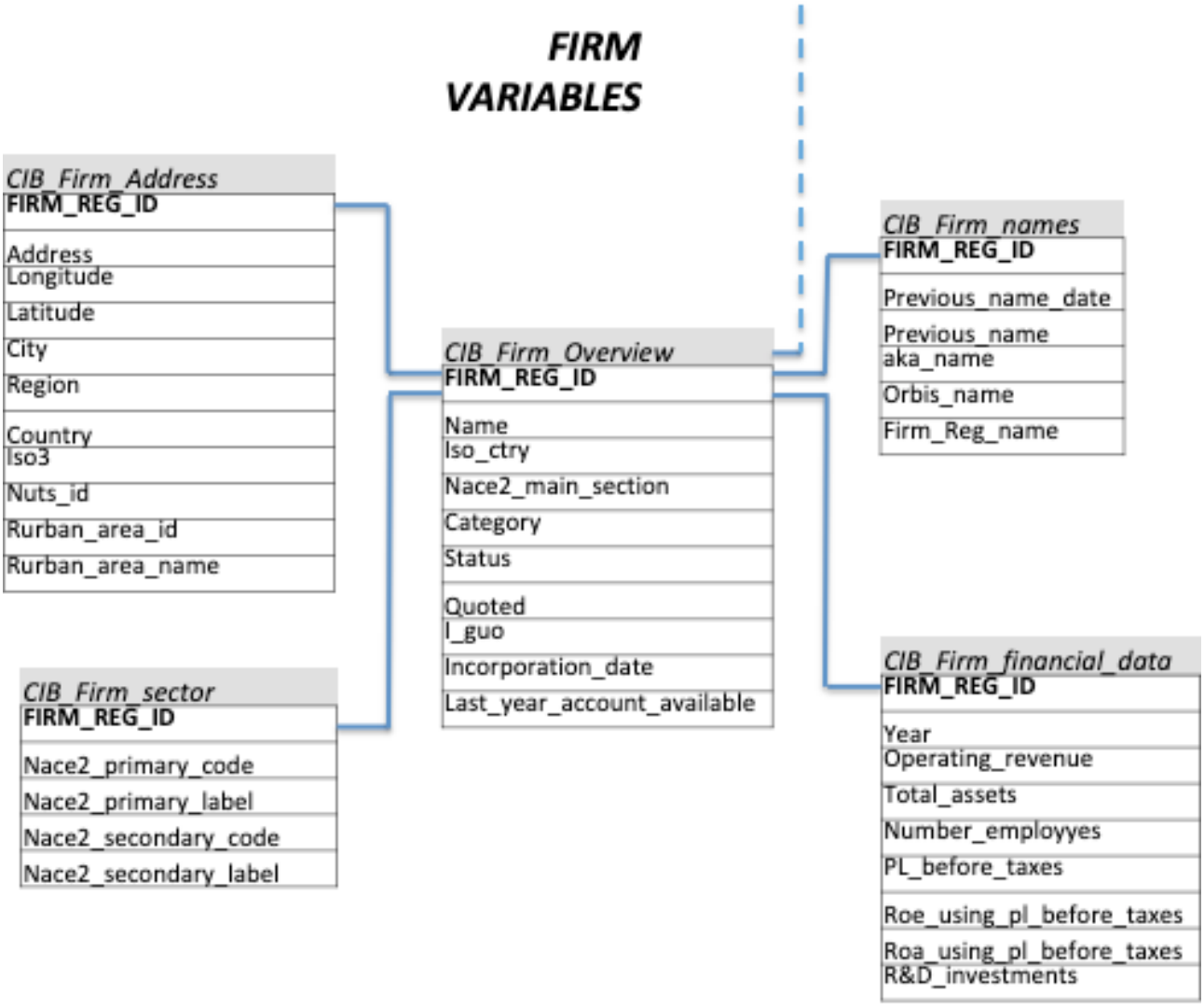


Figure 5: Data model of RISIS Patent -View of firm variables

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091



3.4 Interfaces for access to other infrastructures

CIB2 Database is inter-linked with the firm register (FirmReg). In a following edition, it should link with the register of public institutions (OrgReg). Furthermore, the database is designed to comply with RISIS integrative dimensions (actors, space) as described in the documentation of RISIS Patent Database.

The addresses (firm headquarter, applicants and inventors) are geocoded and allocated to rurban areas (using CORTEXT geo services). The coverage of the address geocoding and allocation to a rurban area is quite good but depends on the quality of the initial geographical information. Table 16 shows the share of geocoded and allocated addresses of inventors. Shares are above 70% in North America and most of the European countries.

Table 15: Completeness of the geocoding and Rurban Area allocation of the inventors addresses.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Country of inventors	Number of addresses	Share of geocoded addresses	Share of addresses allocated to a Rurban area
US	977,602	94,7%	94,1%
JP	964,198	90,0%	68,9%
DE	347,400	72,2%	72,0%
Unknown	303,173	1,0%	1,0%
KR	182,619	47,8%	47,2%
CN	131,952	57,4%	57,4%
FR	121,932	90,3%	89,9%
TW	103,306	43,5%	43,4%
GB	73,790	73,3%	73,2%
IN	59,825	80,3%	79,8%
NL	50,258	92,9%	92,9%
CA	45,446	83,8%	83,6%
CH	40,463	87,9%	87,8%
SE	37,818	91,1%	90,1%
IT	34,801	90,7%	90,7%
FI	26,697	92,2%	92,1%
DK	22,244	85,3%	85,1%
BE	20,999	90,1%	90,1%
IL	18,276	75,1%	74,8%
AT	17,714	77,1%	77,0%
ES	16,122	79,8%	79,7%
AU	10,525	89,0%	87,4%
SG	9,549	90,9%	90,8%

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

BR	7,459	83,3%	82,9%
NO	6,912	68,9%	68,6%

Actors

We deal with actors that are legal applicants (individual applicants were discarded from the RISIS Patent Database. In the current CIB2 database, CIB2 companies are identified with a unique FIRM_REG_ID. Variables related to the companies are given to describe their location, industrial sectors, financial situation, etc. For others legal actors (legal entities that co-apply patents with CIB2 companies), we only rely on the sectorial information provided in the raw PATSTAT database[15].

4 Integration with RCF

The current CIB2, which is made available for access to researchers in RISIS is foreseen to be fully incorporated in RCF, under the condition of controlled access and provided that security of usage is given (i.e. access for selected users with a concrete research project to the parts of the dataset needed for the research). Linking to other datasets in the RCF will be realized via the RISIS registers (providing the respective identifiers to the registers in RISIS Patent Database). Technical issues for incorporation of CIB2 database into RCF (e.g. database system, how can a user access which parts of the dataset, etc.) are to be defined in close cooperation with the developers of the RISIS Core facility that should open in Autumn 2020.

5 Scientific use and main references

CIB2 database is a rich data source accessible via RISIS for research activities in the production of knowledge using patent data. It allows studying the dynamics of knowledge creation along different dimensions: space, actors and technologies.

Thanks to its links with other RISIS facilities, CIB2 database enables to access to these dimensions at a coarse level or at a fine grained level using either usual classification (for technologies, geography) or designing ad-hoc data subsets of patents for specific topics of inventions, for a particular type of institutions and/or in given geographical spaces.

The previous version of the CIB has been used to:

- Analyse the **exploitation of new knowledge** in specific industries (pharmaceutical and chemical industries), done by researchers from Universitй Paris-Est Marne-la-Vallee
- Explore the **Inventive Productivity of Multinational Firms** using non-parametric modelling (Conditional Efficiency Analysis)
- Analyse of **the internationalisation of applied knowledge** production with a focus on special countries (Israel, central European countries)
- Analyse the internationalisation of the IP protection in large R&D corporate performers

[15] Harmonizing names and allocation of assignee sectors in Patstat raw data was done by ECOOM (K.U. LEUVEN; http://www.ecoom.be/en/EEE-PPAT)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Recent References

Laurens, P., Le Bas, C., Schoen, A. (2018), Worldwide IP coverage of patented inventions in large pharmaceutical firms: to what extent do the internationalisation of R&D and firm strategy matter?

Submitted to the International Journal of Technology Management

Laurens, P., Le Bas, C., Lhuillery S., Schoen, A., (2018) Firm specialisation in clean energy technologies: the influence of path dependence and technological diversification.

Revue d economie Industrielle, n°164 (4eme trimestre 2018)

"Evolving technological capabilities of firms; Complexity, divergence, and stagnation" Antoine Schoen[2], Patricia Laurens[1], Alfredo Yegros[3], Philippe Larйdo

STI conference Paris, September 2017

Gaston Heimeriks, Antoine Schoen, Patricia Laurens, Alfredo Yegros and Dieter Franz Kogler

Knowledge, networks and proximities - An analysis of knowledge dynamics in the Chemical and Pharmaceutical and Biotechnology sectors

EUSPRI 2018 conference Paris 2018

Laurens Patricia, Antoine Schoen, Pierluigi Toma and Cinzia Daraio

Exploring the Innovative Efficiency of Big Multinational Firms through Conditional Efficiency Analysis

STI 2018, Leiden (Netherlands) 12-14 September

The RISIS Patent Database documentation can be found on:

https://zenodo.org/record/3342454/files/Documentation_RISIS%20Patstat_Final.pdf?download=1

https://hal.archives-ouvertes.fr/ (id: **hal-02361266, v1**)

https://gitlab.com/cortext/risis-patents-database/-/tree/master/

Information on CIB2 database is also available on: https://github.com/cortext/cib-database.