HARMONIZING HARMONIZED PATENTEE NAMES:

AN EXPLORATORY ASSESSMENT OF TOP PATENTEES

Bert Peeters[2] - Xiaoyan Song[3] - Julie Callaert[3] - Grouwels Joris[2] - Bart Van Looy[1, 2, 3]

[1]Managerial Economics, Strategy and Innovation

Faculty of Economics and Applied Economics

K.U.Leuven

[2]Research Division INCENTIM, Faculty of Economics & Applied Economics, K.U.Leuven

[3]ExpertiseCentrum Onderzoek & Ontwikkeling Monitoring (ECOOM)

For further information please contact: bert.peeters@econ.kuleuven.be, Xiaoyan.song@econ.kuleuven.be or bart.vanlooy@econ.kuleuven.be

Acknowledgments: The authors want to thank Tom Magerman for providing feedback regarding methodological options and Caro Vereyen for providing support in data processing and verification of outcomes.

In cooperation with

Content

1. Introduction

When measuring the innovative performance of countries, regions or organizations, patent statistics are one of the most commonly used data sources. Griliches (1990) paraphrased this as "*In spite of all the difficulties, patent statistics remain a unique source 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".* At the same time, information on patents that is accessible through different on- and offline databases has limitations in terms of the completeness and accuracy of the available data.

When working with patent data on a patentee level, one of the most important issues relates to the observation that the name of one and the same organization or individual can appear in different forms across and within patent databases. This is due to the different channels through which an organization can apply for a patent (different patent authorities; different persons within the organization that fulfill the application process) and to variations in practices over time. Patent databases thus host a variety of names for one and the same organization or individual, which complicates arriving at accurate counts and analyses at the level of the patentee.

Several name harmonization approaches have been developed in the past to correct for different name variants occurring for one organization or individual. Each of these methods however has limitations regarding coverage and/or accuracy. In this paper, we explore a methodology to complement automated harmonizing efforts by inspecting outcomes of harmonized name efforts. The emphasis is put on a high coverage in terms of patent volumes, on high accuracy and on completeness (all person names of the PATSTAT person table[1] that are patentees). The approach developed by Magerman et al. (2009) serves as a starting point. Before discussing this 'manual' procedure, we will briefly discuss previous harmonization attempts as well as the approach followed by Magerman et al. (2006, 2009).

2. Previous harmonization effort

2.1. Derwent WPI and USPTO co-name patentee harmonization

Several name harmonization approaches have been developed in the past to correct for different name variants occurring for one and the same organization or individual, such as the USPTO co-name patentee harmonization, the Derwent World Patent Index company name harmonization (2002) and the methodology of Magerman et al. (2006, 2009). The Derwent WPI is probably the most comprehensive one, but the applied rules are currently not transparent. Both the first stage of the USPTO co-name patentee harmonization and the methodology developed by Magerman et al. (2006, 2009) are automated methods that first clean the names and then harmonize them. Although these methods lead to a considerable improvement, it remains unclear whether all different name variants of an organization/individual are captured. The USPTO co-name assignee harmonization methodology

[1] The 'person' table in PATSTAT includes all names of applicants and inventors. In spite of what its title may suggest, this 'person' table not only include individuals: also companies, universities, research institutes, hospitals, governmental agencies and non-profit organizations are involved.

addresses this problem by introducing a second ('manual') stage, leading to a high level of completeness. Its coverage is however limited to the first assignee names and only the assignee names of USPTO patents are taken into account.

2.2. Automated harmonization method of Magerman et al.

The methodology of Magerman et al. (2009) consists of a comprehensive method to arrive at the harmonization of patentee names in an automated way and has been applied to all names that occur in the person table[1] of the PATSTAT database as patentee of at least one patent. Names are harmonized in several consecutive steps (by removing legal forms, common words, spelling variations, etc). As a result, a harmonized name is attributed to each original name. Before applying the method, the original names from the PATSTAT person table are converted to upper case, reducing the number of unique patentee names by 3,8% from 9.674.722 to 9.310.595 unique names. Further processing and harmonizing of patentee names, reduces the number of unique names to 7.536.191, resulting in an overall reduction of unique patentee names of 22,1%.

While this is a considerable reduction, automated harmonization methods have their limitations in terms of coverage or recall, i.e. the number of names retrieved and harmonized. This has different reasons. First of all, spelling, grammatical or language variations often occur in the patentee names. In the methodology developed by Magerman et al. (2009), this is only corrected for the plain English words that occur in patentee names, but not for proper names. For proper names, the Levenshtein distance (for further information: see Navarro, 2001) between patentee names can give an indication of the closeness / similarity of the names under consideration, but it is dangerous to harmonize names based on this criterion solely in an automated manner (e.g. "International Busines Machines" = "International Business Machines", but "Imtech" <> "Amtech"). Second, organizations sometimes occur as applicants both under their full name and their abbreviated name (e.g. "International Business Machines" and "IBM"). In addition, organizations can change their name over time (e.g. "Minnesota Mining and Manufacturing" became "3M"; 'Tokyo Shibaura Denki' became 'Toshiba'; Alcatel derived its name from 'Alsace Cable & Téléphonie').

Correcting for these name variants can be addressed by introducing more complex algorithms or can be handled by inspecting harmonized names and making appropriate decisions on a case-by-case base. Within this paper, we explore the feasibility and impact of the latter methodology. Engaging in such an analysis at the same time allows to consider the feasibility of automated procedures in the future. Possible future developments – towards automation – might focus further on legal forms. While legal form removal is an important step in the methodology of Magerman et al. (2009), this might create some problems when individual applicant names with more than one initial enter the stage. Second, numerous country references have been removed in an automated way, but not all geographical suffices have been automated yet (e.g. names of cities, such as 'Armonk', which is found in several name variants of IBM). At the same time, it should be noted that country suffices should not always be automatically removed (e.g. it is appropriate to remove "France" in "ABB France" but not in "France Telecom"). Further investigating the frequency and nature of such cases implies, in a first phase, an

assessment of their frequency and impact. In conclusion, the automated harmonization method of Magerman et al. (2009) succeeds in a considerable improvement in terms of harmonizing patentee names. The question that we address in this paper is to what extent this approach can become complemented by harmonizing 'harmonized' names?

3. Harmonizing Harmonized Names: an empirical assessment.

For drawing patent statistics on the patentee level, it is of prevailing importance that all different name variants under which an organization or individual applies for a patent are captured. Building on the automated harmonization method developed by Magerman et al. (2009), we strive at achieving this by further harmonizing the obtained harmonized patentee names. The emphasis is put on a high coverage in terms of patent volumes, on high accuracy ('conservative' rules) and on completeness.

3.1. Harmonization rules

In this method, our harmonization is based on name similarity only. This implies that no consolidation efforts have been undertaken and that no information from financial databases was used to assess whether a different legal entity is involved (for an example of such an approach, we refer to the OECD HAN database[2], Grid forthcoming). The different rules applied are outlined below and will be illustrated by examples.

When a geographical suffix occurs in a patentee name, this can refer to the address - country, city, street or combination - of the company (e.g." IBM Armonk" = "IBM") or it can indicate another legal entity (e.g. "Bayer Antwerpen" <> "Bayer"). While in Magerman et al.'s (2006, 2009) automated procedures, country codes are removed for purposes of harmonizing, other geographical references remain included. For our purposes, these variants are visually inspected and if considered appropriate, additional harmonization is done. Note that a distinction between different entities can still be made by using the address field present in the person table or by using the legal field of the methodology of Magerman et al. (2009). Note also that when other meaningful (non-geographical) words are present in conjunction with a name, we refrained from harmonizing as this might signal co-patenting or a different legal entity (including joint ventures). So for instance "Bayer Cropscience" has not been harmonized to "Bayer".

When a company changed its name over time, or when it took over another company, we harmonized their names if we were able to identify these cases after a brief online search (e.g." Minnesota Mining and Manufacturing" = "3M"). Notice that this approach excludes mergers and acquisitions followed by a name change (e.g. "GlaxoSmithKline" = 'GSK' <> "GlaxoWellcome" <> "SmithKline Beecham"). When an organization applied for patents both under its full and abbreviated name, these names were harmonized (e.g. "BASF" = "Badische Anilin- und Soda-Fabrik"). Finally, for Japanese companies, often

[2] http://www.oecd.org/dataoecd/52/17/43846611.pdf

both the Japanese name as well as the English translation occur. This has been taken into account to the extent that the Japanese words do not signal a different entity or division (e.g. “Toyota Motor Company” = “Toyota Jishoda”; but ‘SANYO ELECTRICAL MACHINERY CORPORATION’ <> ‘SANYO ELECTRIC MEDICA SYSTEMS COMPANY‘).

3.2. Dataset

As mentioned in the introduction, our manual harmonization approach starts from the outcomes of the methodology developed by Magerman et al. (2009). All names present in the person table[1] of the PATSTAT database that are patentee of at least one patent are included. This includes patentee names of patents as well as utility models, from EPO, as well as over 7,5 million harmonized names (Magerman et al., 2009). The total related patent volume sums up to over 51 million patents. To make the volume of manual work feasible to conduct, a selection was made of patentees (organizations) for which harmonization effort was undertaken. The aim here was to maximize impact in terms of patent volume. Moreover, the emphasis was put on EPO/USPTO/WIPO publication authorities because most research is performed on patents applications filed in or patents granted by these publication authorities. This was achieved by selecting the top 500 players based on cumulative counts for EPO/USPTO/WIPO patent documents.

There were however several organizations that occurred multiple times in the top 500 (e.g. “IBM” and “International Business Machines”, “Celanese Corporation” and “Celanese Corporation of America”, “Corning” and “Corning Glass Works” (Corning Glass Works changed its name to Corning Incorporated in 1989). After removing these cases, 453 organizations remained in the list (see Appendix 1 for full list of 453 organizations and Appendix 2 for list of organizations that occur multiple times). This already signals the importance of further harmonization of previously harmonized patentee names. Adopting the sector allocation methodology developed by Du Plessis et al. (2009) reveals that the 453 organizations count 427 companies, 15 governmental non-profit organizations, 10 universities and 1 hospital.

3.3. Identifying matches

In this phase, an effort was undertaken to identify all the different name variants under which an organization applied for a patent, with an emphasis on completeness. To search for all possible name variants of an organization, approximate string searching (Navarro, 2001) was applied. The Levenshtein distance gives an indication of the distance between two strings by calculating the number of transformations needed to arrive from one string to the other (e.g. the Levenshtein distance between “Novartis” and “Novartes” is 1). We used condensed names to calculate such distance as they eliminate already some ‘noise’ (e.g. the distance between the harmonized names “AgfaGevaert” and “Agfa-Gevaert” is 1, while their condensed counterpart in uppercase equals “AGFAGEVAERT”, hence distance zero).

a) Defining search keys and selection of new harmonized name

For the 453 organization that were withheld for the additional harmonization efforts, search keys were developed by removing all common words from the condensed names (e.g. SK for "Celanese Corporation" is "Celanese"). Common words were removed because they result in a considerable extension of the appropriate search perimeter with often very low levels of relevance. The proper names of the company names are always written in full. For company names that (also) occur as an abbreviation, the abbreviation was included as an extra search key (e.g."IBM" was added for "International Business Machines"). Also, for company names that consist of multiple proper names, multiple search keys were defined (e.g. "Agfa" and "Gevaert" for "Agfa-Gevaert"). Changes in organization names were identified by an online search. Consequently, search keys were developed for both the old and current names (e.g. "3M" and "Minnesota Mining and Manufacturing").

b) Approximate string searching

Before applying approximate string searching, a crucial decision had to be made with respect to the Levenshtein distances to include for consideration. It is obvious that, for longer search keys, the allowed Levenshtein distance between the search keys and the matching part in the harmonized names can and should be higher. At the same time, working without an upper boundary on the Levenshtein distance would result in an explosion of the number of potential names requiring inspection. Working with a too small Levenshtein distance might on the other hand result in less coverage. The appropriate balance in this tradeoff was achieved by inspecting a limited number of cases exhaustively, looking for thresholds beyond which false hits constitute the vast majority (> 95%) of additionally identified names. The findings are presented in Table 1.

Table 1: Levenshtein distances included by length of the search keys.

Length of search key	***Levenshtein distances allowed***	
	absolute	relative
0-4 *	0	
5-6	0	
7-8	1	
over 8		20%

*Extra condition besides only exact matches (LD = 0): when search key is at beginning/end of the patentee name or surrounded by non-alphanumerical characters.

Some examples on the amount of potential names generated for different lengths of search keys and for different Levenshtein distances are presented in Table 2. The search key 'Bayer' results in 2.206 potential names for a Levenshtein distance equal to zero. These names include name variants of the company Bayer, but also individuals with Bayer as a surname. The number of hits explodes to 21.606

for Levenshtein distance 1. Here, many patentee names occur that are not related to Bayer (e.g. "TOSHIBA CERAMICS COMPANY" which includes the sequence 'BACER' and "KARL MAYER TEXTILMASCHINENFABRIK" which includes the sequence 'MAYER'). For "INTERNATIONAL BUSINESS MACHINES" on the contrary, higher Levenshtein distances do reveal patentee names which are relevant for harmonizing purposes (e.g. the Levenshtein distance for "INTERNATIOANL BUSINESS MACHINES CORPORATION" is 3).

Table 2: Examples of number of hits per Levenshtein distance per search key for 3 harmonized names.

Harmonized Name	*Search key*	*Abs. LD **	*# hits*
BAYER	BAYER	0	2.206
BAYER	BAYER	1	21.606
INTERNATIONAL BUSINESS MACHINES CORPORATION	IBM	0 **	99
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	0	2
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	1	125
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	2	92
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	3	31
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	4	6
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	5	5
INTERNATIONAL BUSINESS MACHINES CORPORATION	INTERNATIONALBUSINESSMACHINES	6	22
IMPERIAL CHEMICAL INDUSTRIES	IMPERIAL	0	985
IMPERIAL CHEMICAL INDUSTRIES	IMPERIAL	1	82
IMPERIAL CHEMICAL INDUSTRIES	IMPERIAL	2	2.056

*Levenshtein Distance

** Extra condition besides only exact matches (LD = 0): when search key is at beginning/end of the patentee name or surrounded by non-alphanumerical characters.

After determining the relevant Levenshtein distances, approximate string searching was applied with defined search keys on the full set of condensed names.

3.4. Validation and quality control

a) Validation

A validation table was constructed by combining the output of approximate string searching with the observed number of related patent documents. Table 3 shows an example of the distribution of the patent counts associated with the number of retrieved names for "Deutsche Thomson-Brandt".

Table 3: Number of harmonized names per patent count for Deutsche Thomson-Brandt.

Patent count	*Number of Retrieved Names*
1	50
2	8
3	4
5	4

7	1
9	2
17	1
21	1
694	1
708	1
6.816	1

This example illustrates the skewness of the distribution that we observed also in the other cases. A further analysis of this distribution for a sample of firms (n=50) showed that 90% of the patent volume is attached to a limited number of retrieved names (12%) with patent count > 10. Considering only correctly retrieved names (excluding false hits), one observes that retrieved names with a patent count > 10 represent 99,6% of the patent volume (19 % of all considered names).

Based on these observations, inspection efforts were limited to retrieved names associated with 10 or more patent documents, leading to a severe reduction (> factor 5) in the manual validation effort at the cost of only 0,4% recall in terms of patent volume.

Following the harmonization rules outlined in 3.1., all retrieved harmonized names were inspected and, if appropriate, were additionally harmonized. In case of doubt about the validity of harmonizing two names, a brief online search was performed.

b) Quality control

Several quality controls were performed after the manual validation, including verification of multiple or conflicting allocations. Most importantly, we engaged in an analysis of inter-rater reliability. For 22 harmonized names (i.e. 6% of the total number of names), two persons independently engaged in harmonizing harmonized names. The inter-rater correlation was calculated by a kappa-score. The results in Table 4 show a very satisfying kappa score of 95%, signaling consistent scoring.

Table 4: Kappa scores.

		Value	*Approx. Sig.*
Measure of Agreement	**Kappa**	,952	,000
N of Valid Cases		2.915	

For this sample of firms, recall and precision data have been calculated as well. Table 5 shows the obtained results: we observe a precision[3] rate of 99,5% (proportion of correct validations by initial rating) and a recall rate of 99,8% (number of correctly identified hits by the initial rating).

[3] Precision and recall rates are calculated including cases where both rates give a 0. Excluding these cases would lower the rates.

Table 5: Initial Rating * Validated Rating Cross tabulation (harmonized names).

		Validated Rating		*Total*
		0	*1*	
Initial Rating	**0**	2.638	16	2654
	1	7	254	261
Total		2645	270	2915

When we calculate the same statistics based on patent volumes, a precision rate of 99,9% and a recall rate of 99,7% is obtained (see Table 6).

Table 6: first_rater * second_rater Cross tabulation (patent count).

		Validated Rating		*Total*
		0	*1*	
Initial Rating	**0**	909.446	3.359	912.805
	1	1.411	297.550	298.961
Total		91.0857	300.909	1.211.766

3.5. Results

3.5.1. Harmonizing Harmonized Patentee Names

The impact of further harmonizing patentee names is considerable. This is clearly shown in Table 7, which reports on the top ten companies in terms of underlying unique person names and automated harmonized names. For "F. Hoffmann-La Roche", there are 1431 unique person names in the PATSTAT person table that were harmonized. The automated procedure of Magerman et al. (2009) resulted in 132 harmonized names, which can be all grouped under the heading 'F. Hoffmann-La Roche'.

Table 7: Top 10 organizations in terms of underlying unique person names after harmonization.

rank	*Harmonized Name (after second round of Harmonizing)*	*# Person Names (after first round of Harmonizing)*	*# Person names (after second round of Harmonizing)*
1	F. HOFFMANN-LA ROCHE	132	1.431
2	E.I. DU PONT DE NEMOURS & COMPANY	223	948
3	KONINKLIJKE PHILIPS ELECTRONICS	108	865
4	3M INNOVATIVE PROPERTIES COMPANY	475	806
5	BASF	157	743
6	INTERNATIONAL BUSINESS MACHINES CORPORATION	340	702
7	HOECHST	66	493
8	GENERAL ELECTRIC COMPANY	80	490

9	TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)	70	438
10	MATSUSHITA ELECTRIC INDUSTRIAL COMPANY	73	437

On average, for the 453 organizations that were involved in this additional harmonization effort, the number of person names per organization equals 106.

3.5.2. Impact – 453 organizations

Selecting the top players for this manual harmonization already showed that among the top 500 patenting organizations, several occur multiple times under a different name. This illustrates the importance of harmonization. For the 453 unique organizations, the methodology of Magerman (2009) (Level 1) succeeds in allocating 16.670 extra name variants to these companies. This raises the aggregated patent volume of these companies from 7.854.128 to 10.328.128 patents, implying an augmentation by 31,5% (Table 8). Additional harmonization efforts result in allocating an extra 30.960 names to these 453 organizations, which raises the aggregated patent volume from 10.328.128 to 13.251.949 patents: an augmentation by 28,31%. Overall, an overall increase of 68,73% in terms of patent volume is reached.

Table 8: Impact of harmonization for the 453 organizations in terms of names and patent volume.

	# names	*# patents*	*Additional improvement*	*Total improvement*
Level 0	453	7.854.128		
Level 1	17.123	10.328.128	31,50%	
Level 2	48.083	13.251.949	28,31%	68,73%

If we conduct the same analysis for the EPO, USPTO and WO patent documents separately, the overall increase amounts to respectively 13,72%, 21,98% and 18,06% (See tables 9,10 and 11 for detailed results on the number of name variants and patent volumes for EPO, USPTO and WO separately). Results for the publication authorities separately are lower, because name variants associated with high patent volumes occur between publication authorities, rather than within one publication authority.

Table 9: Impact of harmonization for the 453 organizations in terms of names and patent volume (EPO).

	# names	*# patents*	*Additional improvement*	*Total improvement*
Level 0	453	717.743		
Level 1	1.130	757.408	5,53%	
Level 2	2.033	816.192	7,76%	13,72%

Table 10: Impact of harmonization for the 453 organizations in terms of names and patent volume (USPTO).

	# names	*# patents*	*Additional improvement*	*Total improvement*

Level 0	453	1.825.243		
Level 1	4.326	2.026.081	11,00%	
Level 2	13.822	2.226.452	9,89%	21,98%

Table 11: Impact of harmonization for the 453 organizations in terms of names and patent volume (WO).

	# names	*# patents*	*Additional improvement*	*Total improvement*
Level 0	453	442.432		
Level 1	1.557	483.650	9,32%	
Level 2	3.894	522.342	8,00%	18,06%

Results for the top 10 patenting organizations are presented in Table 12. This allows assessing the evolution of their ranking before and after harmonization. "NEC Corporation" for example occupies the 7th place before harmonization and the 2nd place after harmonization. "Canon" on the contrary evolves from the 4th place before harmonization to the 5th place after harmonization.

Table 12: Top 10 patenting organizations with patent count and ranking before and after harmonization.

Harmonized Name (after second round of Harmonizing)	*After harmonization*		*Before harmonization*		*Improvement*
	# patents	*Rank*	*# patents*	*Rank*	
MATSUSHITA ELECTRIC INDUSTRIAL COMPANY	442.211	1	326.425	1	35,47%
NEC CORPORATION	347.687	2	184.195	7	88,76%
HITACHI	342.476	3	260.455	2	31,49%
TOSHIBA CORPORATION	336.649	4	236.744	3	42,20%
CANON	334.891	5	202.820	4	65,12%
MITSUBISHI ELECTRIC CORPORATION	305.575	6	187.569	6	62,91%
SAMSUNG ELECTRONICS COMPANY	274.666	7	201.932	5	36,02%
FUJITSU	270.722	8	158.045	8	71,29%
SONY CORPORATION	258.811	9	144.891	9	78,62%
SIEMENS	256.874	10	104.848	15	145,00%

The overall correlation, based on the total number of patents for the 453 organizations before and after harmonization, is 0,92 (rank order correlation: 0,97).

3.5.3. Impact – 453 organizations - Overall

The impact of harmonizing the patentee names of the 453 organizations, in terms of patent volume, is considerable. As mentioned in the previous paragraph, the 453 organizations have a total patent volume of 13.251.949 patents. This represents 26% of the total patent volume that is available within Patstat (October 2009). Respective shares for EPO, USPTO and WO are 36%, 35% and 11%.

Table 13: Patent volume of the 453 organizations overall, for the EPO, USPTO and WO.

	Overall	*EPO*	*USPTO*	*WO*
Total patent count of the 453 organizations	13.251.949	816.192	2.226.452	522.342
Total patent count	51.225.255	2.242.878	6.328.427	4.678.955
Coverage	25,87%	36,39%	35,18%	11,16%

As previously mentioned, 427 of the 453 organizations are companies. They hold over 98% of the patent volume of the 453 organizations (13.004.136 patents). Using the sector allocation of Du Plessis et al. (2009) points out that the total patent volume of all companies amounts to 34.941.230 patents. So the 427 companies represent 37,22% of the total patent volume of all companies. Overall results for the companies under study (n=427) as well as separate results for EPO, USPTO and WO are presented in table 14.

Table 14: Patent volume of the 427 companies overall, for the EPO, USPTO and WO.

	Overall	*EPO*	*USPTO*	*WO*
Patent volume of the 427 companies	13.004.136	794.721	2.148.669	496.210
Total patent volume	34.941.230	1.936.274	5.118.970	1.377.425
Patent volume of the 427 companies as % of total	37,22%	41,04%	41,97%	36,02%

The impact in terms of reduction of the number of unique person names is of course less significant, as the focus is now on coverage in terms of patent volume. The manual harmonization effort has additionally reduced the number of unique patentee names by 0,16% (from 7.536.191 to 7.523.564 unique names).

4. Conclusions

When creating patent statistics on the patentee level, it is of prevailing importance to identify all the different name variants under which an organization applies for a patent. Automated harmonization methods achieve a considerable improvement in terms of identifying name variants of patentees. But they have limitations and they focus mainly on accuracy. Therefore, we explored a complementary methodology to further harmonize harmonized names, starting from the results of the automated harmonization method developed by Magerman et al. (2009).

By additionally harmonizing patentee names of 453 top patenting organizations, approximately 99,6% of the total patent volume of these organizations has been allocated with a precision rate of 99,9% and a recall rate of 99,7%. In total, 30.920 additional original names have been harmonized to the 453 organizations, thereby augmenting their patent volume by over 28,3% (from over 10,3 million to almost 13,3 million). If we conduct the same analysis for the EPO, USPTO and WO separately, the additionally allocated names augment the total patent volume of the 453 organizations with respectively 7,8%, 9,9% and 8%.

The impact of harmonizing the patentee names of 453 top patenting organizations in patent volume is especially outspoken. The patents of these organizations account for almost 26% of the total patent volume. If we only take into account the 427 companies of the 453 top patenting organizations, and if we compare the associated patent volume with all patent documents where the assignee is a company, it becomes clear that these 427 companies account for over 37% of the total patent volume for which firms act as an applicant.

In the future, this manual harmonization effort can be further improved in various ways. Using address information can increase the precision and recall rates. In addition, including the patentee names with counts lower than 10 can further raise the allocation rate. Finally, applying this additional harmonization effort to a higher number of organizations (e.g. Top 1000 most active patentees) can improve its overall impact.

5. References

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.

Magerman T, Grouwels J., Song X. & Van Looy B. (2009) Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. EUROSTAT Working Paper and Studies, Luxemburg

Magerman, T., Van Looy, B., Song, X. (2006).Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. EUROSTAT Working Paper and Studies, Luxemburg.

Navarro, G. (2001). A guided tour to approximate string matching. ACM Computing Surveys, Vol. 33, No. 1, March 2001.

Thoma, G. et al. (2009) – Harmonizing and Combining Large Datasets – An Application to Patent and Finance Data. Forthcoming STI Working Paper, OECD.

Van Looy, B., Du Plessis, M., Magerman, T. (2006). Data Production Methods for Harmonized Patent Statistics: Patentee Sector Allocation. EUROSTAT Working Paper and Studies, Luxemburg.

APPENDICES

Appendix 1: List of 453 organizations

ABB
ABB RESEARCH
ABBOTT LABORATORIES
ADVANCED MICRO DEVICES
ADVANTEST CORPORATION
AGERE SYSTEMS
AGFA-GEVAERT
AGILENT TECHNOLOGIES
AIR PRODUCTS AND CHEMICALS
AISIN AW COMPANY
AISIN SEIKI COMPANY
AJINOMOTO COMPANY
AKZO NOBEL
ALCATEL
ALCATEL LUCENT
ALLERGAN
ALLIED CHEMICAL CORPORATION
ALLIED CORPORATION
ALLIEDSIGNAL
ALLIS-CHALMERS MANUFACTURING COMPANY
ALPS ELECTRIC COMPANY
ALTERA CORPORATION
ALUMINUM COMPANY OF AMERICA
ALZA CORPORATION
AMERICAN CAN COMPANY
AMERICAN CYANAMID COMPANY
AMERICAN HOME PRODUCTS CORPORATION
AMERICAN OPTICAL COMPANY
AMGEN
AMOCO CORPORATION
AMP
ANALOG DEVICES
APPLE COMPUTER
APPLERA CORPORATION
APPLIED MATERIALS
ASAHI GLASS COMPANY
ASAHI KASEI KOGYO

ASAHI KOGAKU KOGYO
ASML NETHERLANDS
ASTRAZENECA
ATLANTIC RICHFIELD COMPANY
AVERY DENNISON CORPORATION
BAKER HUGHES
BASF
BATTELLE MEMORIAL INSTITUTE
BAUSCH & LOMB
BAXTER INTERNATIONAL
BAYER
BAYER CROPSCIENCE
BAYER HEALTHCARE
BAYER MATERIALSCIENCE
BAYERISCHE MOTOREN WERKE
BECKMAN INSTRUMENTS
BECTON, DICKINSON & COMPANY
BEHR & COMPANY
BEIERSDORF
BELL TELEPHONE LABORATORIES
BENDIX AVIATION CORPORATION
BLACK & DECKER
BOARD OF REGENTS, THE UNIVERSITY OF TEXAS SYSTEM
BOEHRINGER INGELHEIM PHARMA & COMPANY
BOEHRINGER MANNHEIM
BORG-WARNER CORPORATION
BOSTON SCIENTIFIC
BOSTON SCIENTIFIC SCIMED
BRAUN
BRIDGESTONE CORPORATION
BRISTOL-MYERS SQUIBB COMPANY
BRITISH TELECOMMUNICATIONS
BROADCOM CORPORATION
BROTHER KOGYO
BRUNSWICK CORPORATION
BSH BOSCH UND SIEMENS HAUSGERAETE
BURROUGHS CORPORATION
CALIFORNIA INSTITUTE OF TECHNOLOGY
CANON
CARDIAC PACEMAKERS
CARRIER CORPORATION

CASIO COMPUTER COMPANY
CATERPILLAR
CATERPILLAR TRACTOR COMPANY
CELANESE CORPORATION
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)
CHEVRON RESEARCH COMPANY
CHISSO CORPORATION
CHRYSLER CORPORATION
CIBA
CIBA SPECIALTY CHEMICALS HOLDING
CIBA-GEIGY
CISCO TECHNOLOGY
CITIZEN WATCH COMPANY
COLGATE-PALMOLIVE COMPANY
COMBUSTION ENGINEERING
COMMISSARIAT A L'ENERGIE ATOMIQUE
COMPAQ COMPUTER CORPORATION
CONTINENTAL TEVES
CORNELL RESEARCH FOUNDATION
CORNING
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH
DAEWOO ELECTRONICS COMPANY
DAICEL CHEMICAL INDUSTRIES
DAIKIN INDUSTRIES
DAIMLER-BENZ
DAIMLERCHRYSLER
DAINIPPON PRINTING COMPANY
DANA CORPORATION
DEERE & COMPANY
DEGUSSA
DELPHI TECHNOLOGIES
DENSO CORPORATION
DEUTSCHE THOMSON-BRANDT
DIGITAL EQUIPMENT CORPORATION
DOW CORNING CORPORATION
DOW GLOBAL TECHNOLOGIES
DRESSER INDUSTRIES
DR.ING.H.C. F. PORSCHE
DSM
DSM IP ASSETS
E. R. SQUIBB & SONS

EASTMAN CHEMICAL COMPANY
EASTMAN KODAK COMPANY
EATON CORPORATION
EBARA CORPORATION
E.I. DU PONT DE NEMOURS & COMPANY
ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE
ELI LILLY & COMPANY
EMC CORPORATION
ESSO RESEARCH AND ENGINEERING COMPANY
ETHICON
ETHICON ENDO-SURGERY
ETHYL CORPORATION
EXXON CHEMICAL PATENTS
EXXON RESEARCH AND ENGINEERING COMPANY
EXXONMOBIL CHEMICAL PATENTS
EXXONMOBIL RESEARCH AND ENGINEERING COMPANY
F. HOFFMANN-LA ROCHE
FANUC
FARBENFABRIKEN BAYER
FMC CORPORATION
FORD GLOBAL TECHNOLOGIES
FORD MOTOR COMPANY
FRANCE TELECOM
FRAUNHOFER-GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG
FREESCALE SEMICONDUCTOR
FUJI ELECTRIC COMPANY
FUJI JUKOGYO
FUJI PHOTO FILM COMPANY
FUJI XEROX COMPANY
FUJISAWA PHARMACEUTICAL COMPANY
FUJITSU
FUNAI ELECTRIC COMPANY
G.D. SEARLE & COMPANY
GENENTECH
GENERAL ANILINE & FILM CORPORATION
GENERAL DYNAMICS CORPORATION
GENERAL ELECTRIC COMPANY
GENERAL INSTRUMENT CORPORATION
GENERAL MOTORS CORPORATION
GLAXO GROUP
GM GLOBAL TECHNOLOGY OPERATIONS

GTE PRODUCTS CORPORATION
GULF RESEARCH & DEVELOPMENT COMPANY
HALLIBURTON COMPANY
HALLIBURTON ENERGY SERVICES
HAMAMATSU PHOTONICS
HARRIS CORPORATION
HEIDELBERGER DRUCKMASCHINEN
HENKEL
HERCULES
HERCULES POWDER COMPANY
HEWLETT-PACKARD COMPANY
HEWLETT-PACKARD DEVELOPMENT COMPANY
HILTI
HINDUSTAN LEVER
HITACHI
HITACHI CHEMICAL COMPANY
HOECHST
HOECHST CELANESE CORPORATION
HON HAI PRECISION IND. COMPANY
HON HAI PRECISION INDUSTRY COMPANY
HONDA GIKEN KOGYO
HONEYWELL INTERNATIONAL
HUAWEI TECHNOLOGIES COMPANY
HUGHES AIRCRAFT COMPANY
HUMAN GENOME SCIENCES
HYUNDAI ELECTRONICS INDUSTRIES COMPANY
I. G. FARBENINDUSTRIE
IDEMITSU KOSAN COMPANY
IGT
ILLINOIS TOOL WORKS
IMPERIAL CHEMICAL INDUSTRIES
INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE
INFINEON TECHNOLOGIES
INGERSOLL-RAND COMPANY
INSTITUT FRANCAIS DU PETROLE
INTEL CORPORATION
INTERDIGITAL TECHNOLOGY CORPORATION
INTERNATIONAL BUSINESS MACHINES CORPORATION
INTERNATIONAL HARVESTER COMPANY
INTERNATIONAL STANDARD ELECTRIC CORPORATION
INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

ISIS PHARMACEUTICALS
JANSSEN PHARMACEUTICA
JSR CORPORATION
JTEKT CORPORATION
KANEKA CORPORATION
KAO CORPORATION
KAWASAKI STEEL CORPORATION
KIMBERLY-CLARK WORLDWIDE
KOBE SEIKOSHO
KOENIG & BAUER
KOMATSU
KONICA CORPORATION
KONINKLIJKE PHILIPS ELECTRONICS
KURARAY COMPANY
KYOCERA CORPORATION
KYOWA HAKKO KOGYO COMPANY
LAM RESEARCH CORPORATION
LEXMARK INTERNATIONAL
LG CHEM.
LG ELECTRONICS
LG.PHILIPS LCD COMPANY
LINDE
LITTON SYSTEMS
LOCKHEED MARTIN CORPORATION
L'OREAL
LSI LOGIC CORPORATION
LUCAS INDUSTRIES
LUCENT TECHNOLOGIES
MAN ROLAND DRUCKMASCHINEN
MANNESMANN
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MATSUSHITA ELECTRIC INDUSTRIAL COMPANY
MATSUSHITA ELECTRIC WORKS
MATTEL
MAZDA MOTOR CORPORATION
MEDTRONIC
MERCK & COMPANY
MERCK PATENT
METSO PAPER
MICHELIN RECHERCHE ET TECHNIQUE
MICRON TECHNOLOGY

MICROSOFT CORPORATION
MINNEAPOLIS-HONEYWELL REGULATOR COMPANY
MINOLTA CAMERA COMPANY
MINOLTA COMPANY
MITA INDUSTRIAL COMPANY
MITSUBISHI CHEMICALS CORPORATION
MITSUBISHI ELECTRIC CORPORATION
MITSUBISHI GAS CHEMICAL COMPANY
MITSUBISHI HEAVY INDUSTRIES
MITSUBISHI JUKOGYO
MITSUBISHI RAYON COMPANY
MITSUI CHEMICALS
MITSUI TOATSU CHEMICALS
MOBIL OIL CORPORATION
MOLEX
MONSANTO CHEMICAL COMPANY
MONSANTO COMPANY
MOTOROLA
MURATA MANUFACTURING COMPANY
NATIONAL RESEARCH DEVELOPMENT CORPORATION
NATIONAL SEMICONDUCTOR CORPORATION
NAVY USA
NEC CORPORATION
NGK INSULATORS
NGK SPARK PLUG COMPANY
NIKE
NIKON CORPORATION
NIPPON SHEET GLASS COMPANY
NIPPON SHOKUBAI COMPANY
NIPPON STEEL CORPORATION
NIPPON TELEGRAPH AND TELEPHONE CORPORATION
NISSAN MOTOR COMPANY
NITTO DENKO CORPORATION
NOKIA CORPORATION
NOKIA MOBILE PHONES
NOKIA SIEMENS NETWORKS & COMPANY
NORDSON CORPORATION
NORTEL NETWORKS
NORTHERN TELECOM
NORTHROP GRUMMAN CORPORATION
NORTON COMPANY

NOVARTIS
NOVARTIS PHARMA
NOVO NORDISK
NSK
NTN CORPORATION
NTT DOCOMO
NXP
OKI ELECTRIC INDUSTRY COMPANY
OLIN CORPORATION
OLIN MATHIESON CHEMICAL CORPORATION
OLYMPUS CORPORATION
OLYMPUS OPTICAL COMPANY
OMRON CORPORATION
ORACLE INTERNATIONAL CORPORATION
OTIS ELEVATOR COMPANY
OWENS-CORNING FIBERGLAS CORPORATION
OWENS-ILLINOIS
PANASONIC CORPORATION
PEUGEOT CITROEN AUTOMOBILES
PFIZER
PFIZER PRODUCTS
PHARMACIA & UPJOHN COMPANY
PHILIPS NORDEN
PHILLIPS PETROLEUM COMPANY
PIONEER CORPORATION
PIONEER ELECTRONIC CORPORATION
PIONEER HI-BRED INTERNATIONAL
PITNEY BOWES
POLAROID CORPORATION
PPG INDUSTRIES
PRAXAIR TECHNOLOGY
QUALCOMM
RADIO CORPORATION OF AMERICA
RAYCHEM CORPORATION
RAYTHEON COMPANY
RENAULT
RENESAS TECHNOLOGY CORPORATION
RESEARCH IN MOTION
RHODIA CHIMIE
RICOH COMPANY
ROBERT BOSCH

ROCHE DIAGNOSTICS
ROCKWELL INTERNATIONAL CORPORATION
ROHM AND HAAS COMPANY
ROHM COMPANY
ROLLS-ROYCE
SAMSUNG ELECTRO-MECHANICS COMPANY
SAMSUNG ELECTRONICS COMPANY
SAMSUNG SDI COMPANY
SANDOZ
SANYO ELECTRIC COMPANY
SAP
S.C. JOHNSON & SON
SCHAEFFLER
SCHERING
SCHLUMBERGER TECHNOLOGY
SCIMED LIFE SYSTEMS
SEAGATE TECHNOLOGY
SEIKO EPSON CORPORATION
SEIKO INSTRUMENTS
SEMICONDUCTOR ENERGY LABORATORY COMPANY
SERVICES PETROLIERS SCHLUMBERGER
SGS THOMSON MICROELECTRONICS
SHARP CORPORATION
SHELL OIL COMPANY
SHIN ETSU HANDOTAI COMPANY
SHIN-ETSU CHEMICAL COMPANY
SHOWA DENKO
SIEMENS
SILVERBROOK RESEARCH
SMITHKLINE BEECHAM CORPORATION
SOCIETE DES PRODUITS NESTLE
SONY COMPUTER ENTERTAINMENT
SONY CORPORATION
SONY ELECTRONICS
SONY ERICSSON MOBILE COMMUNICATIONS
SPERRY RAND CORPORATION
SQUARE D COMPANY
STANDARD OIL COMPANY
STANDARD OIL DEVELOPMENT COMPANY
STAUFFER CHEMICAL COMPANY
STMICROELECTRONICS

SUMITOMO CHEMICAL COMPANY
SUMITOMO ELECTRIC INDUSTRIES
SUMITOMO RUBBER INDUSTRIES
SUMITOMO WIRING SYSTEMS
SUN MICROSYSTEMS
SUN OIL COMPANY
SUNDSTRAND CORPORATION
SYLVANIA ELECTRIC PRODUCTS
SYMBOL TECHNOLOGIES
TAIWAN SEMICONDUCTOR MANUFACTURING COMPANY
TAKEDA CHEMICAL INDUSTRIES
TDK CORPORATION
TEIJIN
TEKTRONIX
TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)
TERUMO CORPORATION
TETRA LAVAL HOLDINGS & FINANCE
TEXACO
TEXAS INSTRUMENTS
THE B. F. GOODRICH COMPANY
THE BABCOCK & WILCOX COMPANY
THE BENDIX CORPORATION
THE BOC GROUP
THE BOEING COMPANY
THE DOW CHEMICAL COMPANY
THE FIRESTONE TIRE & RUBBER COMPANY
THE FURUKAWA ELECTRIC COMPANY
THE GENERAL HOSPITAL CORPORATION
THE GILLETTE COMPANY
THE GOODYEAR TIRE & RUBBER COMPANY
THE JOHNS HOPKINS UNIVERSITY
THE LUBRIZOL CORPORATION
THE PROCTER & GAMBLE COMPANY
THE REGENTS OF THE UNIVERSITY OF CALIFORNIA
THE REGENTS OF THE UNIVERSITY OF MICHIGAN
THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
THE UNION SWITCH & SIGNAL COMPANY
THE UNITED STATES OF AMERICA AS REPRESENTED BY THE ADMINISTRATOR OF THE
THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF THE AIR
THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF THE ARMY
THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF THE NAVY

THE UNITED STATES OF AMERICA AS REPRESENTED BY THE UNITED STATES
THE UPJOHN COMPANY
THE WHITAKER CORPORATION
THOMSON CSF
THOMSON LICENSING
TOKYO ELECTRON
TOKYO SHIBAURA DENKI
TORAY INDUSTRIES
TOSHIBA CORPORATION
TOSHIBA TEC CORPORATION
TOYOTA JIDOSHA
TRW
UBE INDUSTRIES
UNI-CHARM CORPORATION
UNILEVER
UNION CARBIDE CORPORATION
UNION OIL COMPANY OF CALIFORNIA
UNISYS CORPORATION
UNITED AIRCRAFT CORPORATION
UNITED MICROELECTRONICS CORPORATION
UNITED SHOE MACHINERY CORPORATION
UNITED STATES RUBBER COMPANY
UNITED STATES STEEL CORPORATION
UNITED TECHNOLOGIES CORPORATION
UNIVERSAL OIL PRODUCTS COMPANY
U.S. PHILIPS CORPORATION
VARIAN ASSOCIATES
VICTOR CO OF JAPAN
VOLKSWAGEN
W. R. GRACE & COMPANY
WACKER-CHEMIE
WARNER-LAMBERT COMPANY
WESTERN ELECTRIC COMPANY
WESTINGHOUSE AIR BRAKE COMPANY
WESTINGHOUSE ELECTRIC & MANUFACTURING COMPANY
WESTINGHOUSE ELECTRIC CORPORATION
WHIRLPOOL CORPORATION
WISCONSIN ALUMNI RESEARCH FOUNDATION
WYETH
XEROX CORPORATION
XILINX

YAMAHA CORPORATION
YAMAHA HATSUDOKI
YAZAKI CORPORATION
ZENECA
ZF FRIEDRICHSHAFEN
3M INNOVATIVE PROPERTIES COMPANY

Appendix 2: List of organizations that appear multiple times in top 500 patenting organizations

Harmonized Name (after first round of harmonizing)	Harmonized Name (after second round of harmonizing)
AMERICAN TELEPHONE AND TELEGRAPH COMPANY	BELL TELEPHONE LABORATORIES
ASAHI KASEI CHEMICALS CORPORATION	ASAHI KASEI KOGYO
ASTRAZENECA UK	ASTRAZENECA
AT&T BELL LABORATORIES	BELL TELEPHONE LABORATORIES
AT&T CORPORATION	BELL TELEPHONE LABORATORIES
BADISCHE ANILIN- & SODA-FABRIK	BASF
BELL TELEPHONE LAB.	BELL TELEPHONE LABORATORIES
CELANESE CORPORATION OF AMERICA	CELANESE CORPORATION
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)
CIBA HOLDING	CIBA
CIBA SPECIALTY CHEMICALS CORPORATION	CIBA SPECIALTY CHEMICALS HOLDING
CORNING GLASS WORKS	CORNING
ERICSSON	TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)
EXXON RESEARCH & ENGINEERING COMPANY	EXXON RESEARCH AND ENGINEERING COMPANY
HENKEL KGAA	HENKEL
HEWLETT-PACKARD COMPANY (A DELAWARE CORPORATION)	HEWLETT-PACKARD COMPANY
HONDA MOTOR COMPANY	HONDA GIKEN KOGYO
HONEYWELL	HONEYWELL INTERNATIONAL
IBM	INTERNATIONAL BUSINESS MACHINES CORPORATION
IBM UNITED KINGDOM	INTERNATIONAL BUSINESS MACHINES CORPORATION
ITT CORPORATION	INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION
KIMBERLY-CLARK WORLDWIDE	KIMBERLY-CLARK WORLDWIDE
MINNESOTA MINING AND MANUFACTURING COMPANY	3M INNOVATIVE PROPERTIES COMPANY
MITSUBISHI DENKI	MITSUBISHI ELECTRIC CORPORATION
MITSUBISHI KASEI CORPORATION	MITSUBISHI CHEMICALS CORPORATION
NIPPON ELECTRIC COMPANY	NEC CORPORATION
NIPPONDENSO COMPANY	DENSO CORPORATION
NORTH AMERICAN PHILIPS COMPANY	KONINKLIJKE PHILIPS ELECTRONICS
NORTH AMERICAN ROCKWELL CORPORATION	ROCKWELL INTERNATIONAL CORPORATION
PHILIPS	KONINKLIJKE PHILIPS ELECTRONICS
PHILIPS ELECTRONICS	KONINKLIJKE PHILIPS ELECTRONICS
PITTSBURGH PLATE GLASS COMPANY	PPG INDUSTRIES
PPG INDUSTRIES OHIO	PPG INDUSTRIES
RADIO CORPORATION	RADIO CORPORATION OF AMERICA
RCA CORPORATION	RADIO CORPORATION OF AMERICA
REGIE NATIONALE DES USINES RENAULT	RENAULT
ROHM & HAAS COMPANY	ROHM AND HAAS COMPANY
SONY DEUTSCHLAND	SONY CORPORATION

TELEFONAKTIEBOLAGET LM ERICSSON	TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)
THE PROCTER AND GAMBLE COMPANY	THE PROCTER & GAMBLE COMPANY
THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF	THE UNITED STATES OF AMERICA AS REPRESENTED BY THE ADMINISTRATOR OF THE
THE WESTINGHOUSE AIR BRAKE COMPANY	WESTINGHOUSE AIR BRAKE COMPANY
TOKYO SHIBAURA ELECTRIC COMPANY	TOKYO SHIBAURA DENKI
TOYOTA JIDOSHA KOGYO	TOYOTA JIDOSHA
UOP	UNIVERSAL OIL PRODUCTS COMPANY
UOP INC.DES PLAINES	UNIVERSAL OIL PRODUCTS COMPANY
W.R. GRACE & CO.-CONN.	W. R. GRACE & COMPANY