

**Methods and software for the harmonization and combination of datasets:
A test based on IP-related data and accounting databases
with a large panel of companies at the worldwide level**

Grid Thoma¹

Salvatore Torrisci²

Alfonso Gambardella³

Dominique Guellec⁴

Bronwyn H. Hall⁵

Dietmar Harhoff⁶

Draft version 3 June 2009

Affiliations: ¹ Corresponding Author, University of Camerino, and KITES - L. Bocconi University, via Sarfatti 25 20136 Milano, grid.thoma@unibocconi.it; ² University of Bologna and KITES - L. Bocconi University; ³ L. Bocconi University; ⁴UC Berkeley and University of Maastricht; ⁵OECD; and ⁶ Ludwig-Maximilians-Universität München;

Acknowledgments: We thank Jim Bessen, Hélène Dernis, Megan MacGarvie, Paola Giuri, Myriam Mariani, Kazu Motohashi, Teruo Okazaki, James Rollinson, Philipp Sander, Georg Van Graevenitz, Stephan Wagner, Norihiko Yamano, Maria Pluvia Zuniga, and all the participants at the PATSTAT Users' Meeting in Paris in March 2008 and seminars at the Ludwig-Maximilians-Universität München and "L. Bocconi" University in Milan for very fruitful discussions during the preparation of this paper. We also thank Armando Benincasa and Luisa Quarta from Bureau Van Dijk for clarifications about the structure of the Amadeus database and its changes over time. The usual disclaimer applies.

Table of Contents

1. Introduction.....	3
2. Background on data and sources for innovation studies	4
2.1 Online databases	5
2.2 Off-line databases	7
3. Methods for combining patent data with other sources	8
3.1 The dictionary-based approach.....	9
3.2 The rule-based approach	10
3.3 Additional uses of dictionaries	12
4. Software creation.....	14
5. Dataset Creation.....	15
5.1 Dictionary creation.....	15
5.2 Integration with business directories	15
5.3 Propagation of the matching into the USPTO dataset	16
5.4 Propagation of the matching into the JPO dataset	16
6. Conclusions.....	17
7. References.....	17

1. Introduction

The lack of firm-level data on innovative activities has always constrained the development of empirical studies on innovation. More recently, the availability of large datasets on indicators such as R&D expenditures and patents has relaxed these constraints and spurred the growth of a new wave of research. However, measuring innovation still remains a difficult task for reasons linked to the quality of available indicators and the difficulty of integrating innovation indicators to other firm-level data.

Data on R&D expenditures have traditionally been used as a proxy for innovation effort or value, but because they measure input they are unable to tell us much about the 'success' of innovative activities, nor do they give much detail about their direction. Moreover, especially in the case of European firms, data on R&D expenditures are often missing because reporting these expenditures is not required by accounting and fiscal regulations in a number of countries. Therefore, an increasing number of studies have used patents counts as a measure of inventive output. However, crude patent counts are an extremely noisy indicator of inventive output because they do not account for differences in the value of patented inventions. For this reason, many innovation scholars have introduced various patent-related indicators as a measure of the importance or 'quality' of the inventive output.

To make good use of patent data at the micro (firm) level, it is necessary to merge these data with other firm-level information, such as accounting and financial data. This can be a challenging task due to the difficulty of matching different data sources, something which frequently must be done using only the names of the firms in question. Inaccuracy in data merging and integration can lead to measurement errors and results that are biased. This is a particularly important issue in studies of patenting at the firm level because patent data never comes with firm-level identifiers that match to other sources of data, and researchers much rely only on the names of the firms to combine datasets.

Earlier approaches to this problem have relied on manual matching of firm names across datasets (e.g., the small samples of US patenting firms used by Griliches, 1981) and partially computer-based ad hoc methods combined with manual matching (e. g., Bound et al., 1984 and Hall et al., 2005). But methods involving even some manual matching have limits when confronted with datasets of the size that are common today.

For these reasons, the authors undertook the development of a more comprehensive and automated methodology for company name standardization and the matching of two data sources using the resulting standardized names: IP-related databases and company business directories. Our methodology is based on recent research on automatic Named Entity Recognition (NER) systems. NER systems have been applied successfully in bioinformatics, while their deployment in social sciences is still at an early phase. Thus this study is among the first attempts to deploy these methods in empirical studies in economics and management, contributing to this body of research by comparing two different approaches to the problem.

The first group of approaches are dictionary-based methods, essentially based on large collections of names that serve as examples for a specific entity class, such as the DERWENT Patentee Index, and the USPTO and EPO standard patenter codes. More recently, automatic methods have been suggested for generating a dictionary (Magerman, Van Looy and Song, 2006). The second group are rule-based approaches that build up a set of rules for the comparison of similar names. A pioneering exercise was performed by Thoma and Torrisi (2007) using approximate matching based on string similarity functions. Their

analysis was based on two data sources: the PATSTAT patent database (USPTO and EPO patents) and the Amadeus accounting and financial dataset which contains 2,197 parent firms and their 151,979 subsidiaries.

The present study relies on and improves significantly the software prototype developed by Thoma and Torrisi (2007). It implements a new NER system that consists in providing automatic, reliable and time-saving procedures to combine data whose identifiers are names rather than id numbers for empirical studies in economics and management. In the implementation phase of the software prototype we do not rely on a single NER approach, but experiment with several techniques in parallel. Our goals for these software prototypes are accurate matching, fast computation time, efficient implementation, and ease in using the results.

The paper is organized as follows. We start by describing the data and sources often engaged in innovation studies, such as microdata on patents and other IP information. In particular, we will consider their content, time coverage, mode access, complementary search and management tools and potential integration with other sources.

Secondly, using insights coming from the NER methods that have been experimented with in bioinformatics, we will discuss the two different approaches to name-matching and data combination: the dictionary based approach, which relies on the collection of large datasets of names and their variants, and the rule-based method, which builds on the articulation of rules to establish a similarity link across different entity names. Additionally, we will discuss how the value of existing dictionaries could be enhanced by using other methods to query their entries. We propose a novel method relying on priority links among the patents that enables the combination of distinct dictionaries of entity names originating from patent data of different offices.

Thirdly, we discuss the criteria by which we developed the software prototypes for the different NER approaches analyzed in the section 3. Finally, we conclude by documenting the harmonization and matching that results from the methodology suggested in this paper.

2. Background on data and sources for innovation studies

Scholars of technical change have addressed the lack of data in two ways. The first approach is to collect firm-level information through surveys based on representative samples of the population of innovators, and the second is to rely on publicly available administrative databases such as patent and accounting information, or in some cases on confidential firm level data that resides in National Statistical Offices or Central Banks.

With respect to U.S. data, two widely cited surveys are the Yale Survey (Levin et al 1987) administrated in the early 1980s and the sequel conducted by scholars at the Carnegie Mellon University in the 1990s (Cohen et al 2000); both covered the sources and strategies of innovation at the firm level. In the European context, European National Statistical Offices have conducted a series of Community Innovation Surveys (CIS), collecting detailed data on innovation and other firm characteristics (Arundel, 2001).¹ More recently, European, U.S.,

¹ A large number of countries outside of Europe and North American have followed this lead, with innovation surveys now having been conducted in Asia, Latin America, and other locations.

Japanese, and Korean scholars have conducted inventor surveys based on the individuals named in patent applications which provide very detailed information on the factors driving innovation at the level of the individual inventor (Harhoff et al. 1999; Gambardella et al., 2008; Giuri, Mariani et al., 2007; Nagaoka and Tsukada, 2007; Nagaoka and Walsh, 2008).

The integration of CIS and other survey data with information from other databases, such as patents and accounting data is made difficult by the limitations to the use of CIS data imposed by privacy laws in many countries. Although innovation survey data has produced a number of findings, the difficulty of matching CIS databases to other databases have limited their use for the purpose of research in economics, management and public policy.

Another research line has focused on the collection from secondary sources of information on different qualitative dimensions of innovation such as prizes as a measure of successful inventive races, trademarks as a measure of the new product introduction, newswires as a paper trail of patterns of collaborations among firms such as M&A, licensing and R&D agreements etc. (Moser, 2004; Giarratana and Torrisi, 2006; Fosfuri and Giarratana, 2007; Powell et al., 2000; Arora et al. 2001)

A third line of exploration is centered on innovation counts and R&D. R&D expenditures are a measure of input and do not tell much about the 'success' of innovative activities. An increasing number of studies have used patent counts and patent-related indicators to measure the quantity and the 'quality' of inventive output. Patents as a measure of inventive success have their own drawbacks too but they are the most direct, detailed, and objective measure of innovation (Griliches, 1981 and 1990; Pavitt, 1988).

In this section we briefly review the available sources of microdata on patents and other IP information. In particular, we will consider their content, time coverage, mode access, complementary search and management tools and potential integration with other sources.

2.1 Online databases

US granted patents

Freely available from www.uspto.gov, this database includes information on all US patents (including utility, design, reissue, plant patents and etc) from the first patent issued in 1790 to the most recent issue week.

Full searchable text is offered for patents issued from January 1976 to the present, including all bibliographic data, such as the inventor's name, the patent's title, and the patentee's name (called the assignee at the USPTO), the abstract, the full description of the invention, and the claims. Patents issued prior to December 1975 are only searchable through the patent number, issue date, and current US patent classification.

US published applications

As in the case of granted patents, the application database is freely searchable at the USPTO and consists of the full text of US applications that have been published since its inception in March 2001. After that date patent applications could be kept secret if protection was requested in the US only; otherwise the application is published within 18 months from filing.

The full text of a published application includes all bibliographic data, such as the inventor's name, the published application's title, and the applicant, as well as the abstract, the full description of the invention, and the claims. All of the textual words in the publication are searchable.

USPTO Trademarks

The USPTO website at <http://www.uspto.gov/main/trademarks.htm> provides complete electronic information about trademarks since the birth of the USPTO. The database contains more than 4 million pending, registered and dead trademarks and it provides complete free searchable access to the text and image database of trademarks.

ESPACE on-line

The ESPACE database contains freely searchable information on published patent applications from over 80 different countries and regions. It is based on the PCT minimum documentation, which is defined by WIPO as the minimum requirement for patent collections used to search for prior-art documents for the purpose of assessing novelty and inventiveness. As of March 2007, esp@cenet@ held data on 60 million patents. A total of 30.5 million of these patents have a title, 19.5 million have an abstract in English, and 29.5 million have an ECLA class.²

**Table 1:
ESP@cenet coverage: Starting year of availability for the main patent offices**

Patent Office	Facsimiles	Full Text	ECLA
DE	1877	1970	1877
EP	1978	1978	1978
FR	1900	1970	1902
GB	1859	1893	1859
US	1836	1970	1836
WO	1978	1978	1978

CTM – on line

CTM–ONLINE provides free access to information on EU Community trade mark applications and Community Trademarks, updated on a daily basis regarding: the trademark number, name, type, owner, nice class number, status, filing date, registration date, date of international registration, publication date, expiry date etc.

Delphion

Delphion is a proprietary database of Thomson Corporation that allows the user to search contemporaneously inside different patent databases at the worldwide level. The patent collections of Delphion contain the following:

² ECLA is a European Patent Classification that is about twice as detailed as the IPC (International Patent Class) classification.

- (i) US granted patents since 1971 and US patent applications since March 2001
- (ii) European patent applications since the inception of the EPO in 1978
- (iii) Patent Abstracts of Japan database since 1976
- (iv) Patent application and grants from the German national patent office since 1968
- (v) INPADOC and WIPO/PCT patent information.

2.2 *Off-line databases*

While the on-line databases provide real time and constantly updated information, researchers are often more interested in off-line databases in spite of higher costs and difficulties of updating. Off-line databases allow easier generation and manipulation of innovation indicators for statistical analysis. Moreover, ex-post scalability and integrability with other sources of information is significantly higher. The most important current sources of such data that are easy and low cost are the NBER patent citation database and the EPO-OECD PATSTAT database. However, before describing those two sources we will mention at least two other earlier efforts to generate such data, which contain historical data and are still available.

The pioneering work using patent data in economic studies can be found in Jacob Schmookler's (1966) major book entitled *Invention and Economic Growth*. Schmookler classified patents manually by the industry of their potential use, finding that the top three user industries of patents during the first half of the 20th century were the railroad, petrochemical and building sectors.

The seminal work of Schmookler was followed by that of Griliches and co-workers at the NBER (Bound et al., 1984, which constitutes the first major effort to combine patent counts with economic and financial data, such as sales, capital stocks, research and development, income, at the firm level. The accounting information is drawn from Standard and Poor's Compustat files, which contain data for all firms traded in the major U.S. stock markets. The linking was done mostly manually and it involved about 2,700 US corporation and their subsidiaries as reported in the year 1976. The authors used this dataset to estimate patent production functions and valuation equations for patent portfolios.

The NBER patent database

The NBER patent dataset on USPTO data represents a path-breaking effort of providing additional bibliographic information that could be used to account for differences in the 'value' of patents (Hall, Jaffe and Trajtenberg 2001 and 2005). Hall and colleagues have made these data freely available via the NBER website. The database comprises detailed information on almost 3 million U.S. patents granted between January 1963 and December 1999, all U. S. patent citations made to these patents between 1975 and 1999, constituting over 16 million citation links. Moreover, the database is accompanied by a file containing the link between the names of USPTO patent assignees and the names of US companies listed in the Compustat dataset. This match is a more complete and updated version of the one

described in Bound et al. (1984).³ Recently these data have been updated to 2006 by Cockburn and co-workers at the NBER (Cockburn et al., 2009).⁴

Although useful for the analysis of US-based questions, the downside of using the NBER data is that it does not currently include information on citations to and from other patent databases, so citation counts are likely to be downward biased, especially for foreign-owned patents.

PATSTAT

The EPO Worldwide Patent Statistical Database (PATSTAT), which is available under a no-cost license from the OECD-EPO Task Force on Patent Statistics, includes not only data on patent and utility models indicators such as citations and IPC codes, but also on patent families based on priority date links. The database contains documents from more than 80 patent offices worldwide since the 1970s. Moreover, PATSTAT includes data from the World International Patent Office, which relies on the Patent Cooperation Treaty (hereafter also PCT-WIPO). Having PCT-WIPO patents is important because their counts are the only single indicator which allows accurate international comparisons. Indeed, even triadic families may be a biased indicator because of the publication lag for USPTO applications filed prior to March 2001.

The main elements of PATSTAT are the title and abstract of application; filing, priority and publications dates of the application; patenters and inventors and detailed addresses, IPC classification symbol, priority applications. Moreover, PATSTAT provides also complementary information on citation links such as the type of the citation, citation identification, origin of the citation, non-patent literature bibliography, etc. Currently around 100 institutions have subscribed to PATSTAT and this database is expected to be widely used for innovation studies.

3. Methods for combining patent data with other sources

A major obstacle to the integration of patent data with other indicators of firm performance in large samples is represented by the difficulty of uniquely matching the names of patent patenters with the corresponding legal entity in business directories such as Compustat, Who Owns Whom or Amadeus. In this section we describe some recent developments in methodologies for integrating different IP databases and company directories, methods that have been inspired by some interesting insights from bioinformatics.

Over the past years, biology has become a science increasingly concerned with the analysis of large amounts of information. Consequently, the way that information is stored, managed, visualized and searched is of the most importance to this research field. Moreover, named entity recognition (NER) for biomedical applications, i.e. the task of identifying gene, protein, diseases, and other names in natural text has become a crucial means to extract highly valuable and sometimes hidden information. The NER approach may have interesting

³ The match of patents to Compustat for the 1999 database is based on the 1989 universe of companies. For more details see Hall et al (2001) and <http://www.nber.org/patents/>

⁴ Beta versions of the new datasets are available at <https://sites.google.com/site/patentdataprotect/Home> .

applications to economics and management science, especially in the area of the information integration of company-level data.

In the following sections we will discuss two different approaches to data combination: the dictionary-based approach, which relies on the collection of large datasets of names and names variants, and the rule-based approach which builds a set of rules for similarity links across different entity names.

3.1 The dictionary-based approach

Dictionaries essentially are large collections of names, serving as examples for a specific entity class. Matching dictionary entries exactly against text is a simple and very precise NER method, but typically yields a low level of match when applied to firm names. To compensate, one can either use approximate matching techniques, or try to ‘fuzzify’ the dictionary by automatically generating typical spelling variants for every entry. The extended dictionary is then used for exact matches against the text.

Previous attempts have addressed this issue by implementing ad-hoc matching procedures to reduce the cost of data standardization and integration. For example, Thomson Scientific's Derwent World Patent Index (2002) is constructed by assigning a code to about 21,000 patenters. This index accounts for legal links between parent companies and subsidiaries thus achieving a legal entity standardization. This task requires substantial manual, labor-intensive work and some loss of accuracy in name matching thus giving rise to a potentially large number of false positives.

Drawing on the Derwent methodology, Rachel Griffith and colleagues at the Institute of Fiscal Studies (IFS) have standardized the names of a sample of UK patenters of Triadic patents and matched them with the standardized names of companies contained in Bureau van Dijk's Amadeus database (Griffith et al. 2006). Only identical standardized names found in the two datasets are matched by the IFS using the Derwent semi-manual standardization procedure.

Another example of a dictionary of patenter names is constituted by the USPTO CONAME file compiled by the USPTO. It is a semi-automatic standardization procedure which focuses on the first-named patenter reported in the patent document. For patents granted after July 1992 the patenter name is standardized and matched automatically with other standardized names in the same dataset. New patenters that are not matched automatically with standardized names in the dataset are matched manually. For instance, the entry of a new patenter whose standardized name does not match any previously standardized names is examined by looking at the names of inventors. The CONAME file accounts for changes in patenter names but does not account for legal links between patenter names. Moreover, similar names with a different legal form or the same legal entity from different countries are not matched.

The EPO has elaborated its own dictionary by assigning a standard code to each patenter filing a patent to the office. This index is created by taking into account not only the patenter name and country but also her postal address. According to some interviews we did with EPO representatives this dictionary tries to maximize precision vis-à-vis recall rate for each entry: for example two patenters with the same name but having different addresses will constitute separate entries in the EPO dictionary and they are linked to two different standard codes that identify them.

More recently, a group of researchers from the Katholieke Universiteit Leuven (KUL) have developed an automatic methodology based on the detailed standardization of patent names and perfect matching of names. This methodology, like the CONAME file and EPO standard codes, does not try to establish legal links among patenters. The main advantage of this procedure is high precision, i.e., a limited number of false matches. The KUL methodology has been used to standardize and match patent names from EPO patent applications published between 1978 and 2004 and USPTO granted patents published between 1992 and 2003 (Magerman, Van Looy and Song, 2006).

According to the KUL methodology the creation of a dictionary for company names can be articulated in preprocessing and names standardization. Names standardization requires a series of tasks like punctuation standardization (e.g., from FERRARI ,& C. to FERRARI, & C.) and company name standardization (from FERRARI, & C. to FERRARI, AND COMPANY). The main standardization operations suggested by Magerman, Van Looy and Song (2006) can be summarized as follows: i) character cleaning; ii) punctuation cleaning; iii) legal form indication treatment; iv) spelling variation standardization; v) umlaut standardization; vi) common company name removal; vii) creation of an unified list of patenters.

For U.S. patent assignee names, a major effort to update the existing NBER patent citations database and match to the Compustat files is now underway (Cockburn et al. 2009). The matching of patent assignee names with the names of firms on the Compustat files is part of this project.⁵ A number of enhancements to the original (1999 and 2002) databases have been made. First, the semi-automatic standardization procedure of this file has been extended to all the assignee names in the case of multi-owner patents. Second, using external sources originating from business directories information was collected on the timing of name and ownership changes of the assignee. The data now provided contains information that allows tracking of the assignee changes of ownership over time. Third, there is progress on standardizing the firm names supplied by the USPTO and correcting cases where the USPTO had coded the type of entity (individual, firm, government) incorrectly. The list of entity types has been expanded to include universities, non-profit research institutions, and medical institutions including hospitals. However, much of the standardization work is still incomplete, especially that involving non-U.S. patent assignees.

3.2 *The rule-based approach*

Rule-based approaches build on the definition of rules to compare the similarity of names. Early systems used hand-crafted rules to describe the composition of named entities and their context. For instance, some core words and components of words might be used to extract candidates for more complex names. These core terms are expanded according to a set of syntactic rules. Similarly, starting from more complex names one could invert the process to identify some discriminating core words using the same rules.

⁵ The original 1999 database is at <http://www.nber.org/patents> with 2002 updates at <http://www.econ.berkeley.edu/~bhhall/patents.html> . The latest (2006) version is at <https://sites.google.com/site/patentdatapoint/Home> . The match documentation is at <http://www.nber.org/~jbessen/matchdoc.pdf> .

In the following, we will focus our discussion on the potential usefulness of names similarity functions based on the so-called approximated string matching (ASM) algorithms (Thoma and Torrisi, 2007). However, it is worth remembering that the ASM method constitutes only a specific class of similarity rules. The applicability of other matching methods to company name matching should be analyzed in future research.

The first category of ASM similarity functions is based on the edit distance. For instance, the Levenshtein distance between two strings is defined as the minimum number of operations needed to transform a string into another one. The transformation of a string can be obtained by character inserting, substituting, swapping or substitution (Levenshtein, 1966). An extension of the Levenshtein edit distance was developed by Smith and Waterman (1981). The main difference with the Levenshtein distance is that character mismatches at the beginning and the end of strings are ignored in the calculation of distance. For instance, two companies 'Dr Michal White Plc' and 'Michael White Plc, Dr' has a short distance using the Smith-Waterman distance.

The similarity between two strings x and y of length n_x and n_y can be computed as $1-d/N$, where 1 is the maximum similarity, d is the distance between x and y and $N = \max\{n_x, n_y\}$. To calculate the distance between two strings we need to assign a cost c to each operation required to transform the string x into string y (or vice versa). The cost is assumed to be 1 for substitution and deletion of a character and 0 for perfect matching characters. For instance, the edit distance between IBM and INTEL is the following:

$$1 - [c(I, I) + c(B, N) + c(M, T) + c(\phi, E) + C(\phi, L)]/5 = 1 - 4/5 = 1/5.$$

The second category of ASM similarity functions relies on token-based distance. Measures of token distance, like the J similarity index, are based on the division of strings into tokens or sequences of characters. Token-based distance functions account for differences due to the position of the same tokens between otherwise identical strings (e.g., Peter Ross and Ross Peter). In particular, the J token distance computes the fraction of common tokens, after breaking the strings on white spaces. The J token distance is simply given by the number common tokens in two names and the count of total number of tokens in those names.

To test the performance of the second category of string similarity functions we used the J token distance after breaking the strings at blanks and computing the fraction of common tokens:

$$J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} \quad (1)$$

where $X \cap Y$ measures the number of common tokens between strings X and Y while $X \cup Y$ measures the total number of distinct tokens.

To account for common tokens, we multiply each token by a weight that is inversely proportional to its frequency in the dataset. Formally, each token i has a weight w_i given by

$$w_i = \frac{1}{\log(n_i) + 1}$$

where n_i is the frequency of the token in the dataset. This weighting method is a simplified version of the *tf-idf* weight (term frequency–inverse document frequency) of Salton and Buckley (1988).

To reduce the computational complexity of the J similarity index we approximate the second term of equation (1) as follows:

$$\frac{2|X \cap Y|}{|X|+|Y|} \quad (2)$$

where the denominator is the sum of all tokens, including those tokens that are contained in both strings. This may result in some double counting. On other hand, it would be extremely costly from a computation viewpoint to find tokens common to two strings (company names). To maintain the same approximate scale we have multiplied the index by a factor of 2.

Thus, the weighted J^w distance is equal to the following expression:

$$J^w(X, Y) = 1 - 2 \frac{\sum_{k|x_k \in X \cap Y} w_k}{\sum_{i|x_i \in X} w_i + \sum_{j|y_j \in Y} w_j} \quad (3)$$

Where $x_i \in X$ and $y_i \in Y$ and w_i and w_j are the weights inversely correlated with the frequency of tokens x_i and y_i in the dataset; the terms x_k and w_k are respectively the k^{th} token and relative weight belonging to the intersection set $X \cap Y$.

3.3 Additional uses of dictionaries

In the previous sections we discussed the drawbacks of the dictionary approach due to the low match rate when perfect matching is implemented using the dictionary entries. One suggestion for overcoming the drawbacks of perfect matching is the use of approximate matching based on string similarity functions (see Thoma and Torrisi, 2007). In this section we discuss an additional source of standardization of patenter names that relies on the priority links across patent offices.

A priority link emerges when a patenter claims a priority date antecedent to the filing date of a given patent. Typically priority links refer to patent documents in other patent offices and the set of patents (or applications) filed in several countries which are related to each other by one or several common priority filings, is generally known as a patent family.⁶ It is also often considered that a patent family comprises all patents protecting the same invention. Indeed, the rapid growth in the number of patent documents in the last years have been accompanied by the growth of priority links, and hence the total number of inventions has grown less than the total number of patent documents.

⁶ Patents that refer to earlier patents in the same patent office as their priority are called continuation (at the USPTO) or divisional patents (at the EPO and the USPTO). Because patents are sometimes divided in different ways at different offices and members of a family at one office may claim different priorities elsewhere, there is more than one definition of a patent family. See Harhoff (2008) for a fuller discussion of this issue. We have used the INPADOC definition, which has been included in PATSTAT for the first time in the release of October 2008.

Thus, if a patenter has filed a document in two or more offices claiming a common priority date, it is possible to trace a link from an entry in the patenter names dictionary in one patent office to the corresponding entry in another patenter names dictionary in the other patent offices, on the assumption that the ultimate owner of the patent will be the same at both offices. Based on this assumption, in this section we describe an additional harmonization method for patenter names using priority links across USPTO and EPO patent databases. The objective of the analysis is to assess whether the priority links between US and EPO patents can improve the accuracy of harmonization of two existing dictionaries of patenter names: the USPTO CONAME file and EPO standard names. This methodology may allow us to propagate the matching done with one dictionary to the other, reducing the cost of implementation of such matching across the two dictionaries.

Figure 1 shows how we link these two dictionary files. In TASK 1 we start from the USPTO CONAME file made up of 237,666 distinct patenter names. The file has information on all the patenters that have been granted at least one patent by the USPTO over the period 1963-2007. The USPTO CONAME file can be easily interfaced with the PATSTAT database through the patent publication number (TASK 2). Subsequently in TASK 3, with the PATSTAT database we can rebuild the priority links being generated from and into the EPO patent database; in particular we rely on the INPADOC patent family definition. In TASK 4 we used the priorities compiled from PATSTAT by linking each EPO application with the priority date patent in the US patents via the publication number. Finally, we deal with the identification of the proper link to the EPO patenter names: TASK 5 is designed to take care of the number of priority links and number of patenters per EPO patent. In particular we used a string similarity algorithm and also manual checking to ensure the proper association across the USPTO assignee codes and EPO applicant codes.

<p>TASK 1: Standardized assignee names</p> <p>File Source: USPTO assignee names</p>	<p>TASK 2: USPTO patents 1975-2008</p> <p>Source: PATSTAT Table t1s211</p>	<p>TASK 3: Priority across EPO & USPTO 1978-2008</p> <p>Source: PATSTAT Table t1s219</p>	<p>TASK 4: EPO patent applications 1978-2008</p> <p>Source: PATSTAT table t1s201</p>	<p>TASK 5: EPO standard names 1978-2008</p> <p>Source: EPO source files</p>
--	---	---	---	--

Figure 1 Harmonization process based on priority links: data and sources

The final list of EPO applicant codes with a US standardized name includes 158,972 patenters corresponding to 70,319 names on the USPTO CONAME file. This USPTO/EPO dictionary contains about 72.6% of the EPO patent applications and 77.7% of the US granted patents filed by business organizations. The overall gain in the harmonization of the EPO applicant code is about 55.8%. Thus this approach significantly increased the quality of the EPO standard codes file using the USPTO CONAME file.

4. Software creation

Following an approach similar to that in Magerman et al (2006), we created a dictionary using the following operations:

1. Each of the accented characters is replaced by its unaccented version.
2. The conjunction "and" and its translations into other languages are standardized as "&"
3. Removal of common company words like INC and AB in descending order of their length.
4. Removal of frequent comma and period irregularities
5. Removal of double quotation mark irregularities
6. Proprietary character codes replaced by the ASCII/ANSI equivalent.
7. SGML and HTML codes replaced by the ASCII/ANSI equivalent
8. Replacement of spelling variations with their harmonized equivalent for some frequent words
9. Removal of the round parentheses and cleaning their content
10. Umlaut harmonization
11. Removal of the alphanumeric characters
12. Generation of a unique list of patenters
13. Linkage to the USPTO and EPO standard codes

We then created a software prototype to implement the cleaning of the patenter names dictionary as defined above together with a rule-based approach based on the approximate string-matching algorithms discussed here and in Thoma and Torrisi (2007). The dictionary approach was the first software processing phase. This was followed by rule-based post-processing procedures for the refinement of predicted matching candidates, the resolution of abbreviations and of multiple matching occurrences of the same patenter.

The implementation of the software prototype followed four basic criteria. The first was the requirement that the program run on a partition of the input data into smaller subsets. The time complexity of the first phase is quadratic in the input size, so the partition reduces execution time to the sum of the squares of smaller data sets. The second objective was efficient implementation of the distance functions: in particular, we did not need to evaluate the distance between each pair of names, because we did not consider pairs with distance greater or equal to one. Third, we used a parallel implementation, so that many computers could work at the same time, reducing the time required to that needed to match the largest partition. Finally, the output files have been formatted in a general format that can be easily processed by the most widely diffused statistical and econometrics software packages, such as STATA.

5. Dataset Creation

We used our software prototype to create and integrate a large dataset of patenters originating from a number of countries and their patents obtained in various patent offices worldwide. The sources of the patent data were the following:

- EPO standard name codes from EPOLINE files for EPO and PCT applicant names. <http://ebd2.epoline.org/jsp/ebd1.jsp> up to July 2008
- USPTO CONAME file (CD version March 2007) for U. S. assignee names.
- PATSTAT for priority links across patents and patenters.

Then we retrieved business and ownership information for the companies from Amadeus, which collects information from approximately 10 million European firms and their subsidiaries at the worldwide level.

5.1 Dictionary creation

The final results of the procedure for the creation of a patenter names dictionary for EPO and PCT/WIPO dataset are depicted in the accompanying tables:

- Table 3 reporting the country distribution of the business applicants and applications in the EPO and PCT dataset.
- Table 4 reporting the country distribution of the non-business organization (NBO) applicants and applications in the EPO and PCT dataset.
- Table 5 reporting the country distribution of the individual applicants and applications in EPO and PCT dataset.
- Figure 2 reporting the distribution across the top 20 countries of the gain in the name harmonization with the software prototype described in the previous section. The overall reduction of the size of the dictionary is about 28.4%

5.2 Integration with business directories

In this section we report the results of the merge of patenter names with business directories. We start with the EPO/PCT patenter names for two reasons: First, thanks to the US/EP dictionary described in the previous section we can transfer the matches to a large share of the patenters at the USPTO. Second, exploiting the PCT links we can propagate this dictionary to a significant number of patenters, those holding a large majority of patent documents in PATSTAT.

In the task of matching EPO/PCT patenter names to the business directories we focused only on the business patenters, which constitute about 88.1% of the patenters overall. They encompass 347,206 original names that have been harmonized to 248,772 names according to the dictionary described in the previous section. About 55.5% of the patenters have just one application in the EPO and about 1.1% have more than 100.

The results of the matching to the Amadeus business directories are depicted in the following figures:

- Figure 3 reports the share of the business applicants in the EPO and PCT dataset that have been matched to Amadeus.
- Figure 4 reports the share of the business applicants in the EPO and PCT dataset that have been matched to Amadeus, weighted by their number of patent applications.

For every matched applicant we computed a quality of match score based on the similarity of the name and location across the patenter and the company in the Amadeus business directory. The name similarity is measured as share of the total tokens over the total number of tokens in the two names, whereas the location is given either by the city or zip code correspondence. Table 2 shows the match score definitions.

Table 2
The quality of entity match score

Score	Name	Location
0	Manual check	Manual check
1	Similarity $\geq 50\%$	Same
2	$30\% \leq \text{Similarity} \leq 50\%$	Same
3	Similarity $\geq 50\%$	Unknown
4	Similarity $\geq 50\%$	Different
5	$30\% \leq \text{Similarity} \leq 50\%$	Unknown
6	$10\% \leq \text{Similarity} \leq 30\%$	Same
7	$30\% \leq \text{Similarity} \leq 50\%$	Different
8	$10\% \leq \text{Similarity} \leq 30\%$	Unknown
9	$10\% \leq \text{Similarity} \leq 30\%$	Different

The distribution of the match quality scores is reported in Figure 5. 90% of the matched applicant are characterized by a high matching score, that is a value less than or equal to 4.

5.3 Propagation of the matching into the USPTO dataset

Because of the priority links included in PATSTAT we can directly propagate these matching results to the assignee names in USPTO dataset. The results of this matching to the Amadeus business directories are shown in Figures 6 and 7

- Figure 6 shows the share of the business assignees in the USPTO dataset that are matched to EPO/PCT, unweighted and weighted by their number of patent applications.
- Figure 7 shows the distribution of the match quality scores.

5.4 Propagation of the matching into the JPO dataset

[not yet completed]

6. Conclusions

In this paper we drew on NER methods experimented in bioinformatics to analyze two different approaches to data integration in the context of patent information. First, the Dictionary based approach relies on the collection of large datasets of names and their variants, Second, the rule-based approach builds on the articulation of rules to establish a similarity link across different entity names. Additionally, we discussed how the value of existing dictionaries could be enhanced by using other methods to retrieve original data. Then we applied our methodology to several data sources, including major patent databases and business directories such as AMADEUS.

[to be completed]

7. References

- Arundel, A. (2003), *Patents in the Knowledge-Based Economy, Report of the KNOW Survey*, MERIT, University of Maastricht.
- Arora, A., Fosfuri, A. and Gambardella, A. (2003), "The Division of Inventive Labor: Functioning and Policy Implications", Paper presented at the CREST conference in honour of Zvi Griliches, Paris August 25-27, 2003
- Belenzon, S., Berkovitz, T. and J. M. Van Reenen (2007), AmaPat - Innovation, Ownership and Financials for European Firms: Data Overview, Presentation at the 2007 Kauffman Symposium on Entrepreneurship and Innovation. Data Available at SSRN: <http://ssrn.com/abstract=1022044>
- Bound, J., Cummins, C., Griliches, Z., Hall, B. H., Jaffe, A. B. (1984), Who Does R&D and Who Patents? In Griliches Z. (ed.) *R&D, Patents, and Productivity*. Chicago: University of Chicago Press, 21-54.
- Cohen, W. M., R. R. Nelson, et al. (2000), Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). Cambridge, MA: NBER Working Paper No. 7552.
- Cockburn, I. M., A. Agrawal, J. Bessen, J. H. S. Graham, B. H. Hall, and M. MacGarvie (2009), The NBER Patent Citations Datafile Update. Data available at <https://sites.google.com/site/patentdataproject/Home>
- Curran, J. R. and Clark, S. (2003), Language independent NER using a maximum entropy tagger, in *Proceedings of the 7th Conference on Natural Language Learning*, 31st May-1st June, Edmonton, Canada, 164-167.
- Derwent (2000), *World Patents Index - Derwent Patentee Codes*, Revised Edition 8 ISBN: 0 901157 38 4, Thomson Publishers.
- Fosfuri, A., and Giarratana, M.S. (2007), Product Strategies and Survival in Schumpeterian Environments: Evidence from the US Security Software Industry. *Organization Studies* 28 (6): 909-929.
- Gambardella, A., D. Harhoff, and B. Verspagen (2008), The Value of European Patents, London, UK: CEPR Working Paper No. 6848 , available at http://www.creiweb.org/activities/sc_conferences/23/papers/gambardella.pdf
- Giarratana, M. and Torrisi, S. (2004), Entry and Survival in Foreign Markets: Technology, Brand Building and International Linkages, Social Science Research Network - Electronic Paper Collection, SSRN_ID577401_code386435.pdf (<http://papers.ssrn.com>).

- Giuri, P., Mariani, M. et al. (2005), Everything You Always Wanted to Know about Inventors (but Never Asked): Evidence from the PatVal-EU Survey". LEM Papers Series 2005/20, Sant'Anna School of Advanced Studies, Pisa, Italy.
- Griffith, R., R. Harrison, and G. Macartney (2006), Matching patents to firm accounting data for European countries. Paper presented at the EPIP Workshop on Patent Data, Università L. Bocconi, Milan, Italy, February.
- Griliches, Z. (1990), Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature* XXVIII (Dec.): 1661-1707.
- Griliches, Z. (1981), Market Value, R&D, and Patents, *Economic Letters* 7: 183-187.
- Griliches, Z., Hall, B. H. and Pakes, A. (1991), R&D, Patents. And Market Value Revisited: Is There a Second (Technological Opportunity) Factor?. *Economics of Innovation and New Technology* 1: 183-202.
- Hall, B. H., Jaffe, A. B., and M. Trajtenberg (2005), Market Value and Patent Citations. *Rand Journal of Economics* 36: 16-38.
- Hall, B. H., Jaffe, A. B., and M. Trajtenberg (2001), The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools. Cambridge, MA: NBER Working Paper No. 7741.
- Harhoff, D.(2008) Patent Families, Equivalents and Patent Value, Paper Presented at the Meeting of the NBER Program on Technological Change and Productivity Measurement Dec. 5th, 2008
- Jaccard, P. (1901), *Bulletin del la Société Vaudoisedes Sciences Naturelles* 37, 241-272.
- Leser U. and J. Hakenberg (2005), What makes a gene name - Named entity recognition in the biomedical literature. *Briefings in Bioinformatics* 6 (4): 357-369.
- Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversal. *Soviet Physics Doklady* 10(8): 707-710.
- Levin, R. C., A. K. Klevorick, et al. (1987), Appropriating the Returns from Industrial Research and Development. *Brooking Papers on Economic Activity* 3: 783-831.
- Magerman, T. Van Looy B., and Song X. (2006), Data production methods for harmonized patent statistics: Patentee name standardization. Technical report, K.U. Leuven FETEW MSI.
- Moser, P. (2005), How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World Fairs. *American Economic Review* 95 (4): 1215-1236.
- Nagaoka, S. and N. Tsukada (2007), Innovation process in Japan from inventors' perspective: results of RIETI inventor survey. Tokyo, Japan: RIETI Discussion Paper07-J-046 (in Japanese).
- Nagaoka, S. and J. P. Walsh (2008), How do the innovation systems of US and Japan differ? What are the potential implications?: Evidence from the RIETI-Georgia Tech inventor surveys. Paper presented at the RIETI Brown Bag lunch, Tokyo, Japan, July.
- Navarro, G. (2001), A guided tour to approximate string matching. *ACM Computing Surveys* 33 (1): 31--88.
- Patel, P. and K. Pavitt (1991), Large firms in the production of the world's technology: an important case of 'non-globalisation'. *Journal of International Business Studies* 22 (1): 1-21.
- Pavitt, K. (1985), Patent Statistics as an Indicator of Innovative Activities: Possibilities and Problems. *Scientometrics* 7 (1-2): 77-99.
- Pavitt, K. (1988), Uses and abuses of patent statistics. In van Raan, A. (ed.), *Handbook of Quantitative Studies of Science Policy*, Amsterdam: North Holland.
- Pavitt, K., Robson, M. and Townsend, J. (1987), The Size Distribution of Innovating Firms in the UK: 1945-1983. *Journal of Industrial Economics* 35 (March): 291-316.

- Powell, W. W., D. R. White, et al. (2005), Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. *American Journal of Sociology* 110 (4): 1132-1205.
- Salton, G. and Buckley, C. (1988), Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5): 513-523.
- Schmookler J. (1966), *Invention and Economic Growth*, Cambridge, MA: Harvard University Press.
- Smith, T. F. and Waterman, M.S. (1981), Identification of common molecular subsequences. *Journal of Molecular Biology* 147: 195-197.
- Thoma G. and Torrisi S. (2007), Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases. Paper presented at the Conference on Patent Statistics for Policy Decision Making, 2-3 October 2007, San Servolo, Venice. Milan, Italy: CESPRI-Bocconi University WP 211 (December), available at <http://www.cespri.unibocconi.it/>

Figure 2: Gain in the name harmonization in the EPO/PCT dataset
(business applicants only, top 20 countries)

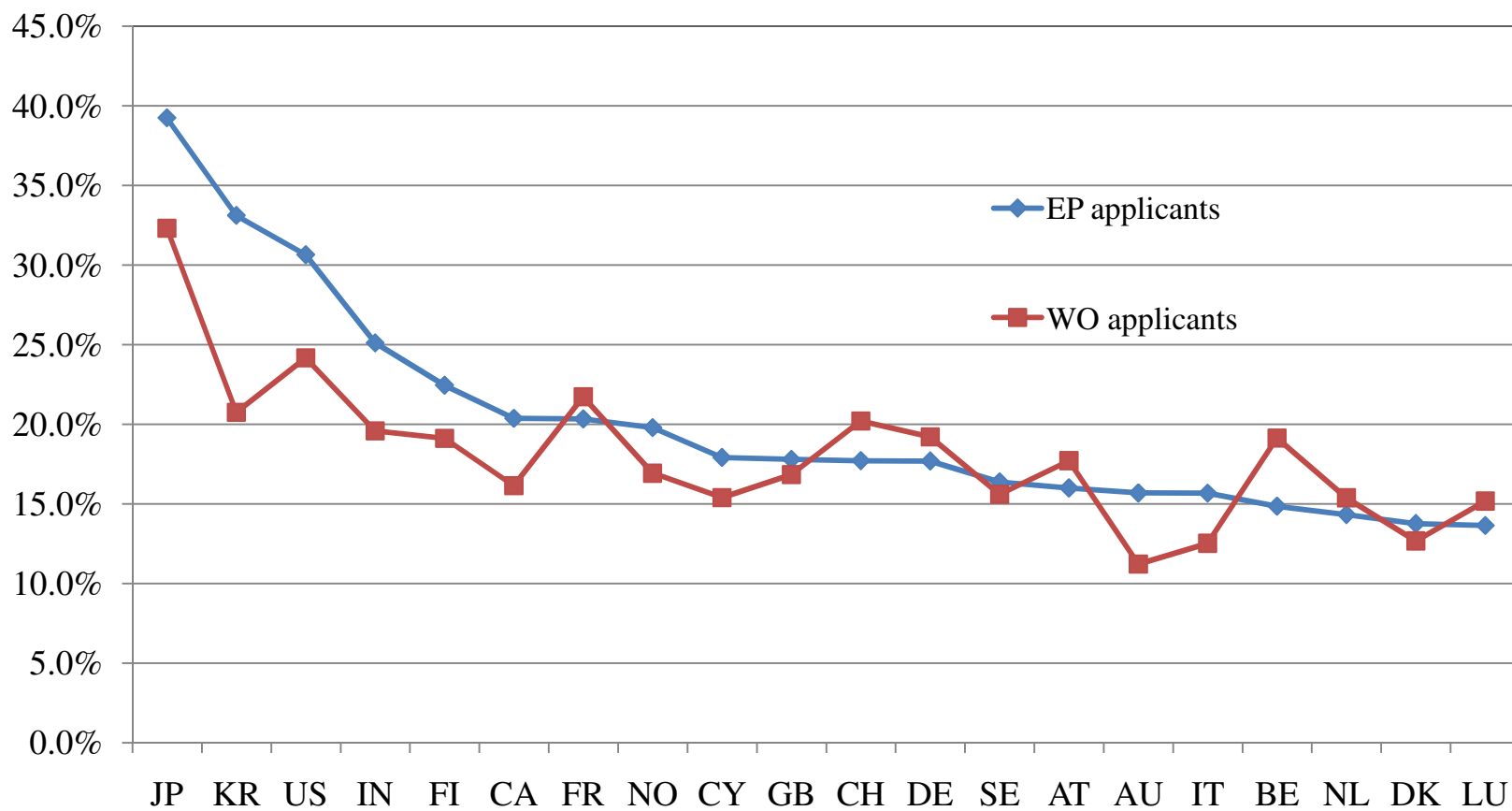


Figure 3: Percentage of matched business applicants
(top 17 EU countries)

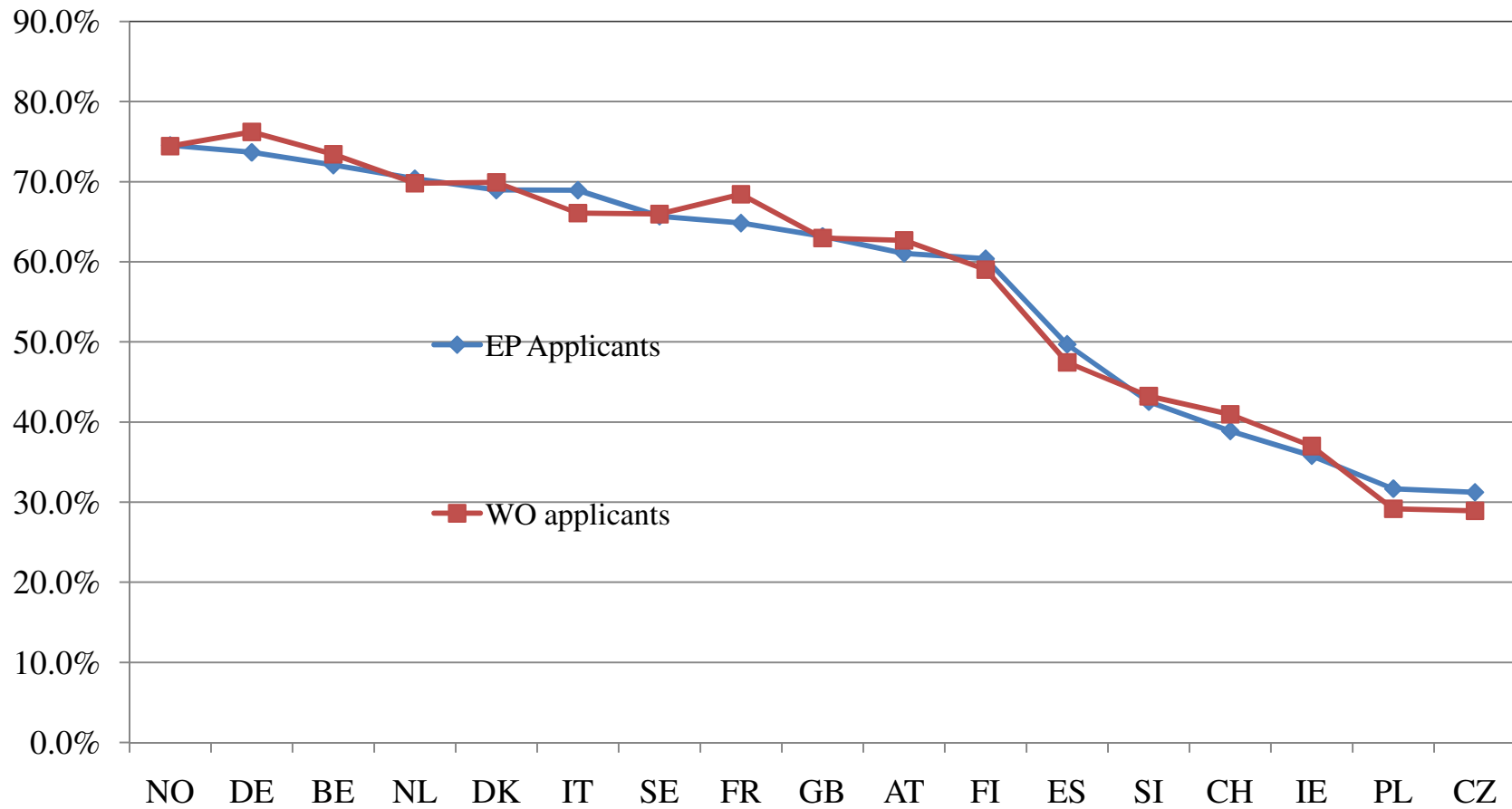


Figure 4: Percentage of matched business applications
(top 17 EU countries)

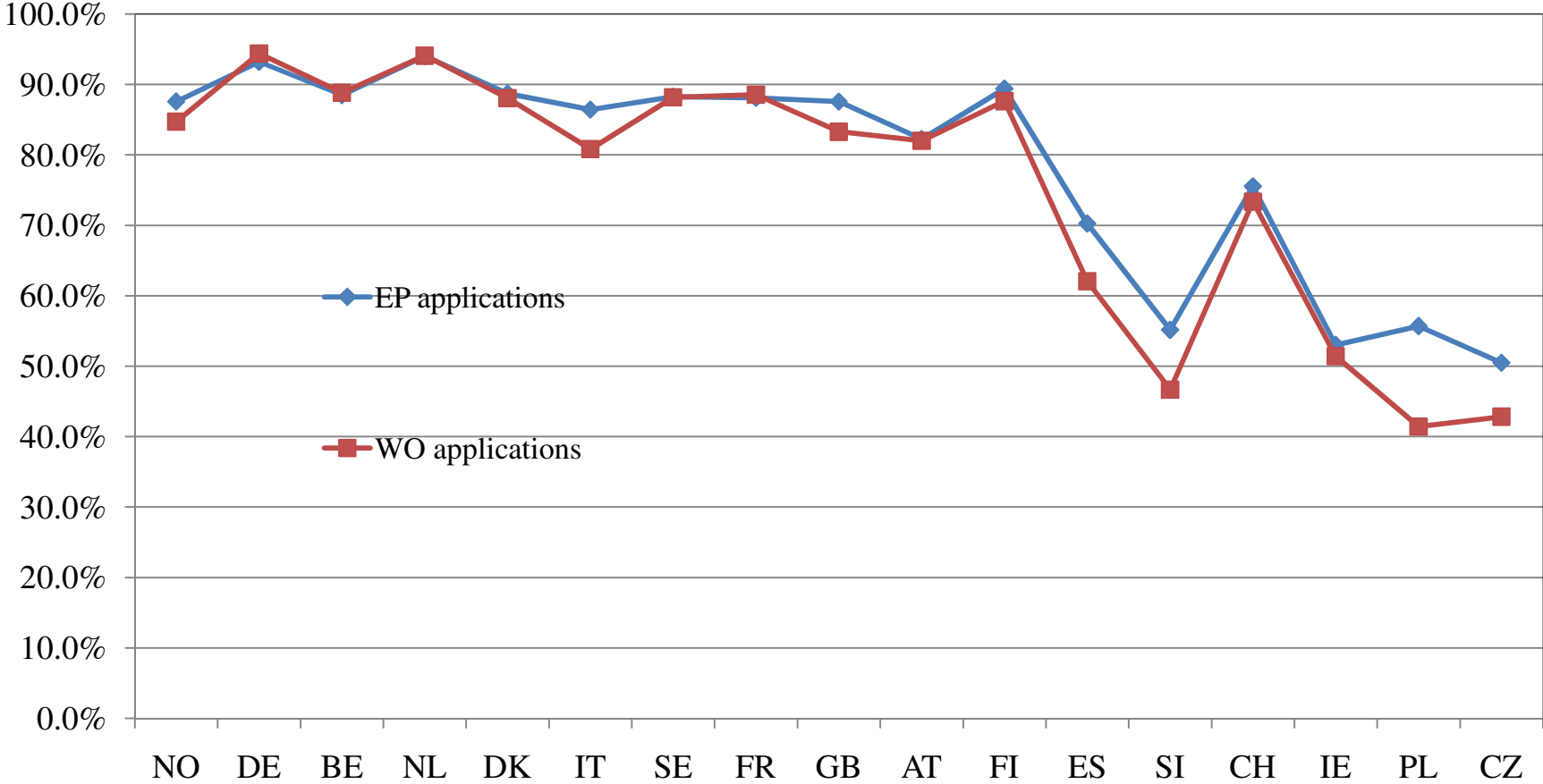


Figure 5: Distribution of the match score for EPO/PCT patentee names matched to the Amadeus business directory

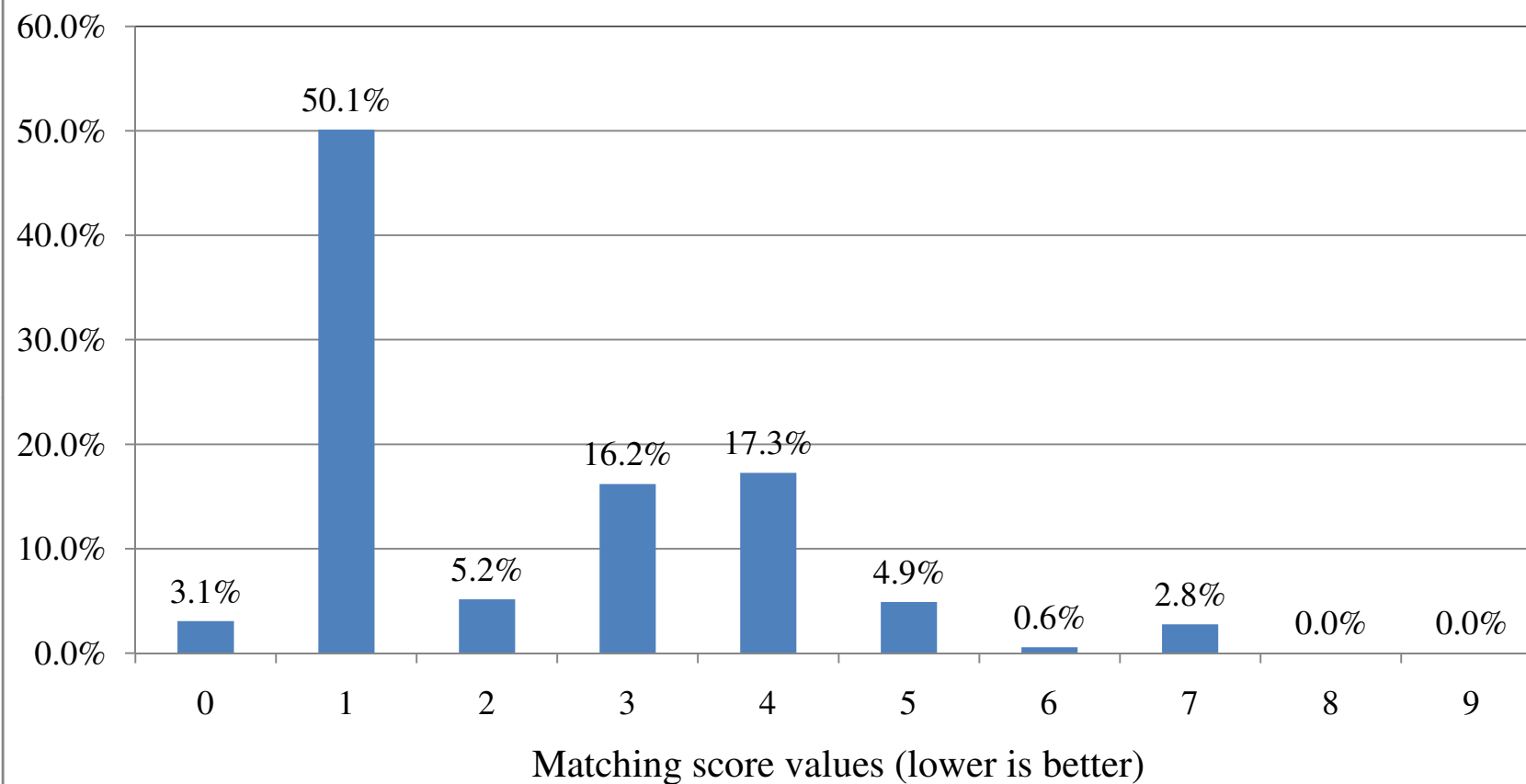


Figure 6: Share of matched business assignees in the USPTO dataset
(top 18 countries)

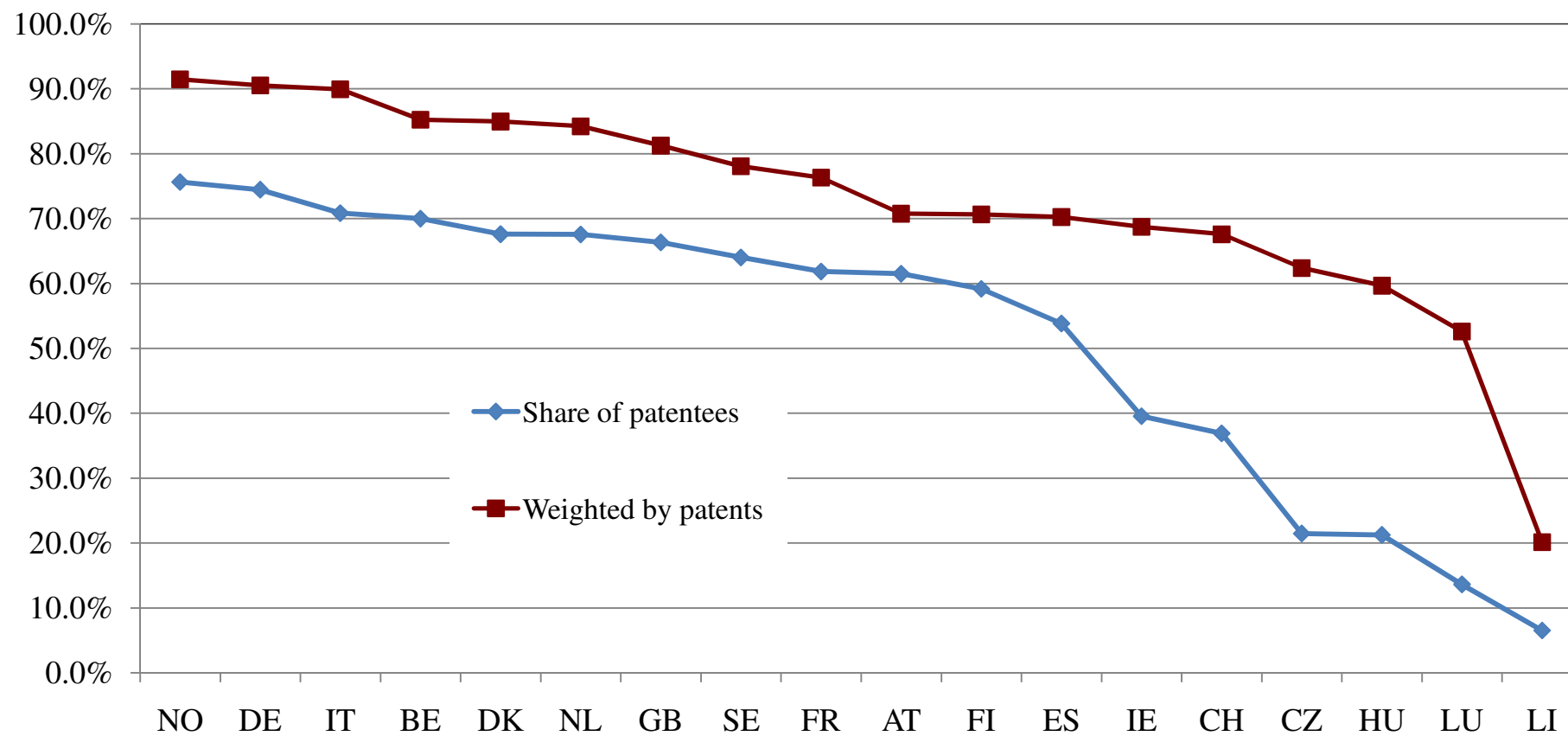


Figure 7: Distribution of the match score for the USPTO assignee names matched to the Amadeus business directory

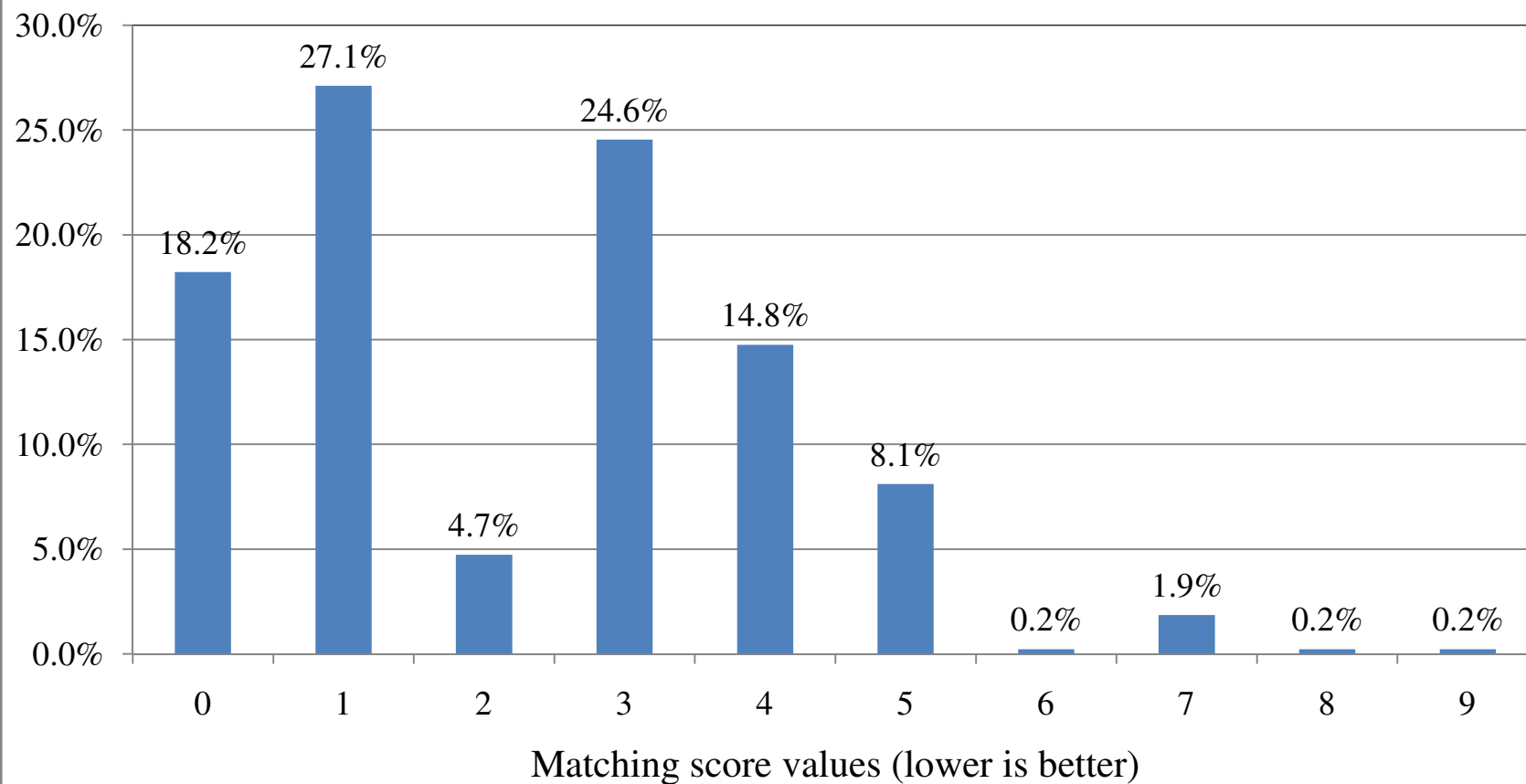


Table 3 Business applicants and applications in EPO and PCT dataset
(distinct original names)

Country	EP Applicants		WO applicants		EP applications		WO applications		Average	Average
	N	%	N	%	N	%	N	%	EP portfolio	WO portfolio
AT	3625	1.4%	2118	0.9%	16118	0.9%	8151	0.7%	4.45	3.85
AU	4999	1.9%	8937	3.8%	9561	0.5%	15897	1.4%	1.91	1.78
BE	3104	1.2%	1673	0.7%	15723	0.9%	7190	0.7%	5.07	4.30
BG	132	0.1%	121	0.1%	165	0.0%	148	0.0%	1.25	1.22
BR	0	0.0%	1	0.0%	0	0.0%	1	0.0%	#DIV/0!	1.00
CA	5720	2.2%	7166	3.1%	18595	1.0%	18467	1.7%	3.25	2.58
CH	9971	3.9%	5735	2.5%	65134	3.7%	28788	2.6%	6.53	5.02
CN	1217	0.5%	2364	1.0%	3388	0.2%	8536	0.8%	2.78	3.61
CY	134	0.1%	156	0.1%	281	0.0%	327	0.0%	2.10	2.10
CZ	303	0.1%	375	0.2%	526	0.0%	560	0.1%	1.74	1.49
DE	38742	15.0%	19736	8.5%	346196	19.5%	148602	13.5%	8.94	7.53
DK	3387	1.3%	3577	1.5%	12017	0.7%	11890	1.1%	3.55	3.32
EE	31	0.0%	51	0.0%	33	0.0%	55	0.0%	1.06	1.08
ES	3719	1.4%	2786	1.2%	7578	0.4%	4897	0.4%	2.04	1.76
FI	3234	1.3%	3475	1.5%	18651	1.0%	18182	1.7%	5.77	5.23
FR	22511	8.7%	11925	5.1%	126707	7.1%	48198	4.4%	5.63	4.04
GB	21182	8.2%	18128	7.8%	85724	4.8%	58103	5.3%	4.05	3.21
GR	316	0.1%	252	0.1%	433	0.0%	354	0.0%	1.37	1.40
HK	0	0.0%	1	0.0%	0	0.0%	1	0.0%	#DIV/0!	1.00
HR	50	0.0%	91	0.0%	169	0.0%	205	0.0%	3.38	2.25
HU	775	0.3%	847	0.4%	1763	0.1%	1525	0.1%	2.27	1.80
IE	1202	0.5%	1087	0.5%	2928	0.2%	2659	0.2%	2.44	2.45
IN	470	0.2%	756	0.3%	1539	0.1%	2926	0.3%	3.27	3.87
IT	17957	6.9%	7272	3.1%	55912	3.1%	17231	1.6%	3.11	2.37
JP	23392	9.0%	18020	7.8%	349595	19.7%	163540	14.9%	14.95	9.08
KR	2779	1.1%	5732	2.5%	22353	1.3%	17323	1.6%	8.04	3.02
KY	0	0.0%	1	0.0%	0	0.0%	3	0.0%	#DIV/0!	3.00
LI	1	0.0%	0	0.0%	8	0.0%	0	0.0%	8.00	#DIV/0!
LT	16	0.0%	24	0.0%	19	0.0%	26	0.0%	1.19	1.08
LU	667	0.3%	422	0.2%	2492	0.1%	1451	0.1%	3.74	3.44
LV	22	0.0%	38	0.0%	25	0.0%	49	0.0%	1.14	1.29
MT	51	0.0%	34	0.0%	116	0.0%	50	0.0%	2.27	1.47
NL	7858	3.0%	4768	2.1%	67422	3.8%	41602	3.8%	8.58	8.73
NO	2173	0.8%	2920	1.3%	4984	0.3%	6258	0.6%	2.29	2.14
NZ	2	0.0%	2	0.0%	15	0.0%	25	0.0%	7.50	12.50
PL	324	0.1%	363	0.2%	498	0.0%	555	0.1%	1.54	1.53
PT	253	0.1%	197	0.1%	441	0.0%	269	0.0%	1.74	1.37
RO	47	0.0%	94	0.0%	56	0.0%	103	0.0%	1.19	1.10
RU	0	0.0%	1	0.0%	0	0.0%	1	0.0%	#DIV/0!	1.00
SE	7901	3.1%	8278	3.6%	36166	2.0%	36583	3.3%	4.58	4.42
SI	162	0.1%	154	0.1%	439	0.0%	372	0.0%	2.71	2.42
SK	63	0.0%	102	0.0%	93	0.0%	150	0.0%	1.48	1.47
TR	180	0.1%	257	0.1%	471	0.0%	818	0.1%	2.62	3.18
TW	1	0.0%	1	0.0%	1	0.0%	16	0.0%	1.00	16.00
UA	1	0.0%	0	0.0%	1	0.0%	0	0.0%	1.00	#DIV/0!
US	70010	27.1%	92100	39.7%	502878	28.3%	429052	39.0%	7.18	4.66
VG	1	0.0%	1	0.0%	5	0.0%	18	0.0%	5.00	18.00
ZA	0	0.0%	2	0.0%	0	0.0%	2	0.0%	#DIV/0!	1.00
Overall	258685	100.0%	232141	100.0%	1777219	100.0%	1101159	100.0%	6.87	4.74

Table 4 Not business organization applicants and applications in EPO and PCT dataset

(distinct original names)*

Country	EP Applicants		WO applicants		EP applications		WO applications		Average	
	N	%	N	%	N	%	N	%	EP portfolio	WO portfolio
AT	125	0.7%	97	0.5%	414	0.4%	257	0.2%	3.31	2.65
AU	494	2.9%	670	3.6%	2524	2.6%	3709	3.2%	5.11	5.54
BE	326	1.9%	244	1.3%	2192	2.3%	1184	1.0%	6.72	4.85
BG	21	0.1%	8	0.0%	41	0.0%	8	0.0%	1.95	1.00
CA	702	4.1%	848	4.6%	2470	2.6%	3498	3.0%	3.52	4.13
CH	409	2.4%	355	1.9%	1757	1.8%	1227	1.1%	4.30	3.46
CN	329	1.9%	576	3.1%	668	0.7%	1776	1.5%	2.03	3.08
CY	3	0.0%	3	0.0%	3	0.0%	3	0.0%	1.00	1.00
CZ	34	0.2%	50	0.3%	66	0.1%	92	0.1%	1.94	1.84
DE	1994	11.6%	1379	7.4%	12682	13.2%	9414	8.1%	6.36	6.83
DK	126	0.7%	152	0.8%	417	0.4%	530	0.5%	3.31	3.49
EE	6	0.0%	10	0.1%	11	0.0%	20	0.0%	1.83	2.00
ES	300	1.7%	392	2.1%	898	0.9%	1652	1.4%	2.99	4.21
FI	66	0.4%	90	0.5%	283	0.3%	448	0.4%	4.29	4.98
FR	1541	8.9%	1180	6.3%	14372	14.9%	9120	7.8%	9.33	7.73
GB	1295	7.5%	1369	7.3%	6649	6.9%	7267	6.2%	5.13	5.31
GR	47	0.3%	32	0.2%	70	0.1%	48	0.0%	1.49	1.50
HR	8	0.0%	8	0.0%	20	0.0%	32	0.0%	2.50	4.00
HU	49	0.3%	48	0.3%	73	0.1%	66	0.1%	1.49	1.38
IE	55	0.3%	57	0.3%	231	0.2%	294	0.3%	4.20	5.16
IN	122	0.7%	188	1.0%	723	0.8%	1093	0.9%	5.93	5.81
IT	476	2.8%	353	1.9%	1870	1.9%	1384	1.2%	3.93	3.92
JP	1657	9.6%	1482	8.0%	7455	7.7%	9551	8.2%	4.50	6.44
KR	380	2.2%	554	3.0%	1452	1.5%	2583	2.2%	3.82	4.66
LT	4	0.0%	5	0.0%	5	0.0%	6	0.0%	1.25	1.20
LU	44	0.3%	36	0.2%	293	0.3%	132	0.1%	6.66	3.67
LV	7	0.0%	7	0.0%	15	0.0%	13	0.0%	2.14	1.86
MT	2	0.0%	1	0.0%	3	0.0%	1	0.0%	1.50	1.00
NL	432	2.5%	394	2.1%	2498	2.6%	1895	1.6%	5.78	4.81
NO	53	0.3%	78	0.4%	85	0.1%	146	0.1%	1.60	1.87
PL	117	0.7%	87	0.5%	240	0.2%	170	0.1%	2.05	1.95
PT	52	0.3%	57	0.3%	91	0.1%	114	0.1%	1.75	2.00
RO	13	0.1%	14	0.1%	18	0.0%	18	0.0%	1.38	1.29
SE	140	0.8%	150	0.8%	271	0.3%	284	0.2%	1.94	1.89
SI	20	0.1%	25	0.1%	32	0.0%	59	0.1%	1.60	2.36
SK	3	0.0%	6	0.0%	9	0.0%	10	0.0%	3.00	1.67
SU	1	0.0%	1	0.0%	4	0.0%	7	0.0%	4.00	7.00
TR	9	0.1%	9	0.0%	13	0.0%	13	0.0%	1.44	1.44
US	5773	33.5%	7615	40.9%	35405	36.8%	58235	50.0%	6.13	7.65
Overall	17235	100.0%	18630	100.0%	96323	100.0%	116359	100.0%	5.59	6.25

Notes: *It includes also those individual applicants having the suffix "Prof." in their name.

Table 5 Individual applicants and applications in EPO and PCT dataset
(distinct original names)

Country	EP Applicants		WO applicants		EP applications		WO applications		Average	Average
	N	%	N	%	N	%	N	%	EP portfolio	WO portfolio
AT	2858	2.6%	1842	1.3%	4506	3.0%	2682	1.5%	1.58	1.46
AU	2542	2.3%	5892	4.2%	2974	2.0%	6868	3.8%	1.17	1.17
BE	1438	1.3%	855	0.6%	1837	1.2%	1006	0.6%	1.28	1.18
BG	95	0.1%	225	0.2%	107	0.1%	260	0.1%	1.13	1.16
CA	2553	2.3%	4509	3.2%	3258	2.2%	5436	3.0%	1.28	1.21
CH	4167	3.7%	2783	2.0%	6190	4.1%	3690	2.1%	1.49	1.33
CN	990	0.9%	4555	3.2%	1153	0.8%	5715	3.2%	1.16	1.25
CY	26	0.0%	38	0.0%	28	0.0%	44	0.0%	1.08	1.16
CZ	194	0.2%	408	0.3%	224	0.1%	475	0.3%	1.15	1.16
DE	23600	21.2%	14223	10.1%	36550	24.2%	19892	11.1%	1.55	1.40
DK	1229	1.1%	1598	1.1%	1569	1.0%	1997	1.1%	1.28	1.25
EE	19	0.0%	53	0.0%	19	0.0%	56	0.0%	1.00	1.06
ES	2208	2.0%	2273	1.6%	2656	1.8%	2588	1.4%	1.20	1.14
FI	1075	1.0%	1696	1.2%	1347	0.9%	2171	1.2%	1.25	1.28
FR	11042	9.9%	7782	5.5%	14614	9.7%	9887	5.5%	1.32	1.27
GB	7409	6.7%	8752	6.2%	9100	6.0%	10556	5.9%	1.23	1.21
GR	466	0.4%	476	0.3%	556	0.4%	588	0.3%	1.19	1.24
HR	86	0.1%	296	0.2%	99	0.1%	351	0.2%	1.15	1.19
HU	661	0.6%	1218	0.9%	786	0.5%	1530	0.9%	1.19	1.26
IE	532	0.5%	534	0.4%	656	0.4%	652	0.4%	1.23	1.22
IL	0	0.0%	1	0.0%	0	0.0%	3	0.0%	#DIV/0!	3.00
IN	263	0.2%	917	0.7%	318	0.2%	1329	0.7%	1.21	1.45
IT	6985	6.3%	4027	2.9%	9068	6.0%	4944	2.8%	1.30	1.23
JP	5784	5.2%	7254	5.1%	9076	6.0%	10515	5.9%	1.57	1.45
KR	1929	1.7%	7740	5.5%	2333	1.5%	9657	5.4%	1.21	1.25
LT	13	0.0%	48	0.0%	13	0.0%	54	0.0%	1.00	1.13
LU	72	0.1%	44	0.0%	107	0.1%	64	0.0%	1.49	1.45
LV	26	0.0%	86	0.1%	30	0.0%	100	0.1%	1.15	1.16
MT	17	0.0%	15	0.0%	20	0.0%	19	0.0%	1.18	1.27
MX	0	0.0%	1	0.0%	0	0.0%	2	0.0%	#DIV/0!	2.00
NL	2012	1.8%	1466	1.0%	2593	1.7%	1787	1.0%	1.29	1.22
NO	874	0.8%	1376	1.0%	1045	0.7%	1693	0.9%	1.20	1.23
PL	205	0.2%	512	0.4%	237	0.2%	608	0.3%	1.16	1.19
PT	123	0.1%	105	0.1%	132	0.1%	124	0.1%	1.07	1.18
RO	63	0.1%	161	0.1%	73	0.0%	192	0.1%	1.16	1.19
SE	4196	3.8%	5279	3.7%	5419	3.6%	6719	3.8%	1.29	1.27
SI	140	0.1%	208	0.1%	178	0.1%	251	0.1%	1.27	1.21
SK	66	0.1%	156	0.1%	72	0.0%	185	0.1%	1.09	1.19
SU	1	0.0%	0	0.0%	1	0.0%	0	0.0%	1.00	#DIV/0!
TR	104	0.1%	242	0.2%	124	0.1%	285	0.2%	1.19	1.18
TW	2	0.0%	3	0.0%	2	0.0%	3	0.0%	1.00	1.00
UA	0	0.0%	1	0.0%	0	0.0%	1	0.0%	#DIV/0!	1.00
US	25226	22.7%	51274	36.4%	32225	21.3%	63944	35.7%	1.28	1.25
Overall	111291	100.0%	140924	100.0%	151295	100.0%	178923	100.0%	1.36	1.27

Table 6 Matched business applicants and applications in EPO and PCT dataset

Country	EP Applicants		WO applicants		EP applications		WO applications		Average	
	N	%	N	%	N	%	N	%	EP portfolio	WO portfolio
Not available	5	0.0%	4	0.0%	637	0.1%	66	0.0%	0.01	0.06
AN	1	0.0%	0	0.0%	2	0.0%	0	0.0%	0.50	4.08
AT	2502	1.9%	1343	1.8%	13462	1.3%	6744	1.2%	0.19	1.52
AU	226	0.2%	203	0.3%	565	0.1%	776	0.1%	0.40	3.26
BE	2526	1.9%	1238	1.6%	14081	1.3%	6423	1.1%	0.18	1.46
BG	8	0.0%	5	0.0%	5	0.0%	5	0.0%	1.60	13.04
BM	1	0.0%	1	0.0%	3	0.0%	4	0.0%	0.33	2.72
BR	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
CA	437	0.3%	325	0.4%	2415	0.2%	2039	0.4%	0.18	1.48
CC	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
CH	4329	3.3%	2372	3.1%	49654	4.6%	21334	3.8%	0.09	0.71
CN	10	0.0%	7	0.0%	46	0.0%	83	0.0%	0.22	1.77
CS	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
CY	2	0.0%	2	0.0%	3	0.0%	4	0.0%	0.67	5.43
CZ	159	0.1%	109	0.1%	267	0.0%	241	0.0%	0.60	4.85
DD	3	0.0%	0	0.0%	16	0.0%	0	0.0%	0.19	1.53
DE	31779	24.2%	15247	20.0%	326515	30.6%	142407	25.3%	0.10	0.79
DK	3016	2.3%	2518	3.3%	10702	1.0%	10534	1.9%	0.28	2.30
EE	11	0.0%	11	0.0%	7	0.0%	12	0.0%	1.57	12.81
ES	2455	1.9%	1334	1.8%	5371	0.5%	3078	0.5%	0.46	3.73
FI	2583	2.0%	2059	2.7%	16700	1.6%	15951	2.8%	0.15	1.26
FO	10	0.0%	9	0.0%	4172	0.4%	874	0.2%	0.00	0.02
FR	16627	12.7%	8244	10.8%	115648	10.8%	44582	7.9%	0.14	1.17
GB	17260	13.2%	11551	15.2%	76166	7.1%	49473	8.8%	0.23	1.85
GR	105	0.1%	71	0.1%	159	0.0%	132	0.0%	0.66	5.38
HK	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
HR	24	0.0%	15	0.0%	138	0.0%	112	0.0%	0.17	1.42
HU	187	0.1%	150	0.2%	766	0.1%	581	0.1%	0.24	1.99
IE	545	0.4%	404	0.5%	1556	0.1%	1372	0.2%	0.35	2.86
IL	1	0.0%	1	0.0%	1	0.0%	1	0.0%	1.00	8.15
IN	34	0.0%	25	0.0%	25	0.0%	64	0.0%	1.36	11.09
IT	14100	10.8%	4837	6.4%	48675	4.6%	14106	2.5%	0.29	2.36
JP	1890	1.4%	1114	1.5%	55438	5.2%	21637	3.8%	0.03	0.28
KE	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
KR	265	0.2%	165	0.2%	9190	0.9%	2950	0.5%	0.03	0.24
KY	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
LI	1	0.0%	0	0.0%	8	0.0%	0	0.0%	0.13	1.02
LT	5	0.0%	1	0.0%	6	0.0%	2	0.0%	0.83	6.79
LU	106	0.1%	58	0.1%	859	0.1%	460	0.1%	0.12	1.01
LV	4	0.0%	3	0.0%	2	0.0%	4	0.0%	2.00	16.30
MT	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
NL	6446	4.9%	3365	4.4%	63863	6.0%	39386	7.0%	0.10	0.82
NO	2427	1.9%	2187	2.9%	4386	0.4%	5325	0.9%	0.55	4.51
NZ	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
PL	191	0.1%	114	0.1%	353	0.0%	263	0.0%	0.54	4.41
PT	95	0.1%	48	0.1%	189	0.0%	91	0.0%	0.50	4.10
RO	18	0.0%	14	0.0%	15	0.0%	14	0.0%	1.20	9.78
RU	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
SE	6561	5.0%	5502	7.2%	32009	3.0%	32342	5.8%	0.20	1.67
SI	94	0.1%	67	0.1%	242	0.0%	174	0.0%	0.39	3.17
SK	27	0.0%	24	0.0%	29	0.0%	44	0.0%	0.93	7.59
SL	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
TR	2	0.0%	1	0.0%	2	0.0%	3	0.0%	1.00	8.15
TW	1	0.0%	0	0.0%	5	0.0%	0	0.0%	0.20	1.63
UA	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
US	13985	10.7%	11305	14.9%	214052	20.0%	138353	24.6%	0.07	0.53
VG	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
YU	1	0.0%	1	0.0%	2	0.0%	3	0.0%	0.50	4.08
ZA	0	0.0%	0	0.0%	0	0.0%	0	0.0%		
Overall	131065		76054		1068407		562049		0.12	0.14