

Methodological Annex – Chapter 5

Patenting activity in Brazil and abroad

1. Data furnished by IPEA and INPI

The information base used for Chapter 5 was a compilation of two different databases. The first was furnished by Instituto de Pesquisas Econômicas Aplicadas (IPEA), which forwarded a plain-text (“.txt”) file supplied by Instituto Nacional da Propriedade Industrial (INPI), Brazil’s patent office, with information on patent applications filed until 2003. This file contained information such as patent application numbers and filing dates, first-named and other inventors, international classification (WIPO),¹ patentee taxpayer numbers (CNPJ for corporate entities, CPF for individuals), and patentee countries of origin.

The information in this first database presented a number of problems. First, some patent applications were dated 1900, evidently owing to form completion errors or data entry errors. The most recent applications were dated 2000, marking the end of the period covered by the information contained in the file, which referred to a total of 299,749 patent applications.

Data treatment was carried out, and the types of grant applied for were identified on the basis of the application numbers. The types available are: Invention Patent (PI); Utility Model (MU); Industrial Design (DI); and Certificate of Addition (C). The types used for the analysis in this chapter were PI and MU, which amounted to 252,515, or 84.2% of the total filed in the period. INPI’s website explains the different types as follows: “Based on the differences between inventions, grants can be classified as one of the following types: Invention Privilege or Invention Patent (PI), which must meet the requirements of novelty, inventive activity and industrial application; Utility Model (MU), involving a new shape or arrangement and an inventive act that results in a functional improvement to an object. There is also the Certificate of Addition to an Invention, used to protect an improvement to an invention for which a claim has been filed or a patent already granted. An industrialisable creation relating to

the plastic ornamental shape of an object or set of lines and colours, applicable to a product to produce a new and original visual result in its external appearance and serving as a type of fabrication, requires an application for registration as an Industrial Design and is not considered patentable.”

Given these definitions, the decision to work with PIs and MUs reflected the need to deal only with claims resulting from more significant innovative activities. Industrial Designs were excluded because they relate only to form and are not considered patents by INPI. Certificates of Addition were excluded because their inclusion would have required considering any alteration to an existing claim or patent as a significant innovation regardless of its content.

Information regarding the applicant’s country of origin enabled claims to be grouped into resident and non-resident filings. The number of INPI filings for PIs and MUs in which Brazilian residents were first-named inventors (both corporate entities and individuals) amounted to 91,372, or 35.2% out of a total of 252,515, with non-resident claims amounting to 161,143 or 63.8%.

Taxpayer numbers (CNPJ for corporate entities, CPF for individuals) were available for 77,875 or 85.2% of the claims filed by residents. Of these, 55,032 (70.7%) had a structure compatible with a CPF number (nine digits plus two check digits), characterizing the first-named inventor as an individual, while 22,843 (29.3%) had a CNPJ structure (12 digits plus two check digits), characterizing the first-named inventor as a corporate entity. There remained 13,497 resident claims (14.8%) containing insufficient information to classify the first-named inventor as a corporate entity or individual.

The problem of missing or inaccurate information was even more serious in respect of the postal code field (CEP in Brazil). This field is important to associate patent applications with geographic locations. Two-thirds of resident claims (60,662 out of 91,372, or 66.4%) contained no CEP, while the rest (30,710)

1. The World Intellectual Property Organization (WIPO) is a U.N. agency established in 1967 to promote intellectual property protection on a world scale through cooperation among countries and collaboration with other international organizations. It currently has 184 member states. For more information, go to www.wipo.org.

required additional treatment (checking against the Ministry of Labor's RAIS Annual Employee Register, searching the internet etc.) in order to determine a geographic location based on this field.

The first restriction on the information received from INPI via IPEA related to periodization. For the purposes of the analysis discussed in this chapter, patent applications filed between 1980 and 1999 inclusive were used. Although this dataset contained data up to 2003, the data for the last three years (2000-03) were discarded because data of better quality were later received from INPI for the period 2000-07.²

As a result, 179,692 applications for PIs and MUs (71.16% of the total) were used: of these, 66,609 (37.1%) were filed by Brazilian residents and 113,083 (62.9%) by non-residents. Over half the filings by residents (38,686 or 58.1%) had an individual as first-named inventor, while most of the rest (16,371 or 26.6%) had a corporate entity. The CNPJ or CPF field was left blank in 11,552 or 17.34%.

The second dataset resulted from a request sent by FAPESP to INPI, exclusively for use in this chapter. In response, INPI furnished two plain-text (".txt") files with a total of 1,321,320 records, corresponding to 158,594 claims filed between 2000 and 2007 for PIs, MUs, Industrial Designs and Certificates of Addition.

The number of applications for PIs and MUs totaled 126,345 (79.67%). The first-named inventor was a Brazilian resident in 44,689 (35.4%) and a non-resident in 81,656 (64.6%).

This dataset for 2000-07 also provided taxpayer numbers (CNPJ or CPF), enabling the applications to be classified as corporate or individual. Thus 39,411 (88%) of the applications filed by residents contained information in this field, with 27,593 (61.2%) displaying a CPF structure and 11,818 (26.4%) displaying a CNPJ structure. The remaining 5,278 (11.8%) contained nothing in the CNPJ or CPF field.

Because of the 18-month confidentiality requirement, and given that the INPI database was generated in mid-2007, the data for 2006 and 2007 did not correspond to the total number of applications actually filed in those years. Thus only applications for the period 2000-05 were used, since the confidentiality requirement for these did not restrict the information available.

The number of applications filed for PIs and MUs in the period 2000-05 amounted to 118,648. Of these, 40,173 (33.9%) named residents as first inventors and 78,475 (66.1%) named non-residents. Individuals were identified

as patentees in 25,046 resident filings (62.3%), corporate entities in 10,615 (26.5%). The remaining 4,513 (11.2%) contained nothing in the CNPJ or CPF field.

In contrast with the initial dataset for the period 1980-1999, the dataset for 2000-05 did not include postal codes for applicants or inventors, although it did include some information on states and cities. However, 606 resident filings did not specify the state and 4,653 did not specify the city.

2. Treatment of data furnished by INPI

To create the database used for this chapter, the data furnished by INPI were treated with three aims in mind. The first was to produce the largest amount of information possible by filling in the blanks in the data received. The second was to improve the quality of the information furnished. The third was to add external information from the RAIS statistics and the internet so as to enrich the analysis of the technological activities performed in Brazil between 1980 and 2005.

The first goal was achieved by harmonising the data and completing blank fields wherever possible. This involved comparing records that contained blank fields with others from which the requisite information could be gleaned. For example, if two patent applications had the same corporate taxpayer number (CNPJ) and one contained no information on the first-named inventor's geographic location, this information was copied from the more complete record based on the assumption that both taxpayer number and location were the same.

Work done toward the second goal entailed refining the database so that it could be used to construct more reliable statistics, given that the original dataset contained many spelling and other errors. For example, records with identical CNPJ numbers but different first-named inventors were harmonized using the name in the most complete record. When the disparities were very significant, a search of the Federal Revenue's website was conducted to retrieve the official information for the CNPJ concerned.

Another important procedure involved obtaining information on the first-named inventor's state and city from the postal codes (CEPs) for patent applications filed between 1980 and 1999. The first step was the treatment described in the preceding paragraph, to correct the information supplied so as to obtain a number compatible

2. The last two years of this database at least do not accurately reflect all patenting activity in the years concerned because (i) INPI safeguards the confidentiality of applications for 18 months after the date of filing, and (ii) the analysis in this chapter is based not on patents granted but on claims or applications filed. A patent application filed on January 1, 2001, would not appear in the data until July 2002. Because the database was extracted from INPI's system at some point in 2003, there was no guarantee that it contained data on all patent applications filed since 2001. Moreover, the new database furnished later (with data for 2000-07) contained more consistent information. This second database was therefore prioritized as a basis for the analysis up to the admissible limit, i.e. 2000, instead of the same year in the previous database.

with the structure of a CEP code in as many records as possible. This treatment was applied to all 91,732 patent applications, 60,662 of which did not contain any CEPs while the other 30,710 required checking. When such treatment proved impossible, the information was discarded and the code “NA” (not available) was inserted in the fields relating to location (state and city).

The second step was to use a CD (“Endereçador”) supplied by the Brazilian Post Office (ECT) to convert CEPs into states and cities. The above procedure was then repeated to complete blanks in records with the same CNPJ.

This part of the database construction process was especially laborious. The second database supplied by INPI contained many inconsistencies in the fields for state and city. The state was blank in many cases, requiring a search for the state in which the city is located in order to complete the missing information. In other cases the city was blank, making it difficult to complete the missing information, especially for individual filings because of the impossibility of obtaining detailed CPF records. Other inconsistencies include fields completed with the wrong information, e.g. neighbourhood in the city field, or city followed by neighbourhood in the same field. This required examination and cleansing of the entire database record by record. In some cases the information in the city field was incompatible with the information in the state field, in that no such city exists in the state in question. In other cases it was possible to correct the name of the state because no other city with the same name exists anywhere else in Brazil. Otherwise the information in the city field had to be deleted.

It is important to clarify a difference between the two datasets relating to information on the location of first-named inventors. The first dataset, for the period 1980-99, did not contain information directly specifying this location, which was initially identified by converting CEPs (once they had been corrected or completed, as described above) into cities and states using the “Endereçador” search engine. The second dataset, covering the period 2000-05, specified cities and states, although this information frequently had to be corrected or completed. Thus the 4,653 records that did not specify the first-named inventor’s city and the 606 that did not specify the state were compared with other records that had the same CNPJ or CPF to complete the missing information.

Once treated, the two datasets were combined and harmonized so that the information retrieved was consistent for the entire period between 1980 and 2005.

To achieve the third goal, two other sources were

used. The first was the Annual Employment Survey (RAIS) produced by the Ministry of Labor & Employment (MTE). This identifies every corporate entity in Brazil by federal taxpayer number (CNPJ) and a specific RAIS identification number (Rais_ID). The resulting database included a classification of patent applications by CNPJ according to the National Economic Activity Classification (CNAE) adopted by IBGE, Brazil’s national bureau of statistics and population census.

A number of precautions and some corrections were required, however, in using CNAE categories. When the CNPJs supplied in patent applications had been matched against RAIS data, some CNPJs were found in the CNAE category “Corporate head offices & local administrative units”, which did not correspond to the actual activity of the corporate entities that filed the claims concerned. In such cases a search was again made of the Federal Revenue’s database of CNPJs to check the correct classification of the patent claims concerned or to classify them when this information was missing from the RAIS data. The first-named inventor was then classed in the CNAE category assigned by the Federal Revenue, using CNAE 1.0 from IBGE.³

When it proved impossible to obtain reliable information from the Federal Revenue, first-named corporate entities were classified according to the main economic activity of the group to which they belonged, based on the root code in the CNPJ as filed, and this was checked against the other information in the database, including location and year of filing.

In addition, the Rais_ID was also used to find cities and states for first-named inventors, completing blank fields and correcting inaccurate information in the original databases.

The second source was *Science & Technologie: indicateurs 2006* (OST, 2006), which supplied the algorithm for aggregating the technologies cited by patent applications into technology domains and subdomains. Thus, for example, it was possible to identify a claim originally classified under “Peptides and proteins” according to the International Patent Classification (IPC) used by WIPO (class C07K) as a patent application in biotechnology. The end-result was the aggregation of hundreds of classes into seven technology domains and 30 technology subdomains.

Following treatment of the two files received from INPI, the data were compiled into two databases for patent applications filed between 1980 and 2005. The first was more complete and referred to filings by Brazilian residents. The second contained information on applications filed by non-residents in the same period.

3. The Federal Revenue uses CNAE 2.0. The National Classification of Economic Activities (CNAE 2.0), developed by the Brazilian Institute of Geography and Statistics (IBGE) and used in this chapter, corresponds to the International Standard Industrial Classification (ISIC), in its fourth revision, which is adopted by the United Nations (UN).