Data production methods for harmonized patent statistics: Patentee name harmonization

Tom Magerman[2, 3], Joris Grouwels[2], Xiaoyan Song[3], Bart Van Looy[1, 2, 3],

[1]*Managerial Economics, Strategy and Innovation, Faculty of Ecomomics & Applied Economics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*

[2]*Research Division INCENTIM, Faculty of Economics & Applied Economics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*

[3]*Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium*

Leuven, December 2009

Acknowledgments: The authors want to thank Julie Callaert, Bert Peeters and Caro Vereyen for providing support in data processing and verification of outcomes.

Corresponding author E-mail address: Tom.magerman@econ.kuleuven.be; Xiaoyan.song@econ.kuleuven.be and/or Bart.vanlooy@econ.kuleuven.be

Recommended citation: Magerman T., Grouwels J., Song X., Van Looy B. (2009), "Data production methods for harmonized patent statistics: Patentee name harmonization", Eurostat Working Paper.

1 INTRODUCTION

Patent documents are one of the most comprehensive data sources on technology performance. Although technology indicators based on patent documents have certain limitations[1], Griliches' observation of two decades ago still seems to hold: "*In spite of all the difficulties, patent statistics remain a unique resource for the analysis of the process of technical change. Nothing else even comes close in the quantity of available data, accessibility, and the potential industrial, organizational and technological detail*." (Griliches, 1990). Patent indicators are now used by companies and by policy and government agencies[2] to assess technological progress on the level of regions, countries, domains[3], and even specific entities such as companies, universities and individual inventors. However, with respect to the latter (i.e. analysis on the level of the patentee), specific concerns can be discerned.

These concerns stem from the heterogeneity of patentee names found in patent documents. The same organization or individual can appear in different guises when patentees apply for patents through different channels over extended time periods. While this poses no real challenge to the functioning of the patent system itself – where patent documents are used on a recurrent basis to assess prior art – it complicates analyses on the level of patentees. The analyst is confronted with inconsistencies such as spelling mistakes, typographical errors and name variants, which often reflect idiosyncrasies in the organization of research and intellectual property right activities at particular moments within one and the same organization.

These discrepancies in the naming of identical patentees in current patent databases justify efforts to achieve name harmonization so that analysis at the level of patentees can be facilitated. Quality, in terms of both completeness and accuracy, is a crucial issue in this respect. We refer to 'completeness' as the extent to which the name-harmonization procedure is able to capture all name variants of the same patentee. 'Accuracy' relates to the extent to which the name-harmonization procedure correctly allocates name variants to a single, harmonized patentee name. Unfortunately, completeness and accuracy do not go hand in hand. Efforts directed to maximizing the number of identified name variants will ultimately lead to decreasing accuracy, while maximizing accuracy inevitably leads to an increase in missed or unidentified name variants, or to a decrease in completeness.

In 2006, we developed a comprehensive method to achieve harmonization of patentee names in an automated way (Magerman et al., 2006) with an emphasis on accuracy. This method was developed and applied on an extensive set of all patentee names found in all EPO patent applications published between 1978 and 2004 and in all granted USPTO patents published between 1991 and 2003. The current update builds on the complete person table provided by PATSTAT, implying an additional extension in terms of names (450.000 versus 11.000.000+ names). While the applied methodology basically has remained the same, significant steps have been undertaken to improve the name harmonization method. First, efforts have been made to improve completeness by adding extra rules (see infra) - mainly addressing legal forms and country indications – while remaining true to the philosophy of emphasizing accuracy. Second, in order to cope with the considerable increase of treated data, it became necessary to engage in a complete code overhaul and to port the existing methods to a more powerful environment. Whereas the 2006 version of the method was conveniently implemented in MS-Access, this platform proved utterly inappropriate for the current volume of data. Therefore, an implementation environment based on Java and Oracle SQL has been developed.

---

1 Propensities to patent differ among industries, firms and countries.

2 Patent indicators are now to be found in recurrent publications of the National Science Foundation (US), the European Commission (Science and Technology Indicator Reports) and the OECD alike.

3 Analysis by domains is feasible by using the WIPO International Patent Classification or aggregation schemas like the 'Systematic of OST/INPI/FhG ISI of 5 technology areas and 30 sub-areas'; analysis in relation to industries is enabled by concordance schemes based on patent classification, like the MERIT concordance table (Verspagen, 1994), the OECD Technology Concordance (Johnson, 2002), or the EC DG Research and FhG ISI/OST/SPRU concordance table (Schmoch, Laville, Patel, Frietsch, 2003).

Apart from these major steps, some modifications have been introduced (e.g. possibility of restricting rules to country codes.)

Before discussing in detail the methodology and its effects as applied to the patentee name list, we will first clarify the difference between patentee name harmonization and legal entity identification. In addition, and in order to shed light on our specific contribution, we will briefly expand on the methods and approaches that have been developed previously to address the issue of patentee name harmonization (section 3). In section 4 we outline the different steps undertaken; the obtained results are presented and discussed in section 5.

2 PATENTEE NAME HARMONIZATION AND LEGAL ENTITY HARMONIZATION

The focus of the methodology outlined in this paper is on patentee name harmonization. This is not the same as harmonization on the level of the legal entity. Legal entity harmonization is concerned with the identification of all patents owned by one and the same legal entity. In this respect, legal entity harmonization is not only concerned with name inconsistencies but takes into account mergers and acquisitions, name changes, and subsidiaries. For instance, when aiming at legal entity harmonization, all patents held by Hewlett Packard, Digital Equipment Corporation and Compaq might be considered as belonging to one and the same legal entity; likewise, "ANDERSEN CONSULTING" would become harmonized to "ACCENTURE" (name change).

In other words, when harmonizing legal entities, every patentee name needs to be checked against historical information on naming practices and ownership in order to address the following issues:

- Identification of entities (business units, departments, subsidiaries) that may have a different name but that belong to the same legal entity;
- Identification of name changes over time;
- Identification of mergers and acquisitions;
- Identification of joint ventures;
- Identification of mother and daughter relationships / subsidiary companies.

It is clear that this type of information is not available in current patent databases. External information is needed - on ownership, changes of ownership, and organizational practices with regard to names - to arrive at a comprehensive methodology for legal entity harmonization. Given the absence of databases providing exhaustive coverage of information needed to achieve legal entity harmonization[4], such efforts are not included in the name-harmonization methodology outlined in this paper.

Accordingly, our methodology focuses on the identification of name variations by comparing each patentee name with all other patentee names. The objective is to match names that appear to be similar but that differ because of spelling or language variations. The same patentee name can appear in a different form in the patentee name list for the following reasons:

- Spelling variations, e.g. "IBM" and "I.B.M.", or "BAIN & CO" and "BAIN AND COMPANY";
- Typographical errors, e.g. "INTERNATIONAL BUSINESS MACHINES" and "INTERATIONAL BUSINESS MACHINES";
- Addition of the legal form (again with possible acronyms, spelling variations, mistakes, and typographical errors in the legal form), e.g. "IBM", "IBM CORP.", "IBM CORPORATION" and "IBM COPRORATION", or "BAYER", "BAYER A.G." and "BAYER AG";

[4] While information providers like Graydon, Dunn & Bradstreet, Bureau Van Dijk and Thomson Scientific offer data on mergers and acquisitions and subsidiaries, this information is limited to larger entities and/or is confined to more recent years.

- Errors, e.g. "INTERNATIONAL BUSINESS MACHINES" and "INTELLIGENT BUSINESS MACHINES";
- Addition of establishment, business unit, department, subsidiary name or geographic identifier, e.g. "IBM" and "IBM JAPAN";
- Acronyms, e.g. "IBM" and "INTERNATIONAL BUSINESS MACHINES".

All of these issues will be analyzed in a systematic manner in order to develop an appropriate methodology. It will become apparent that spelling variations, typographical errors and the additions of legal forms can be addressed in an automated manner while for errors, acronyms and business unit or department extensions, additional validation efforts will be required in order to be accurate. However, before discussing in detail the methods and their impact in detail, it can be noted that name harmonization efforts concerning patentee names have been undertaken in the past, notably by USPTO and by Derwent (Thomson Scientific). Before discussing the development of the name cleaning and harmonization procedures proposed in this paper, these other approaches will be briefly discussed.

3 EXISTING NAME HARMONIZATION APPROACHES

3.1 USPTO CONAME ASSIGNEE NAME HARMONIZATION

As part of the USPTO TAF database, first-named assignee names of organizational entities are harmonized for utility patents granted since 1969.

The USPTO harmonization rules are conservative, as further consolidation of names is considered far easier than separating combined names. Harmonization efforts do not address subsidiary ownership, but are limited to identify assignee name variations. In addition, organizations with similar names but associated with different countries or a different legal form are not harmonized.

In the case of patents granted prior to July 1992, harmonization is primarily based on a manual process of comparing names. For patents granted after July 1992, harmonization is largely based on an automated procedure. This procedure can be summarized as follows:

- Extract name of first-named assignee;
- Condense assignee name by removing spaces and non-alphanumeric characters;
- Convert to uppercase characters;
- Match condensed name with existing list of condensed and harmonized names;
- Manual review of all new assignee names not yet matched to an existing name in previous step (e.g. by looking at assignees of other patents granted to the same inventor or inventors);
- Annual large scale manual review to verify integrity of the entire assignee file.

The partial manual approach of USPTO offers potential to achieve high levels of completeness. Especially the 'staging' approach, whereby new names not yet matched are compared with previously harmonized names, allows for a complete harmonization solution.

On the other hand, the USPTO harmonization has several shortcomings:

- The partial manual approach implies significant resources every time new patentee names appear in the database;
- Only the first assignee is processed;

- Names reflecting different legal forms or associated with different countries are not combined[5];
- The manual review process is not transparent and might cause rule variation since harmonization is performed by different persons, jeopardizing the reproduction on a broader set of names (e.g. EPO applicant names, second assignee) [6].

3.2 DERWENT WPI COMPANY NAME HARMONIZATION

The DERWENT WORLD PATENT INDEX provides patentee codes for all patentees. One can summarize the DERWENT WPI method to produce these patentee codes as follows[7]:

- Take the name and replace commonly occurring words with a standardized version or abbreviation, as listed in the DERWENT abbreviated word list (Russian and Japanese words are first translated to English);
- Select the first significant word(s) of the resulting name, ignoring 'common' words listed in the DERWENT list of common descriptors;
- Replace frequently occurring words recorded in the DERWENT list of general descriptors with a two-letter abbreviation;
- Replace continent, country, region and town names with a two-letter abbreviation (some commonly used names are replaced with three-letter abbreviations);
- Replace points of the compass with one- or two-letter abbreviations;
- Take the first four letters of the remaining word.

This results in a long list of so called non-standard patentee codes consisting of four letters. These codes are not necessarily unique; several unrelated patentees can have the same automatically generated patentee code[8].

Next, a selection of these patentees is analyzed in depth to arrive at unique standard patentee codes. The emphasis in this phase shifts towards legal entity harmonization. This objective is achieved by incorporating additional information on companies derived from secondary financial sources. These efforts are however limited to patentees applying for larger numbers of patent applications. This reduction is understandable since arriving at standard patentee codes in the WPI approach implies legal entity harmonization: mergers and acquisitions, name changes and subsidiaries.

At present, the index of standard patentee codes provided by WPI contains 21,000 entities and can be considered the most comprehensive harmonized index currently available, as it includes legal entity harmonization. At the same time, the process to arrive at standard names is not transparent and case specific (for example: standard codes are retained for company name changes.In case of mergers and acquisitions however, either one of the codes is retained and the others abandoned, either a new code is created). The precise rules that have been applied

---

5 For example, in the USPTO harmonization, the following name variations of "BURR-BROWN" can be found in the list of harmonized names: "BURR-BROWN CORPORATION", "BURR-BROWN INC." and "BURR-BROWN LIMITED".

6 For instance, this can be observed in the list of original assignee names harmonized to "AT&T CORP.": "Bell Telephone Laboratories Inc.", "AT&T Corp/CSI Zeinet (A Cabletron Co.)", "ATT Corp--Lucent Technologies Inc" and "AT&T Middletown". It is clear that some of these names are associated with "AT&T Corp." based on criteria other than name similarity. However, it remains unclear which additional rules have been applied and to what extent.

7 For a more detailed description, see: http://www.thomsonscientific.com/media/scpdf/patenteecodes.pdf

8 For example, the non-standard code "HUSS" is associated with "HUSSMANN CORP", "HUSSOR SA", "HUSSOR ERECTA SA", "HUSS MASCHFAB GMBH & CO KG", "HUSS UMWELTTECHNIK GMBH" and "HUSSMANN DO BRASIL LTDA".

in each case are only evident after analysis of the names associated to a certain standard patentee code (information which is not publicly available)[9].

For companies for which a standard code is not available (because of a limited number of patents), or for companies not recognizable as a subsidiary of a company that already has a standard code, the automatically generated non-standard code cannot be considered appropriate to achieve harmonization of the complete list of patentee names. The rules for arriving at the non-standard code result in numerous false matches and a low level of accuracy[10].

4 A CONTENT-DRIVEN NAME HARMONIZATION APPROACH FOCUSING ON ACCURACY

As indicated in the introduction, name harmonization involves a trade-off between completeness and accuracy. It has been a deliberate choice in the methodology outlined here to favor accuracy over completeness for reasons of transparency, as it is easier to combine additional names than to separate combined names. An accurate but somewhat incomplete set of harmonized names provides users with ample opportunities to extend the methodology and its results to a broad range of applications. Given an accurate set of harmonized names, additional name matches that are considered relevant can be identified and added in a straightforward way. Reverse operations, starting with a more complete set, are much more complicated since previous steps undertaken to achieve a more complete result might need to be undone or 'reverse engineered'. In practice, this would prove to be a much more complicated endeavor than combining disaggregated names. Hence, this methodology, conceived as a transparent and accurate set of harmonized names in which completeness can be gradually improved, is considered far more appealing than a more complete set which contains the risk of not being accurate or being unsuited to specific analytical purposes.

As a result, the development of the methodology is based on the underlying principle that every step in the cleaning and harmonization process must increase completeness without decreasing accuracy. Every action that jeopardizes accuracy will ultimately be excluded from the methodology, because combining two names that belong to two different legal entities has to be avoided at all cost. Moreover, in order to achieve sufficient levels of accuracy, several of the procedures and rules that have been developed take into account the specificities of the full original name list. This content-driven approach results in a partly manual, and hence labor-intensive, development process.

The final procedure can be completely automated in a modular approach to allow further refinements and improvements. The entire procedure is organized as a series of generic steps and sub-steps that are implemented by taking into account the nature of the source data. It should be noted that, while the more generic parts of the procedure can be used for all kinds of name-harmonization applications, some procedures are highly content-specific and additional analysis and refinements might be needed to apply the methodology to a different set of organization names.

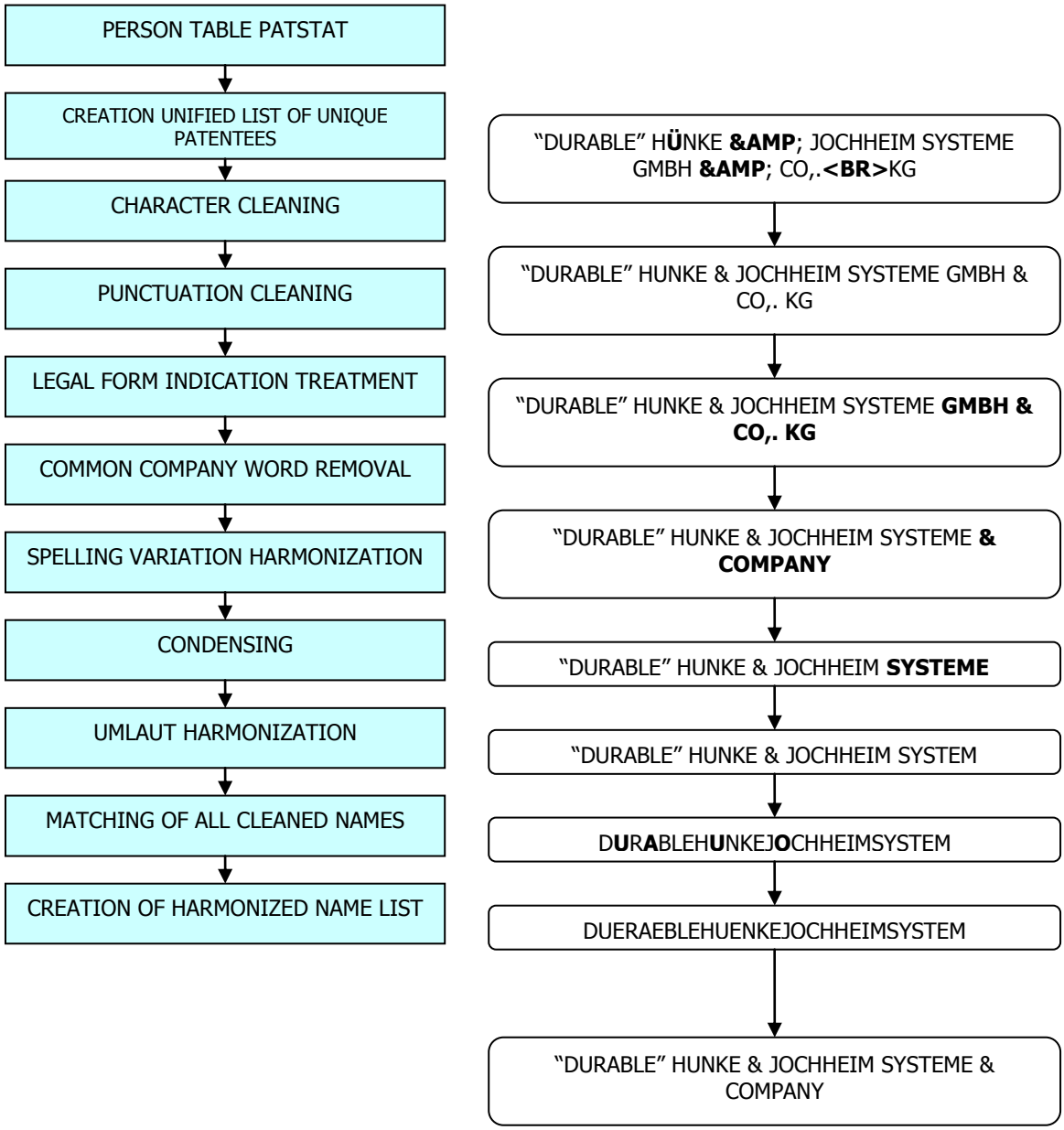
Figure 1 contains an overview of the developed methodology, consisting of a sequence of steps that include both data pre-processing and name-harmonizing activities. An example patentee name is included to illustrate the results of each step (string parts that will be affected in the next processing step are highlighted in bold).

---

[9] For example, the standard code "CANO" is associated with "CANON CAMERA", "CANON KK", "CANON PRECISION INC", "CANON PRECISION MAC" and "CANON SEIKI KK". Another standard code "CAND" is associated with "CANON DENSHI KK", "CANON ELECTRONICS CO LTD" and "CANON ELECTRONICS INC".

[10] These non-standard codes are however useful because they provide a high level of completeness, resulting in a maximum set of names that might be combined.

Figure 1: Overview schema name cleaning and harmonization



4.1 DATA PRE-PROCESSING

In the pre-processing steps, data are prepared for processing to facilitate actual name cleaning and harmonization. The individual impact of each step on the number of unique patentee names is limited but it smoothes progression through consecutive steps and it considerably increases the overall impact. Data pre-processing is highly dependent on the content of the underlying data. Consequently, extensive refinements or adaptations may be needed when processing names from a different data source.

4.2 CHARACTER CLEANING

Depending on the data source, non-letter (A to Z) and non-digit (0 to 9) characters can be coded or represented in a variety of ways (e.g. ANSI, SGML), inducing additional name variations. Data can also contain codes that bear no relation to the real data and that merely represent formatting issues, again inducing additional name variations.

Character cleaning removes different types of character representations and formatting codes or converts them to genuine standard ASCII characters. For instance, HTML formatting codes such as "
" are removed or replaced by spaces and SGML codes such as "&OACUTE;" are removed or replaced by their ASCII equivalent whenever possible.

In this step, names are also scanned for proprietary coded characters like "{UMLAUT OVER (A)}" in USPTO data. These codes are also removed or replaced whenever possible. Accented characters like "É" are replaced with their unaccented ASCII equivalents. Particular problems with alternative spellings of the umlaut in German (and some other languages) are treated at a later stage (see section 4.3.5 – Umlaut harmonization).

4.2.1 Punctuation cleaning (pre-parsing)

Names may not only contain letters and digits but also characters such as ",", ";", and "-", used to separate words or to indicate abbreviations and combinations. These characters might complicate the separation or parsing of names into individual words, which is necessary in further cleaning steps (e.g. identifying the legal form). Punctuation cleaning aims to harmonize all of these punctuation characters, and to thereby facilitate the parsing of names in individual words at a later stage.

Firstly, double spaces are replaced with single spaces. Quotation marks followed by a space appearing at the beginning of a name, or preceded by a space appearing at the end of a name, are replaced with quotation marks without a trailing or leading space. Quotation marks are removed from names tha have only quotation marks at the beginning and at the end of the name. Next, names are scanned for non-alphanumerical characters at the beginning and at the end of the name, and these characters are removed if appropriate. Finally, comma and period irregularities are harmonized, so that commas are not preceded by spaces but followed by a space (unless acting as decimal or thousand separators) and so that periods are only preceded by letters or digits.

4.3 NAME CLEANING

In the name cleaning steps, the actual name cleaning and harmonization is performed. As mentioned previously, our approach is based on the specific data content. Extensive refinements or adaptations might be needed when names from a different data source are processed.

4.3.1 Legal form indication treatment

A lot of patentee names contain some kind of legal form indication (e.g. "INC.", "LIMITED", and "LTD."). These legal form indications are responsible for a considerable number of name variations due to the variety of abbreviations and spellings used. In this step, legal form indications are harmonized and moved to a separate field, thereby considerably reducing name variations.

4.3.2 Common company word removal

Legal form indications are separated out since they do not constitute a distinctive part of the name; this logic applies to some other words as well. In the case of companies especially, additional words like "COMPANY", "CORPORATION", "GESELLSHAFT" and "SOCIETE" add nothing to the distinctive character of a company name. When two names are found to be identical except for the presence of such words, the underlying patentee name will be considered as referring to one and the same organization. Examples include "3COM" and "3COM CORPORATION", "AMIC" and "AMIC COMPANY", "BAUR SPEZIALTIEFBAU" and "BAUR SPEZIALTIEFBAU GESELLSCHAFT", and "SOCIETE NOVATEC" and "NOVATEC".

4.3.3 Spelling variation harmonization

Typographical errors and spelling mistakes are responsible for considerable name variations. These kinds of error can be identified by assessing word similarities. Whereas this type of analysis is straightforward for common English words, proper names usually require manual validation efforts in order to ensure accuracy. For example, "AMTECH" and "IMTECH" only differ in a single character but it would be incorrect to automatically assume that the names refer to one and the same patentee. For common words, spelling and language variations can be identified without ambiguity and, therefore, harmonized effortlessly. For example, "SYSTEM", "SYSTEMS", "SYSTEMEN", and "SYSTEMES" can all be harmonized to "SYSTEM" or "SYSTEMS". Spelling variation harmonization replaces all variants of common words with one harmonized variant that will be used to match name variants.

4.3.4 Condensing

Significant name variations are also caused by word separation, punctuation, and non-alphanumerical characters, which clearly have no relevance in identifying the distinctive characteristics of a name (e.g. "3 COM" and "3COM", and "AAF-MCQUAY", "AAF MCQAY" and "AAF – MCQAY"). Condensing removes all non-alphanumerical characters so that a harmonized variant can be used to match names.

4.3.5 Umlaut harmonization

Although accented characters have already been replaced (see section 4.2 – Character cleaning), German characters with a diacritic mark (umlaut: "ä", "ö", "ü") still generate spelling variations because words containing them can occur in three varieties, one with an umlaut (e.g. "für"), an alternative spelling without an umlaut but with an additional "e" (e.g. "fuer"), and a simplified form without both an umlaut and an additional "e" (e.g. "fur"). Umlaut harmonization identifies and matches different variants of words including "ä", "ö" and "ü".

4.4 IMPROVEMENTS 2009

4.4.1 Extending legal form coverage by country (language)

When the original algorithm was applied to the extended dataset, analysis revealed a bias towards bigger countries. This was due to the fact that the legal form discovery in 2006

occurred on indexes (first word, last words, full text) of patentees irrespective of their country codes. Discovery of new legal forms is a tedious manual process, and this approach guarantees the best overall yield, but at the same time it introduces a bias in favor of bigger countries and more commonly used languages (USA, Germany, UK, France, Japan…; English, Japanese, German, French,..). Legal forms that occur less frequently (e.g. because in Greek or Bulgarian) are less easily noticed, with frequency counts being the main criterion to guide the manual search and validation process. The same problem holds for less common legal forms from bigger countries.

In order to remedy this issue, an environment was created that made it possible to explore the data on a per-country code base. It offers a way to search the last words and their respective names in a hierarchical way (1 last word > 2 last words > 3 last words) and it reveals the frequency of occurrence. Names without any country code were not included. Every time a potential candidate for a new legal form rule was identified, its validity was checked on the complete data set to make sure that objectives of precision are not jeopardized.

The country codes that were visually inspected are: US, JP, DE, GB, FR, CA, KR, IT, AU, SE, NL, CH, TW, IL, CN, RU, ES, FI, DK, AT, BE, IN, NO, SU, NZ, SG, IE, BR, HU, HK, PL, CZ, GR, MY, LU, SI, PT, LI, BG, SK, RO, EE, CY, LV, LT and MT.

These efforts resulted in 292 new legal form rules, of which the majority was not present before in any variation. In addition, we have been able to identify 630 new rules, stemming from variants of legal forms that had already been identified in the methodology of 2006. These new variants stem from extending the data to all patent offices (<> USPTO, EPO, WO).

Note that some candidate rules generated a considerable number of hits for certain language groups/country codes, but at the same time yielded errors in other cases. To minimize this risk for errors, a feature was introduced to restrict a rule to one or more country codes. E.g. rules concerning the legal form A.G. (Aktiengesellschaft) became restricted to companies from German speaking countries (country codes DE, DT, DD, CH) in order to limit the probability of making mistakes (A.G. could also be name initials, for example). This logic was also applied with respect to spelling variation of legal forms and/or common words.

4.4.2 Discovery of spelling variations of already known legal forms and common words using approximate string searching

Patent data from PATSTAT are full of spelling variations. This is also the case for legal forms in assignee names. Most of the spelling variations occur only in a limited number of cases and are therefore not captured by inspecting frequency occurrences of certain (combined) words. Finding such spelling variations becomes feasible by using approximate searching[11]. The patterns for legal forms in the beginning and at the end of a name were matched approximately with the dataset. The resulting matches were verified manually, as they include false positives. The retained correct matches were then converted to rules that are now integrated in the name harmonization algorithm. A similar procedure was followed for common words. The total amount of new rules for legal forms and common words together was 2227.

4.4.3 Country Code Correction

It was found that a large number of names in the PATSTAT person table have a country code added to them as a suffix in the name field. This prevents parts of the harmonizing algorithm from working correctly. An analysis revealed that it was possible to remedy this problem for certain authorities in a preprocessing step (see infra, 5.1.). In total, 956,479names were hence corrected.

[11] Several approximate string searching tools were tried, and eventually the TRE library (http://laurikari.net/tre/) has been used within this exercise.

4.4.4 Porting toolbox to UTF8 and Java/Oracle

The April 2008 dataset already stretched the MS Access-platform to its limits with 2.8 million records. So, for applying the methodology to the even larger complete person table of PATSTAT, the toolbox needed to be ported to a more powerful platform. A combination of Java and Oracle was chosen as a solution. This allows to process large amounts of records and at the same time to program the application in a generic way, allowing for new rule mechanisms that can be used in future releases (e.g. full support for regular expressions.).

The harmonization algorithms were adapted for the processing of UTF8-data.

5 APPLICATION OF THE METHOD ON PATSTAT PATENTEES (EDITION OCTOBER 2009)

The complete name cleaning and harmonization procedure has been applied to all 11,100,882 patentee records present in the PERSON table of PATSTAT (edition October 2009). These 11,100,882 patentee records contain 9,310,595 distinct names (ignoring uppercase/lowercase variations in names). The cleaning and harmonization procedure reduces this number of names to 7,536,191, a reduction of 19%.

In the following sections, we will describe the practical application and results of intermediate steps, validation of the method, and final results and impact for all PATSTAT patentees.

5.1 PATSTAT DATA PREPROCESSING

A particular observation for the PATSTAT patentee records is the presence of country codes in the patentee name field for a considerable number of records (e.g. 'WELDON TOOL AND ENG CO,US'). This phenomenon hampers the implementation of the name harmonization method. Before processing PATSTAT patentee names, these country code suffixes had to be removed.

Analysis revealed that these country code suffixes are mainly present for a limited number of patent authorities. The phenomenon is particularly present for German and Soviet patents, and to a lesser extent for US patents, patents from some East-European and Scandinavian countries, and France.

To solve the problem, two-character strings preceded by a comma, appearing at the end of patentee names, that correspond with a valid country code (ISO 3166 country code standard) were removed from patentee names that are linked to patents of relevant patent authorities (with some additional constraints: the potential country code is not different from the country code of the address and the potential country code can not be confused with a common legal form abbreviation that is relevant for the country).

In total 956,479 person names from 10 patent authorities were corrected for country code suffixes. This correction reduced the number of distinct original names from 9,310,595 to 9,206,551 names, a reduction of 1,1%.

One should notice that we only removed country code suffixes in the name field of patentee records in PATSTAT. We observed more general problems with address information being added to the patentee name field (e.g. postal codes and city names). Especially patentee records linked to German patent office patents seem to suffer from this (e.g. 'MOCO MASCHINEN- UND APPARATEBAU HUBER GMBH, 6806 VIERNHEIM, DE'). Here we only dealt with country codes because of the more general nature of the problem and to avoid 'false removals'. Dealing with all address information present in name fields would require a far more elaborated approach.

5.2 IMPACT OF INTERMEDIATE STEPS

The more than 4,000 rules of the name harmonization method were executed on all patentee names.

Table 1 contains the impact of intermediate name harmonization steps (numbers in the column 'DISCTINCT NAMES' represent the number of distinct names after the name harmonization step mentioned in column 'STEP').

Table 1 : Impact of intermediate name harmonization steps

STEP	DISTINCT NAMES	DROP	RATE
Distinct original patentee names	9,310,595		
Country code suffix removal	9,206,551	104,044	1.1%
Character cleaning	9,193,182	13,369	0.1%
Punctuation cleaning	9,123,685	69,497	0.7%
Legal form indication removal	8,588,630	535,055	5.7%
Common company word removal	8,504,278	84,352	0.9%
Spelling variation harmonization	8,493,979	10,299	0.1%
Condensing	7,548,399	945,580	10.2%
Umlaut harmonization	7,536,191	12,208	0.1%
Total		**1,774,404**	**19.1%**

Condensing has by far the biggest impact (about 50% of the total impact of the name harmonization), followed by legal form indication removal (about 25% of total impact).

5.3 VALIDATION

Before discussing final results and impact, we want to focus on the validation of the method.

We did both a precision and recall validation, based on samples. In the precision validation, we verified whether an original name was harmonized correctly. In the recall validation, we checked for the presence of missed names, i.e. names that should have been harmonized, but that were not. As the method was designed with maximum accuracy in mind, favouring accuracy over precision, we expect better precision results compared to recall results.

5.3.1 PRECISION VALIDATION

In the precision validation, we verified to what extent the method correctly harmonizes names. The precision rate is calculated by counting the correct number of linked names over the total number of linked names. We calculated precision rates based on sample sets that were validated by two independent raters. Each of the two human raters checked two sample sets, a small set of 250 harmonized names, and a big set of 1000 harmonized names. The harmonized names were randomly selected from the full population of harmonized names being linked to at least two different original patentee names. For all harmonized names in the sample sets, all original patentee names were retrieved, and for each pair of original name – harmonized name, a validation score was given by the human raters (Y, the original name is correctly linked to the

harmonized name; N, the original name is wrongly linked to the harmonized name, or there is doubt whether the original names is correctly linked to the harmonized name).

Table 2 contains the results of the precision validation. The number of errors and error rates in this table are presented on the level of harmonized names, i.e., they indicate the number and rate of harmonized names that have at least one original name that is wrongly linked to that original name.

Table 2 : Precision validation results

SAMPLE	RATER	HRM NAMES	ORIG NAMES	ERRORS	ERROR RATE
1	1	250	726	2	0.8%
2	2	250	703	1	0.4%
3	1	1,000	3,119	9	0.9%
4	2	1,000	2,988	6	0.6%
Total		**2,500**	**7,536**	**18**	**0.7%**

On average, 0.7% of the harmonized names have at least one original name that is incorrectly linked, with small variations between individual sample sets. On the level of patent volume, the number is slightly lower: the harmonized names having at least one original name wrongly linked to them represent about 0.5% of the patent volume of all harmonized names involved in the precision validation.

However, most reported errors are not clear errors, but doubtful cases for which it is difficult to determine whether the harmonized name is correct. Of the 18 errors reported, 10 cases are doubtful cases, 7 are real errors, and one case depends on the interpretation of the results. Examples of these are provided in the following paragraphs.

Doubtful cases are mostly cases in which names with legal form indications are mixed up with names without legal form indication, making it unclear whether all names belong to the same legal entity, or whether one of them belongs to a legal entity and the other belongs to an individual.

For example: 'GERHARD GEIER GMBH & CO. KG' and 'GERHARD GEIER' are both harmonized to 'GERHARD GEIER & COMPANY', although it is unclear whether the latter is an individual or the legal entity (no address information available to further inquire).

Another source of doubtful cases is the ambiguity between abbreviations of legal forms and initial of individuals.

Example: 'MATIMAR, S.A.' and 'MATIMAR, S. A.' (with space in between 'S.' and 'A.') are both harmonized to 'MATIMAR' and 'S.A.' is moved to the field containing the legal forms, assuming that 'S.A.' is an abbreviation of a legal form. But 'S.A.' may just as well represent the initials of an individual with surname 'MATIMAR'.

An example of a clear error is the harmonized name 'PAUL, S.' to which the original names 'PAULS LIMITED' and 'PAUL, S.' are linked. The likelihood is high that the first one refers to a company and the latter refers to an individual having nothing to do with the former. The combination of legal form removal and condensing erroneously brought the two names together.

An example of a case where the error is dependent on the interpretation is the harmonized name 'SCHNELL' with original names 'SCHNELL & CO.', 'SCHNELL S.P.A.' and 'SCHNELL 'S.R.L.'. All names are harmonized to 'SCHNELL' because the variation is in the legal forms. However, address information and information found on the internet reveals that 'SCHNELL S.P.A.' and 'SCHNELL S.R.L.' are indeed the same company (from Italy), but are different from 'SCHNELL & CO.' (from Switzerland). If the harmonized name would be used to identify unique entities, 'SCHNELL & CO.' would be incorrectly taken together with the others. If the combination of harmonized name and removed legal form would be used to identify unique entities, 'SCHNELL S.R.L.' and 'SCHNELL S.P.A.' would not be taken as the same company. Hence, both interpretations lead to mistakes. Making use of the country in the address information to identify unique entities can resolve such problems (take names having different legal forms

within one country together, but keep them separated if they have different countries). This approach might be limited in practice because of the lack of address information for many patentee records in PATSTAT.

To conclude, we observe a precision rate beyond 99%, and presumably beyond 99,5% taking into account doubtful cases and interpretation problems that can be resolved by making use of country information.

5.3.2 RECALL VALIDATION

The objective of the recall validation is to estimate how many names were missed by the name harmonization method, i.e., how many original names are not harmonized while they ideally should have been? The recall rate is calculated by counting the number of harmonized names over the number of names that should have been harmonized. The recall rate was calculated based on a sample set of harmonized names. In practice, this exercise implied three activities. First, a list of relevant keywords was constructed for every harmonized name in the sample (containing all relevant parts of the name to be used in a broader search for all similar names). Second, all original names that match the keywords are automatically retrieved using a approximate string search algorithm based on the Levenshtein distance (using the TRE-AGREP tool[12]). Finally, all retrieved original names are verified to confirm whether or not they should be linked to the harmonized name.

The 500 top patentees (after name harmonization) were used for the recall validation sample. All details on the sample and validation of the sample can be found in Peeters et al.(2009) as this recall validation sample is the basis of the exploratory assessment of top patentees elaborated in that paper.

Table 3 contains the results of the recall validation. Results are expressed at the level of names (how many name variants are captured by the harmonization method compared to all name variants that should have been captured for the sample) and at the level of patent volume (how many patents are linked to the name variants captured by the harmonization method compared to the total patent volume of all name variants that should have been captured for the sample). Overall figures for name variants and patent volume for all patent authorities/offices present in PATSTAT are also broken down by three major patent authorities/offices (WIPO, EPO and USPTO)

Table 3 : Recall validation results

	OVERALL	EPO	USPTO	WO
Recall rate at the level of names	35.6%	55.6%	31.3%	40.0%
Recall rate at the level of patent volume	77.9%	92.8%	91.0%	92.6%

These figures show that although recall rates are rather low at the level of name variants captured, recall rates in terms of patent volume are higher than 90% at the level of specific patent authorities/offices. The overall results are calculated on the number of names and the patent volume summed over all application authorities present in PATSTAT. The overall recall rate in terms of patent volume is 13% lower than the recall rates for the individual authorities; this signals that different names are used within different patent systems in a consistent manner.

[12] TRE-agrep (http://laurikari.net/tre/)

5.4 FINAL RESULTS AND IMPACT

Overall, harmonization has reduced the number of unique patentee names by 19.1%, from 9,310,595 to 7,536,191 names.

The average number of patents per patentee name increases from 5.5 before to 6.8 after harmonization. 13.4% of the harmonized names are related to more than one original name, ranging from 2 to 418 original names. Table 4 displays the harmonization impact on the number of patentee names, overall and broken down by three major patent authorities/offices (WIPO, EPO and USPTO)[13].

Table 4 : Harmonization impact on number of patentee names

	OVERALL	WIPO	EPO	USPTO
Original names	9,310,595	1,560,738	349,765	1,250,384
Patent count	51,225,255	10,303,722	2,243,681	15,334,250
Harmonized names	7,536,191	1,462,437	325,704	1,072,540
Name reduction (rate)	19%	6%	7%	14%
Average patent count by original name	5.5	6.6	6.4	12.3
Average patent count by harmonized name	6.8	7.0	6.9	14.3
(rate)	23%	6%	7.8%	16.2%
Harmonized names linked to multiple original names	1,011,531	79,909	19,007	100,607
(rate)	13.4%	5.5%	5.8%	9.4%
Maximum number of linked original names	418	36	16	106
Original names affected by harmonization	2,785,935	178,210	43,068	278,451
(rate)	30%	11%	12%	22%
Patent volume affected by harmonization	36,488,733	1,346,028	1,096,339	4,536,443
(rate)	71%	13%	49%	30%

Notice that the overall impact for the person table is considerably higher compared with the rates obtained for specific patent offices separately. This signals a higher rate of consistency – in terms of the use of similar names – within patent systems than between patent systems.

While only 13.4% of harmonized names are related to multiple original names (overall), they cover 30% of all original names, representing 71% of the total patent volume. For EPO and USPTO these latter figures amount to 49 and 30% respectively.

Notice finally that while the average impact on the level of the patentee in terms of patent count might seem modest for certain patent systems (e.g. WIPO, 6%; EPO, 7,8% versus USPTO 16,2%; overall 23%), one also observes considerable variation within each system. Table 5 provides an overview of the most extreme cases overall and for EPO, USPTO and WIPO separately. It becomes clear that for a number of organizations, name harmonizing is an essential requirement to create a more accurate view of the relevant patent portfolio.

[13] The patent count in this table is based on the number of patents linked to all patentee names involved. Patents having multiple patentees will be fully counted for every patentee, hence to overall total patent count as present in the table is higher than the total number of patents present in PATSTAT.

Table 5 : Highest harmonization impact (patentee names)[14]

	HRM_NAME	NBR NAMES	NBR PAT	MAX PAT	ADD PAT	ADD SHARE
Max additional patents	CANON	96	317663	202820	114843	36%
Max. additional patents EPO	UNILEVER	43	62314	21738	40576	65%
Max. additional patents USPTO	E.I. DU PONT DE NEMOURS & COMPANY	274	103134	40225	62909	61%
Max. additional patents WO	UNILEVER	43	62314	21738	40576	65%
Max. additional share	DEUTSCHE GOLD- UND SILBER-SCHEIDEANSTALT	28	36	3	33	92%
Max. additional share EPO	BIOSCAN	10	63	15	48	76%
Max. additional share USPTO	COMET	30	173	44	129	75%
Max. additional share WO	ADTECH COMPANY	16	60	21	39	65%
Max. matched names	F. HOFFMANN-LA ROCHE	393	36874	14833	22041	60%
Max. matched names EPO	ABB	54	6690	3787	2903	43%
Max. matched names USPTO	SAMSUNG ELECTRONICS COMPANY	195	265656	201932	63724	24%
Max. matched names WO	SIEMENS	195	204387	104848	99539	49%

[14] Figures in table refer to overall PatStat data (Number of names, Patents,…)

6 CONCLUSION

In this contribution, we have updated a comprehensive approach oriented towards name harmonizing. Emphasis has been placed on maximizing the accuracy of procedures that can be implemented automatically, i.e. without additional - and time consuming - validation efforts that require secondary information sources. Name variations are not combined if there is any doubt that the names relate to different legal entities.

This has resulted in a transparent method, the outcome of which is a reduced set of harmonized names.

EUROSTAT and its partners deliberately opted for a transparent method so that all interested parties will be able to build further on the results obtained. In the belief that the procedures described in this methodology can be further enriched and refined – and this also applies to legal entity normalization - we would encourage activities in this direction.

We are fully aware that improving the recall levels for the methodology as a whole is feasible by introducing expert assessments in a systematic manner. Engaging in such an effort for all patentees in the PatStat Database, goes beyond the current resources of EUROSTAT and its partners who developed this methodology (INCENTIM/ECOOM, K.U.Leuven, and SOGETI). However, for assessing the feasibility and the accuracy of such an approach, we engaged in an exercise implying the Top 500 of patent assignees. The results (Peeters et al., 2009) clearly indicate additional recall rates that justify further efforts. At the same time, numerous researchers and analysts are currently working on name harmonizing efforts with specific samples (e.g. technological fields, countries/regions, and sectors). For researchers engaged in such efforts, building on this methodology might be helpful. At the same time, the insights obtained by these researchers and analysts might be beneficial for further refinement of the current methodology. In other words, by sharing this methodology among the different communities involved in patentee analysis, further improvements could be envisaged. Consequently, EUROSTAT and its partners decided to make the complete methodology available into the public domain. Furthermore, given its continuous involvement in the PATSTAT Taskforce activities, EUROSTAT, in collaboration with the researchers at K.U.Leuven who developed this methodology, is committed to making freely available all future improvements in this methodology, including those obtained from other researchers and analysts[15].

7 REFERENCES

- DERWENT WORLD PATENTS INDEX Patentee Codes, Revised Edition 8, 2002, Thomson Scientific, United Kingdom, ISBN 0 901157 38 4 (www.thomsonscientific.com/media/scpdf/patenteecodes.pdf)
- Du Plessis, M., Van Looy B., Song X. Magerman, T. (2009). Data Production Methods for Harmonized Patent Statistics: Patentee Sector Allocation. EUROSTAT Working Paper and Studies, Luxemburg.
- Griliches, Z. (1990). "Patent statistics as economic indicators: A survey." Journal of Economic Literature, 28, 1661-1707
- Johnson, D. K. N. (2002). "The OECD Technology Concordance (OTC): patents by industry of manufacture and sector of use." STI Working Papers 2002/5
- Magerman, T., Van Looy, B., Song, X. (2006).Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. EUROSTAT Working Paper and Studies, Luxemburg.
- Peeters B., Song X., Callaert J., Grouwels J., Van Looy B. (2009) . "Harmonizing harmonized patentee names: an exploratory assessment of top patentees" EUROSTAT working paper

[15] A toolbox is available from the authors with the automated procedure described in this paper.

- Schmoch U., Laville F., Patel P., Frietsch R. (2003). "Linking Technology Areas to Industrial Sectors" Final Report to the European Commision, DG Research
- Verspagen B., van Moergastel T., Slabbers M. (1994). "MERIT concordance table: IPC – ISIC (rev. 2)" MERIT Research Memorandum 2/94-004
- Wu S. and Manber U. (1992). "AGREP -- A fast approximate pattern-matching tool" Proc. Winter 1992 USENIX Technical Conference, 153-162