NBER WORKING PAPER SERIES

# HARMONIZING AND COMBINING LARGE DATASETS – AN APPLICATION TO FIRM-LEVEL PATENT AND ACCOUNTING DATA

Grid Thoma
Salvatore Torrisi
Alfonso Gambardella
Dominique Guellec
Bronwyn H. Hall
Dietmar Harhoff

Working Paper 15851
http://www.nber.org/papers/w15851

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2010

We thank Jim Bessen, Hélène Dernis, Megan MacGarvie, Paola Giuri, Stine Grodal, Myriam Mariani, Kazu Motohashi, Teruo Okazaki, James Rollinson, Philipp Sander, Georg von Graevenitz, Stefan Wagner, Norihiko Yamano, Maria Pluvia Zuniga, and the participants at the PATSTAT Users' Meeting in Paris in March 2008, seminars at the Ludwig-Maximilians-Universität München and Università L. Bocconi in Milan for very fruitful discussions. 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. Thoma and Hall are grateful to the Kauffman Foundation for support of some of this work. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Harmonizing and Combining Large Datasets – An Application to Firm-Level Patent and Accounting Data

Grid Thoma, Salvatore Torrisi, Alfonso Gambardella, Dominique Guellec, Bronwyn H. Hall, and Dietmar Harhoff

NBER Working Paper No. 15851
March 2010
JEL No. C81,O34

## ABSTRACT

This paper discusses methods for the harmonization and combination of large-scale patent and trademark datasets with each other and other sources of data. Dictionary- and rule-based approaches to the consolidation of applicant names in patent data are presented and shown to have both benefits and drawbacks in isolation. We combine the two methods and develop a set of rules and dictionaries to consolidate European, Patent Cooperation Treaty (PCT) and US patent data with firm accounting data. The resulting data encompass about 131,000 patent applicant names from 46 countries, covering 58.8 percent of EPO applications and 50.6 percent of PCT applications by business organizations during the time period from 1979 to 2008. For US data, the resulting dataset includes around 54,000 assignee names and 51.3 percent of US granted patents during approximately the same time period.

Grid Thoma
School of Science and Technology
University of Camerino, Italy
Via Madonna delle Carceri, 9
62032 Camerino (MC)
and KITES-Bocconi University
grid.thoma@unibocconi.it

Salvatore Torrisi
Department of Management
University of Bologna, Italy
Via Capo di Lucca 34 - 40126 Bologna
and KITES-Bocconi University
torrisi@unibo.it

Alfonso Gambardella
Department of Management and KITeS
Bocconi University
Via Roentgen 1
20136 Milan
Italy
alfonso.gambardella@unibocconi.it

Dominique Guellec
OECD/OCDE
2, rue Andre Pascal
75016 Paris France
Dominique.Guellec@oecd.org

Bronwyn H. Hall
Dept. of Economics
549 Evans Hall
UC Berkeley
Berkeley, CA 94720-3880
and NBER
bhhall@nber.org

Dietmar Harhoff
Institute for Innovation Research
Munich School of Management
University of Munich
Kaulbachstrasse 45
D-80539 Munich, Germany
harhoff@bwl.uni-muenchen.de

# 1 Introduction

While innovation is frequently touted to be one of the most important drivers of economic growth, it is also one that is not observed in particular detail in standard statistics. The knowledge economy still largely escapes the grasp of statisticians, and researchers in this field have therefore experienced serious data constraints for many decades. Recently, the availability of large datasets containing information on R&D expenditures, patents and trademarks has relaxed these constraints considerably and has spurred the growth of a new wave of research.

However, the availability of such large-scale datasets has led to a new embarrassment of riches. Many researchers have put effort into the matching and consolidation of names of applicants of patents or trademarks, or have combined such data with firm-level financial data. Predictably, there is currently much duplication of such efforts. While the matching task has proven quite feasible for focused industry and technology studies, there is still no reliable and proven standard approach for larger datasets. Moreover, replication of studies is made very difficult by the lack of acknowledged standard approaches. This paper seeks to fill this gap, building on the experiences and results that the authors have achieved in separate and joint efforts. We present a methodological discussion and make the results of our data consolidation efforts available to the innovation research community via a website.[1]

Innovation studies have made good use of a variety of data types in order to evade the aforementioned data constraint. A first avenue is the collection from secondary sources of information on different qualitative dimensions of innovation. Examples include prizes as a measure of successful innovation, newswire (e.g., ad-hoc) announcements as a paper trail of collaboration among firms or of acquisitions, licensing and R&D agreements (Moser, 2005; Giarratana and Torrisi, 2006; Greenhalgh and Rogers, 2007; Fosfuri and Giarratana, 2007; Powell et al., 2000; Arora et al., 2001). By choosing data from particular contexts, sometimes using quasi-experimental features, researchers have been able to sort out competing explanations of particular innovation phenomena and to generate results with a high degree of internal validity and reliability. Recent attempts to search for well-defined quasi-experiments in particular areas (Stern and Furman, 2009; Murray et al., 2008) also belong to this group.

The second approach is based on the collection of information through surveys of innovating entities (organizations or individuals) and has produced many important datasets and results. 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 by eliciting information from R&D managers. These surveys were followed by the development of various innovation surveys in Europe, culminating in the Community Innovation Survey (Mairesse and Mohnen 2010). More recently, European 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 and within invention processes (Gambardella et al., 2008; Giuri, Mariani et al., 2007). Their example has been followed by Japanese, Korean, US and

[1] See http://www.epip.eu/datacentre.php

Australian researchers (Nagaoka and Tsukada, 2007; Nagaoka and Walsh, 2008; Um 2005; Heong et al. 2007).

Finally, the third approach is to rely on publicly available administrative and financial databases such as patent and accounting information, or in some cases on confidential firm level data that resides in National Statistical Offices, Central Banks or institutions which have been commissioned by government agencies to perform surveys and data collections on a regular basis.[2] In our application, we focus on this third type of data, which are useful because it provides comprehensive coverage of industries, technologies and countries. In addition, this kind of data is usually collected on a regular basis allowing for replication and updating of studies.

The distinction made here may be overly stylized. Indeed, data of different types and from different sources are often combined. For example, the inventor surveys undertaken in various countries depend critically on inventor information as contained in official patent data. Moser's 2005 study combines historical patent data, entries in exhibition catalogues and innovation award data. Moreover, it is frequently the case that innovation-related data is complemented with information from accounting sources or other forms of financial data. In most cases, a large number of observations in each dataset need to be merged in the absence of a common identifying code, which means that researchers have to rely only on the entity names to do the match. The problems that arise in this case and possible ways to solve them are the subject of this paper.

The matching and harmonization problem we address is usually present for any of the datasets described in the above, and it becomes more complex as the size of the dataset (the number of entities) increases. It is important to note that the problem may also exist even if just one type of data is being used. Variations in the spellings of names routinely occur within large-scale datasets such as patent and trademark data.[3] The problem can usually be handled manually in smaller datasets, but requires some form of automated approach in larger datasets. The exact nature of the problems to be addressed is described in section 2.3.

Turning to the types of data of particular interest to us, we note that the most often encountered ones are i) data on R&D expenditures as collected by statistical agencies or revealed in accounting data and stock market reports; ii) data from regularly performed innovation surveys; iii) data on intellectual property rights, such as from patent and trademark offices; iv) data on firm inputs and value added, such as is collected by various national statistical agencies; and v) corresponding data on the financial status of firms, such as P&L data, balance sheet data and so forth. We discuss these here before turning to IP and financial data as our particular focus in this article. This list is not complete, but presumably covers the most frequently used data for which a name matching problem occurs.

---

[2] What distinguishes the second and third approach is that in the second case, data collection often represents a singular research effort yielding a cross-sectional database. The results are of considerable value, but typically not suited to provide representative data on innovation over time.

[3] For example, uncleaned USPTO patent data contain several hundred versions of the name of IBM Corporation.

**R&D data**. Data on R&D expenditures have been collected for many decades by statistical agencies and other entities commissioned by government agencies.[4] The foundation for these surveys is the Frascati Manual (OECD 2002) which contains definitions of R&D and related terms and detailed hints for the surveys. The results of R&D surveys are summarized on an annual basis in national and in OECD reports, but the firm-level data are usually not available publicly for use in micro-econometric studies. Under US and UK accounting rules, firms listed in the stock market have to report their R&D expenditures if they are "material" and under revisions to the International Accounting Standards Board, the practice of reporting such data is spreading. However, especially in the case of European firms, data on R&D expenditures are often missing because reporting these expenditures is not currently required by accounting and fiscal regulations. The definition of R&D used for accounting purposes can also differ from that in the Frascati Manual; for the US example, see Hall and Long (1999).

**Innovation survey data**. In many European and non-European countries these are now being undertaken on a regular (usually annual) basis. Innovation surveys seek to elicit information not only on R&D, but also on the broader innovation process as described in the OECD/Eurostat - Oslo Manual (2005).[5] Although R&D is a good indicator of the commitment of a firm to inventive activities because it is chosen by the firm, it does not tell us much about their 'success', technical and economic, which is rather reflected in the notion of innovation (which is an invention reaching the stage of implementation, either as a new process, a new product or a new marketing approach). Moreover, not all inputs into innovation processes are covered by the classical Frascati Manual definition.

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.[6] 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 confidentiality laws in all countries. Innovation survey data continue to generate important findings, but the difficulty of matching CIS databases to other databases have limited their use for the purpose of research in economics, management and public policy (Mairesse and Mohnen, 2010).[7]

---

[4] In Germany, the R&D surveys are performed by Wissenschaftsstatistik GmbH, a branch of Stifterverband, an association of German industry which supports research and science. In this case, the performer of the survey is not an official statistical agency. In most other countries such as Austria, the UK, France, Italy and the Netherlands, the statistical agencies carry out the surveys. In the US, the survey is carried out by the Bureau of the Census under a contract from the National Science Foundation.

[5] Available at the following link http://www.oecd.org/dataoecd/35/61/2367580.pdf

[6] See Arundel (2001) for details. 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. The United States does not perform a government innovation survey, although similar private NSF-funded surveys have been done in the past (Levin et al. 1987; Cohen et al. 2000) and a major new initiative is now underway, directed by Wesley Cohen.

[7] There are a few exceptions to this rule. The German government has commissioned innovation surveys on a regular (annual) basis (of which the CIS survey is a part) and the resulting cross-sectional and panel data have been used in a large number of studies. This dataset has also been combined with patent data and other external information. See the survey by Janz et al. (2001) for a detailed description.

**Patent data**. Due to the wider availability of computer-ready datasets, an increasing number of studies use 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, but they constitute a detailed measure of innovation (Griliches, 1981 and 1990; Pavitt, 1988). 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 (Harhoff et al. 2003, Hall et al. 2005).

**Trademark data**. Comprehensive studies on the economic role of trademarks are still rare, but recent studies confirm that trademarks can be economically important, explaining a significant share of the market capitalization of firms (Greenhalgh and Rogers, 2006; Sandner 2009). Trademarks have also been used as proxy for firm's diversification activities (Mendoça et al, 2004), processes of market entry and survival (Fosfuri and Giarratana, 2007), appropriability strategies (Graham, and Somaya, 2004) and firm's reputation and advertisement efforts (von Graevenitz, 2004).

**Accounting and other economic data**. In some cases, the isolated use of any of the data may be sufficient to undertake a study on innovation. But frequently, R&D, innovation survey, or IP data need to be combined with accounting or financial data for the firm in question, especially if we wish to measure outcomes such as productivity or profitability. Databases that contain such data (Amadeus, Compustat, Reuters and others) usually have their own system of entity identification that does not match up with the sources of IP and innovation data.

Therefore, studies that seek to employ data from different sources (for example, financial data jointly with patent and trademark data) have to merge multiple types of firm identification. Since inaccuracies in data merging and integration can lead to measurement errors, biased results or lack of sufficient statistical power, correct matching is an important but neglected issue. 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 must rely only on the names of the firms to combine datasets.

**Earlier solutions** 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 a small amount of manual matching have limits when confronted with the large datasets that are common today.

Researchers are therefore forced to apply one or several automated methods. 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 patent-holder 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. Doing a good match requires using a combination of the two approaches, as neither one is sufficient on its own.

The **contribution of this paper** is 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 utilizes recent advances in 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 use these methods in empirical studies in economics and management. We do not rely on a single NER approach, but experiment with several techniques in parallel. We find that a combination of dictionary-based and rule-based approaches produces the most favorable results in our application.

The paper is organized as follows. In section 2, we start by describing the data sources currently available for large-scale studies on patenting and other IP information. Using examples from these data, we then describe various types of problems researchers are facing. In section 3, we then describe two fundamental approaches to solving the problem: 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 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.

In section 4, we describe the software prototypes for the different approaches analyzed in the section 3. In this section, we combine the different approaches in order to achieve reliable results for the problem at hand. Finally, in section 5, we conclude by documenting the harmonization and matching that results from the methodology suggested in this paper and assessing the type I and type II errors thus achieved. Section 6 concludes.

## 2 Frequently Used Data Sources and Matching Problems

In this section we briefly review the sources of patents and other IP data frequently used in economic studies. In particular, we will consider their content, time coverage, mode of access, complementary search and management tools and potential integration with other sources. We also provide a description of typical matching problems arising from the use of these data.

### *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 others) 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 patenter'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.

*IIP Japan database*

This database contains information on all 9 million published Japanese patent applications between 1964 and 2004, along with 2.6 million patent registrations (grants). The information provided consists of the application and registration numbers, dates of application, exam request, grant, and expiration, the number of claims, IPC codes, applicant and rights holder information including geographic location, and citation links. The dataset is provided freely online and documented in Goto and Motohashi (2007).[8]

*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.[9]

---

[8] See http://www.iip.or.jp

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

**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 |

Source: Own analysis based on http://ep.espacenet.com/?locale=en_EP

*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[10] classification codes, status, filing date, registration date, date of international registration, publication date, expiry date etc.

### *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 US stock markets. The linking was done mostly manually and it involved about 2,700 US corporations and their subsidiaries as reported in the year 1976. The authors used this dataset to estimate patent production functions and valuation equations.

[10] See http://www.wipo.int/classifications/nice/en/classifications.html (last download Oct. 15th, 2009) for details on the classification.

*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 allowing a surge of new wave of research in innovation studies. For a partial set of contributions, see Jaffe and Trajtenberg (2002) and Cockburn et al. (2004).

The database comprises detailed information on almost 3 million US patents granted between January 1963 and December 1999. Available data fields in the NBER database consist of four main building blocks. First, it includes application and publication numbers and dates and technological classifications. Second, detailed and harmonized information is supplied on inventor names, address, city, zip code, and state and country code. 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).[11] Recently these data have been updated to 2006 by Cockburn and co-workers at the NBER (Cockburn et al., 2009).[12]

The fourth building block consists of the citation links, in particular all US patent citations made to these patents between 1975 and 1999, constituting over 16 million citation links. Although useful for the analysis of US-based questions, the drawback to using the NBER citation 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*

Creation of a worldwide statistical patent database was initiated by the OECD task force on patent statistics; in response, PATSTAT was developed by the EPO in 2005 on the basis of the DOCDB database. It includes bibliographic details on patents filed to 80 patent offices worldwide, covering more than 60 million documents. Hence filings in all major countries and the PCT filings at the World International Patent Office are covered.[13] Available fields of PATSTAT are listed below and a fuller discussion of these indicators can be found in (OECD, 2009):[14]

- Application and publication details such as authority, number, kind and dates;
- Technical information and descriptions such as title, abstract, international and national classification;

[11] 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/

[12] Beta versions of the new datasets are available at https://sites.google.com/site/patentdataproject/Home .

[13] Having complete coverage of PCT-WIPO filings is important because their counts allow for interesting international comparisons.

[14] A comprehensive data catalog is provided along with the tables, describing the fields' codes. An additional document lists the currently available fields and the time period covered for each country.

- Applicant and inventor name, address, and country code;
- Identification of claimed priority, designating international application, parent application and technically related application;
- Identification of cited publication including patent and non-patent prior art, category and origin of the citation.

PATSTAT is released by the EPO twice a year, early spring and early autumn. Each version presents a snapshot of the source databases at a single point in time.[15] However, the content and design of PATSTAT is not intended to be static: a "change management procedure" has been put in place by the EPO to allow task force members to request changes in the data catalog (i.e. including additional information, variables, etc.) within a reasonable time before each release. Currently around 100 institutions have subscribed to PATSTAT and this database is expected to be widely used for innovation studies.

### *2.3 Typical Matching Problems*

The matching and consolidation problems faced by researchers typically fall into three broad classes:

1) variations in spelling, some of which are simply typographical errors, *within a given list or database*; see the examples in Box 1.
2) variations in the way names appear *in two or several different lists*, in many cases caused by different naming conventions; see the examples in Box 2.
3) the problem of matching firm subsidiaries in one list to ultimate owners in another (Box 3). This problem is especially important when matching firm financial data (reported at a consolidated level) to patent data (which is often at the subsidiary level).

**Box 1. Different spellings and misspellings**

SYNRES INTERNATIONAL B.V.
SYRNES INTERNATIONAL B.V.

BSH BOSCH UND SIEMENS AKTIENGESELLSCHAFT
BSH BOSCH UND SIEMENS AKTINGESELLSCHAFT
BSH BOSCH UND SIEMENS HANSGERAETE GMBH
BSH BOSCH UND SIEMENS HAUS-GERAETE GMBH
BSH BOSCH UND SIEMENS HAUSERATE GMBH

MINNESOTA MINING AND MANUFACTURING COPANY
MINNESOTA MINING AND MANUFACTURING COPMANY

[15] Much of the data is extracted from the EPO's master bibliographic database – DocDB, also known as the EPO Patent Information Resource. However, depending on the patent office, the coverage of data may be partial or delayed over time.

MINNESOTA MINING AND MANUFACTURING CORP

INTERNATIONAL BUSINESS MACHINES, CORPORATION
INTERNATIONAL BUSINESS MACHINES CORPORATION
INTERANTIONAL BUSINESS MACHINES CORPORATION
INTERANATIONAL BUSINESS MACHINES CORPORATION
INTERANTIONAL BUSINESS MACHINES CORPORATION
INTERNAIONAL BUSINESS MACHINES CORPORATION
INTERNAITONAL BUSINESS MACHINES CORPORATION
…..and so forth

RÜTGERSWERKE AKTIENGESELLSCHAFT
RUTGERSWERKE AKTIENGESELLSCHAFT
RUETGERSWERKE AKTIENGESELLSCHAFT
R_TGERSWERKE AKTIENGESELLSCHAFT

REGIE NATIONALE DES USINES RENAULT Société Anonyme
REGIE NATIONALE DES USINES RENAULT (Societe Anonyme)
REGIE NATIONALE DES USINES RENAULT Societé Anonyme
REGIE NATIONALE DES USINES RENAULT (SociTtT Anonyme dite)

**Box 2. Variations in naming conventions**

ABITIBI PRICE CORPORATION
ABITIBI PRICE INC

SMITH (JOHN) LTD
JOHN SMITH LTD

INTERNATIONAL BUSINESS MACHINES – IBM
IBM CORP. (INTERNATIONAL BUSINESS MACHINES)
IBM CORPORATION (INTERNATIONAL BUSINESS MACHINES)

MINNESOTA MINING & MFG CO
3M CORP
MINNESOTA & MINING MANUFACTURING

FLUID COMPONENTS INTL.
FLUID COMPONENTS INTERNATIONAL

SENETAS CORP. LTD. (USA)
SENETAS CORP.

**Box 3. Assignment to aggregate entities (ownership issues)**

| Subsidiary | Ultimate Owner |
|---|---|
| ADHESIVE TECHNOLOGIES INC | MINNESOTA MINING & MFG CO |
| AVI INC | MINNESOTA MINING & MFG CO |
| D L AULD CPY | MINNESOTA MINING & MFG CO |
| DORRAN PHOTONICS INCORPORATED | MINNESOTA MINING & MFG CO |
| EOTEC CORPORATION | MINNESOTA MINING & MFG CO |
| NATIONAL ADVERTISING CPY | MINNESOTA MINING & MFG CO |
| RIKER LABORATORIES INC | MINNESOTA MINING & MFG CO |
| TRIM LINE INC | MINNESOTA MINING & MFG CO |

# 3 Methods for name matching

In this section we describe 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, biological science has become increasingly concerned with the analysis of large amounts of information. Consequently, the way that information is stored, managed, visualized, and searched has increased in importance. 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 is hard-to-find information. The NER approach has the potential for interesting 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 the two different approaches to dataset merging: the dictionary-based approach, which relies on the collection of large datasets of names and name variants, and the rule-based approach which builds a set of rules for similarity links across different entity names. The latter uses some of the methods gleaned from bioinformatics.

## *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. Moreover, the matching is only available to subscribers of the Derwent service.

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 this procedure.

Another example of a dictionary of patenter names is the USPTO CONAME file compiled by the USPTO. This uses 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 can be matched by investigating the names and locations of the inventors. The CONAME file accounts for changes or variations 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 developed 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 Katholike Universiteit Leuven (KUL) have developed an automatic methodology based on the detailed standardization of patenter 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 patenter 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 US 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.[16] 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 in the current (2009) version of the database is still incomplete, especially that involving non-US 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.

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 deleting (Levenshtein, 1966). An extension of the Levenshtein edit distance was developed by Smith and Waterman (1981). The main difference is that character mismatches at the beginning and the end of strings are ignored in the calculation of distance. For instance, the two company names 'Dr Michal White Plc' and 'Michael White Plc, Dr' have 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:

---

[16] 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/patentdataproject/Home .
The match documentation is at http://www.nber.org/~jbessen/matchdoc.pdf .

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

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 up the strings into words at the blank spaces. The *J* token distance is simply given by the number of common tokens in two names and the count of total number of tokens in those names, that is:

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

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:

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

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 the 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:

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

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 Enhancing the value of existing 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 the string and token similarity functions described earlier. In this section we discuss an additional source of standardization of patent holder 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.[17] A patent family is sometimes also defined as all patents that protect the same basic invention. The rapid growth in the number of patent documents in recent years have been accompanied by the growth of priority links and resulting patent family size, and hence the total number of inventions has grown a than the total number of patent documents.

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.[18] 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 patent-holders 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, using the PATSTAT database we can identify the priority links from and into the EPO patent database; in particular we rely on the INPADOC patent family definition. In TASK 4 we use the priorities compiled from PATSTAT by linking each EPO application to the US priority date

---

[17] 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.

[18] It is important to note that this assumption will not always hold. A patent-holder with headquarter in one region may decide to sell patents held in another region. We have no firm data on the extent to which this is the case. However, our manual inspections of the data have shown that in the vast majority of cases, the patents are under the ownership of the same entity or at least the same multinational firm (with subsidiaries in the respective regions).

patent via the publication number. Finally, we deal with the identification of the proper link to the EPO patent-holder names in TASK 5, which takes account of the number of priority links and number of patenters per EPO patent. We used a string similarity algorithm and also manual checking to ensure the proper association across the USPTO assignee codes and EPO applicant codes.

| **TASK 1:** Retrieval of USPTO assignee names<br><br>Source: USPTO standardized assignee names file<br><br>*3,758,421 patents 238,344 names* | **TASK 2:** Retrieving the_USPTO patent publication numbers<br><br>Source: PATSTAT April 2009 Table tls211<br><br>*11,132,167 patent documents* | **TASK 3:** Merging using priority links across EPO & USPTO patents<br><br>Source: PATSTAT April 2009 Table tls219<br><br>*4,492,601 priority links* | **TASK 4:** Retrieving the EPO patent publication numbers<br><br>Source: PATSTAT April 2009 table tls201<br><br>*2,309,079 patent application documents* | **TASK 5:** Merging to obtain EPO/PCT standard names<br><br>Source: EPOLINE source files April 2008<br><br>*3,480,628 patent documents 588,525 names* |
|---|---|---|---|---|

**Figure 1.** Harmonization tasks based on priority links: data and sources

The final list of EPO applicant codes with a US standardized name includes around 158 thousands patenter names corresponding to about 70 thousands names on the USPTO CONAME file. This USPTO/EPO dictionary contains about 77.1% of the EPO patent applications filed by business organizations 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 name codes file by using the USPTO CONAME file.

# 4 Implementation

## *4.1 Software prototype*

In this section we summarize the structure of the software prototype for the creation of the dataset.

The **first software module** regards the cleaning phase, i.e., the development of a dictionary. We built on the previous contributions by Magerman et al. (2006) and Cockburn et al. (2009),

adding some additional cleaning operations. The final list of operations is summarized by the following sequence:[19]

1. Transformation into upper case to simplify matching. Addition of a blank space at the beginning and end of the string to facilitate word-based tests.

2. SGML and HTML codes substituted by the ASCII/ANSI equivalent, such as for example "&OACUTE;" replaced by "O" etc.

3. Proprietary character codes replaced by the ASCII/ANSI equivalent.

4. Each of the accented characters is replaced by its unaccented version.

5. Removal of frequent comma, double quotation mark irregularities and other period irregularities and non-alphanumeric characters.

6. The conjunction "and" and its translations into other languages are standardized as "&".

7. Umlaut harmonization by reducing variations such as "ue", "ae", and "oe" to respectively "u", "a", and "o".

8. Removal of common company words like INC and AB in descending order of their length.

9. Replacement of spelling variations with their harmonized equivalent for some frequent words (such as INTL for INTERNATIONAL and its variants).

10. Removal of the round parentheses and cleaning their content; typically this content consists of geographical information or former company names.

11. Removal of multiple blank spaces, replacing with a single space.

12. Generation of a unique list of patenters by removing duplicates after cleaning.

13. Linkage and indexing to the USPTO[20] and EPO standard name codes.[21] The goal of this operation is to achieve consistency and interoperability of the harmonization and matching files generated in this paper and future updates of the USPTO and EPO patent files that rely on these codes.

The **second software module** deals with the matching phase. We relied on a rule-based approach based on the approximate string-matching algorithms discussed above and in Thoma and Torrisi (2007). In particular we adopted the Jaccard weighted distance operator as

---

[19] We implemented these operations in a Java software prototype. An equivalent SQL query including these operations is available under request.

[20] We extracted USPTO assignee names the so-called CONAME file. Source: USPTO, CD version March 2007.

[21] The source of the EPO standard name codes for EPO and WIPO/PCT patent dataset is the weekly EPOLINE files (up to July 2008 in our case). http://ebd2.epoline.org/jsp/ebd1.jsp

described by equation (3). Compared to the edit distance this operator has several improvements. Firstly, it is easier to deploy in software and it is characterized by a faster computation time. Secondly, it is more conservative than edit distance in establishing a similarity link across two different names because it allows for variation across tokens but not those within the tokens: we think that in the case of business entity names variation across tokens are more frequent with respect of those within tokens. Indeed we operated a deeper cleaning phase in the first software module – compared to by Magerman et al. (2006) and Cockburn et al. (2009) - in order to solve the business entity name variations within tokens. Thirdly, the operator described by equation (3) ensures an interesting trade-off of statistical relative weights for discriminating and non-discriminating tokens which makes it more suitable than edit distance for solving name variations as those described by Box 1 and 2.

The **third software module** consists of some fine tuning operations included the refinement of predicted matching candidates, the resolution of abbreviations and of multiple matching occurrences of the same patenter, and formatting of the output files to be easily processed by electronic spreadsheets and statistical software packages.

### *4.2 Dictionary creation*

We used our software prototype to create and integrate a large dataset of patenters originating from a high number of countries. 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 2 reports the country distribution of the business applicants and applications in the EPO/PCT dataset.
- Table 3 reports the country distribution of the non-business organization (NBO) applicants and applications in the EPO/PCT dataset.
- Table 4 reports the country distribution of the individual applicants and applications in EPO/PCT dataset.[22]
- Figure 2 reports the distribution across the top 18 countries of the reduction in the number of applicants after name harmonization using the software prototype described in the previous section. The overall reduction of the size of the dictionary is about 28.8%.

*[Tables 2-4 about here]*

*[Figure 2 about here]*

[22] Unlike the USPTO, the EPO does not provide a standardized identifier for this kind of applicant. Typically the names of individual applicants are supplied in the format "Surname", "First name" "Middle name(s)" if any. Hence, the identification of the individual applicants has been based on a stepwise heuristic procedure. Starting from all the applicants that include a "," in their name and have more than one token, we first excluded those applicants which had a typical common company word in their name such as INC, LTD and AB. Second, we removed those applicants with tokens related to non-business organizations. Third, we removed applicants with generic words such "technology", "system", etc. Fourth, we inspected all the applicants with more than two tokens and a case sensitive format such as "Smith, John" (for example) by hand.

### *4.3 Matching with business directories*

In this section we report the results of the merge of patenter names with business directories. In particular, we retrieved business and ownership information for the companies from Amadeus by Bureau Van Dijk, which collects information from approximately 10 million European firms and their subsidiaries at the worldwide level.

We extracted all the historical information included in the Amadeus CD version files during the period 1998-2006. For each December release we retrieved information on:

- company demographics such as country, city, region, zip code, date of incorporation, industry 3 digits core code.
- unique identification number, which relies on national company identifiers such as VAT number, Chamber of Commerce numbers, etc.
- ownership structure such as subsidiaries and shareholders
- changes in the company names and some additional information.[23]

We start with the EPO/PCT applicant 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 the patent documents in PATSTAT.

In matching EPO/PCT patenter names to the business directories we focused only on the business patenters, which constitute about 63.4% of the patenter names in the EPO and about 54.8% in the PCT system. Overall they encompass about 337 thousands original names that have been harmonized to about 240 thousands names according to the dictionary described in the previous section. About 43.3% of the patenters have just one application in the EPO and about 0.6% have more than 100.

We matched patenter names only if they also came from the same country, that is, the same nationality of the patenting entity in the EPO/PCT dataset and company in Amadeus. The results of the matching to the Amadeus business directories are depicted in the following figures:

- Table 5 and Figures 3a and 3b report the share of the business applicants in the EPO and PCT dataset that have been matched to Amadeus.
- Figures 4a and 4b report 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.

We also matched the USPTO patenters to the Amadeus dataset. This task was in two steps: in the first we identified the matched EPO/PCT patenters (Table 5) that have been active also in the USPTO using equivalents link file discussed in section 3.3. In the second step, for the

[23] Additional information has included website, emails, telefax numbers, and sales size class.

residual USPTO patenter names not matched using this file we matched directly to Amadeus using the software prototype described in section 4.1.

- Figure 5 shows the share of the business assignees in the USPTO dataset that have been matched to Amadeus, unweighted and weighted by their number of patent applications.

*[Table 5 about there]*

*[Figures 3, 4, 5 about here]*

# 5 Robustness checks and quality analysis

In this section we present some quality measures for the dataset developed in the previous section in the hope of quantifying potential false positives (Type I error) and false negatives (Type II error). A full-fledged analysis of this topic is beyond the scope of this paper and will be addressed by future research. Indeed, it requires not only a broader theoretical discussion about the entity matching process but also a deeper investigation of the generation of business directories such as Amadeus. Phenomena that potentially could influence the matching process include: relocation and reincorporation of the business activities, a firm's ownership structure, mergers and acquisitions, and others. In this direction we limit the discussion to some aspects that could enable a more efficient use of the dataset developed in section 4.

## *5.1 False positives*

For every matched pair we computed a quality of match score based on the similarity of the name and location of the information in our patent data and the data 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 6**
**The quality of entity match score**

| Score | Name | Location |
|---|---|---|
| 0 | Manual check | Manual check |
| 1 | Similarity >= 50% | Same |
| 2 | 30% <= Similarity <= 50% | Same |
| 3 | Similarity >= 50% | Unknown |
| 4 | Similarity >= 50% | Different |
| 5 | 30% <= Similarity <= 50% | Unknown |
| 6 | 10% <= Similarity <= 30% | Same |
| 7 | 30% <= Similarity <= 50% | Different |
| 8 | 10% <= Similarity <= 30% | Unknown |
| 9 | 10% <= Similarity <= 30% | Different |

For the EPO/PCT dataset the distribution of the match quality scores is reported in Figure 6. 90% of the applicants are characterized by a high matching score, that is a value less than or equal to 4. Similarly, Figure 7 shows the distribution of the match quality scores for the USPTO dataset, where a similar proportion of 89% have a matching score less than or equal to 4.

*[Figures 6 and 7 about here]*

A second quality measure is given by the patenting lag, that is the number of years since the birth date of the patenting firm before it files its first patent application. We would expect this lag to be greater than or equal to zero. A large negative patenting lag could be a symptom of potential false positive matching. However, performing this check is hampered by the fact that the company birth date information is not usually reported in business directories such as Amadeus. Often business directories give the date of incorporation, which is when the company took the limited liability legal form. This date can be later than the founding date if the business venture started with self-employment or other legal company forms.

In spite of this limitation we created a patenting lag distribution relying on the incorporation date. This lag was computed as the difference between the filing year of the earliest patent and the incorporation year in Amadeus plus one year. Table 7 reports the distribution of the patenting lag for the different levels of the match score. Overall we have a negative patenting lag for about 7.4% of all the matched patenters and for 6.1% of the patenters that are active after year 2000, which is when Amadeus began its very wide pan-European coverage. Moreover, for patenters with a low match score there are visibly smaller proportion of negative patenting lags (values 1, 2 and 3). On the one hand, this finding shows the validity of the matching performed in section 4.3 and usefulness of the match score indicator. On the other hand, it invites further investigation of the matched names having a negative lag in order to check for the presence of potential false positives.[24]

**Table 7**
**Negative patenting lag by different levels of the match score**

| | **All Patenters** | | **Patenters active after 2000** | |
|---|---|---|---|---|
| **Score** | **obs** | **%** | **obs** | **%** |
| 0 | 1,385 | 13.9% | 1,188 | 10.9% |
| 1 | 32,748 | 5.2% | 29,748 | 4.6% |
| 2 | 3,819 | 9.1% | 3,470 | 8.5% |
| 3 | 1,189 | 10.0% | 1,081 | 6.8% |
| 4 | 11907 | 19.9% | 9,846 | 7.4% |
| 5 | 164 | 25.0% | 138 | 18.1% |
| 6 | 390 | 13.1% | 360 | 11.4% |
| 7 | 2,098 | 15.2% | 1,745 | 12.6% |
| 8 | 4 | 0.0% | 4 | 0.0% |
| 9 | 0 | 0.0% | 0 | 0.0% |
| Overall | 53,704 | 7.4% | 47,580 | 6.1% |

A third and final analysis of the Type I error problem was performed by analyzing manually a sample of a hundred firm names drawn randomly from among the European business applicant names at the EPO.[25] The dataset developed in section 4 identified 76 matched

[24] A preliminary analysis revealed that negative patenting lag was often associated with the re-incorporation of a company after the decision to relocate its business activities or with a merger or acquisition.

[25] The harmonization and matching operations were performed only within the sample of business applicant names at the EPO. Hence we used the same population to draw a random sample.

applicants to Amadeus, whereas 24 names were not matched. Out of these 76, three were false matches, that is 3.9%. In particular, they were the following:

i) Patenter DRALORIC ELECTRONIC GMBH matched with company ELECTRONICS (DE715000111). Location zip or city is unknown (match score 3).

ii) Patenter COPRECI S COOP LTDA matched with company BATZ S COOP (ESF48037600). Location zip or city is matched (match score 1).

iii) Patenter VAN BUUREN VAN SWAAY matched with company VAN SWAAY BEHEER (NL09085054 ) Location zip or city does not match (match score 4).

Example i) is characterized by a high Jaccard similarity although the company name consists of only one non-discriminating token. In addition, the correspondence location zip and city were unknown and hence the value of the match score was equal to 3. In this example, adding further quality criteria, the user of the dataset could reduce the ambiguity of the name similarity. In particular names with no discriminating tokens might need to be treated separately.

Example ii) is a false positive because the match is identified by one common company word ("COOP") and one non-discriminating token ("S"). Moreover the correspondence of the location was verified so the match score was 1. This example is similar to the previous in that the names matched only on non-discriminating tokens.

Example iii) shows that the match is made up by one discriminating token ("SWAAY") and one not discriminating ("VAN"). However, the non-discriminating token is repeated two times in the patenter name which increases artificially the similarity of the two names. This example suggests that it might be useful to remove duplicate tokens in the names before matching.

### *5.2 False negatives*

In general, the assessment of the Type II error (failing to match an applicant that should be matched) is more difficult because it requires broader definition of matching criteria. In this section we describe two approaches.

The first approach deepens the analysis of the one hundred patenter names that were examined in the previous section. We reported that 24 patenter names out of 100 were not matched to any company in Amadeus. First, we found that among them were 2 individual applicant names and one non-business organization name, that is about 10% of the unmatched applicants. On the one hand, this finding shows some drawbacks in the identification method of the institutional type of patenter names used in Figure 3.[26] On the other, it suggests that the real coverage of the matching depicted in Figure 4 could be even higher; in particular for the European patenters the actual matching to Amadeus could be as high as 80% of the EPO standard codes.

[26] Indeed we noted that the format "Surname, First name Middle name(s)" has not been strictly followed by the EPO. A full-fledged methodology could therefore be based on the matching of the individual applicant names with inventors of the same invention. This is left for future development.

Second, we found only one case of a false negative, that is:

- Patenter name DISTEC should have been matched with DISTEC VERTRIEB VON ELEKTRONISCHEN BAUELEMENTEN (DE8330189835)

This example could be explained by the presence of the non-discriminating token "DISTEC". This common word generates a low similarity score in the expression of eq. (3) because company name is made up of four other tokens, two of them discriminating and hence with high statistical score. This example suggests that it might be worth exploring other token based similarity functions.

Thirdly, there were 20 patenter names that could not be found in Amadeus business directory. The full-fledged identification of the origin of these applicants in terms of sector, age and size is beyond the scope of this paper. There are several explanations for the fact that these firms are not found in Amadeus (1998-2008). First, they may have been included in Amadeus after the date of our version. Second, some of these patenters are probably not limited liability firms. Third, their names may have changed since the patent was applied for but before they incorporated. Fourth, they have been incorporated in a different country from the one given on the patent. Fifth, they could have exited, merged and been acquired by another firm before the year 1998.

The second approach for the assessment of Type II errors analyzes the names of some large EU firms which have reported R&D activities. Typically, the R&D process is accompanied by patenting activities and a large share of patenting is done by large R&D performers. Thus focusing on firms with R&D reduces the cost of searching for false negatives. This analysis involved extracting some uniquely identifying keywords from the names of some large R&D performers and checking whether they were matched.

Table 8 reports a list of discriminating tokens that identify uniquely about a hundred large R&D firms from Europe. These tokens have been extracted from the company names of the top 2,197 European R&D doers.[27] We selected these tokens randomly from the single token company names that were highly discriminating.

The fact that these tokens are discriminating does not necessarily mean that the similarity measure of equation (3) scores high. For example the token "ABB" is discriminating but it is also statistically very frequent in the dataset, which dwarfs the weighted Jaccard similarity distance of equation (3). Secondly, while these words are extracted from single token company names in Amadeus, in the patenter names they could occur jointly with other words – such as generic, geographical, etc – that lower the size of the similarity index of equation (3).

We report the statistical distribution of these tokens in the EPO/PCT matched dataset and the dataset that contains the remaining unmatched applicants – confining ourselves to patenters from European countries. Among the unmatched patenter names, about 6.5% of the names could be associated to these discriminating tokens. In other words, for about 6.5% unmatched patenter names, these discriminating tokens are in their words or subwords.

*[Table 8 about here]*

[27] For more information about the dataset from which they were extracted, see Hall et al. (2007).

Broadly speaking, this percentage could be a potential proxy for Type II error. However, when we restricted the search only to location in the same country this percentage falls to less than 1%. Similarly, focusing on the patentees that have more than 2 patents the percentage is about 2.1%.

Put differently, if we condition for the same country location – as we did in all the matching tasks – false negatives are probably a small number. In general, some false negatives could be associated with a discrepancy in the nationality of the patenter and the country location of the company in the business directory. One solution to overcome this drawback could be the relaxation of the matching condition requiring the same geographic location. In the current version of the dataset we opted for a conservative approach, assuming that there is one-to-one correspondence of a patenter from a given country and the legal business entity recorded in Amadeus in that country.

An additional robustness check weighted the false negatives by patent counts (see again Table 8). The overall percentage is very small about 0.5%. In some few cases of companies the share of missing patent counts due to false negatives is high such as "SAFEWAY", "ALPHAFORM", "INGENTA", "PENNON", "SOGEFI". However these companies are characterised by a limited patenting activity, that is a patent portfolio of about a dozen patents.

In conclusion, the two tests conducted in this section point to consistent estimates and a small relative share of false negatives – less than 5% in general. Further analysis could be done by adjusting the distance measure described in equation (3) and complementing it with other similarity functions in order to improve the false negative rate further. In addition, a deeper investigation of the generation of business directories database such as Amadeus could contribute to the understanding of the unmatched patenter names. In particular, phenomena that have not been explored here are the relocation and reincorporation of business activities, firms' ownership structures, the effects of mergers and acquisitions, and other name changes or reorganizations.

## 6 Conclusions

In this paper we drew on NER methods from bioinformatics and applied two different approaches to data integration in the context of patent information. The dictionary-based approach relies on the collection of large datasets of names and their variants, while the rule-based approach articulates a set of rules to establish similarity links 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.

The resulting data contains around 131 thousands patent applicant names from 46 countries, covering approximately 58.8 percent of EPO applications and 50.6 percent of PCT applications by business organizations during the time period 1979 to 2008. For the US, the resulting dataset includes about 54 thousands assignee names and 51.3 percent of US granted patents during approximately the same time period.

There are several novel elements associated with this paper and dataset. First, this paper is among of the first attempts to adopt insights from other domains such as bioinformatics in the field of the information integration of company-level data. Second, we focus on patenter names of major patent offices, including the EPO and the USPTO, but also the WIPO/PCT

database which has been rarely employed in innovation studies. Third, we relied on a pan-European business directory (Amadeus by Bureau Van Dijk) to trace the patent-holder names to owning entitities. Fourth, we created a more comprehensive dataset than has been available hitherto that constains all the patenting activity of European firms at the EPO, USPTO and PCT since the 1970s. Fifth, we developed a matching methodology that relies not only on the name similarity of business entities, but also on such criteria as location information, age, etc.

There are also limitations to this work that will have to be addressed in future research. First, our focus has been on European patent-holders and their subsidiaries whereas a large share of patents are filed by North American firms – US and Canada – and those from Asia – such as Japan, Korea, Taiwan, and increasingly China. The methodology developed in this paper could be extended to patenting entitites in these countries, but there are some caveats. For US and Canadian firms extensive business directories such as Dun & Bradstreet files or Icarus and Orbis by Bureau Van Dijk report data by firm establishments and not by legal business entities: thus for the same legal business entity there could be several establishments. The different structure of the business directories will require a slightly different approach to consolidation to the parent firms. For Asian firms we noticed that often the patenter names are characterised by transliteration errors at the level of a single word, which hampers a direct implementation of the token-based similarity measures. In this case other similarity measures could be used to pre-process the patenter names such as the edit/Levenshtein distance, Hamming distance, and others.

# 7 References


Arora, A., Fosfuri, A. and Gambardella, A. (2001), *Markets for Technology: The Economics of Innovation and Corporate Strategy*. MIT Press, Cambridge MA

Arundel, A. (2003), *Patents in the Knowledge-Based Economy*, *Report of the KNOW Survey*, MERIT, University of Maastricht.

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.

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

Cockburn, I. M., B. H. Hall, W. Powell, and M. Trajtenberg (2004), Patent Data Project – Brief Literature Review, section from the proposal to the National Science Foundation. Available at http://www.econ.berkeley.edu/~bhhall/bhpapers.html

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.

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, *European Management Review* 5 (2), 69-84

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. et al. (2007), Inventors and Invention Processes in Europe. Results from the PatVal-EU Survey, *Research Policy* 36 (8), 1107-1127

Goto, A., and K. Motohashi (2007), Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities. *Research Policy* 36(9), 1431-42.

Graham, S. D. Somaya (2004) "The Use of Patents, Copyrights, and Trademarks in Software: Evidence from Litigation," with. In Patents, Innovation and Economic Performance, OECD Conference Proceedings.

Greenhalgh, C., and M. Rogers (2007), The value of intellectual property rights to firms and society, *Oxford Review of Economic Policy* 23 (4): 541–567

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, Universitá L. Bocconi, Milan, Italy, February.

Griliches, Z. (1981), Market Value, R&D, and Patents, *Economic Letters* 7: 183-187.

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

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. (1993), The Stock Market's Valuation of Research and Development Investment During the 1980s. *American Economic Review 83*: 259-264.

Hall, B. H., and W. F. Long (1999), Differences in Reported R&D Data on the NSF/Census RD-1 Form and the SEC 10-K Form: A Micro-Data Investigation. NBER, UC Berkeley, Oxford University, and BPRA. Available at http://www.econ.berkeley.edu/~bhhall/bhpapers.html

Hall, B. H., Griliches, Z., and J. A. Hausman (1986), Patents and R&D: Is There a Lag? *International Economic Review* 27: 265-83.

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.

Hall, B. H., Jaffe, A. B., and M. Trajtenberg (2005), Market Value and Patent Citations. *RAND Journal of Economics* 36: 16-38.

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

Harhoff, D., F. M. Scherer and K. Vopel (2003). "Citations, Family Size, Opposition and the Value of Patent Rights - Evidence from Germany," *Research Policy*, 32, 1343-1363.

Heong, W. M., D. Page, R. Abello and K. P. Pang, (2007), "Explorations of Innovation and Business Performance Using Linked Firm-Level Data," Research Paper, Australian Bureau of Statistics cat. no. 1351.0.55.020, September, Canberra, Australia.

Jaccard, P. (1901), *Bulletin del la Société Vaudoisedes Sciences Naturelles* 37, 241-272.

Janz, N., G., Ebling, S., Gottschalk, H., and Niggemann (2001), "The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW)," *Schmollers Jahrbuch - Zeitschrift für Wirtschafts- und Sozialwissenschaften* 121, 123-129.

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.

Mairesse, J., and P. Mohnen (2010), Innovation surveys, forthcoming in Hall, B. H., and N. Rosenberg (eds.), *Handbook of the Economics of Innovation*, Elsevier-North Holland.

Mendonca, S. & Pereira, T. S. & Godinho, M. M., (2004). "Trademarks as an indicator of innovation and industrial change," Research Policy, Elsevier, vol. 33(9), pages 1385-1404.

Moser, P. (2005), How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World Fairs. *American Economic Review* 95 (4): 1215-1236.

Murray, F. Aghion P., Dewatrapont M., Kolev J. and S. Stern (2008). "Of Mice and Academics: The Role of Openness in Science". MIT Sloan Working Paper.

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.

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).

Navarro, G. (2001), A guided tour to approximate string matching. *ACM Computing Surveys* 33 (1): 31--88.

OECD (2002), Frascati Manual. OECD, Paris.

OECD (2009), OECD Patent Statistics Manual, OECD, Paris.

OECD and Eurostat (2005), Oslo Manual, The measurement of scientific and technological activities Proposed guidelines for collecting and interpreting technological innovation data (3rd edition), OECD, Paris.

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.

Sandner, P. (2009). The Market Value of R&D, Patents, and Trademarks, LMU working Paper, mimeo, http://ssrn.com/abstract=1469705

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/

Um, M. J, (2005), Korean Innovation Survey: Manufacturing Sector, Policy Research Report 12/2005, Seoul Korea, available at http://www.stepi.re.kr/eng/

von Graevenitz, G. (2007), "Which Reputations Does a Brand Owner Need? Evidence from Trade Mark Opposition", GESY Discussion Paper No. 215.

## Table 2 Business applicants and applications in EPO and PCT dataset

(distinct original names, countries with more than 100 EP applicants)

| | EP Applicants | | WO applicants | | EP applications | | WO applications | | Average | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **EP portfolio** | **WO portfolio** |
| AN | 286 | 0.1% | 66 | 0.0% | 1,096 | 0.1% | 245 | 0.0% | 3.83 | 3.71 |
| AT | 3,404 | 1.3% | 1,936 | 0.9% | 15,707 | 0.9% | 7,893 | 0.7% | 4.61 | 4.08 |
| AU | 4,787 | 1.8% | 8,134 | 3.7% | 9,236 | 0.5% | 14,888 | 1.4% | 1.93 | 1.83 |
| BB | 140 | 0.1% | 69 | 0.0% | 1,134 | 0.1% | 191 | 0.0% | 8.10 | 2.77 |
| BE | 2,892 | 1.1% | 1,513 | 0.7% | 15,421 | 0.9% | 6,992 | 0.6% | 5.33 | 4.62 |
| BG | 111 | 0.0% | 63 | 0.0% | 142 | 0.0% | 80 | 0.0% | 1.28 | 1.27 |
| BM | 107 | 0.0% | 61 | 0.0% | 298 | 0.0% | 168 | 0.0% | 2.79 | 2.75 |
| BR | 343 | 0.1% | 254 | 0.1% | 661 | 0.0% | 352 | 0.0% | 1.93 | 1.39 |
| CA | 5,574 | 2.1% | 6,610 | 3.0% | 18,404 | 1.0% | 17,808 | 1.6% | 3.30 | 2.69 |
| CH | 9,612 | 3.7% | 5,411 | 2.4% | 64,547 | 3.6% | 28,326 | 2.6% | 6.72 | 5.24 |
| CN | 1,337 | 0.5% | 2,200 | 1.0% | 3,501 | 0.2% | 8,292 | 0.8% | 2.62 | 3.77 |
| CY | 129 | 0.0% | 148 | 0.1% | 276 | 0.0% | 318 | 0.0% | 2.14 | 2.15 |
| CZ | 248 | 0.1% | 253 | 0.1% | 452 | 0.0% | 409 | 0.0% | 1.82 | 1.62 |
| DE | 37,564 | 14.4% | 18,733 | 8.4% | 345,386 | 19.3% | 147,941 | 13.5% | 9.19 | 7.90 |
| DK | 3,242 | 1.2% | 3,357 | 1.5% | 11,836 | 0.7% | 11,629 | 1.1% | 3.65 | 3.46 |
| ES | 3,286 | 1.3% | 2,273 | 1.0% | 7,000 | 0.4% | 4,271 | 0.4% | 2.13 | 1.88 |
| FI | 3,096 | 1.2% | 3,209 | 1.4% | 18,467 | 1.0% | 17,846 | 1.6% | 5.96 | 5.56 |
| FR | 21,361 | 8.2% | 10,918 | 4.9% | 125,162 | 7.0% | 46,988 | 4.3% | 5.86 | 4.30 |
| GB | 20,538 | 7.9% | 17,042 | 7.7% | 84,913 | 4.7% | 56,743 | 5.2% | 4.13 | 3.33 |
| GR | 180 | 0.1% | 132 | 0.1% | 274 | 0.0% | 207 | 0.0% | 1.52 | 1.57 |
| HK | 301 | 0.1% | 16 | 0.0% | 436 | 0.0% | 18 | 0.0% | 1.45 | 1.13 |
| HU | 633 | 0.2% | 599 | 0.3% | 1,596 | 0.1% | 1,226 | 0.1% | 2.52 | 2.05 |
| IE | 1,150 | 0.4% | 1,020 | 0.5% | 2,862 | 0.2% | 2,578 | 0.2% | 2.49 | 2.53 |
| IL | 2,500 | 1.0% | 1,704 | 0.8% | 4,845 | 0.3% | 2,703 | 0.2% | 1.94 | 1.59 |
| IN | 439 | 0.2% | 577 | 0.3% | 1,498 | 0.1% | 2,680 | 0.2% | 3.41 | 4.64 |
| IT | 17,024 | 6.5% | 6,542 | 2.9% | 54,688 | 3.1% | 16,345 | 1.5% | 3.21 | 2.50 |
| JP | 23,703 | 9.1% | 17,883 | 8.0% | 350,015 | 19.6% | 163,365 | 15.0% | 14.77 | 9.14 |
| KR | 2,977 | 1.1% | 5,655 | 2.5% | 22,550 | 1.3% | 17,237 | 1.6% | 7.57 | 3.05 |
| LI | 588 | 0.2% | 62 | 0.0% | 2,363 | 0.1% | 78 | 0.0% | 4.02 | 1.26 |
| LU | 659 | 0.3% | 414 | 0.2% | 2,483 | 0.1% | 1,442 | 0.1% | 3.77 | 3.48 |
| NL | 7,502 | 2.9% | 4,460 | 2.0% | 67,101 | 3.7% | 41,232 | 3.8% | 8.94 | 9.24 |
| NO | 2,028 | 0.8% | 2,628 | 1.2% | 4,785 | 0.3% | 5,885 | 0.5% | 2.36 | 2.24 |
| NZ | 598 | 0.2% | 553 | 0.2% | 1,046 | 0.1% | 835 | 0.1% | 1.75 | 1.51 |
| PL | 249 | 0.1% | 195 | 0.1% | 402 | 0.0% | 342 | 0.0% | 1.61 | 1.75 |
| PT | 204 | 0.1% | 142 | 0.1% | 380 | 0.0% | 204 | 0.0% | 1.86 | 1.44 |
| RU | 372 | 0.1% | 467 | 0.2% | 473 | 0.0% | 597 | 0.1% | 1.27 | 1.28 |
| SE | 7,487 | 2.9% | 7,632 | 3.4% | 35,584 | 2.0% | 35,735 | 3.3% | 4.75 | 4.68 |
| SG | 304 | 0.1% | 432 | 0.2% | 856 | 0.0% | 719 | 0.1% | 2.82 | 1.66 |
| SI | 151 | 0.1% | 135 | 0.1% | 428 | 0.0% | 351 | 0.0% | 2.83 | 2.60 |
| SU | 112 | 0.0% | 0 | 0.0% | 227 | 0.0% | 0 | 0.0% | 2.03 | #DIV/0! |
| TR | 181 | 0.1% | 219 | 0.1% | 468 | 0.0% | 776 | 0.1% | 2.59 | 3.54 |
| TW | 924 | 0.4% | 101 | 0.0% | 1,826 | 0.1% | 220 | 0.0% | 1.98 | 2.18 |
| US | 70,194 | 26.9% | 87,431 | 39.3% | 503,399 | 28.1% | 423,571 | 38.8% | 7.17 | 4.84 |
| VG | 267 | 0.1% | 124 | 0.1% | 1,158 | 0.1% | 500 | 0.0% | 4.34 | 4.03 |
| ZA | 649 | 0.2% | 390 | 0.2% | 1,211 | 0.1% | 579 | 0.1% | 1.87 | 1.48 |
| Others | 1,565 | 0.6% | 868 | 0.4% | 3,129 | 0.2% | 1,366 | 0.1% | 2.00 | 1.57 |
| Overall | 260,997 | 100.0% | 222,628 | 100.0% | 1,789,721 | 100.0% | 1,092,169 | 100.0% | 6.86 | 4.91 |

**Table 3 Non-business organization applicants and applications in EPO and PCT dataset**

(distinct original names, countries with more than 100 EP applicants)*

| | EP Applicants | | WO applicants | | EP applications | | WO applications | | Average | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **EP portfolio** | **WO portfolio** |
| AT | 130 | 0.7% | 97 | 0.5% | 419 | 0.4% | 257 | 0.2% | 3.22 | 2.65 |
| AU | 506 | 2.7% | 671 | 3.5% | 2,535 | 2.6% | 3,708 | 3.2% | 5.01 | 5.53 |
| BE | 332 | 1.8% | 244 | 1.3% | 2,198 | 2.3% | 1,184 | 1.0% | 6.62 | 4.85 |
| CA | 711 | 3.8% | 850 | 4.4% | 2,478 | 2.6% | 3,498 | 3.0% | 3.49 | 4.12 |
| CH | 419 | 2.3% | 355 | 1.9% | 1,767 | 1.8% | 1,227 | 1.0% | 4.22 | 3.46 |
| CN | 345 | 1.9% | 578 | 3.0% | 684 | 0.7% | 1,778 | 1.5% | 1.98 | 3.08 |
| DE | 2,014 | 10.9% | 1,377 | 7.2% | 11,494 | 11.9% | 8,718 | 7.5% | 5.71 | 6.33 |
| DK | 129 | 0.7% | 152 | 0.8% | 420 | 0.4% | 530 | 0.5% | 3.26 | 3.49 |
| ES | 311 | 1.7% | 393 | 2.1% | 909 | 0.9% | 1,653 | 1.4% | 2.92 | 4.21 |
| FR | 1,558 | 8.4% | 1,182 | 6.2% | 14,389 | 14.8% | 9,122 | 7.8% | 9.24 | 7.72 |
| GB | 1,323 | 7.1% | 1,370 | 7.1% | 6,624 | 6.8% | 7,254 | 6.2% | 5.01 | 5.29 |
| IL | 204 | 1.1% | 127 | 0.7% | 956 | 1.0% | 444 | 0.4% | 4.69 | 3.50 |
| IN | 126 | 0.7% | 188 | 1.0% | 727 | 0.8% | 1,093 | 0.9% | 5.77 | 5.81 |
| IT | 491 | 2.6% | 354 | 1.8% | 1,885 | 1.9% | 1,385 | 1.2% | 3.84 | 3.91 |
| JP | 1,683 | 9.1% | 1,485 | 7.7% | 7,320 | 7.6% | 9,541 | 8.2% | 4.35 | 6.42 |
| KR | 409 | 2.2% | 557 | 2.9% | 1,481 | 1.5% | 2,586 | 2.2% | 3.62 | 4.64 |
| NL | 436 | 2.4% | 393 | 2.1% | 2,327 | 2.4% | 1,892 | 1.6% | 5.34 | 4.81 |
| PL | 119 | 0.6% | 87 | 0.5% | 242 | 0.2% | 170 | 0.1% | 2.03 | 1.95 |
| RU | 125 | 0.7% | 87 | 0.5% | 226 | 0.2% | 145 | 0.1% | 1.81 | 1.67 |
| SE | 143 | 0.8% | 151 | 0.8% | 274 | 0.3% | 285 | 0.2% | 1.92 | 1.89 |
| SU | 159 | 0.9% | 2 | 0.0% | 350 | 0.4% | 9 | 0.0% | 2.20 | 4.50 |
| US | 5,892 | 31.8% | 7,618 | 39.8% | 34,777 | 35.9% | 58,106 | 49.7% | 5.90 | 7.63 |
| Others | 980 | 5.3% | 845 | 4.4% | 2,429 | 2.5% | 2,371 | 2.0% | 2.48 | 2.81 |
| Overall | 18,544 | 100.0% | 19,162 | 100.0% | 96,910 | 100.0% | 116,956 | 100.0% | 5.23 | 6.10 |

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

## Table 4 Individual applicants and applications in EPO and PCT dataset

(distinct original names, countries with more than 100 EP applicants)

| | EP Applicants | | WO applicants | | EP applications | | WO applications | | Average | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **EP portfolio** | **WO portfolio** |
| AR | 243 | 0.2% | 69 | 0.0% | 291 | 0.2% | 73 | 0.0% | 1.20 | 1.06 |
| AT | 3,165 | 2.4% | 2,029 | 1.2% | 5,002 | 2.9% | 2,943 | 1.4% | 1.58 | 1.45 |
| AU | 2,991 | 2.3% | 6,713 | 4.1% | 3,530 | 2.0% | 7,878 | 3.9% | 1.18 | 1.17 |
| BE | 1,702 | 1.3% | 1,017 | 0.6% | 2,191 | 1.3% | 1,206 | 0.6% | 1.29 | 1.19 |
| BG | 120 | 0.1% | 283 | 0.2% | 134 | 0.1% | 328 | 0.2% | 1.12 | 1.16 |
| BR | 302 | 0.2% | 506 | 0.3% | 329 | 0.2% | 557 | 0.3% | 1.09 | 1.10 |
| CA | 2,895 | 2.2% | 5,072 | 3.1% | 3,645 | 2.1% | 6,101 | 3.0% | 1.26 | 1.20 |
| CH | 4,687 | 3.6% | 3,121 | 1.9% | 6,937 | 4.0% | 4,166 | 2.1% | 1.48 | 1.33 |
| CN | 1,232 | 0.9% | 4,727 | 2.9% | 1,401 | 0.8% | 5,966 | 2.9% | 1.14 | 1.26 |
| CZ | 267 | 0.2% | 530 | 0.3% | 316 | 0.2% | 626 | 0.3% | 1.18 | 1.18 |
| DE | 25,515 | 19.3% | 15,279 | 9.3% | 39,302 | 22.5% | 21,297 | 10.5% | 1.54 | 1.39 |
| DK | 1,434 | 1.1% | 1,832 | 1.1% | 1,810 | 1.0% | 2,272 | 1.1% | 1.26 | 1.24 |
| ES | 2,791 | 2.1% | 2,791 | 1.7% | 3,384 | 1.9% | 3,219 | 1.6% | 1.21 | 1.15 |
| FI | 1,270 | 1.0% | 1,967 | 1.2% | 1,588 | 0.9% | 2,512 | 1.2% | 1.25 | 1.28 |
| FR | 12,552 | 9.5% | 8,814 | 5.4% | 16,519 | 9.5% | 11,121 | 5.5% | 1.32 | 1.26 |
| GB | 8,481 | 6.4% | 9,863 | 6.0% | 10,392 | 5.9% | 11,955 | 5.9% | 1.23 | 1.21 |
| GR | 619 | 0.5% | 597 | 0.4% | 732 | 0.4% | 736 | 0.4% | 1.18 | 1.23 |
| HK | 134 | 0.1% | 13 | 0.0% | 182 | 0.1% | 13 | 0.0% | 1.36 | 1.00 |
| HR | 109 | 0.1% | 353 | 0.2% | 125 | 0.1% | 422 | 0.2% | 1.15 | 1.20 |
| HU | 837 | 0.6% | 1,472 | 0.9% | 986 | 0.6% | 1,833 | 0.9% | 1.18 | 1.25 |
| IE | 602 | 0.5% | 602 | 0.4% | 740 | 0.4% | 734 | 0.4% | 1.23 | 1.22 |
| IL | 1,265 | 1.0% | 1,141 | 0.7% | 1,523 | 0.9% | 1,283 | 0.6% | 1.20 | 1.12 |
| IN | 384 | 0.3% | 1,099 | 0.7% | 449 | 0.3% | 1,578 | 0.8% | 1.17 | 1.44 |
| IT | 8,263 | 6.3% | 4,780 | 2.9% | 10,637 | 6.1% | 5,853 | 2.9% | 1.29 | 1.22 |
| JP | 6,023 | 4.6% | 7,436 | 4.5% | 9,367 | 5.4% | 10,748 | 5.3% | 1.56 | 1.45 |
| KR | 2,135 | 1.6% | 7,842 | 4.8% | 2,540 | 1.5% | 9,768 | 4.8% | 1.19 | 1.25 |
| MX | 132 | 0.1% | 441 | 0.3% | 142 | 0.1% | 506 | 0.2% | 1.08 | 1.15 |
| NL | 2,495 | 1.9% | 1,784 | 1.1% | 3,216 | 1.8% | 2169 | 1.1% | 1.29 | 1.22 |
| NO | 1,077 | 0.8% | 1,671 | 1.0% | 1,302 | 0.7% | 2068 | 1.0% | 1.21 | 1.24 |
| NZ | 350 | 0.3% | 414 | 0.3% | 401 | 0.2% | 433 | 0.2% | 1.15 | 1.05 |
| PL | 298 | 0.2% | 680 | 0.4% | 351 | 0.2% | 821 | 0.4% | 1.18 | 1.21 |
| PT | 181 | 0.1% | 160 | 0.1% | 202 | 0.1% | 189 | 0.1% | 1.12 | 1.18 |
| RU | 923 | 0.7% | 2,683 | 1.6% | 1,101 | 0.6% | 3472 | 1.7% | 1.19 | 1.29 |
| SE | 4,758 | 3.6% | 5,937 | 3.6% | 6,149 | 3.5% | 7577 | 3.7% | 1.29 | 1.28 |
| SI | 157 | 0.1% | 231 | 0.1% | 195 | 0.1% | 276 | 0.1% | 1.24 | 1.19 |
| SU | 288 | 0.2% | 5 | 0.0% | 348 | 0.2% | 5 | 0.0% | 1.21 | 1.00 |
| TR | 132 | 0.1% | 281 | 0.2% | 156 | 0.1% | 328 | 0.2% | 1.18 | 1.17 |
| TW | 1,431 | 1.1% | 144 | 0.1% | 2,056 | 1.2% | 180 | 0.1% | 1.44 | 1.25 |
| US | 27,442 | 20.8% | 56,118 | 34.1% | 34,846 | 19.9% | 69714 | 34.3% | 1.27 | 1.24 |
| YU | 159 | 0.1% | 119 | 0.1% | 218 | 0.1% | 130 | 0.1% | 1.37 | 1.09 |
| ZA | 602 | 0.5% | 1,439 | 0.9% | 681 | 0.4% | 1628 | 0.8% | 1.13 | 1.13 |
| Others | 1484 | 1.1% | 2453 | 1.5% | 1823 | 1.0% | 2954 | 1.5% | 1.23 | 1.20 |
| Overall | 131,923 | 100.0% | 164,503 | 100.0% | 174,732 | 101.4% | 203,054 | 102.3% | 1.32 | 1.23 |

## Table 5 Matched business applicants and applications in EPO and PCT dataset

| | EP Applicants | | WO applicants | | EP applications | | WO applications | | Average | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **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 | 2,502 | 1.9% | 1,343 | 1.8% | 13,462 | 1.3% | 6,744 | 1.2% | 0.19 | 1.52 |
| AU | 226 | 0.2% | 203 | 0.3% | 565 | 0.1% | 776 | 0.1% | 0.40 | 3.26 |
| BE | 2,526 | 1.9% | 1,238 | 1.6% | 14,081 | 1.3% | 6,423 | 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 |
| CA | 437 | 0.3% | 325 | 0.4% | 2,415 | 0.2% | 2,039 | 0.4% | 0.18 | 1.48 |
| CH | 4,329 | 3.3% | 2,372 | 3.1% | 49,654 | 4.6% | 21,334 | 3.8% | 0.09 | 0.71 |
| CN | 10 | 0.0% | 7 | 0.0% | 46 | 0.0% | 83 | 0.0% | 0.22 | 1.77 |
| 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 | 31,779 | 24.2% | 15,247 | 20.0% | 326,515 | 30.6% | 142,407 | 25.3% | 0.10 | 0.79 |
| DK | 3,016 | 2.3% | 2,518 | 3.3% | 10,702 | 1.0% | 10,534 | 1.9% | 0.28 | 2.30 |
| EE | 11 | 0.0% | 11 | 0.0% | 7 | 0.0% | 12 | 0.0% | 1.57 | 12.81 |
| ES | 2,455 | 1.9% | 1,334 | 1.8% | 5,371 | 0.5% | 3,078 | 0.5% | 0.46 | 3.73 |
| FI | 2,583 | 2.0% | 2,059 | 2.7% | 16,700 | 1.6% | 15,951 | 2.8% | 0.15 | 1.26 |
| FO | 10 | 0.0% | 9 | 0.0% | 4,172 | 0.4% | 874 | 0.2% | 0.00 | 0.02 |
| FR | 16,627 | 12.7% | 8,244 | 10.8% | 115,648 | 10.8% | 44,582 | 7.9% | 0.14 | 1.17 |
| GB | 17,260 | 13.2% | 11,551 | 15.2% | 76,166 | 7.1% | 49,473 | 8.8% | 0.23 | 1.85 |
| GR | 105 | 0.1% | 71 | 0.1% | 159 | 0.0% | 132 | 0.0% | 0.66 | 5.38 |
| 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% | 1,556 | 0.1% | 1,372 | 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 | 14,100 | 10.8% | 4,837 | 6.4% | 48,675 | 4.6% | 14,106 | 2.5% | 0.29 | 2.36 |
| JP | 1,890 | 1.4% | 1,114 | 1.5% | 55,438 | 5.2% | 21,637 | 3.8% | 0.03 | 0.28 |
| KR | 265 | 0.2% | 165 | 0.2% | 9,190 | 0.9% | 2,950 | 0.5% | 0.03 | 0.24 |
| 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 |
| NL | 6,446 | 4.9% | 3,365 | 4.4% | 63,863 | 6.0% | 39,386 | 7.0% | 0.10 | 0.82 |
| NO | 2,427 | 1.9% | 2,187 | 2.9% | 4,386 | 0.4% | 5,325 | 0.9% | 0.55 | 4.51 |
| 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 |
| SE | 6,561 | 5.0% | 5,502 | 7.2% | 32,009 | 3.0% | 32,342 | 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 |
| 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 |
| US | 13,985 | 10.7% | 11,305 | 14.9% | 214,052 | 20.0% | 138,353 | 24.6% | 0.07 | 0.53 |
| YU | 1 | 0.0% | 1 | 0.0% | 2 | 0.0% | 3 | 0.0% | 0.50 | 4.08 |
| Overall | 131,065 | | 76,054 | | 1,068,407 | | 562,049 | | 0.12 | 0.14 |

**Table 8 Discriminating tokens from the company name of the top EU R&D performers**

| | | Number of names in various datasets | | | | | Weighted by patent counts | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | Unique identifying token for company | All | Matched | Unmatched | Unmatched; same country | Unmatched; with 3+ pats | Matched | Unmatched | % Unmatched |
| CH | ABB | 599 | 504 | 95 | 6 | 27 | 6,839 | 350 | 4.9% |
| DE | SIEMENS | 382 | 371 | 11 | 1 | 6 | 37,452 | 69 | 0.2% |
| CH | ROCHE | 196 | 165 | 31 | 9 | 14 | 8,554 | 71 | 0.8% |
| FR | ALCATEL | 185 | 178 | 7 | 0 | 1 | 10,229 | 10 | 0.1% |
| GB | SMITHKLINE | 120 | 116 | 4 | 0 | 0 | 5,249 | 5 | 0.1% |
| DE | BASF | 118 | 118 | 0 | 0 | 0 | 16,161 | 0 | 0.0% |
| FI | NOKIA | 117 | 114 | 3 | 5 | 3 | 9,521 | 9 | 0.1% |
| FR | AVENTIS | 113 | 113 | 0 | 0 | 0 | 2,680 | 0 | 0.0% |
| FR | ALSTOM | 110 | 110 | 0 | 0 | 0 | 1,670 | 0 | 0.0% |
| FR | THALES | 96 | 93 | 3 | 0 | 2 | 1,374 | 11 | 0.8% |
| FR | RENAULT | 83 | 82 | 1 | 1 | 0 | 2,744 | 1 | 0.0% |
| BE | SOLVAY | 83 | 76 | 7 | 0 | 4 | 1,792 | 30 | 1.6% |
| IT | PIRELLI | 78 | 78 | 0 | 0 | 0 | 743 | 0 | 0.0% |
| DE | THYSSENKRUPP | 77 | 77 | 0 | 0 | 0 | 617 | 0 | 0.0% |
| SE | VOLVO | 74 | 72 | 2 | 1 | 1 | 1,559 | 8 | 0.5% |
| DE | DEGUSSA | 63 | 60 | 3 | 0 | 3 | 3,112 | 13 | 0.4% |
| SE | SANDVIK | 49 | 40 | 9 | 4 | 3 | 1,115 | 25 | 2.2% |
| CH | CIBA | 48 | 38 | 10 | 0 | 3 | 6,217 | 22 | 0.4% |
| DK | DANISCO | 46 | 46 | 0 | 0 | 0 | 294 | 0 | 0.0% |
| CH | NOVARTIS | 42 | 40 | 2 | 0 | 2 | 4,553 | 55 | 1.2% |
| GB | RECKITT | 41 | 39 | 2 | 0 | 0 | 632 | 3 | 0.5% |
| FR | PEUGEOT | 38 | 37 | 1 | 1 | 1 | 2,416 | 21 | 0.9% |
| IT | FIAT | 37 | 36 | 1 | 0 | 0 | 1,109 | 1 | 0.1% |
| DK | DANFOSS | 36 | 35 | 1 | 0 | 0 | 322 | 2 | 0.6% |
| CH | SERONO | 30 | 29 | 1 | 0 | 1 | 325 | 3 | 0.9% |
| GB | UNILEVER | 30 | 30 | 0 | 0 | 0 | 9,135 | 0 | 0.0% |
| GB | INVENSYS | 29 | 29 | 0 | 0 | 0 | 106 | 0 | 0.0% |
| GB | KENWOOD | 29 | 28 | 1 | 0 | 0 | 273 | 1 | 0.4% |
| CH | CLARIANT | 28 | 28 | 0 | 0 | 0 | 1,299 | 0 | 0.0% |
| GB | ASTRAZENECA | 27 | 27 | 0 | 0 | 0 | 2,056 | 0 | 0.0% |
| FR | DASSAULT | 26 | 26 | 0 | 0 | 0 | 182 | 0 | 0.0% |
| GB | COURTAULDS | 23 | 22 | 1 | 0 | 0 | 187 | 1 | 0.5% |
| IT | EDISON | 20 | 18 | 2 | 0 | 0 | 293 | 1 | 0.3% |
| FR | DANONE | 19 | 18 | 1 | 1 | 0 | 125 | 1 | 0.8% |
| DE | VOLKSWAGEN | 19 | 19 | 0 | 0 | 0 | 2,015 | 0 | 0.0% |
| CH | SYNGENTA | 18 | 18 | 0 | 0 | 0 | 1,006 | 0 | 0.0% |
| DE | BEIERSDORF | 16 | 16 | 0 | 0 | 0 | 1,391 | 0 | 0.0% |
| GB | SEVERN | 16 | 16 | 0 | 0 | 0 | 30 | 0 | 0.0% |
| GB | HAMWORTHY | 15 | 13 | 2 | 0 | 0 | 14 | 2 | 12.5% |
| FR | GENESYS | 14 | 14 | 0 | 0 | 0 | 216 | 0 | 0.0% |
| DE | GROHE | 13 | 13 | 0 | 0 | 0 | 443 | 0 | 0.0% |
| GB | DOLPHIN | 12 | 12 | 0 | 0 | 0 | 16 | 0 | 0.0% |
| GB | REUTERS | 12 | 9 | 3 | 0 | 0 | 15 | 4 | 21.1% |
| IT | SAIPEM | 12 | 12 | 0 | 0 | 0 | 74 | 0 | 0.0% |
| FR | SOMFY | 12 | 12 | 0 | 0 | 0 | 260 | 0 | 0.0% |
| GB | INNOVATA | 11 | 11 | 0 | 0 | 0 | 35 | 0 | 0.0% |
| GB | WESSEX | 10 | 10 | 0 | 0 | 0 | 11 | 0 | 0.0% |
| IT | FINMECCANICA | 9 | 9 | 0 | 0 | 0 | 83 | 0 | 0.0% |
| GB | HOZELOCK | 9 | 9 | 0 | 0 | 0 | 30 | 0 | 0.0% |
| BE | BTICINO | 8 | 8 | 0 | 0 | 0 | 131 | 0 | 0.0% |
| DE | HOCHTIEF | 8 | 8 | 0 | 0 | 0 | 48 | 0 | 0.0% |
| GB | PLASMON | 8 | 8 | 0 | 0 | 0 | 32 | 0 | 0.0% |
| IT | DATALOGIC | 7 | 7 | 0 | 0 | 0 | 112 | 0 | 0.0% |
| GB | DOMNICK | 7 | 7 | 0 | 0 | 0 | 27 | 0 | 0.0% |
| IT | INDESIT | 7 | 7 | 0 | 0 | 0 | 111 | 0 | 0.0% |
| GB | LINX | 7 | 6 | 1 | 0 | 0 | 151 | 1 | 0.7% |
| CH | LOGITEC | 7 | 7 | 0 | 0 | 0 | 17 | 0 | 0.0% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| GB | RENISHAW | 7 | 7 | 0 | 0 | 0 | 358 | 0 | 0.0% |
| GB | CHIROSCIENCE | 6 | 6 | 0 | 0 | 0 | 7 | 0 | 0.0% |
| GB | MEDEVA | 6 | 6 | 0 | 0 | 0 | 26 | 0 | 0.0% |
| IT | SOGEFI | 6 | 5 | 1 | 0 | 1 | 15 | 6 | 28.6% |
| GB | TEPNEL | 6 | 6 | 0 | 0 | 0 | 13 | 0 | 0.0% |
| IT | BENETTON | 5 | 5 | 0 | 0 | 0 | 90 | 0 | 0.0% |
| IT | GEWISS | 5 | 5 | 0 | 0 | 0 | 60 | 0 | 0.0% |
| GB | KALAMAZOO | 5 | 5 | 0 | 0 | 0 | 5 | 0 | 0.0% |
| GB | LATCHWAYS | 5 | 5 | 0 | 0 | 0 | 38 | 0 | 0.0% |
| GB | SPIRENT | 5 | 5 | 0 | 0 | 0 | 22 | 0 | 0.0% |
| GB | UMBRO | 5 | 5 | 0 | 0 | 0 | 6 | 0 | 0.0% |
| GB | XENOVA | 5 | 5 | 0 | 0 | 0 | 27 | 0 | 0.0% |
| GB | ACAMBIS | 4 | 4 | 0 | 0 | 0 | 23 | 0 | 0.0% |
| IT | BEGHELLI | 4 | 4 | 0 | 0 | 0 | 37 | 0 | 0.0% |
| GB | INSIGNIA | 4 | 4 | 0 | 0 | 0 | 2 | 0 | 0.0% |
| GB | PHYTOPHARM | 4 | 3 | 1 | 0 | 0 | 16 | 1 | 5.9% |
| GB | PROTHERICS | 4 | 4 | 0 | 0 | 0 | 9 | 0 | 0.0% |
| GB | SAFEWAY | 4 | 3 | 1 | 1 | 0 | 4 | 1 | 20.0% |
| GB | TRANSENSE | 4 | 4 | 0 | 0 | 0 | 24 | 0 | 0.0% |
| GB | VICTREX | 4 | 4 | 0 | 0 | 0 | 30 | 0 | 0.0% |
| FR | VIVENDI | 3 | 3 | 0 | 0 | 0 | 2 | 0 | 0.0% |
| ES | ACERINOX | 2 | 2 | 0 | 0 | 0 | 3 | 0 | 0.0% |
| GB | MINORPLANET | 2 | 2 | 0 | 0 | 0 | 8 | 0 | 0.0% |
| GB | RADSTONE | 2 | 2 | 0 | 0 | 0 | 2 | 0 | 0.0% |
| GB | SKYEPHARMA | 2 | 2 | 0 | 0 | 0 | 27 | 0 | 0.0% |
| SE | SWITCHCORE | 2 | 2 | 0 | 0 | 0 | 3 | 0 | 0.0% |
| IT | TARGETTI | 2 | 2 | 0 | 0 | 0 | 21 | 0 | 0.0% |
| ES | TELEFONICA | 2 | 2 | 0 | 0 | 0 | 56 | 0 | 0.0% |
| GB | ZOTEFOAMS | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 0.0% |
| DE | ALPHAFORM | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 100.0% |
| GB | CASSIDY | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0.0% |
| GB | DATONG | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0.0% |
| NL | GETRONICS | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0.0% |
| GB | INGENTA | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 100.0% |
| IT | NATUZZI | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0.0% |
| GB | NETCALL | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0.0% |
| GB | PENNON | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 100.0% |
| GB | PIXOLOGY | 1 | 1 | 0 | 0 | 0 | 4 | 0 | 0.0% |
| FR | AVERYS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| DE | AZEGO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | BAKKAVOR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | FLOMERICS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| DE | HOECHST | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | INTERCLUBNET | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | KELDA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | MERANT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | OCTROI | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | ONCOCENE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | QSP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | QUADNETICS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | SENETEK | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | SINOVATION | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | SPECTRIS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | STATPRO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| GB | UTILITEC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| BE | ZENITEL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | n.d. |
| | Overall | 3,475 | 3,264 | 211 | 31 | 73 | 148,148 | 734 | 0.5% |
| | as % of the total | | 93.9% | 6.1% | 0.9% | 2.1% | | | |

Figure 2
Reduction in the size of the name file after harmonization of the EPO/PCT dataset (business applicants only, top 18 countries)

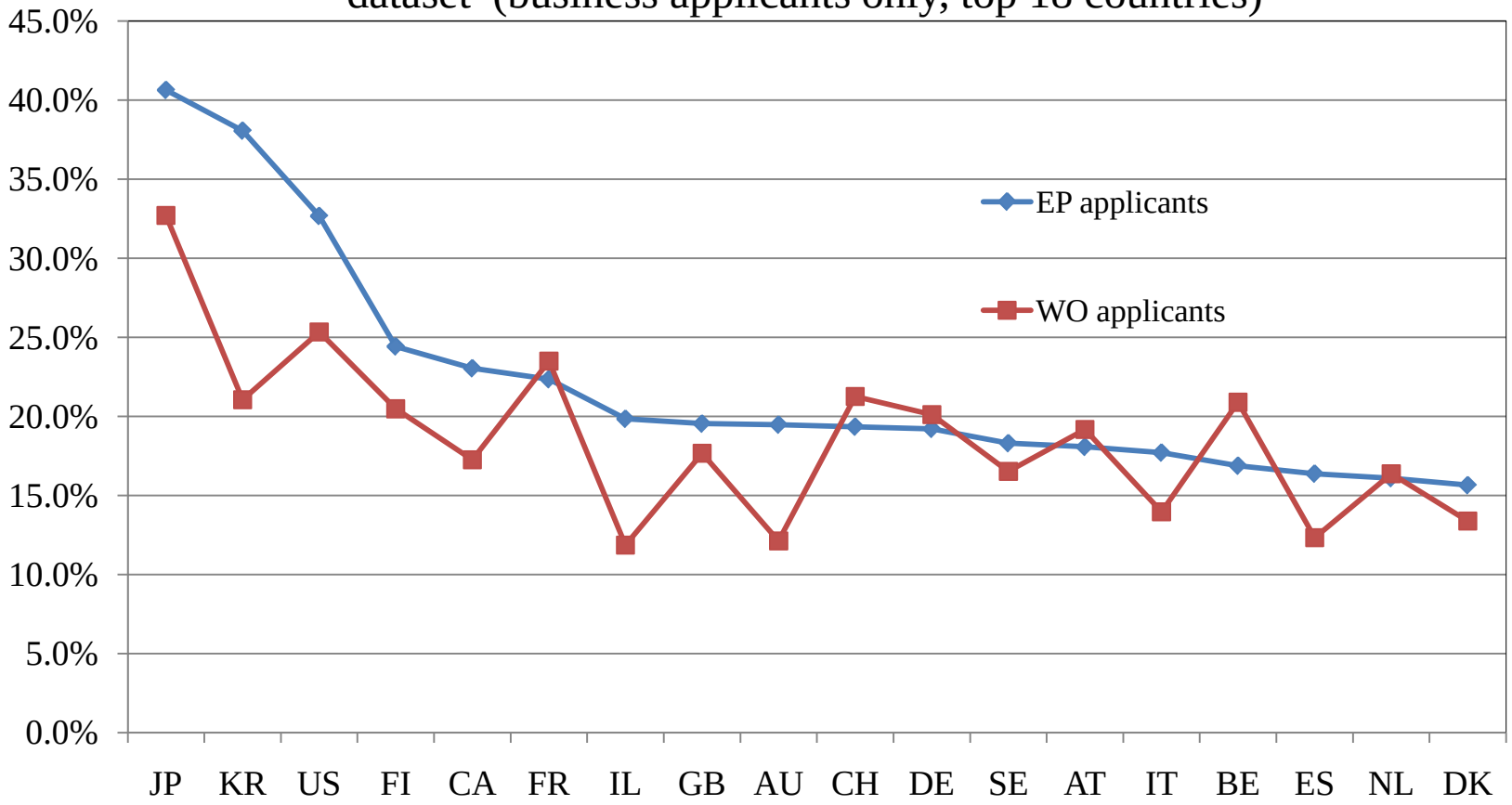

Figure 3a: Share of business applicant names matched
(top 18 EU countries)

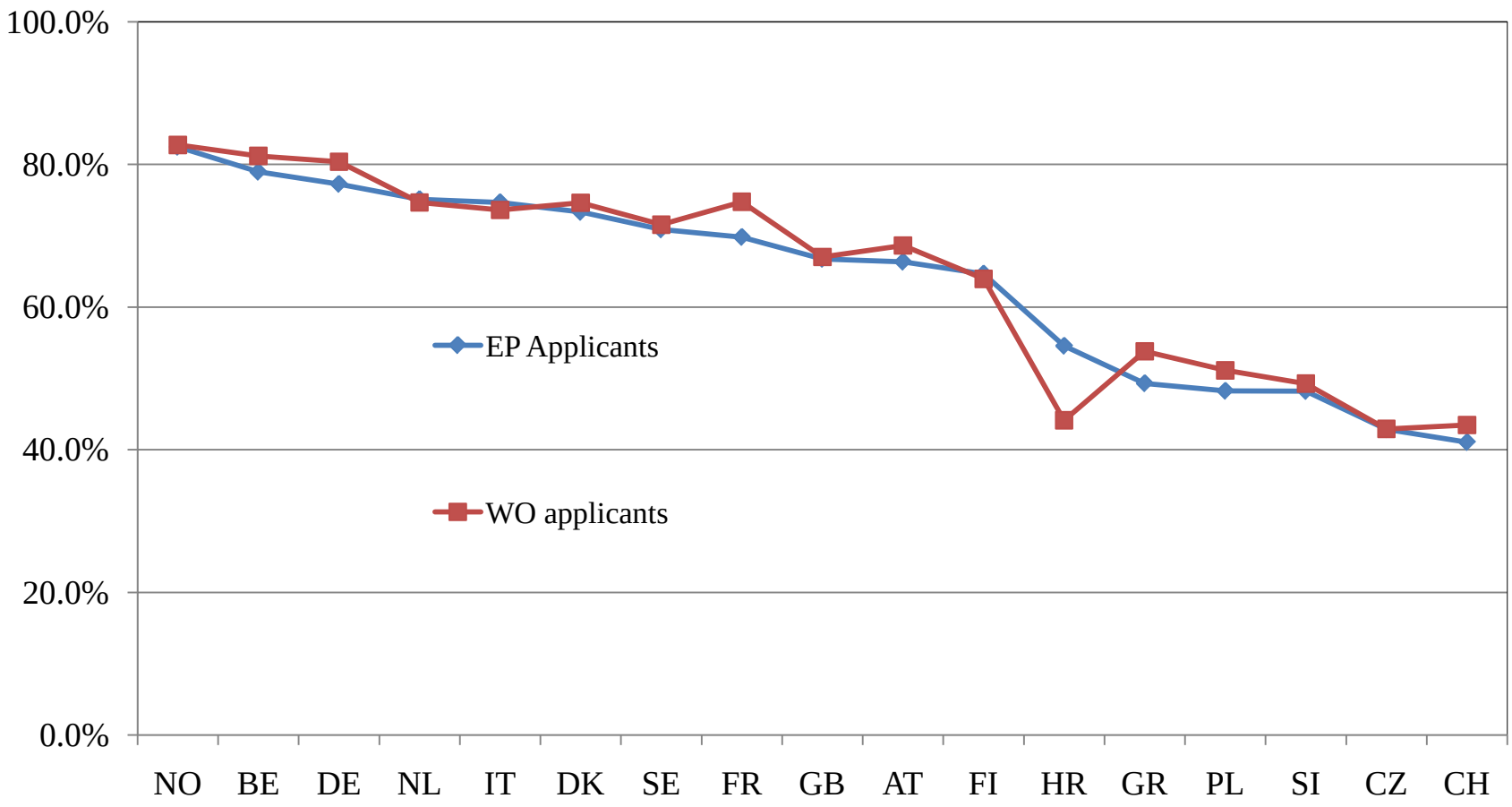

Figure 3b: Share of business applicant names matched
(active in year 2001 or later, top 18 EU countries)

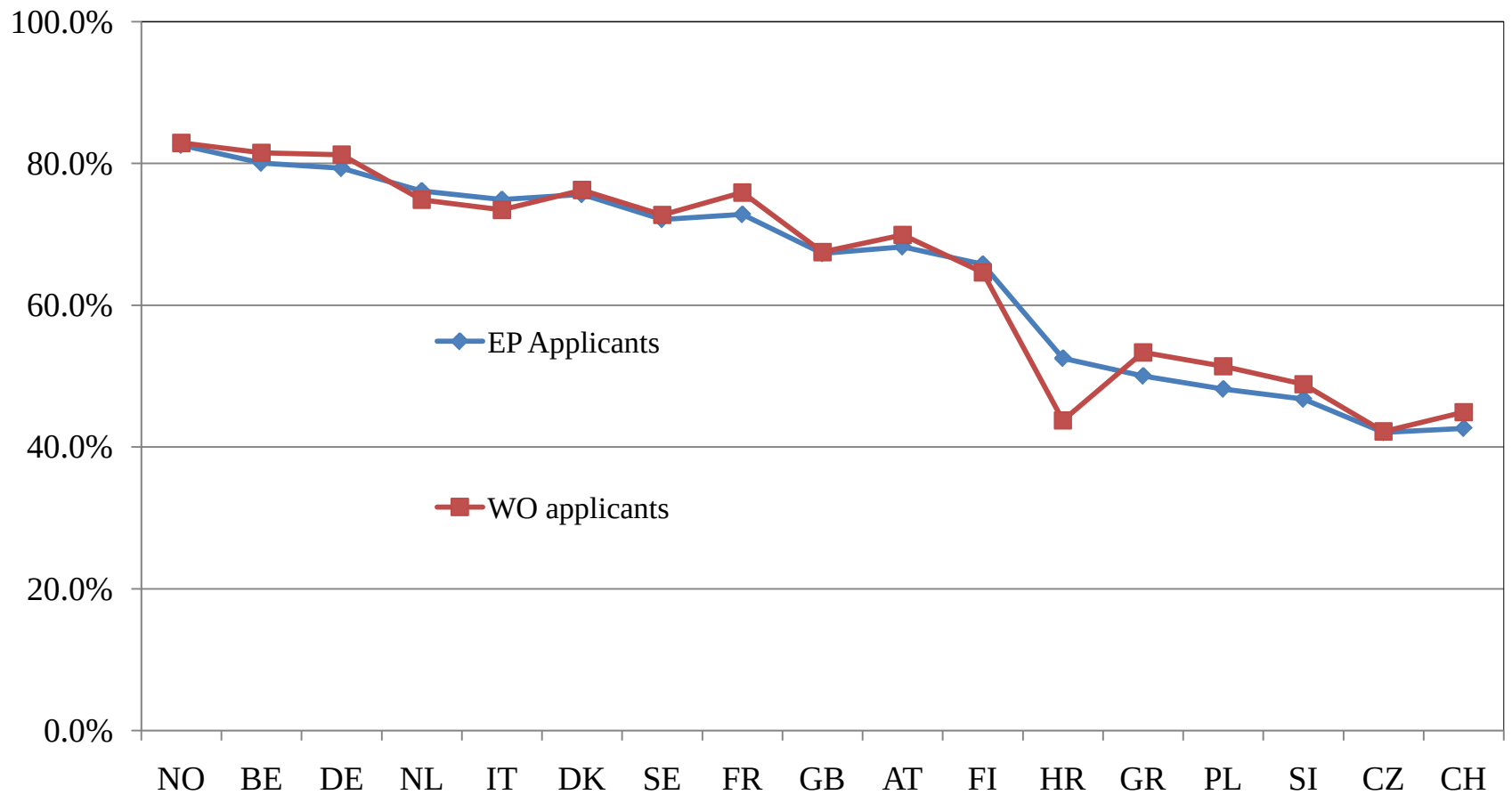

Figure 4a: Share of business applicant names matched, weighted by number of applications (top 18 EU countries)

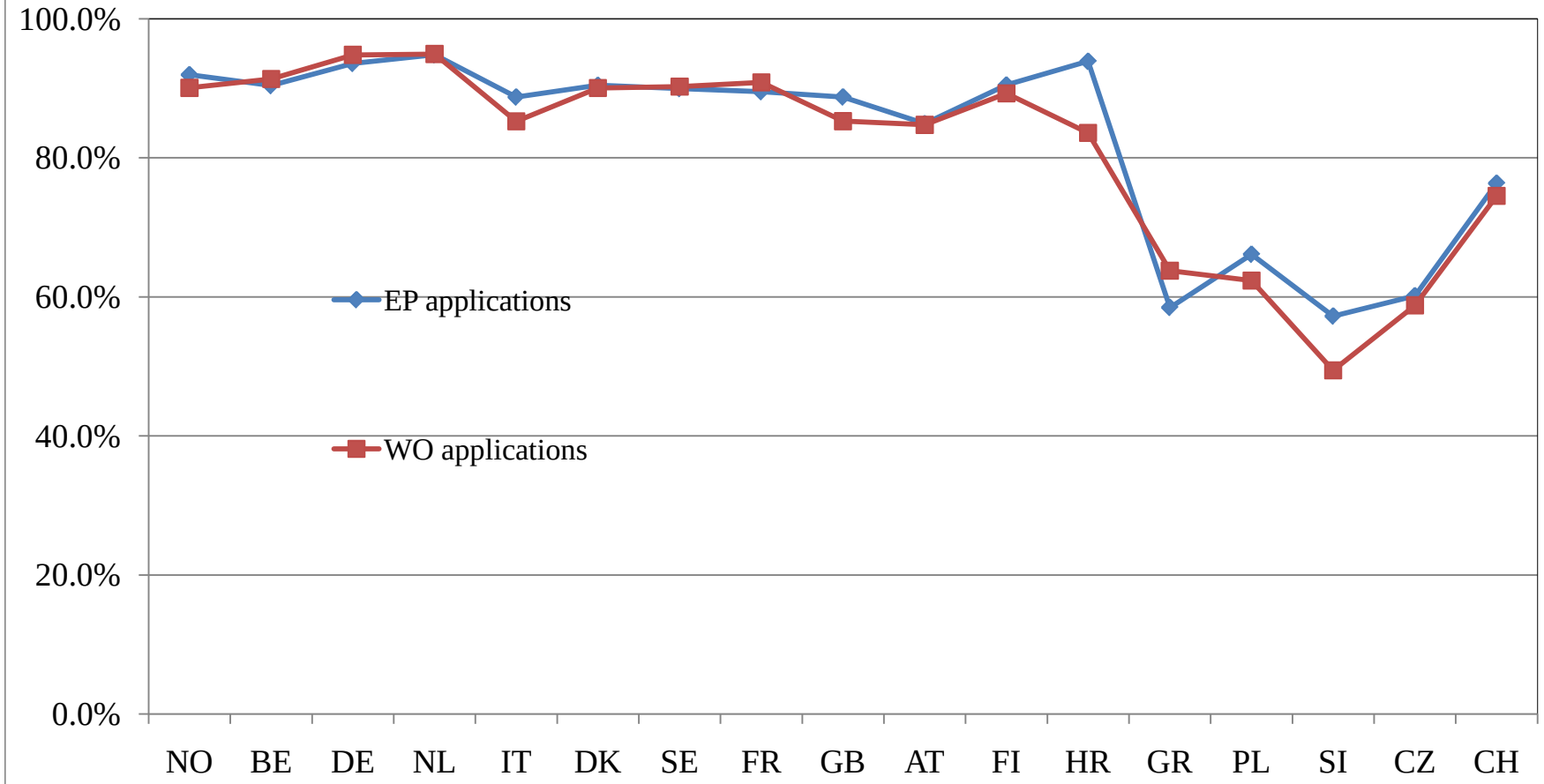

Figure 4b: Share of business applicant names matched, weighted by number of applications (active in year 2001 or later, top 18 EU countries)

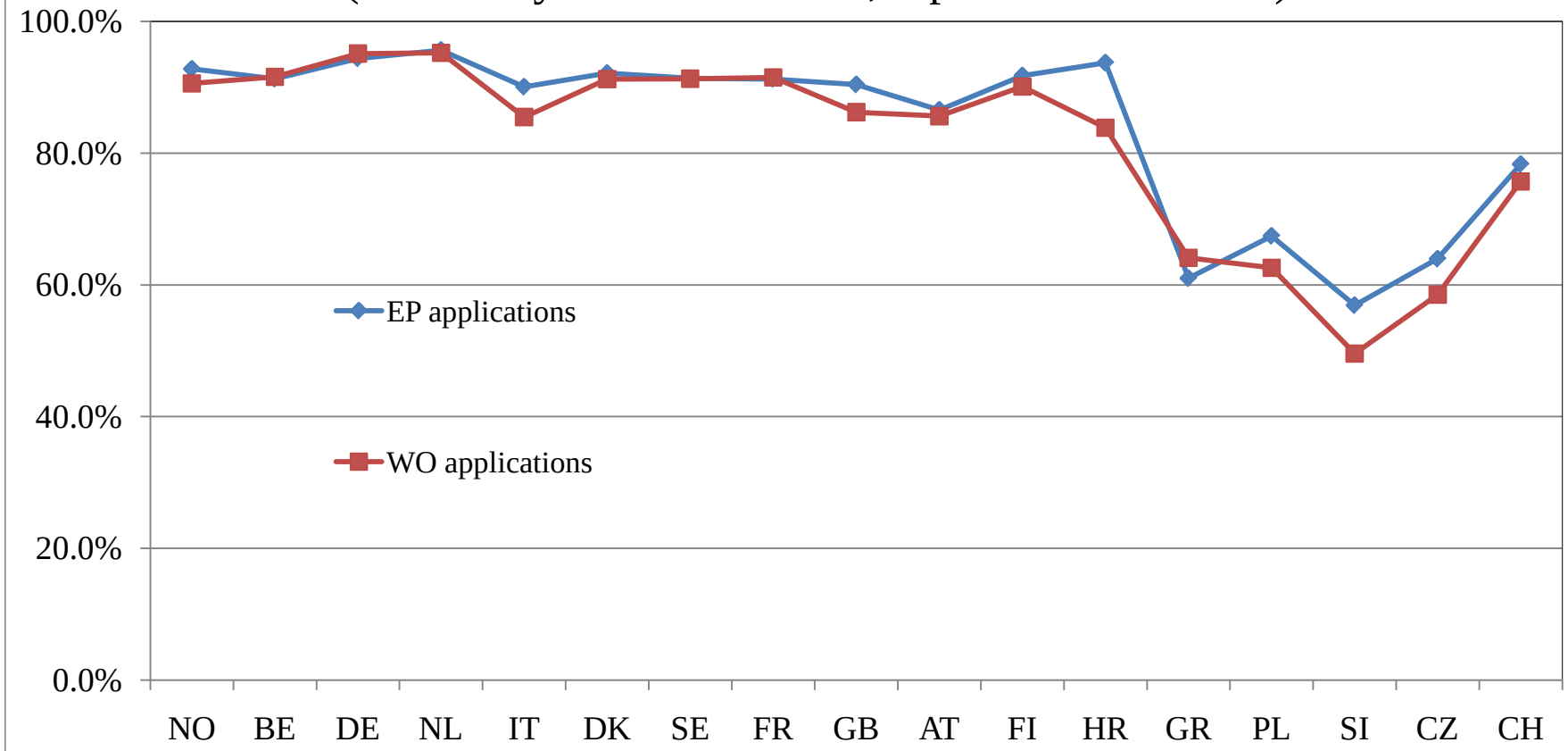

Figure 5: Share of business assignees in the USPTO dataset matched to Amadeus (top 18 countries)

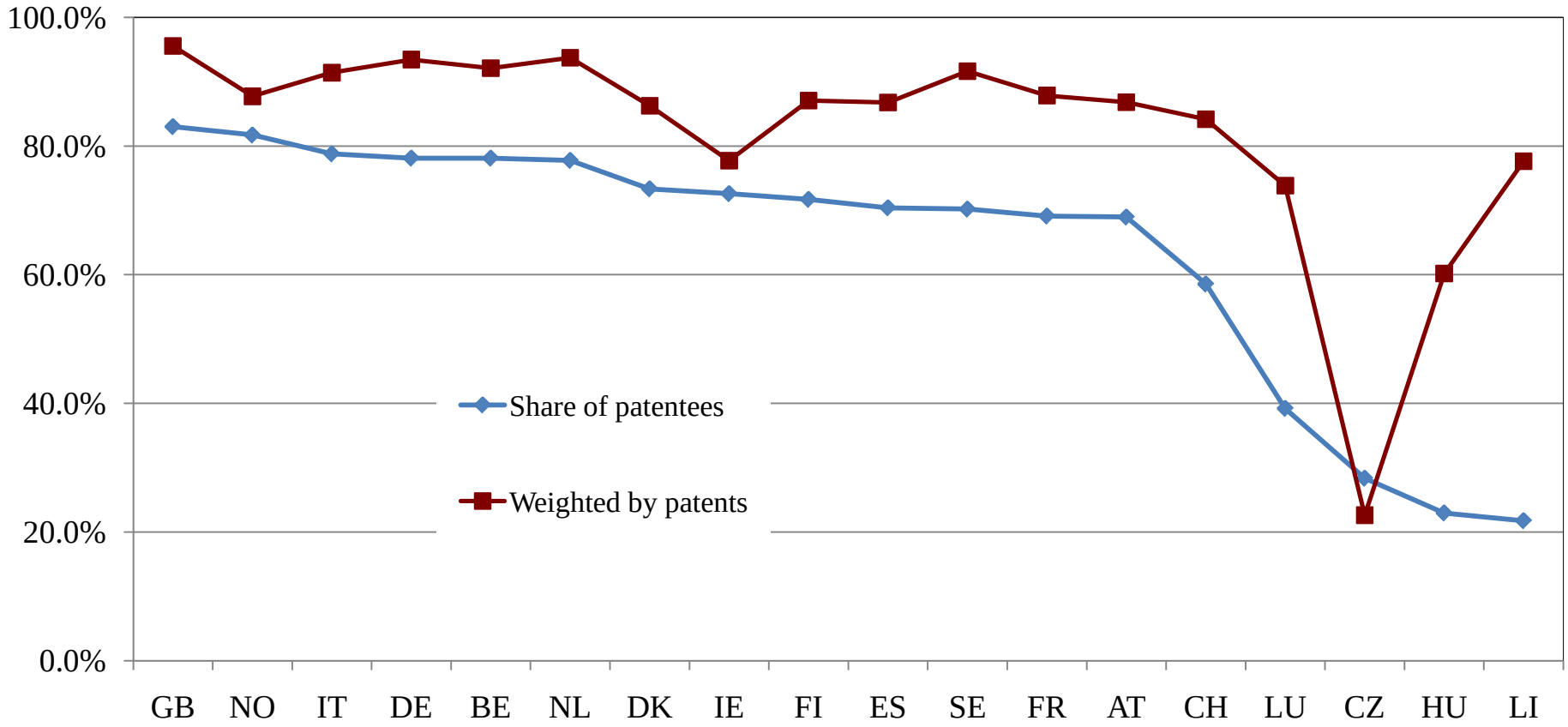

Figure 6: Distribution of the match score for EPO/PCT patenter names matched to the Amadeus business directory

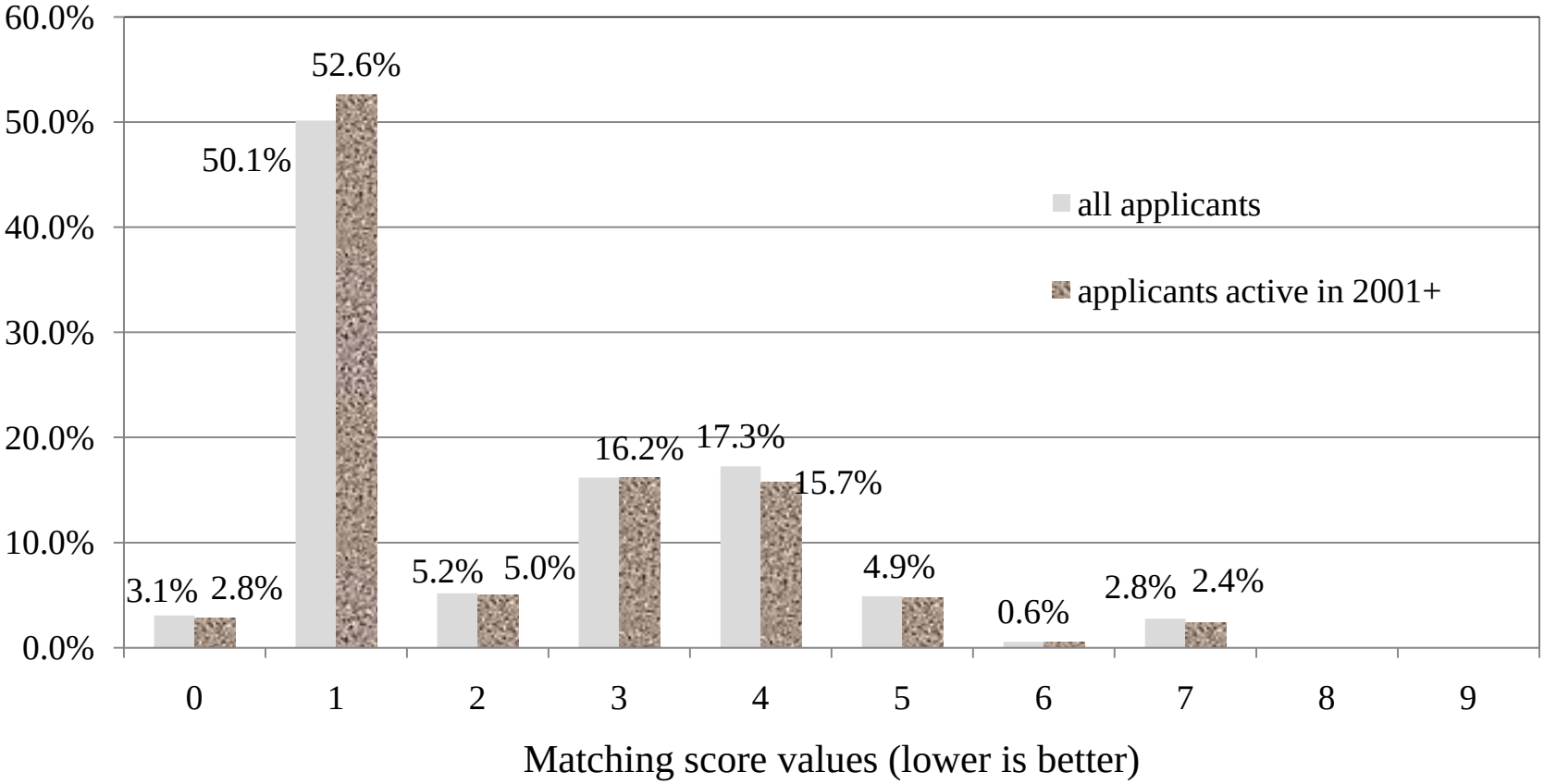

Figure 7: Distribution of the match score for the USPTO assignee names matched to the Amadeus business directory