CRIOS

WORKING PAPER SERIES

CRIOS – Patstat Database: Sources, Contents and Access Rules

Monica Coffano and Gianluca Tarasconi

Working Paper n.1

This Version: February 2014

CRIOS – Center for Research on Innovation, Organization and Strategy
Via Guglielmo Röntgen 1, 20136 Milano –Italy
www.crios.unibocconi.eu

The opinions expressed in the working papers are those of the authors alone, and not those of the Institute, which takes non institutional policy position, nor those of Università Bocconi.

Crios-Patstat Database: Sources, Contents and Access Rules

Monica Coffano [a,1], Gianluca Tarasconi [b,2]

[a] *Ecole Polytechnique Federale De Lausanne, CDM ITPP CEMI, ODY 4 16 (Odyssea) Station 5 CH-1015 Lausanne, Switzerland*

[b] *Bocconi University, CRIOS Via Roentgen 1, Milan, Italy*

Abstract

The CRIOS-PatStat is a patent database made by a team of researchers active at Bocconi University (Milan).

In this database the user can find, for European patent office based applications, disambiguated inventors and applicants' names as well as other standardized information that are often difficult to find in the other patent databases.

This paper was written with the aim to provide a complete explanation of its content and also an overview on the steps that led to the creation of this source of information.

JEL Classification: C81

KEYWORDS: Patents; academic patenting; inventor data; intellectual property; data cleaning

ACKNOLEDGMENTS: We are grateful for helpful comments from Francesco Lissoni, Valerio Sterzi. We acknowledge the valuable efforts of all the researchers that during these years contributed in the creation of the DB, in particular Stefano Breschi and Michele Pezzoni.

[1] *E-mail address :* monica.coffano@epfl.ch (M.Coffano) Tel: +41 21 69 30008

[2] *E-mail address :* gianluca.tarasconi@unibocconi.it (G. Tarasconi) Tel: +39 02 5836 2334

Sommario

Introduction

Patent data have for long been a key source of information for economists of innovation. Early and current criticisms of their limitations and possible biases have done little or nothing to prevent their diffusion in scholarly work (Griliches, 1990; Nagaoka et al., 2010). Such a success is mainly explained by the increasing, and now almost complete, digitalization of patent archives, a change of attitude by patent authorities (more and more open to access requests by the scientific community), and the growth of computational power.

Along with more extensive use, economists have experienced with more intensive use, that is to say they have gone beyond the simple counting of patents or their classification by broad technological classes and geographic units, and started making use of information on applicants, inventors, addresses (especially of inventors) citations, claims, oppositions, priority links across patents from different offices (patent families), and a lot more.

Such information, when obtained directly at the source (that is the patent offices, as opposed to commercial retailers such as Thomson© or Questel©), comes in raw formats, which require various operations of parsing, cleaning, and disambiguation. Name disambiguation of inventors, in particular, it has recently become a key issue, due to widespread interest in themes such networks of inventors, inventors' mobility and other topics that require treatment of individual inventors or applicants over time, across space, or in relational terms (see for example Melamed (2006) or Raffo(2009)).

Name disambiguation of applicants has become important, too, due to the diffusion of micro-econometric studies on innovation, entrepreneurship, and R&D collaboration, which make use and possibly combine different sources of firm-level data, as in Thoma (2007) but also helps in increasing accuracy of patents portfolio analysis so measuring non tangible assets in financial evaluations..

In this paper we provide a technical description of the CRIOS-PatStat database, which contains disambiguated data on inventors, applicants, normalized and reclassified information on technology area(s) where the parent belongs to, registry data of the application document (application date vs. priority date), forward and backward citations and legal status data (like changes of ownership, payment of renewal fees and other), for patent applied at EPO and USPTO.

The CRIOS-PatStat database is the result of such disambiguation efforts, as they have been pursued over time by a team of researchers active at Bocconi University (Milan) since the 1990s, first at a research center called CESPRI, then KITES, and now CRIOS (which explain why, over time, the dataset has taken various names, derived from the center it existed at the time: for instance, Lissoni (2006) cites it as EP-INV and Frazzoni (2011) cites it as KITeS-Patstat). Over time, the database made use of different raw data sources, the latest one being PatStat, the short name for the Worldwide Patent Statistical Database, released periodically by EPO, the European Patent Office. Due to large and increasing size of the scholar community of PatStat users, the CRIOS-PatStat database has become a key research tool, which CRIOS researchers share with an increasing number of colleagues. Hence the necessity to provide a full summary of its contents, as well as of the methodologies used for its creation and the access rules.

Data sources

Before describing in details tables structure and data available in Crios-PATSTAT DB it is necessary to deepen the content of data that originate the tables.

Besides PatStat, the CRIOS-PatStat database makes us of two other data sources, namely: PROFLIST (a collection of country-specific datasets of university researchers, complete with affiliation) and TLS221 (a table produced by EPO, which contains information on the legal status of patent applications over time). Some information for early patents also derives from sources available before the creation of PatStat by EPO, which are worth mentioning for methodological reasons.

We examine all of these sources in turn.

Early Sources (Pre-Patstat)

CRIOS (previously CESPRI and KITeS) started in 1996 building datasets of EPO[3] inventors and applicants.

The first dataset was based on EPO database **ESPACE BULLETIN**[4] , using **Mimosa**[5] Software. The set of data downloaded included only EPO published patents, structured by publication number.

[3] European Patent Office

[4] ESPACE Bulletin offers a rich and comprehensive access to bibliographic and procedural data for all European patent applications. Patent attorneys, patent information professionals and others can use ESPACE Bulletin for a variety of activities.

[5] MIMOSA retrieval software is a Microsoft Windows-based software developed for the Trilateral Offices (the EPO, the Japan Patent Office and the US Patent and Trademark Office). MIMOSA has a user-friendly interface in ten different languages. You can customize various parameters for display, printing and downloading to suit your own needs.

Espace Bulletin was the source for applicants and inventors names and addresses, along with other patent related data (like publication date, international patent classification…)

Aside from such data, another product named REFI[6], provided by EPO in a TXT file, was source data for patent citations table.

In 2001 the dataset was restructured as a ***relational database***[7] where the main key was EPO publication number, since this was the patent key in Espace Bulletin data and in REFI.

The core structure of database remained unchanged until 2008, when EPO released PATSTAT[8] (EPO worldwide patent statistic database) that became the main source of data for Crios due to the increased easiness to import the data and completeness and consistence of the content.

Patstat was, differently from Mimosa, structured by application id, defined as a surrogate key for distinct patent applications. Unfortunately application ids were not stable, changing each edition, until April 2011 edition. At that point it was possible to restructure Crios database using application id as key for core tables, building up tables' structure as described in this article.

Following chapters will describe in detail actual content of database, its sources and some of applied algorithms for data quality and entities disambiguation.

[6] REFI database contains citations from all searches made at the EPO (for BE, CH, EP, FR, GB, NL, TR) and citations for all WO publications are supported by the master database as well as data from a number of other countries (including US and DE).

[7] A relational database is a DB in which data are stored in tables and also the relationships among the data are stored in tables. The data can be accessed or reassembled in many different ways without having to change the table forms.

[8] See http://www.epo.org/searching/subscription/raw/product-14-24.html

Patstat

PatStat is the short name for “EPO worldwide PATent STATistical Database”, a single database covering a large number of patent offices, which has been developed by the European patent Office (EPO) for all non-commercial users interested in advanced, large scale statistical analysis of patent data.

There are also many other organizations that cooperate with EPO to create PatStat, like the World Intellectual Property Organization (WIPO), the OECD , Eurostat and the United States Patent and Trademark Office (USPTO), making it a de facto standard. [9]

Patstat is also the largest source of information about patents, with about 76 million applications, extracted from the original patent documents from over 80 countries. The information are extracted from EPO's master bibliographic database DOCDB, the EPO's worldwide patent information resource[10]. It is updated every 6 months, at the beginning of February and of August.

DOCDB is an ‘examiner centered’ dataset, in the sense that data useful for examination process (i.e. technological classes, citations, etc) are normalized and more complete than those that are marginal for that purpose (like applicant’s name or address).

PATSTAT is built based on PATENT APPLICATION, to whom the full sets of bibliographic variables are added; the most important are:

- Priority date, application and publication number and date

[9] Information on the access and documentation can be found here: http://www.epo.org/searching/subscription/patstat-online.html

[10] For more information about this step you could read “The Global Patent Data Coverage” document http://documents.epo.org/projects/babylon/eponet.nsf/0/2464E1CD907399E0C12572D50031B5DD/$File/global_patent_data_coverage_0711.pdf

- Title and abstract of the application
- Status of application (granted, pending)
- *International Patent technological Classification* (IPC) codes
- Applicant's name and address
- Inventors' names and addresses
- References (citations) to prior-art patents and to non-patent literature
- Patent Family [11]

The information included in Patstat can be organized in three different models:

1) Conceptual model
2) Logical model
3) Physical Model

The most important difference among these three models is the way in which the variables are connected and grouped. In other words, the placement of the attributes is different from a model to another and this means also that the cardinality of the relation changes. In other words, the number of links between 2 variables can change depending from the model. For any reason related to the kind of analysis undergone, it could preferable use one or another.

EPO has decided to deliver to its customers PATSTAT database using the physical model.

The conceptual model is

[11] Explanation of patent families: http://www.epo.org/searching/essentials/patent-families.html

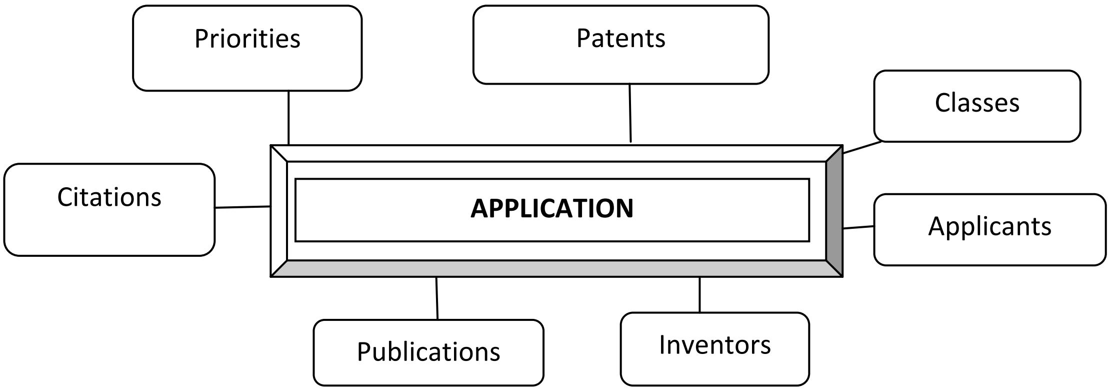


Figure 1: conceptual model Patstat Source: Patstat Manual

The milestone for this model is the APPLICATION (appln_id). This means that for every application you have from 1 to n links with the other information. Some variables could be taken just once, for example the applicant, others more than one, for example citations (one application can have from one to many citations to other patent or non-patent documents).

The logical model instead has two cornerstones: the application and the person conceived as applicant and inventor.

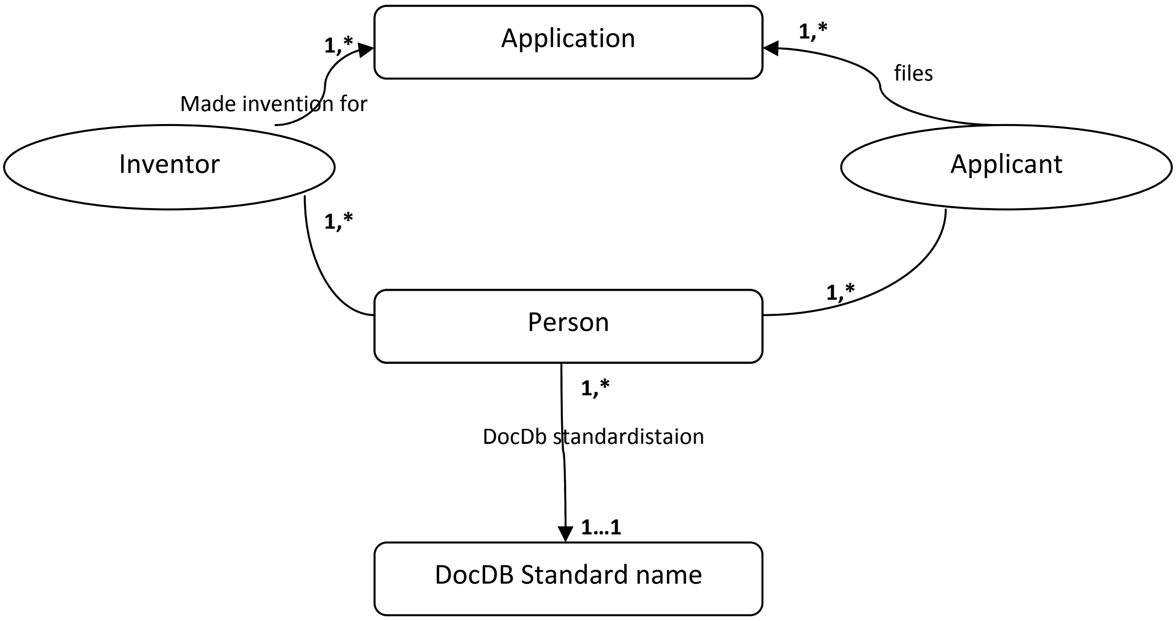


Figure 2: Patstat logical model. Souce: Patstat manual

From the end of December 2007, EPO is committed to the effort of standardization of inventor's names to a standard name. The main problem is that any times an inventor applied for a patent, he/she gives his/her generalities. If those information are not exactly the same as the ones in the previous application, the key (person_id) associated with him/her will be different. In other words, for a single person are associated two different 'person_id'. One of the main aims of the KITES DB is to fix this problem, using an algorithm that can recognize if two people are the same person, based on additional information, like the address or the technological specialization. This will be anyway explained in deep in the next few paragraphs.

The last model, the physical one is the one provided by EPO to its customers. The starting-point is the table tls201, the one in the center of the graph. Its primary key is the 'appln_id', namely the key for the patent application. In this model, all the variables listed before are grouped in tables. Just to give an example that will be useful later on, the table tls206 includes all the information regarding the inventors and the applicants.

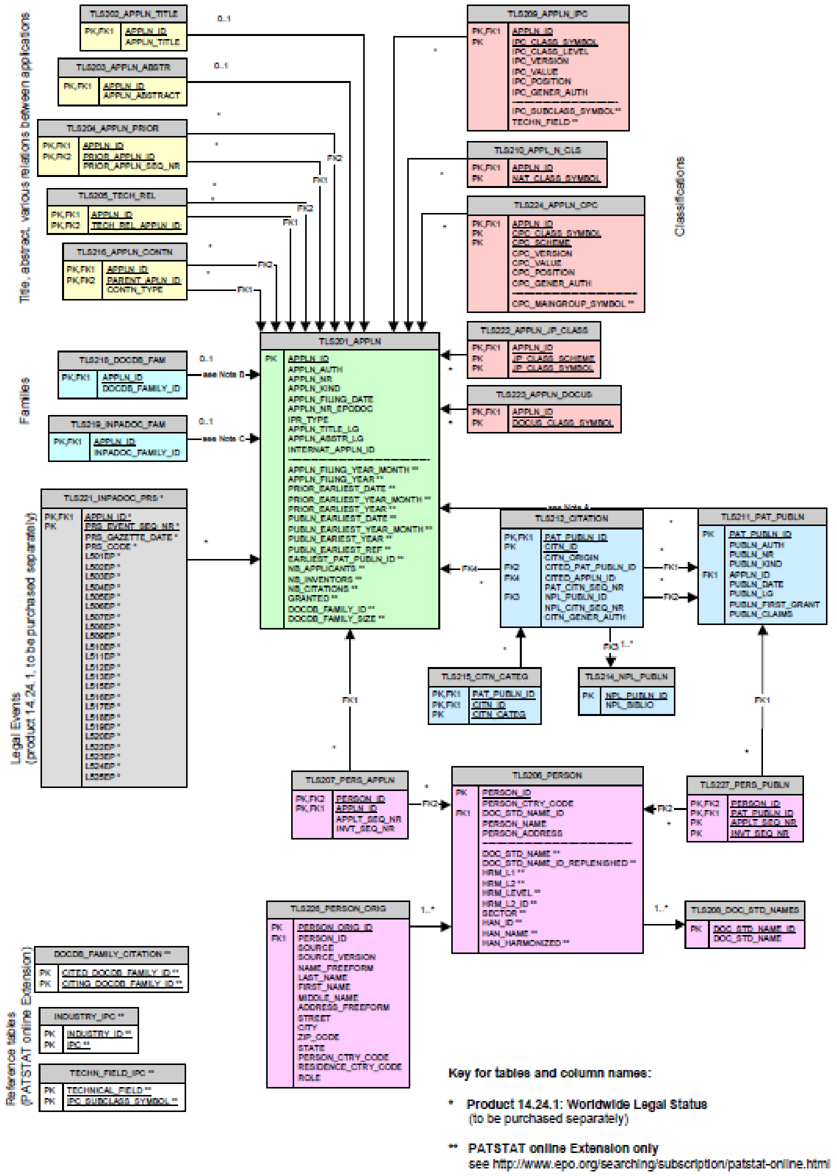


Figure 3: Patstat physical model. Source: Patstat Manual

It is important to underline that all the three models must be considered at the same time, in order to understand completely all the variables and the relationships among them.

Further information on bibliographic data can be retrieved from the EPO GLOBAL PATENT INDEX (GPI)[12]. This is a database designed to enhance easier access to DOCDB, also known as the EPO Patent Information Resource with different level of coverage.

All the data are accessible to everyone, on the condition to purchase a license.[13]

Other information can be found as well on the EPO webpage www.epo.org .

Proflist

The idea of structuring cleaning and harmonizing the data in CRIOS-Patstat DB, began in 2004 when CRIOS in association with BETA (Universitè "Louis Pasteur", Strasbourg) and CHALMERS IMIT-Chalmers University (Gotheborg) decided to collect additional information about academic inventors. The aim was to have a list of professors from 2000 to 2004, with some personal information attached, collected through a direct interview aimed to discover with high accuracy if a professor was the owner of the list of patent that appears in CRIOS-Patstat DB linked with inventor name.

Proflists were created initially for Italy, France, Sweden[14], United Kingdom[15], Denmark[16] and Netherlands[17].

For all countries, the social scientists and humanists were not taken into account, even if there are many social scientists that have an academic training as engineer, just to decrease the incidence of the problem of homonymy.

[12] Global patent index references: http://patenty.bg.agh.edu.pl/graf/globalpatent.pdf

[13] More information on EPO license: http://www.epo.org/searching/subscription/raw/product-14-24.html

[14] More information on Italian, French and Swedish proflis: Lissoni, Sanditov, Tarasconi (2006), Lissoni Pezzoni et al, 2013

[15] More information on UK: Sterzi (2013)

[16] More information on Denmark: Lissoni, Lotz et al., 2009

[17] More information on Netherlands: Lissoni e Montobbio, forthcoming

Even if in each country the regulation regarding professors could differ, the procedure that leads to the creation of the proflists is more or less the same. Just to give an example, in Italy a professor is a civil servant, so the Italian PROFLIST was provided by the Italian Ministry of Education. In Sweden on the contrary, the professors are not civil servant so Ingrid Schild (Dept. of Sociology, Umea Univ.) was the person in charge to collect data of personnel from as many Swedish academic institutions as possible.

Furthermore the APE-INV project[18] has been created with the aim of "sharing experiences for the creation of inventors Database, producing a Database on **Academic** Patenting in Europe, editing joint publications using the Data-set, designing a method to allow users to correct data (NAME GAME COMPETITION) and cooperating with established institutions in the field of patent data". This project led many participants' universities to create their own PROFLIST, whose obviously could be added in future at CRIOS-Patstat DB to improve it.

[18] The APR-INV is a project funded by the European Science fundation. More information on: http://www.esf-ape-inv.eu/

So nowadays the proflists available are the following:

	DENMARK	NETHERLANS	SWEDEN	UK	FRANCE	ITALY
Time interval investigated	1994-2007	2006-2007	1978-2010	1986-2006	1978-2002	1996-2007
Variables included						
Name	YES	YES	YES	YES	YES	YES
Surname	YES	YES	YES	YES	YES	YES
Gender	YES	YES	YES	YES	YES	YES
Date of birth	YES	YES	YES	YES	YES	YES
University affiliation	YES	YES	YES	YES	YES	YES
Public institution	NO	NO	NO	NO	NO	YES
Rank	YES	YES	YES	YES	YES	YES
Honorary Status	NO	NO	NO	NO	NO	YES
Date of Nomination	NO	NO	YES	NO	YES	YES
Disciplinary field	YES	YES	YES	YES	YES	YES
Faculty	YES	YES	YES	YES	NO	YES
Private address	YES	YES	YES	YES	NO	NO
Department	NO	YES	YES	YES	NO	NO

The proflists are not an open source. They must be requested directly from the author.

It is possible to find the complete list of the contacts here: http://www.esf-ape-inv.eu/index.php?page=3#acadpat

Legal status data (TLS221)

The last data source comes directly from EPO, and it contains legal status information, which is best conceived as "events" affecting the patent during its lifetime. Among such events, those of most immediate interest are the payments of renewal fees (from which one can calculate the effective life span of the patents), the oppositions (if any) and the designated contracting states. Different from the previous is the ownership changes of applications, that can happen only when the patent is not granted yet (so it is still an application).

Using a more technical, but also more accurate vocabulary, the legal status events can be classified into several categories: invalidations, reinstatement, oppositions, Assignments, PCT Entries, SPC[19]/Term extension. Each category has a unique code, which allows immediate retrieval of the events of interest. The full list of possible legal status by patent authority can be downloaded from EPO website at following URL: http://archive.epo.org/inpadoc/index_epcodes.htm.

The dataset, also known as PRS (Patent Register Service), is distributed separately from PATSTAT.

It is anyway structured in the same way as the other Patstat tables, with ***appln_id*** as primary key (in this case the Patstat table where the user can find these kind of information is called TLS221).

The records originate from the patent gazettes or register, as provided by 63 patenting authorities (as of October 2011). They concern patents in 39 countries plus the national phases for PCT/EP documents in 24 countries, and include patent number, gazette date announcing the action, the legal status code and the equivalent text description.

It must be noted that time coverage of legal status is different among application authorities: the oldest data are from Suisse patent office (1958) while some offices have no data before 2008 (Croatia and Egypt).

[19] **supplementary protection certificate**

We may consider that the dataset has a full coverage for EP from its foundation up to now, and that data for European patent offices are well covered for the last 20 years.

Other major application authorities have gaps that make the dataset not useful: for Japan FI only PCT entries are listed into the table; for USPTO not all events are covered (FI extension and designated states related is given only for those applications linked to EP or WO publications). In this way an analysis of legal events for such patent offices would be biased because some events / kind of applications would not appear at all.

EEE-PPAT database

ECOOM-EUROSTAT-EPO PATSTAT Person Augmented Table (EEE-PPAT), is a plug and play extension of patstat based on persons table, adding two pieces of information:

Sector assignment – i.e. identifying whether patentees are private business enterprises, universities / higher education institutions, governmental agencies, individuals – is highly relevant for analyzing the constituents and dynamics of technological performance on the level of innovation systems;

Harmonized Assignee Names: ECOOM, in partnership with Sogeti, developed a comprehensive method in 2006 to arrive at harmonized patentee names in an automated way. The number of unique patentee names was reduced by approximately 20% and the average number of patents per patentee increased from 5.5 before to 6.8 after harmonization.

We use such data as additional information in patent applicants tables.

For details and methodology see Du Plessis (2009)

Data Quality Algorithms and Data Enrichments

Address Cleaning and Standardization

As previously written, Patstat data variables that are not directly involved in examination process have a lesser quality than others. This is the case of misspelled names and/or addresses of applicants and inventors, as well as wrongly parsed applicants' or inventors' records. Quite common it is also the case of identical names and addresses written according to different standards in different records (e.g. inversion of names and surnames, inconsistent use of initials, inversion of civic number and road name etc).

Nevertheless such data are fundamental for researchers, for example in order to locate inventions in space or to attribute correctly to the same individual, when it is the case.

The first step in order to build a reliable dataset of inventors and applicants is based on data-mining techniques, according to the following steps:

PARSING: the data are often stored in a unique text field regardless disomogeneity of content; in this stage strings are parsed into other fields according to their content. (FI field address is broken down into street, city, zip code or generic name divided into first name, surname, title, etc)

CLEANING: the most common spelling errors are automatically corrected.

STANDARDIZATION: same contents coming from different sources are expressed in the same way (i.e. MILAN versus MILANO – language disomogeneity).

DEDUPLICATION: inventors or applicants that after previous three stages now contain the same information are collapsed into the same ID.

Due to the strict sequentially of the process, last steps (address standardization and deduplication) results greatly depend from the quality of first two steps.

All above operations rely on a methodology based on 25 dictionaries tables and 950 recursive queries, allowing maximum portability to other context.

MASSACRATOR: further inventors disambiguation[20]

After completing the first cleanup steps it has been clear that a sheer usage of data-mining techniques was not enough for reconciling all possible name and address variations (also taking in count possible changes in address due to inventor mobility).

For example:

Name	Address	City	Zip	codinv2
Tarasconi, Gianluca	Via P. Maspero, 24	Milan		1
Tarasconi, Gianluca	Via Maspero, 24	IT-20137 Milan		2
Tarasconi, G.	c/o university bocconi	Milano	20136	3
Tarasconi, Gianluca	c/o university bocconi	Milano	20136	4
Tarasconi, Gianluca	35, Via Tertulliano	Milan		5

Can be collapsed to

Name	Address	City	Zip	codinv2
Tarasconi, Gianluca	Via Maspero, 24	Milano	20137	1
Tarasconi, Gianluca	c/o university bocconi	Milano	20136	3
Tarasconi, Gianluca	Via Tertulliano, 35	Milano	20135	5

[20] For more detailed information on Massacrator routine please see : http://ideas.repec.org/p/grt/wpegrt/2012-29.html

But from this point onward only using name/address information we cannot say whether these three homonyms are the same person or not.

Inventors data are restructured following a structure person (entity) based (CODINV) vs person@location (CODINV2).

All inventors with similar name and surname are compared in pairs, through the Massacrator SQL routine.

Such algorithm, described in detail in Pezzoni 2012, allows to identify which homonyms are likely to be the same person, based on other information contained in the rest of the dataset (see picture below).

All common factors among two candidate pairs are computed giving eventually a score. Pairs whose score is above a certain threshold are considered to be the same.

Massacrator can calibrate in order to either maximize one out of two data quality indicators (namely, precision and recall) or to choose one among the several Pareto-optimal combinations of the two. At present, the CRIOS-PatStat database make use of the results of a “balanced” calibration, with 70% precision and recall[21] rate between 50% till 65%, when tested against available benchmarks.

[21] **Precision rate** = true positives / (true positives + false positives)
Recall rate = true positives / (true positives +false negatives)

Eventually three dataset are produced choosing different criteria:

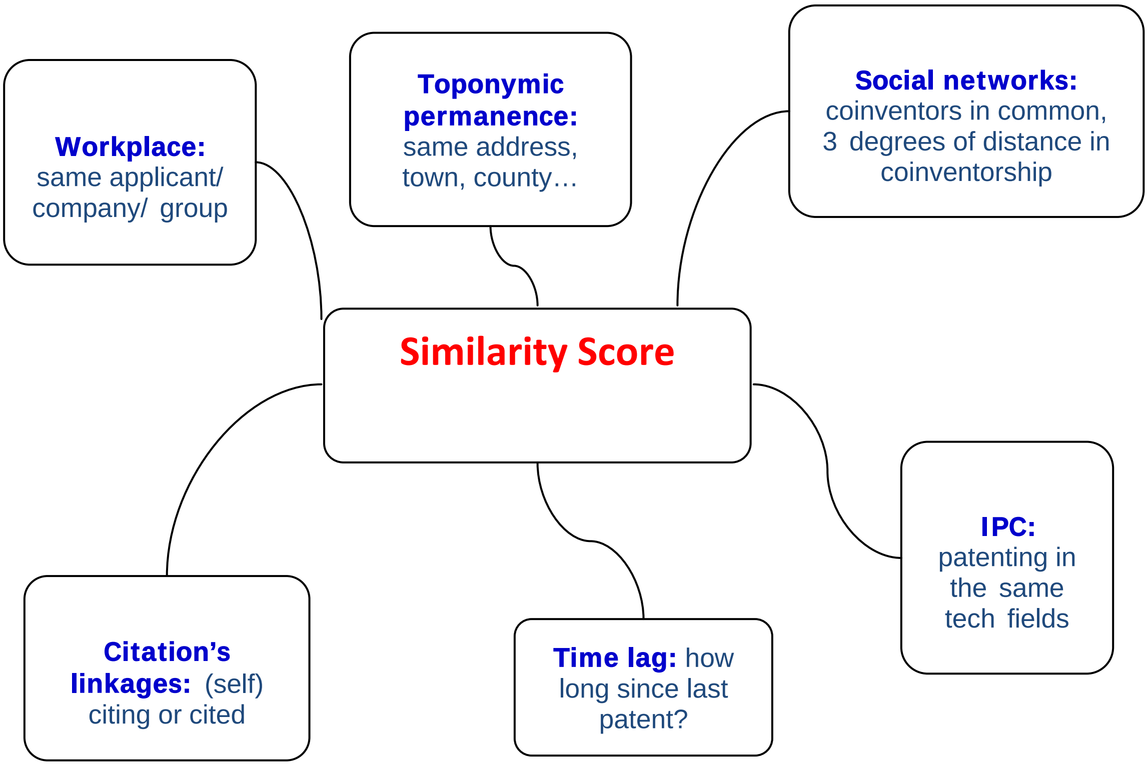

Workplace: same applicant/ company/ group
Toponymic permanence: same address, town, county…
Social networks: coinventors in common, 3 degrees of distance in coinventorship
Similarity Score
Citation's linkages: (self) citing or cited
Time lag: how long since last patent?
IPC: patenting in the same tech fields

Figure 4: disambiguation

Matching Inventor to Academic Scientists

As written before, the matching of inventors and academic scientists (for short, professors) from the PROFLISTs, is meant to identify exactly the person and his or her patents, trying to minimize the negative effect of homonymy. This procedure results from three steps where the first is the Massacrator procedure that is already explained in section 3.b, and the last two can be defined as "matching" (and "filtering").

In the matching step, pairs of inventors and professors are created, through name analysis. In the filtering step, false positives (matched professors and inventors who are not, in fact,

High precision: minimizing the number of false positives
High recall: minimizing the number of false negatives
Balanced: choosing the best compromise among the two rates

the same person) are eliminated, in order to retain only the true positives (professors and inventors who are the same person that is the true and only “academic inventors”).The matching step is, in turn, organized in three sub-steps.

First, the entire professors’ and inventors’ name strings, middle name included, were matched. Even if this is the “narrow” match, before proceeding, there were taken some devices, several general, some according with the country under analysis. For every inventor, the string of the full name is composed by surname+ first name+ middle name and the main goal is to divide correctly the three different parts. To do this, in Massacrator it has been decided to follow the characteristics of the different PROFLIST languages. For example for French, the letter “M.” means Mr. or Ms. and it is deleted to render lighter the matching. Then some common small words like 'BEN', 'DA', 'DEL', 'DI', 'DU', 'EL', 'LA', 'LE', 'VON', 'SAINT', 'SAINTE', 'DAL', 'MC', 'DO', 'DOS', 'DES', ‘DE’, ‘DE LA’, ‘DE SAINT’, ‘VAN’, ‘VAN DE/DEN/DER’, ‘AB DER’ are seen as part of the surname, while others like ‘DE’, ‘DE LA’, ‘DES’, ‘DA’, ‘VAN DER’, ‘VON’ are part of the first name. Then the data are undergone to a manual check to avoid some previous typing errors, for example when the surname is inserted two times instead of one. The same procedure of separation of the fields of the name string is done for the name in the PROFLISTs.

In the second sub-step, just the first name and the surname were taken into account.

In the third sub-step, the more precise one, there were added two filters on the age and on the field of study, in order to avoid the matching between two individuals that could not be the same person for a too large topic difference or easily because they were too young to claim a patent. In fact the age filter is aimed to drop the professors whose the priority date of the matched patents is antecedent the 21th year of the professors age. The field filter is more intuitive: two professor that patenting in two different fields as astronomy and biology cannot be the same person. Unfortunately this filter is based on the common logic and not on an experienced "list of fields that could be linked for patenting", so maybe this procedure could be optimized.

After this procedure we are not sure yet if the matching between professor and inventors is correct. Even adding some additional information like the name of the applicant of the patent and the name of the co-inventors, we cannot reach an indisputable result but these information are useful to create an Access database used by research assistant to perform the last of the three steps: the email/telephone interview. "The record of the database contains information on one professor, on the inventor(s) who have been matched to him/her, and on these inventor(s)' patents (such as IPC codes, applicant, priority date and title); it also contains a few blank fields to be filled in by the mask user (such as e-mail address, phone number, and a number of yes/no fields to indicate whether the professor has been contacted, and whether he/she has confirmed to be the same person as the matched inventor)".

In conclusion, it is interesting to see the count of records in each step, to see how this step series works. For France, for example, the starting point before the Massacrator procedure included 119625 French inventors (this means inventors with a French address). After the assignment of CODINV the inventors were 98227. The French PROFLIST contained 32006 professors, working just in scientific or technical fields. After the matching on the entire string of the full name the result was of 4503 records. After the matching between name and surname the result was of 9270 records that are reduced at 7100 after the filtering of age and discipline. The combination of these two results led to a total of 4731 pairs of inventor professor. The interviews for French professors were too much, so it was decided to contact just those that pairs having their latest patent after 1993, so in total 3951. Information on 2884 pairs was collected through direct contact. The result was that 1324 pairs correspond to 1235 academic inventors. For all the details about the results of the other PROFLIST see Lissoni, Tarasconi, Sanditov (2006).

Company Structure and Data

One of the most important data subset included in Patstat is regarding name and address of applicant.

These data help to understand who and where the ownership of invention belongs to. Obviously raw data coming from Patstat need to be cleaned and standardized, as already explained in previous paragraphs, in order to:

- Distinguish individual applicants from companies
- Disambiguate and classify patenting entities
- Rebuild company relationships
- Match companies with other sources of info (FI economic data etc.)

One more complex topic related to applicants names and data is to keep track of companies (and their IP assets) overtime, also taking in count mergers and acquisition that may (or not) reflect in changes in patents ownership. In a previous version Crios-PATSTAT database had, among applicants tables, a data structure aimed to structure such information but due to the complexity of data collection and the availability of other matches among PATSTAT data and other sources of economic information (see [Thoma 2007] and [Van Looy 2009]) such tables are still available but no more maintained.

Nowadays data structure contains three levels of data for application owners:

- **Applicant**: it is the cleaned data contained in PATSTAT. It is the original name along with geographic data.
- **Company**: it is the aggregation of applicants based on country and name, after cleaning and standardizing it. It may group applicants with small names change overtime or consider also ownership changes where the data are available. Recently also standard names from EEE-PPAT database (see paragraph 2.e) has been also used for helping disambiguation.
- **Group**: it aggregates applicants via ultimate global owner. This variable aggregates companies also across different countries.

Aggregation of data has been done, after removing individual applicants, on the base of name / country considering the same company the pair NAME / COUNTRY CODE, accepting thus a percentage of error due to applicants' homonymy that is anyway very small.

Group data have been collected manually and only for certain technological area (ICT in particular) so the database do not have a 100% coverage for such table.

IPC Classification

One of the most useful codifications to handle patents is the International Patent Classification (IPC), a hierarchical classification system created in 1971, as the results of the Strasbourg Agreement[22], and now widely used to classify the patent based on the inventions' technical or scientific domain and/or destination of use. The IPC system is based on an international multi-lateral treaty administered by WIPO. It is composed of eight sections (indicated with Latin letters from A to H), and approximately 70 000 subdivisions, which are revised periodically (we are, at present, at the 8th revision, or IPC-8). It means that there are 8 big groups that characterize the industry. Then each IPC has an extension of 12 digits, in order to get more into detail. Each subdivision has a symbol consisting of Arabic numerals and letters of the Latin alphabet, for example F23G7/06. More detailed information about the IPC is available on the WIPO website[23].

Sometimes this classification turns out to be redundant. For this reason there are two kinds of level, the core and the advanced one. In other words, the core level is a subgroup of the advance level. Each document has at least the core-level classification, that provides about 20 000 classifications[24], so it goes less in deep than the advanced one. For all the others agents that need to have the deeper classification anyway, the advanced-level classification comes into play to provide the more detailed classification.

When a patent office publishes a patent application or grants a patent, it classifies the invention using the version of the current IPC version at the time. Revisions to the IPC generally take place once a year. Whenever the IPC is changed, older documents published under an earlier version are reclassified. So, the user can search all the documents in PATSTAT using the symbols of the latest IPC version.

[22] Full document: http://www.wipo.int/treaties/en/classification/strasbourg/trtdocs_wo026.html

[23] http://www.wipo.int

[24] Since October 2011 only advanced classification IPC are included in Patstat where both core and advanced are available.

In October 2011 PATSTAT contained the IPC-8 codes, or IPC Reform; symbols from IPC systems versions 1 through 7 were not in. From December 2010 WIPO provides a tool to identify the IPC code (http://www.wipo.int/classifications/ipc/en but in any case this is a tool not included in Patstat too.

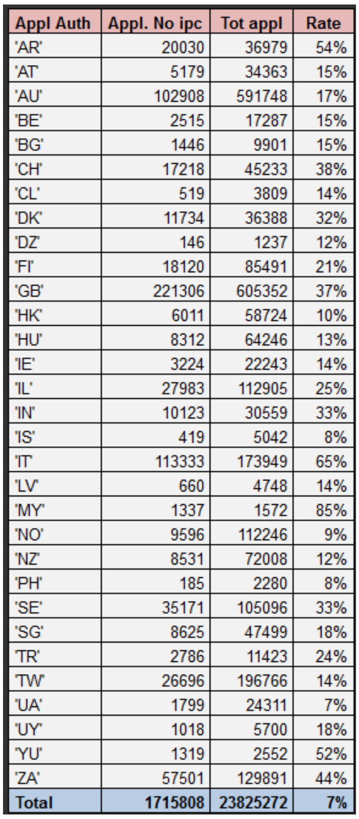
The Patstat table that includes the patents technical classification is the TLS209.

A large number of applications in Patstat have no IPC in TLS209 table. If we take the 2010/10 Patstat edition, about 20% of application ids (12.697.090 out of 66.226.956) have no match in IPC table (but 589.586, an additional 0,8% have a 4 digits only IPC).[25]

This scenario must be cleaned removing D application kinds, 9999 filing years and other oddities, where the application kind shows the kind of document that you are looking at. For example, W identifies a PCT application, U a utility model, A is a patent, D, K, L, M, N identify the dummy for de-duplicating of the artificial publications and F is a design patent.

Anyway if we consider only application kind A and W (patents of invention and PCTs) along with years 1990 to 2008 we still have a 7% of applications without IPC (1.715.808 out of 23.825.272).

[25] Source: http://rawpatentdata.blogspot.ch/



Appl Auth	Appl. No ipc	Tot appl	Rate
'AR'	20030	36979	54%
'AT'	5179	34363	15%
'AU'	102908	591748	17%
'BE'	2515	17287	15%
'BG'	1446	9901	15%
'CH'	17218	45233	38%
'CL'	519	3809	14%
'DK'	11734	36388	32%
'DZ'	146	1237	12%
'FI'	18120	85491	21%
'GB'	221306	605352	37%
'HK'	6011	58724	10%
'HU'	8312	64246	13%
'IE'	3224	22243	14%
'IL'	27983	112905	25%
'IN'	10123	30559	33%
'IS'	419	5042	8%
'IT'	113333	173949	65%
'LV'	660	4748	14%
'MY'	1337	1572	85%
'NO'	9596	112246	9%
'NZ'	8531	72008	12%
'PH'	185	2280	8%
'SE'	35171	105096	33%
'SG'	8625	47499	18%
'TR'	2786	11423	24%
'TW'	26696	196766	14%
'UA'	1799	24311	7%
'UY'	1018	5700	18%
'YU'	1319	2552	52%
'ZA'	57501	129891	44%
Total	**1715808**	**23825272**	**7%**

Table 2: missing IPC

These are detailed results by application authority, filtered for applications authorities with more than 1000 applications and where no IPC rate is equal or greater than average.

We notice some data about some EU national offices like Sweden UK and Italy have 1/3 or more patents with no IPC so we may not rely on such data.

In the CRIOS-Patstat DB there are two tables called “IPCCLASS” and “IPCMAIN” created with the aim to provide a “facilitating tool” in the data handling. In fact, in addition to the data already present in PATSTAT, they report further information. First of all, CRIOS DB offers the concordance tables to convert IPC codes into more aggregated and manageable technological classes as OST7 or OST30 or IPC35 (see Schmoch 2003).

These three different methods divide patents respectively in 7, 30 and 35 groups.

While with the OST7 the patents are grouped in macro-sectors (see appendix B to more detailed information), the OST30 provides a more defined classification, going more into detail. The IPC35 is an updated version of the IPC30, that contains the new subsector like for example digital communications, micro-structural and nano-technologies.

It is easy to understand that, taken the limited precision of these methods, compared with the 700000 IPC classes, one patent could belong to more groups.

The second reason why it is easier work on CRIOS DB to handle data, it is the presence of a variable, called "CLMN", that provide the IPC code in a fix format (for example, the IPC A01C3K becomes A01C003K). This is very useful in matching or comparing data, because all the problems related to the different way to write the code are avoided.

Finally, the last useful element of the CRIOS DB is the possibility to choose the method that better fit with the kind of research undergone or moreover if it is not clear yet which way is the correct one to implement, the possibility to compare the different methods of patent classification and understand which is the favorite one in any case.

Patent Families[26]

The last piece of information derived from Patstat concerns the affiliation of each patent application to one or more "patent families".

The creation of patent family reflects the idea that an invention is better described with a set of applications rather than a single patent, since, due to differences in law, one document applied at one patenting authority may be split into many others when extended at another

[26] More information on patent families: http://www.epo.org/searching/essentials/patent-families/espacenet.html

authority. In other words, one patent in a specific patent authority could be linked with a single application or, if this patent is granted to another patent authority, to n applications. In order to understand the entire innovation, in the second case, the applications are put in the same family, so they will have common information like same priority date, same applicant…

Using the data describing relationships among applications (priorities, continuations and technical relationships) many kind of families or groups equivalents may be built, as described in [Martinez 2010].

Our data add for each application, information regarding two types of families:

DOCDB FAMILY: defined as a group of application with ***all*** priorities in common; these are in reality equivalent patents.

INPADOC FAMILY: defined as a group of application with ***at least one*** priority in common. These are in effect extended families.

Some Articles Using These Data

Bacchiocchi E., Montobbio F. (2010), International Knowledge Diffusion and Home-Bias Effect: Do USPTO and EPO Patent Citations Tell the Same Story?, "Scandinavian Journal of Economics, Vol. 112, Issue 3, pp. 441-470".

Bourelos E., Magnusson M., McKelvey M. (2010), Moving beyond the paradox: Searching for the key factors in research commercialization, To be presented at ESF-APE-INV 3th "Name Game" Workshop-Brussels, on September 5th 2011.

Hall B. H., Thoma G., Torrisi S. (2007), The market value of patents and R&D: Evidence from European firms, "The national bureau of economics research".

Guarisco S., Lissoni F., Sterzi V. (2009), *Academic Patenting in the UK: Evidence from the CID-KEINS Database,* "CESPRI, Università Bocconi".

Huang H., Tang L., Walsh J. (2010), *Disambiguating Patent Inventors: A Non-Name-Matching Approach* Presented at ESF-APE-INV 2nd "Name Game" Workshop- Madrid, 9-10 December 2010.

Lawson C., Sterzi V. (2014) The role of early career factors in the formation of serial academic inventors, *Science and Public Policy*, forthcoming.

Lissoni F., (2008), *Academic inventors as brokers: An exploratory analysis of the KEINS database,* "CESPRI Working Paper 213, Università Bocconi".

Lissoni F., Lotz P., Schovsbo J., Treccani A. (2009), *Academic patenting and the professor's privilege: evidence on Denmark from the KEINS database,* Science and Public Policy, Volume 36, Number 8, pp. 595-607(13)".

Lissoni F., Llerena P., McKelvey M., Sanditov B., (2008), *Academic patenting in Europe: new evidence from the KEINS database*, "Research Evaluation, Volume 17, Number 2, pp. 87-102(16)".

Lissoni F, Maurino A., Pezzoni M., Tarasconi G. (2011), *Ape Inv's "Name Game" Algorithm Challenge: A Guideline for Benchmark Data Analysis & Reporting Version* 1.3, To be presented at ESF-APE-INV 3th "Name Game" Workshop- Brussels, on September 5th 2011.

Lissoni F., Bulat Lissoni F. Sanditov B. (2004), *Networks of Inventors and Academics in France, Italy and Sweden: evidence from the Keins Database,* "CESPRI, Università Bocconi"

Lissoni F., Tarasconi G., Sanditov B. (2006), *The KEINS Database on Academic Inventors: Methodology and Contents*, "CESPRI Working Paper 181, Università Bocconi".

Ljungbergy D., McKelvey M. (2009), *On the relative importance of firms' academic patents,* To be presented at ESF-APE-INV 2th "Name Game" Workshop- Madrid, on December 8th 2010.

Ljungberg D. (2011), *Academic inventors and firm inventiveness: A quesi-experimental analysis of firms'patents.*

Brigid O'Leary G. M., Vecchi M. (2007), *Cross-country analysis of productivity and skills at sector level,* "National Institute of Economic and Social Research, London".

Mejer M. (2010), *Academic Inventors in Belgium. Methodology and Content*, Presented at ESF-APE-INV 2nd "Name Game" Workshop- Madrid, 9-10 December 2010 .

Miguélez E. and Moreno R. (2010), *A Gravity Approach To Cross-Regional Mobility Of Inventors. Evidence From Europe*, Presented at Western Regional Science Association Annual Meeting February-March, 2011, Monterey, California.

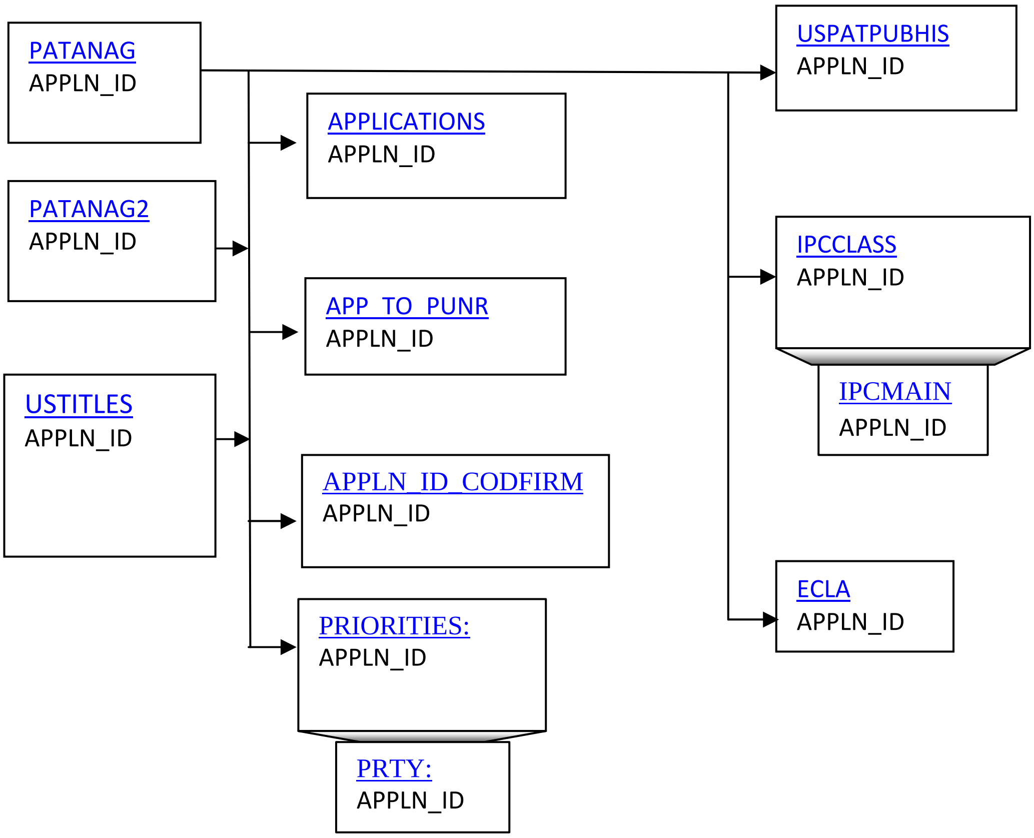
Carayol N., Sterzi V. (2013), Signaling and ownership of academic patents WIP

Zinovyeva N., Cowan R. (2010), *University Effects on Regional Innovation*, To be presented at ESF-APE-INV 3th "Name Game" Workshop- Brussels, on September 5th 2011.

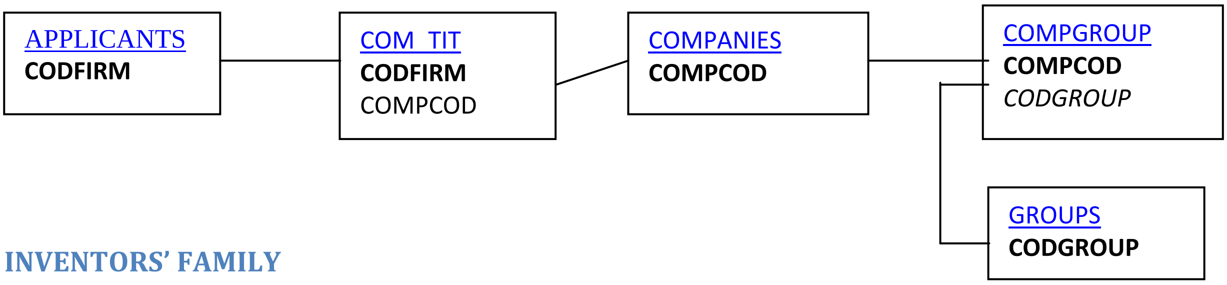
Database Description

ENTITY RELATIONSHIPS DIAGRAMS

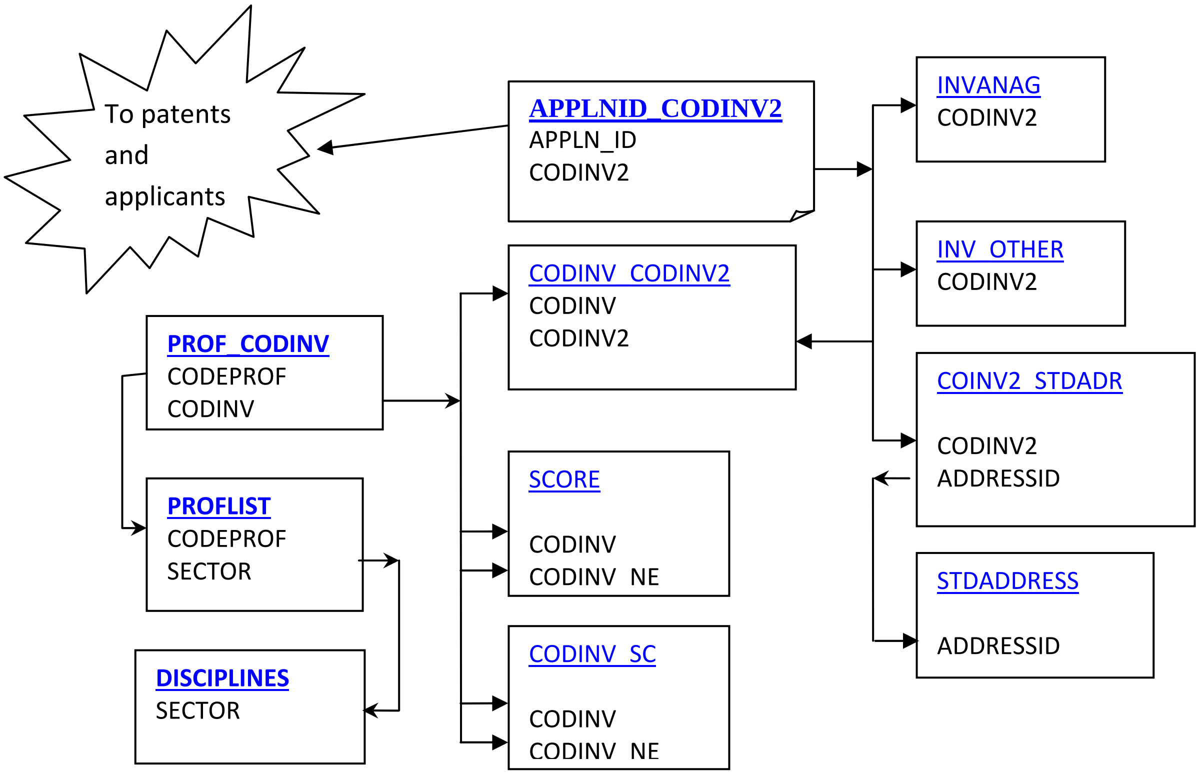
PATENT'S BASIC DATA



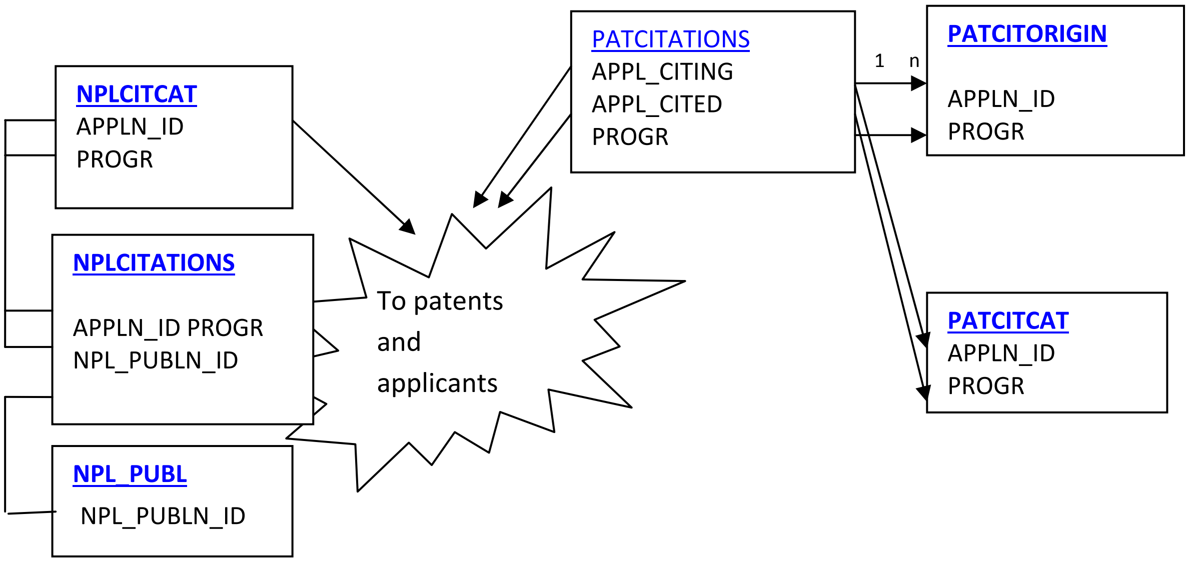
APPLICANT FAMILY



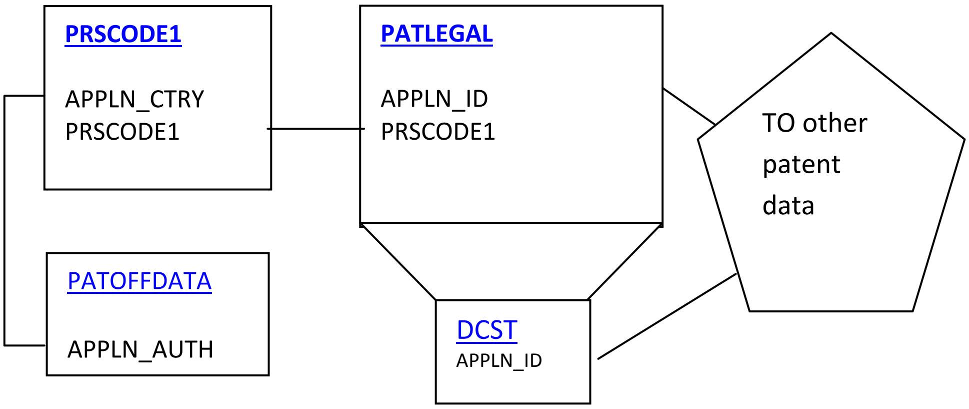
INVENTORS' FAMILY



CITATIONS FAMILY



LEGAL STATUS DATA



TABLES DETAILS

PATENT'S BASIC DATA

APP_TO_PUNR

Bridge table among application and publication numbers

APPLN_ID	9(10)	Patstat application Id
PUBLN_AUTH	$(2)	EP / wo
PUBLN_NR	$(15)	Publication number
PUBLN_KIND	$	Kind of publication

APPLICATIONS

Basic data related to applications

APPLN_AUTH	$(2)	application authority
PUBLN_AUTH	$(2)	publication authority
PUNR	9(10)	publication number
APPLN_ID	9(10)	Patstat application id (stable from 2011/04)
INPADOC_FAMILY_ID	9(10)	INPADOC family ID (depends from patstat ediction)
DOCDB_FAMILY_ID	9(10)	DOCDB family ID (depends from patstat ediction)

ECLA

The European Classification system (ECLA) is used by the EPO for carrying out patent application searches.

APPLN_ID	9(10)	Patstat application Id
EPO_CLASS_SCHEME	$(4)	EC, ICO, IDT or ECNO (see note)
EPO_CLASS_SYMBOL	$(50)	classification

Note: EPO_CLASS_SCHEME can have values:
EC - known as ECLA, the European Classification scheme
ICO- In Computer Only - an internal scheme used as the EPO for classifications which are planned for moving into ECLA at some stage

IDT - Indeling der Techniek , an old Dutch Patent Classification scheme.
ECNO - ECLA symbols which have been allocated to a document by a patent examiner who does not work for the EPO.

IPCCLASS

Patent ipc classification data and other reclassifications

APPLN_ID	9(10)	Patstat application Id
CLMN_OLD	varchar(16)	IPC class
CLMN	varchar(16)	IPC class normalized (1)
IPC_CLASS_LEVEL	char(1)	IPC class Advanced or Core indicator (A/C)
IPC_VERSION	DATE	IPC version
IPC_VALUE	char(1)	Classification value (2)
IPC_POSITION	char(1)	First or later position of symbol (3)
IPC_GENER_AUTH	char(2)	Patent office that generated the IPC classification
NCLAP	smallint(5)	Main class reclassification in 30 classes (4)
OST30	smallint(5)	Main class reclassification in 30 classes (OST/INPI) (4)
OST7	smallint(5)	Main class reclassification in 7 classes(4)
NACE	$(11)	NACE code concorded with IPC
IPC35	smallint(5)	FhG 35 class IPC concordance (see APPENDIX E)

Notes:
(1) Field CLMN is like A99B999/C99 where CLMN_OLD has no fixed format for digits after 4th.
9 could be a filler = 0; For instance: CLMN_OLD = C1B13/D1 , CLMN = C10B130/D10
(2) Indication of the value of the classification i.e. is the class symbol relating to the invention or to aspects not related to the invention (but in the application). Values: I=inventive, N=non-inventive space=unidentified
(3) Indicates the position of the IPC class in the sequence of classes that form the classification. Values: F=first, L=later. space =unidentified; For patent authorities where the law entails the concept of "first class", the first class symbol in a list of class symbols is the main class. For other authorities, like the EPO, there is no meaning in the position - classes may be quoted in alphabetical order for instance.
(4) for values of reclassifications see APPENDIX B: ipc reclassifications

IPCMAIN

Patent ipc classification where main class is indicated; ; note recently EPO has confirmed “The order of appearance of the classes as apparent from IPC_POSITION has a special meaning for some generating offices such as the USPTO but has no special meaning for others offices such as the EPO.” So such table may lead to misguiding results.

APPLN_ID	9(10)	Patstat application Id
CLMN	$(14)	main int class normalized as below
CLMN _OLD	$(14)	Main int class not normalized
IPCV	$(2)	IPC Version (NOT FILLED IN PATSTAT)
NCLAP30	9(2)	Main class reclassification in 30 classes
OST30	9(2)	Main class reclassification in 30 classes (OST/INPI) see (1)
OST7	9(2)	Main class reclassification in 7 classes
NACE	$(11)	NACE code concorded with IPC

IPC35	smallint(5)	FhG 35 class IPC concordance (sse appendix E)

Note: field CLMN is like A99B999/C99

Where 9 could be a filler = 0; Empty fields contain “NON RIL/”;
The field CLMN_OLD will be removed from further db version
For instance: CLMN_OLD = C1B13/D1 , CLMN = C10B130/D10
(1) for values of reclassifications see APPENDIX B: ipc reclassifications
NOTE: 84000 records (@11/2013) applications had no main class but an ipcclass records

PATANAG

Not-repeated data about patent

APPLN_ID	9(10)	Patstat application Id
APNR	$(15)	Application #
PINR	9(10)	International pubbln number (field no more maintained)
AKIND	$(2)	Kind of application (note 3)
APDT	DATE	date of filing
AIDT	DATE	International date of filing (no more maintained)
IAPNR	9	PCT corr. appl id (PATSTAT)
STATUS	$(2)	if A* or blank: applied, if B* granted
CLAIMS	*9*	*Number of claims (data available from 201110 patstat)*
TRIADIC	9	0-not triadic; 1 = inpadoc tr; 2 = docdb tr;. 3 = both

Note: field AIDT is date of WIPO filing;
Field PINR, where not null, is filled with WIPO PUNR.
Note2: links among PRDT, APDT, AIDT
If PRDT is empty there are 2 cases:
if PINR is empty the first filing was in EPO
If PINR contains a value, then AIDT is the priority since contains the date of filing in WIPO.
Note 3: W=PCT application, U=utility model, A=patent D,K,L,M,N=dummy for de-duplicating
Note about triadic patents: Triadic patents are defined here as application belonging to an inpadoc/docdb extended family with at least one EPO application, one JPO application and one USPTO ***grant***. [OECD definition of triadic patents]

PATANAG2

APPLN_ID	9(10)	Patstat application Id
PUBDT	DATE	Publication date
PUBKIND	$(2)	Kind of Publication (1)
PUBLG	$(2)	Publication language
FIRSTGRANT	9(4)	if 1 the date of publication is date of first publication of grant
CLAIMS	9(6)	Number of claims at latest status

NOTE 1: Kind of Publication consists of a letter (typically A or B) followed by a number. A kind codes (e.g., A1, A2) are used for patent applications; B kind codes (e.g., B1, B2) are used for issued patents.

PRIORITIES:

Patent applied and published before application of EP patent

APPLN_ID	9(8)	Patstat application Id
PROGR	9(4)	Progressive number
PRDT	DATE	Priority date
PR_PUBL_AUTH	$(2)	Patent office of published priority
PR_PUNR	$(15)	patent number of publ. priority
PR_APPL_AUTH	$(2)	patent office of applied priority
AP_APNR	$(15)	application number of priority document

NOTE: different priorities publication numbers with same priorities application number result to have same PROGR

PRTY:

Patent applied and published before application of EP patent: compressed table = contains only first date of priority or, if the patents have no priority, application filing date is chosen.

APPLN_ID	9(8)	Patstat application Id
PRDT	DATE	Priority date
KIND	$(1)	Origin of PDT: P = priority A= application filing date

PATPUBHIS

Patent publication history

APPLN_ID	9(8)	Patstat application Id
PUBLN_AUTH	char(2)	publication authority (pat office) of patent
PUNR	varchar(15)	publication number of patent
PUBLN_KIND	char(2)	Kind code (1)
PUBLN_DATE	date	Publication date
CLAIMS	9(4)	Number of claims at given publication state

(1) KIND CODE MEANING (for EP only)
A1 European Patent Application (with search report)
A2 European Patent Application (without search report)
A3 European Patent Application (search report for A2)
B1 European Patent
B2 European Revised Patent

APPLN_ID_ CODFIRM

APPLN_ID	9(8)	Patstat application Id
PROGR	9(4)	Progressive number

CODFIRM	9(8)	Applicant’s progressive number

TITLES

Contains descriptions of patent: title and abstract

APPLN_ID	9(8)	Patstat application Id
TITLE	$(3000)	Patent title
ABSTRACT	$(10000)	Patent abstract

APPLICANTS’ FAMILY

APPLICANTS

Anagraphic data about applicants: societies and individuals

CODFIRM	9(8)	Applicant’s progressive number
TITCY	$(2)	Applicant’s nation
TITNM	$(255)	Applicant’s name
TITTRDNM	$(255)	Applicant’s trade name
TITKIND	$(20)	Kind of society (AG, A/S, BV, SA…)
TITSTR	$(255)	Address
TITSTRALTRO	$(255)	Address other (PO box, zone industrielle..)
TITCIT	$(255)	Town
TITZONE	$(75)	Zone: lesser level of aggregation (county)
TITZONE2	$(50)	Zone: intermediate level of aggregation (region)
TITZONE3	$(50)	Zone: higher level of aggregation (nation in federations)
ZIP_CODE	$(10)	ZIP CODE - string
INDUM	$(1)	Flag: I = individual; C= Society
NUTS3	$(10)	NUTS code level 3 source OECD-REGPAT

Note: the field Indum is filled following below criteria:
Are counted as individuals those records where TITNM (name) is like
NAME, SECONDNAME [...] (seek string “, “) and where acronyms like those listed in table 1 are not present.

COM_TIT

Bridge among companies and applicants

CODFIRM	9(8)	Applicant’s progressive number
COMPCOD	9(8)	Progressive Company number
COMPCODHIS	9(8)	Previous company of applicant (if changed)

COMPGROUP

Crosstable between Companies and groups (m to n)

COMPCOD	9(8)	Progressive company number
CODGROUP	9(8)	Last group
GRPKIND	$(2)	Kind of grouping: JV = joint venture FU = Fusion SO = Spinoff
CODGROUPHIS	9(8)	Former group (reference: **Grouphistory**)
CFID	9(8)	Table key field

COMPANIES

Companies applicants unique by nation

COMPCOD	9(8)	Progressive company number
TITNM	$(255)	Company’s name
TITCY	$(2)	Company’s country
DATEFROM	9(4)	Starting date
DATETO	9(4)	Closing date
FIRSTPATYR	9(4)	First patent’s year
LASTPATYR	9(4)	Last patent’s year
COMPDUN	$(10)	company’s dun number
DOMULTDUN	$(10)	Duns number of ultimate domestic parent
DOMULTNAM	$(100)	Name of ultimate domestic parent
COMPTYPE	$(1)	Area of the company: I=Enterprise U=UNIVERSITY A=Pubblic reasearch center B=Private res. center S=Foundations, NGO C=Consortium X=OTHER
COMPTYPE2	$(1)	other data on Company kind: J = joint venture, F = Corporate Spinoff, D = division of long established company, W = subsidiary of foreign company, Y = individual company
CFLAG	$(1)	how data have been filled (see note)
NOTE	$(500)	internet site, other….
EEPPAT_NAME	$(400)	standardized name from EE_PPAT database
EEPPAT_SECTOR	$(45)	sector of activity assigned from EE_PPAT database
ALIVE	9(1)	if 1 company is still alive in database

Notes about CFLAG

A= dunscodes mated searching on WOW(R) 2001; data for domestic and global have been given by D&B data 6/2003, aside from 145 cases where a search on data has been performed and 151 cases (that can be distinguished from the lack of a domestic parent) searched on WOW 2001 cd;

B = mated using DFpowerstudio and handcheck; C= duans mated via name match with D&B data where we have no domestic and global parent; D= handcheck Breschi/Tarasconi; F= Montobbio data added aug. 2004;

GROUPS

Groups of companies

CODGROUP	9(8)	progressive number for group
GROUPNAME	$(50)	Name of the group
DATEFROM	9(4)	Date activity’s start
DATETO	9(4)	Date activity’s end
GROUPCY	$(2)	Group’s Country
GROUPDUN	$(10)	Group’s Dunsnumber
GRNOTE	$(255)	Notes about group
ALIVE	0/1	1 means group is still in activity

INVENTORS' FAMILY

APPLNID_CODINV2

Bridge table from applications data to inventors data

APPLN_ID	9(8)	Patstat application Id
PROGR	9(3)	progressive number
CODINV2	9(8)	Inventor code unique for address / name

CODINV CODINV2

Cross table from individual to location of the inventor

CODINV	9(8)	Inventor code; unique for name
CODINV2	9(8)	Inventor code; unique for address / name
ORIGIN	Boolean	if 2^0 on then EPO if 2^1 on then uspo

Note: the difference between CODINV and CODINV2 is given by research discovering that the same name in different addresses is the same person; in such case at the same CODINV could correspond more CODINV2.

CODINV_SC

Cross table from individual to location of the inventor, with euristic criteria

CODINV	9(8)	Inventor code; unique for name
CODINV_NE	9(8)	Inventor code; unique for address / name

Following criteria have been applied:

= name AND (= IPC 12 OR 3 degrees of distance OR citing OR same company OR same applicant OR same address OR same IPC6 and same nation OR coinventor in common)

DISCIPLINES

Linkage table for descriptions of professors' sector

SECTOR	$(9)	code of teaching
DESCRIPTION	$(37)	long description

INVANAG

Main data about inventors

CODINV2	9(10)	Inventor code unique for address / name
INCY	$(2)	Country of inventor
INNAME	$(150)	Inventor name
INADDR	$(200)	Address
INADOTH	$(100)	Address other (PO box, zone industrielle..)
INCITY	$(100)	Town

INCOUNTY	$(75)	Zone: lesser level of aggregation (county)
INREGION	$(130)	Zone: intermediate level of aggregation (region)
INSTATE	$(50)	Zone: higher level of aggregation (nation in federations)
INZIP	$(20)	Zip code
NUTS3	$(10)	NUTS code level 3 source OECD-REGPAT

INV_OTHER

Other data about inventors

CODINV2	9(10)	Inventor code unique for address / name
INNM1	$(75)	Inventor first name (surname)
INNM2	$(75)	Inventor second name
INNM3	$(75)	Inventor third name
INNMEXT	$(10)	Inventor name extension (Jr., Sr., III)
INTITLE	$(25)	Academic titles if any
INBYWHO	$(75)	If c/o any labs or society
INLIVE	$(2)	If “X” inventor is deceased

PROF_CODINV

Translation table for innovators who also are professors

CODEPROF	$(15)	Professor unique id
CODINV	9(10)	Inventor code; unique for name

NOTE: Codeprof = TOBESEARCHED means innovator identified as a professors in ***DEALS*** research without a correspondent in prof_list table

PROFLIST

CODEPROF	$(29)	Professor's code AA999999_DDDD where AA = status ; 9999 = progressive; DDD = teaching sector ; for UK is a numeric code
UNI_CITY	$(41)	location of university
UNI_NAME	$(25)	field used to distinguish between university with similar names
UNI_PROV	$(11)	County of university
QUALIFIC	$(12)	RU = researcher; PA = associated professor; PO = ordinary prof
SURNAME	$(36)	surname
NAME	$(27)	name
SECTOR	$(10)	code of scientific sector of teaching (see table DISCIPL for descriptions)
DOB	DATE	date of birth
NOME_IN	$(55)	Name of inventor
ACCENT	9	positions where accent is in the name string
COGN_ACC	$(36)	professor surname without accent
UOA	$(4)	Unit of Assessment - affiliation Number (UK)
UNIVCODE	$(12)	code of university (UK)

SCORE

Elements and weight that concur to give a similarity among 2 codinv with the same name.

CODINV	9(8)	Inventor code unique for individual
CODINV_NE	9(8)	Inventor code unique for individual same name of CODINV
SCORE	9(2)	Score for similarity due to reason
REASON	$(15)	Reason of similarity

More details on reasons, score and methodology in: Lissoni, Francesco, Gianluca Tarasconi & Bulat Sanditov, 2006, The KEINS Database on Academic Inventors: Methodology and Contents, CESPRI Working Paper 181, Universita' Bocconi

STDADDRESS

Inventors addresses standardized via google API

ADDRESSID	9(7)	Progressive id
ADDRESS	$(150)	street address
CITY	$(75)	city
COUNTY	$(75)	county
REGION	$(75)	region / state
ZIPCODE	$(12)	zip code
CTRY	$(2)	country code iso 3166
XCOORD	9	latitude
YCOORD	9	longitude
GPRECISION	9(2)	precision of google result

CODINV2_STDADR

Bridge table among condinv2 and standard addresses

CODINV2	9	Inventor code unique for address / name
ADDRESSID	9	progressive id for standard addresses

CITATIONS FAMILY

NPLCITATIONS

Non patent literature reference

APPLN_ID	integer	application id of citing patent
PROGR	smallint(4)	progressive number of citation
NPL_PUBLN_ID	integer	id for NPL
CITN_ORIGIN	varchar(5)	Origin of the citation (see below)
EE_CITING	integer	EP equivalent (if exists) for citing

NOTE: due to duplications in table NPLPUBL and double publications this table has 16280 duplicates by citing_auth, punr, progr

NPLCITCAT

Citation categories for patent-NPL citations

APPLN_ID	integer	application id of citing patent
PROGR	smallint(4)	progressive number of citation
EE_CITING	integer	EP equivalent (if exists) for citing
CITN_CATEG	char(1)	Category of the citation as mentioned in Search Reports

Values allowed for CITN_CATEG: X, Y, A, D, E, P, L, T, O (see appendix A)

NPL_PUBLN

Publications cited by patents

NPL_PUBLN_ID	integer	id for NPL
NPL_BIBLIO	$(3000)	unparsed publication cited

PATCITATIONS

Patent literature citations

APPLN_CITING	bigint(15	application id of citing patent
APPLN_CITED	bigint(15)	application id of cited patent
PROGR	smallint(4)	progressive number of citation
EE_CITING	bigint(15)	application id of EPO equivalent of citing patent
EE_CITED	bigint(15)	application id of EPO equivalent of cited patent

PATCITCAT

Citation categories for patent-patent citations

APPLN_ID	integer	application id of citing patent
PROGR	smallint(4)	progressive number of citation
CITN_CATEG	char(1)	Category of the citation as mentioned in Search Reports

Values allowed for CITN_CATEG: X,Y,A,D,E,P,L,T,O (see appendix A)

PATCITORIGIN

Cathegory of citations given, indicating the moment of examination n process the citation has been introduced.

APPLN_ID	integer	application id of citing patent
PROGR	smallint(4)	progressive number of citation
CITN_ORIGIN	varchar(5)	Origin of the citation (1)

NOTES:

(1) 0 - SEA- citations introduced during search
1 - APP- citations introduced by the applicant
2 - EXA- citations introduced during examination
3 - OPP - citations introduced during opposition
4 - 115- citations introduced according to Art 115 EPC
5 - ISR -citations from the International Search Report
6 - SUP - citations from the Supplementary Search Report
7 - CH2 - citations introduced during the Chapter 2 phase of the PCT

LEGAL STATUS PATENT DATA

Source of this data is inpadoc where not otherwise indicated

DCST

List of designated contracting states (part of PATLEGAL table)

APPLN_ID	9	Application id
Prs_event_seq_nr	Int(5)	Progr event number ny punr
Prs_gazzette_ date	Date	Date of occurrence
Prs_code	$(4)	Type of legal event
Progr	9	progressive for same punr/event
Cy	$(2)	iso code for contacting state

NOTE: PRS_CODE may have value AK = DESIGNATED CONTRACTING STATES, or AX = EXTENSION OF THE EUROPEAN PATENT TO

PRSCODE1

List of legal status descriptions by application authority

APPLN_CTRY	$(2)	application authority (pat office)
PRSCODE1	$(12)	code for status
FLAG1	$(3)	various meanings

FLAG2	$(1)	+ stands for increasing, - for decreasing (fi appl cys)
DESCRIPTION	$(200)	English description of the status
CATHEGORY	$(50)	macrocathegory of the code

PATLEGAL

List of legal status events from inpadoc DB

field name	type	Description	Corresp. TAG
APPLN_ID	9	Application ID	
Prs_event_seq_nr	Int(5)	Progr event number ny punr	
Prs_gazzette_date	Date	Date of occurrence	
Prs_code	$(4)	Type of legal event	
PRSREFCY	$2	Corresponding country code for PRS code •EP REG••	L501EP
PRSEPCOD	$4	Corresponding EP code 1 for PRS code •EP REG••	L502EP
PATCORR	20	Corresponding patent document	L503EP
PATCORRCY	$2	Country code of corresponding patent document	L504EP
PATCORRPD	DATE8	Publication date of corresponding patent	L505EP
PATCORRKD	$2	Kind of corresponding patent document	L506EP
DCSTLIST	$300	List of designated states	L507EP
EXTCY	$2	Extension state	L508EP
NEWOWNER	$255	New owner name or address if name or address of owner changes; addresses are NOT stored in this tag	L509EP
NOTES	700	Free format text	L510EP
SPCNUMBER	$20	SPC number	L511EP
FILINGDT	DATE8	Filing date	L512EP
EXPIRYDT	8	Expiry date	L513EP
INVNAMES	$255	Inventor name (separated by ;)	L515EP
IPCS	$50	International Patent Classification (comma separated)	L516EP
REPRNM	$255	Representative's name(s)	L517EP
PAYDATE	DATE8	Payment date	L518EP
OPPNAME	$50	Opponent name(s)	L519EP
FEEPAYYR	10	Year of fee payment - contains the xxth year for which the payment was made	L520EP
NEWIPRNR	$30	New kind of IPR, new number; e.g. Brazil utility model - code GA;"MI4601602-3"	L521EP

REQNAME	$50	Name of requester	L522EP
EXTDATE	DATE8	Extension date	L523EP
CTRYLIST2	$100	List of countries concerned with an event L507EP & L508EP have special significance.	L524EP
EFFECTDT	DATE8	Effective date; DATE_IN_FORCE	L525EP
WITHDRDT	DATE8	Date of withdrawal	L526EP
FPFLAG2	$1	Indicator for format of attribute list document number following rules for either (F)iling applications or (P)ublications. If not known, this tag will not be present; refers to the document given in L503EP and L504EP	L527EP

PATOFFDATA

Summary data regarding patent office: source may vary and is different from patstat

APPLN_AUTH	$(2)	Application authority / patent office
YEAR	9(4)	year of reference
PI	9	patent of invention applied
UM	9	utility models applied
TOT	9	PI+UM (where not distinguished is the only data available)
PCT	9	applications to PCT
EP	9	applications to EPO
GRANTED	9	patents granted in the year
EXAM	9	average number of examiners
SOURCE	$(20)	source of data (1)
NOTE	$(255)	notes

(1) Source: PATOFF = email / direct contact with patent office; REPORT: from patent office reports.

Note: date collected for years 1988 1998 2005 and 2006 for 40 patent offices

Access Rules

Data described are accessible to researchers and no profit institutions at three different level, depending from data origin contributors and other issues described in detail below.

FULL ACCESS:

Summary data elaborated from our dataset are freely available at URL:

http://ricercaweb.unibocconi.it/criospatstatdb/

where it is possible to select and download the following data in CSV format:

Patent count by inventor country / year

Patent count by applicant country / year

Patent count by inventor region / year

Patent count by applicant region / year

Patent count by inventor nuts3 / year

Patent count by applicant nuts3 / year

Patent count by applicant name / year

Patent count by main IPC - first 4 digits

Patent count by main IPC class reclassified on OST30

Patent count by applicant, year, OST30

Patent count by applicant country, county, region, OST30, year

Citations count by applicant name, year

Citations count by applicant country, year

Citations count by inventor country, year

Copatenting by inventor country, year

Copatenting by applicant country, year

Applicants by IPC - first 4 digits

Inventors by IPC - first 4 digits

CERTIFIED USERS:

The full set of tables is available on CRIOS ftp server in SAS format. All Crios fellows are authorized to download them; external users should be introduced and authorized by a person of Crios. Since data are an elaboration of PATSTAT users institution should also aquire from EPO a license of Patstat.

All the set of tables are available, except those listed in the following list of data for SPECIAL AUTHORIZATIONS.

In a further period we will increase the amount of data for free download through web facilities including part of the data actually are available only for certified users.

SPECIAL AUTHORIZATIONS:

Data that are part of the PROFLIST family need a special authorization for download and usage since they have been built jointly with other institutions and also contain data which are sensitive from privacy point of view. Please contact the authors for further information.

DATA APPENDIX A: values for citation category[27]

X - particularly relevant if taken alone
Y - particularly relevant if combined with another document of the same category
A - technological background
O - non-written disclosure
P - intermediate document
T - theory or principle underlying the invention
E - earlier patent document, but published on, or after the filing date
D - document cited in the application
L - document cited for other reasons

More on page 235 of "Guidelines for Examination in the European Patent Office (status April 2010)" http://documents.epo.org/projects/babylon/eponet.nsf/0/7ffc755ad943703dc12576f00054cacc/$FILE/guidelines_2010_complete_en.pdf

Categories of documents (X, Y, P, A, D, etc.)
All documents cited in the search report are identified by placing a particular letter in the first column of the citation sheets. Where needed, combinations of different categories are possible. The following letters are used:

(i) particularly relevant documents
Where a document cited in the European search report is particularly relevant, it should be indicated by the letter "X" or "Y". Category "X" is applicable where a document is such that **when taken alone**, a claimed invention cannot be considered novel or cannot be considered to involve an inventive step. Category "Y" is applicable where a document is such that a claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other documents of the same category, such combination being obvious to a person skilled in the art. However, if a document (a so-called "primary document") explicitly refers to another document as providing more detailed information on certain features (see C-IV, 7.1) and the combination of these documents is considered particularly relevant, the primary document should be indicated by the letter "X", i.e. not "Y", and the document referred to should be indicated as "X" or "L" as appropriate;

(ii) documents defining the state of the art and not prejudicing novelty or inventive step
Where a document cited in the European search report represents state of the art not prejudicial to the novelty or inventive step of the claimed invention, it should be indicated by the letter "A" (see, however, III, 1.1);

(iii) documents which refer to a non-written disclosure
Where a document cited in the search report refers to a non-written disclosure, the letter "O" should be entered (see VI, 2). Examples of such disclosures include conference proceedings. In cases where the oral disclosure took place at an officially recognised exhibition (Art. 55(1)(b)), see VI, 5.5. The document category "O" is always accompanied by a symbol indicating the relevance of the document according to (i) or (ii), for example: "O, X"; "O, Y"; or "O, A";

(iv) intermediate documents
Documents published on dates falling between the date of filing of the application being examined and the date of priority claimed, or the earliest priority if there is more than one (see VI, 5.2 and XII, 4), should be denoted by the letter "P". The letter "P" should also be given to a document published on the very day of the earliest date of priority of the patent application under consideration. The document category "P" is always accompanied by a symbol indicating the relevance of the document according to (i) or (ii), for example: "P, X"; "P, Y"; or "P, A";

(v) documents relating to the theory or principle underlying the invention
Where a document cited in the search report may be useful for a better understanding of the principle or theory underlying the invention, or is cited to show that the reasoning or the facts underlying the invention are incorrect, it should be indicated by the letter "T";

(vi) potentially conflicting patent documents

[27] More information of the citations categories see EPO data catalog: http://documents.epo.org/projects/babylon/eponet.nsf/0/830d207d355f3af2c1257aa1002d0cfb/$FILE/data_catalog_v5.00_en.pdf

Any patent document bearing a filing or priority date earlier than the filing date of the application searched (not the priority date – see VI, 3 and XII, 4) but published later than that date and the content of which would constitute prior art relevant to novelty (Art. 54(1)) should be indicated by the letter "E". Where the patent document and the application searched have the same date (see C-IV, 6.4), the patent document should also be identified by the letter "E". An exception is made for patent documents based on the claimed priority under consideration; these documents should not be cited;

(vii) documents cited in the application

When the search report cites documents already mentioned in the description of the patent application for which the search is carried out, these should be denoted by the letter "D" (see IV, 1.3); (viii) documents cited for other reasons Where in the search report any document is cited for reasons (in particular as evidence – see XII, 5) other than those referred to in the foregoing paragraphs, for example:

(a) a document which may throw doubt on a priority claim (see VI, 5.3);

(b) a document which establishes the publication date of another citation (see XII, 5); or

(c) a document relevant to the issue of double patenting (see IV, 2.3(v), and C-IV, 6.4), such document should be indicated by the letter "L". Brief reasons for citing the document should be given. The citation of documents of this type need not be linked to any of the claims. However, where the evidence which they provide relates only to certain claims (for example the "L" document cited in the search report may invalidate the priority claim in respect of certain claims only), then the citation of the document should be linked to those claims, in the manner indicated in X, 9.3.

DATA APPENDIX B: IPC RECLASSIFICATIONS

Source: Schmoch et al. 2003

OST30-code	OST30-name	OST7-code	OST7-name
1	Electrical engineering	1	Electrical engineering; Electronics
2	Audiovisual technology	1	Electrical engineering; Electronics
3	Telecommunications	1	Electrical engineering; Electronics
4	Information technology	1	Electrical engineering; Electronics
5	Semiconductors	1	Electrical engineering; Electronics
6	Optics	2	Instruments
7	Technologies for Control/Measures/Analysis	2	Instruments
8	Medical engineering	2	Instruments
9	Nuclear technology	2	Instruments
10	Organic chemistry	3	Chemicals; Materials
11	Macromolecular chemistry	3	Chemicals; Materials
12	Basic chemistry	3	Chemicals; Materials
13	Surface technology	3	Chemicals; Materials
14	Materials; Metallurgy	3	Chemicals; Materials
15	Biotechnologies	4	Pharmaceuticals; Biotechnology
16	Pharmaceuticals; Cosmetics	4	Pharmaceuticals; Biotechnology
17	Agricultural and food products	4	Pharmaceuticals; Biotechnology
18	Technical processes (chemical, physical, mechanical)	5	Industrial processes
19	Handling; Printing	5	Industrial processes
20	Materials processing, textile, glass, paper	5	Industrial processes
21	Environmental technologies	5	Industrial processes
22	Agricultural and food apparatuses	5	Industrial processes
23	Machine tools	6	Mechanical eng.; Machines; Transport
24	Engines; Pumps; Turbines	6	Mechanical eng.; Machines; Transport
25	Thermal processes	6	Mechanical eng.; Machines; Transport
26	Mechanical elements	6	Mechanical eng.; Machines; Transport
27	Transport technology	6	Mechanical eng.; Machines; Transport
28	Space technology; Weapons	6	Mechanical eng.; Machines; Transport
29	Consumer goods	7	Consumer goods; Civil engineering
30	Civil engineering	7	Consumer goods; Civil engineering

DATA APPENDIX C: Example for patcitations

	citing_auth	punr	progr	cited_auth	punr_cited	EE_citing	EE_CITED
1	DE	10334869	1	AT	4639	1503182	1202025
1	US	6894487	1	AT	4639	1503182	1202025
2	DE	3608665	1	AT	4690	0	510024
3	WO	9519640	1	AT	4694	739536	739536
4	EP	1448803	0	AT	4810	1448803	1263067
5	WO	2004015156	1	AT	4810	1536031	1263067

Examples above from patcitations table gives some cases like:

1: DE and US patents with same EP equivalent citing AT pat with EP equivalent;
In such case in patcitations_ep only 1 record of ep-ep will remain; bytheway the count of citations must be done in a careful way cause DE and US are equivalents

2: DE citing AT with an ep equivalent
Such case in patcitations_ep will be dropped

3 and 5:
In patcitations_ep only ep-ep will remain

4 EP vs AT with an ep equivalent

When sobstituting equivalences some problems arised:

1) ep citing other ep with higher punr
~65000 cases mainly due to patastat/epo error; suche cases in the EP-EP db have been deleted
see example:
http://v3.espacenet.com/searchResults?locale=en_GB&PN=de1042080&compact=false&DB=EPODOC
where DE1042080 exists twice (one for 1958 and one for 2001 and the latter cites ep1042080!!!!!!)

2) EP citing himself
such cases (~15000) are due to extension to other offices (mostly US) where the priority had to be cited for completeness

they have been deleted in EP-EP

DATA APPENDIX D: deeper into data definitions

PRIORITY NUMBERS SPECIFICATIONS
The number assigned by the Patent Office when a patent application is submitted. It is not the number that is assigned to a patent itself.
Usually the application number is mada up of a a country code (two letters) and 11-digit number where the first 4 digits indicate the filing year of the application.
§It may differ for WO where it contains also a country code.

Examples:

EP20060760335
WO2006US20070
US20050165972

PUBLICATION NUMBER SPECIFICATIONS
The publication number is the number assigned to a patent application on publication. The publication number is made up of a country code (two letters) and a serial number (variable, one to ten digits) (eg US4325348).
Sometimes the number is followed by a kind code corresponding to a specific stage in the procedure, that should be omitted.

DATA APPENDIX E: IPC35 description[28]

IPCCLASS	Description	macroclass
1	Electrical machinery, apparatus, energy	Micro-structural and nano-technology
2	Audio-visual technology	Micro-structural and nano-technology
3	Telecommunications	Micro-structural and nano-technology
4	Digital communication	Micro-structural and nano-technology
5	Basic communication processes	Micro-structural and nano-technology
6	Computer technology	Micro-structural and nano-technology
7	IT methods for management	Micro-structural and nano-technology
8	Semiconductors	Micro-structural and nano-technology
9	Optics	Instruments
10	Measurement	Instruments
11	Analysis of biological materials	Instruments
12	Control	Instruments
13	Medical technology	Instruments
14	Organic fine chemistry	Chemistry
15	Biotechnology	Chemistry
16	Pharmaceuticals	Chemistry
17	Macromolecular chemistry, polymers	Chemistry
18	Food chemistry	Chemistry
19	Basic materials chemistry	Chemistry
20	Materials, metallurgy	Chemistry
21	Surface technology, coating	Chemistry
22	Micro-structural and nano-technology	Chemistry
23	Chemical engineering	Chemistry
24	Environmental technology	Chemistry
25	Handling	Mechanical engineering
26	Machine tools	Mechanical engineering
27	Engines , pumps, turbines	Mechanical engineering
28	Textile and paper machines	Mechanical engineering
29	Other special machines	Mechanical engineering
30	Thermal processes and apparatus	Mechanical engineering
31	Mechanical elements	Mechanical engineering
32	Transport	Mechanical engineering
33	Furniture, games	Other fields
34	Other consumer goods	Other fields
35	Civil engineering	Other fields

NOTE: CLASS 0 may mean unclassified; -1 stands for removed due to other ipcs in the applications (FI C12M% with A61K%) (-1 value in IPCMAIN means main ipc cannot give IPC35 reclassification)

[28] http://www.wipo.int/ipstats/en/statistics/technology_concordance.html

Bibliography

Information of CRIOS DB:

http://db.Crios.unibocconi.it/ .

Gesellschaft F. (2009), Technology-Oriented Classification, Institut National de la Propriété Industrielle (INPI, Paris) and Observatoire des Sciences and des Techniques (OST, Paris).

Information on IPC:

http://www.epo.org/law-practice/case-law-appeals/business-distribution/technical.html

http://rawpatentdata.blogspot.com/2010/04/ipc-core-vs-advanced-level-in-patstat.html

http://rawpatentdata.blogspot.com/search/label/IPC

IPC References from EPO:

http://www.epo.org/law-practice/case-law-appeals/business-distribution/technical.html .

IPC WIPO Tool:

http://www.wipo.int/classifications/ipc/en

http://www.wipo.int/treaties/en/classification/strasbourg/index.html

http://www.wipo.int/classifications/ipc/en.

Lissoni F., Tarasconi G., Sanditov B. (2006), *The KEINS Database on Academic Inventors: Methodology and Contents*, "CESPRI Working Paper 181, Università Bocconi".

ODEC Patent Statistics Manual (2009),

http://www.oecd.org/document/29/0,3343,en_2649_34451_42168029_1_1_1_1,00.html.

Patstat data catalog (EPO) version 2009 'EPO Worldwide Patent Statistical Database'.

Catalina Martinez, 2010. "Insight into Different Types of Patent Families," OECD Science, Technology and Industry Working Papers 2010/2, OECD Publishing

Du Plessis, M., Van Looy, B., Song, X & Magerman, T. (2009) Data Production Methods for Harmonized Patent Indicators: Assignee sector allocation. EUROSTAT Working Paper and Studies, Luxembourg

Frazzoni, S., Mancusi, M., Rotondi, Z., Sobrero, M. A. U. R. I. Z. I. O., & Vezzulli, A. (2011). Relationships with banks and access to credit for innovation and internationalization of SMEs. *L'Europa e oltre-Banche e imprese nella nuova globalizzazione*.

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. Journal of Economic Literature, 28(4), 1661-1707.

Melamed, R. , Shiff, G. and Trajtenberg, M. (2006), 'Names Game': Harnessing Inventors Patent Data for Economic Research. CEPR Discussion Paper No. 5833

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent statistics as an innovation indicator. Handbook of the Economics of Innovation, 2, 1083-1127.

Peeters B., Song X., Callaert J., Grouwels J., Van Looy B. (2009). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. EUROSTAT working paper and Studies, Luxembourg

Raffo, J., & Lhuillery, S. (2009). How to play the "Names Game": Patent retrieval comparing different heuristics. Research Policy, 38(10), 1617-1627.

Sterzi, V. (2013) "Patent quality and ownership: An analysis of UK faculty patenting", Research Policy 42(2), 564-576.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, 1.

Thoma G. & Torrisi S., (2007). Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases, KITeS Working Papers 211, KITeS, Centre for Knowledge, Internationalization and Technology Studies, Universita' Bocconi, Milano, Italy, revised Dec 2007

Pezzoni M. Lissoni F. and Tarasconi G. How To Kill Inventors: Testing The Massacrator Algorithm For Inventor Disambiguation. Cahiers du GREThA, 2012