ISSN 1977-0375

eurostat
Methodologies &
Working papers

Patent Statistics at Eurostat: Methods for Regionalisation, Sector Allocation and Name Harmonisation

2011 edition

eurostat
EUROPEAN COMMISSION

eurostat
Methodologies and
Working papers

Patent Statistics at Eurostat: Methods for Regionalisation, Sector Allocation and Name Harmonisation

2011 edition

eurostat
EUROPEAN COMMISSION

Cataloguing data can be found at the end of this publication.

Luxembourg: Publications Office of the European Union, 2011

ISBN 978-92-79-20237-7
ISSN 1977-0375
doi:10.2785/13718
Cat. No KS-RA-11-008-EN-N

Theme: Science and technology
Collection: Methodologies & Working papers

Dear readers,

I am pleased to introduce the first Eurostat Compendium of methodologies for the production and dissemination of enhanced patent statistics, aiming to monitor trends in EU Innovation policies.

Innovation is one of the five cornerstones of the Europe 2020 strategy, which aims to turn the EU into a smart, sustainable and inclusive economy that delivers high levels of employment, productivity and social cohesion. Europe 2020 sets out a vision of Europe's social market economy for the 21st century.

Knowledge creation and innovation dynamics unfold within innovation systems that consist of a variety of actors, including firms, universities, entrepreneurs, and public and private research institutes.

The availability of well-defined and clear indicators — covering inputs and outputs of actors on different levels of analysis — is essential to assess a system’s innovative performance.

Patent statistics play a central role in these efforts. They are recognised as valuable data sources for monitoring, evaluating and even forecasting technological activities. The systematic and widespread availability of patent data has spurred the development and deployment of patent-related indicators among policy-makers, researchers, and practitioners.

This compendium of methodologies developed by Eurostat in the field of patent statistics contributes to the further development of indicators that are instrumental for analysis and policy development. It presents several methodological enhancements to deal with limitations in patent data sources. First, until now, no exhaustive sector allocation was available for identifying the nature of the applicant: individual, firm, university, public research organisation… Eurostat has bridged this gap by developing an exhaustive sector allocation methodology that is now made available for research and policy analysis. Second, regarding applicant names in existing patent data sources, non-uniformity is the rule rather than the exception. Therefore, name harmonisation algorithms have been developed with a considerable impact in terms of coverage, resulting in highly improved indicator accuracy. Third, regionalisation methodologies have been developed to better capture the regional dimension of technology development within the European Research Area (at EU-27 level).

These enhancements allow greater efficiency and accuracy in patent indicator extractions at regional, sectoral and institutional level, and are hence a considerable step forward in monitoring innovation systems in terms of technological activities.

Inna Steinbuka

Director

Eurostat, Directorate F

Social and information society statistics

This publication was managed by Eurostat, Unit F4 — Education, Science and Culture Statistics.

Project leader

Bernard Félix — Bernard.Felix@ec.europa.eu

Eurostat, Unit F4 — Education, Science and Culture Statistics

Eurostat
Statistical Office of the European Union
Bech Building
Rue Alphonse Weicker, 5
L-2721 Luxembourg

Production

This report was prepared by J. Callaert[1,2], M. du Plessis[1], J. Grouwels[2], C. Lecocq[2], T. Magerman[1,2], B. Peeters[2], X. Song[1], B. Van Looy[1,2] and C. Vereyen[2].

This report was prepared in collaboration with Sogeti Luxembourg S.A. (G. Dierickx, G. Châteaugiron and S. Sioen) under a Eurostat contract for the project “Patent Statistics”.

[1] ECOOM, Katholieke Universiteit Leuven, Waaistraat 6 — bus 3536, 3000 Leuven, Belgium.

[2] INCENTIM, Katholieke Universiteit Leuven, Minderbroedersstraat 8A — box 5105, 3000 Leuven.

1. GENERAL INTRODUCTION

Patent documents provide a comprehensive data source to assess and monitor technology performance. Griliches's observation of two decades ago still holds: '*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, organisational and technological detail.*' (Griliches, 1990).[1.] Hence, patent indicators are widely used by researchers, companies, and government agencies to assess technological progress in countries and regions, in technological and industrial domains and at micro-level (i.e. companies, universities and individual inventors).

The use of patent indicators has grown over the last decades as encompassing patent databases have become increasingly available. Such databases contain detailed information on individual patent documents: procedure dates; inventor and patentee names and addresses; technology classifications; patent and non-patent citations; patent family information; etc. This enables the development, comparison and monitoring of patent-related indicators at different levels of analysis.

Moreover, users of large patent databases are faced with several caveats that need to be dealt with in order obtain reliable and/or complete information. Several of these caveats relate to the heterogeneous codification of name and address entries of patentees and inventors. This heterogeneity seriously complicates the exhaustive identification of patentee locations, sectors and identities. There is therefore a need to enhance the information available in patent databases, in particular by 'harmonising' name information and/or by adding fields which convey address or sector information in a consistent manner.

Eurostat actively contributes to such methodological development efforts. The objective of enhancing patent data — in order to provide more comprehensive and accurate technology indicators — is pursued in close collaboration with EPO and the OECD Task force on Patent Statistics. Since 2007, Eurostat's production of EPO and USPTO data has been based almost exclusively on the *EPO Worldwide Statistical Patent Database*. This database, also known as PATSTAT, was developed by the EPO in 2005 using their collection and knowledge of patent data. This compendium outlines several enhancements developed in recent years for the EPO PATSTAT database, in particular regarding the regionalisation (according to the NUTS classification) of patentee and inventor addresses (EU-27), patentee sector allocation and patentee name harmonisation.

In order to deploy the enhancements efficiently, the developed methodologies primarily revolve around automated procedures and algorithms that allow a quick and accurate translation of raw data sources into enhanced information fields. Quality, in terms of both coverage and accuracy, is crucial in this respect. Coverage, or 'completeness', refers to the extent to which the developed procedures are able to target and translate all source data that are eligible for the developed application (e.g. the extent to which the name-harmonisation procedure captures all name variants of the same patentee). 'Accuracy' refers to the extent to which translations and manipulations of source data yield correct results (e.g. the extent to which all name variants allocated to one patentee reflect one and the same organisation). Methodologies aimed at maximising coverage (through automation) generally imply a loss of accuracy. Maximising the coverage of targeted source data requires automated procedures. Quality checks and validation are necessary to ensure accuracy in the results of these procedures, which entails a considerable portion of labour-intensive work. For each of the methodologies outlined in this document, several validation efforts and quality control activities have been performed iteratively. Hence, each methodology — regionalisation, sector allocation and name harmonisation — is the result of a meticulously designed combination of automated procedures and verification efforts in order to maximise both coverage and accuracy.

In addition, further improvements to the developed methodology are considered feasible and relevant. Researchers and analysts worldwide are working on related matters; hence sharing the developed

[1] Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of economic literature, 28, 1661 – 1707.

methodologies would be beneficial for all communities involved in patentee analysis. To encourage this process, the ECOOM-EUROSTAT-EPO PATSTAT Person Augmented Table (EEE-PPAT[2]) was made available in 2010 to present the work carried out on sector allocation and name harmonisation.

The three following sections outline the developed methodologies for regionalisation, sector allocation, and name harmonisation. These methodological outlines are complemented with illustrations on data yielded by the methodologies. Finally, sector allocation and name harmonisation methodologies are used to conduct an analytical study on the evolution of innovation actors and the influence of legislation. This chapter serves as an illustration of the potential applications of the methodologies outlined in this compendium.

[2] The EEE-PPAT table is free of charge, under Eurostat's commitment to making methodological developments publicly available.To obtain it for research and/or academic purposes, please send an e-mail describing the nature of your request to: TechnoInfo@ecoom.be

2. REGIONALISATION OF PATENT DATA

(C. Lecocq; B. Van Looy; C. Vereyen)

2.1. Introduction

Until recently, economic geography has played only a minor role in economic theory, despite the obvious fact that economic activities are not equally distributed over space. Relatively little empirical attention has been paid to the emergence and growth of regional clusters of technological activities. The existing evidence is mostly based on case study research, while large-scale empirical evidence or verification is rather scarce (Lecocq, 2010[3]). One reason for this lack of large-scale empirical evidence on the phenomenon of technology clusters is the low availability of quantitative data at the region-technology level, covering regions worldwide over longer periods. Patent data, which provide information on the date and geographic location of technological development and on the organisations and institutions involved, have become increasingly available at regional level. However, in order to be able to construct patent indicators from them, addresses of inventors and patentees need to be allocated to regions. This section outlines a methodology for achieving this.

Regional patent statistics build on the allocation of inventor and patentee addresses to regions. This allocation or “regionalisation” exercise requires first of all an exhaustive list of postcodes and city names and their respective regions. Within Europe, the NUTS classification (Nomenclature of Territorial Units for Statistics) is a hierarchical system used to divide the economic territory of the EU[4]. It is used in the collection, development and harmonisation of EU regional statistics; in socio-economic analyses of the regions; and in the framing of EU regional policies.

The production of a list of postcodes, city names and corresponding regions is the first step in the regionalisation methodology. A focal point of attention at this stage is the assessment, for each country under study, of the specific characteristics of the postal code system and of the administrative subdivisions, including relevant historical changes and revisions. After the compilation of reference files containing postcodes, city names and NUTS regions, the address information from the patent database (patentee as well as inventor addresses) can be matched to the reference list.

The methodology and results outlined in this section pertain to data stemming from EPO patents in the EPO PATSTAT database (April 2009). A methodology based on matching scripts has been developed and allows allocating the majority of the patentee and inventor addresses of EU-27 Member States to their respective NUTS 2 regions. To obtain the targeted coverage (99 % or more), the matching scripts have been complemented by manual search procedures. To ensure accuracy, different quality control procedures have been incorporated. This minimises misallocations that may stem from typing errors in addresses, historical changes in postcode systems, undetected city homonyms or incorrect parsing of city names. Such quality control procedures were used to monitor the accuracy of the methodology (where obtained levels exceed 99 % for most countries, cf. infra), and were instrumental in adapting and refining the overall methodology. To summarise, Table 2.1 provides a schematic overview of the efforts undertaken.

[3] Lecocq, C. (2010), “Technological Performance of Regions (and Firms) The Case of Biotechnology”, Doctoral Dissertation, K.U.Leuven.

[4] See also http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Table 2.1: Overview of methodological steps undertaken to regionalise patent data

Step 1: Country assessment • Postcodes (PC), City Names (CN), NUTS — Current and historical situation. • Assembling/Creating Reference Files (CN, PC, NUTS region) for each country.
↓
Step 2: Preparing patent address fields for matching Parsing of postcodes and city names (for some countries parsing of region — province/county — names) on a country-by-country basis. (PATSTAT address fields, selection limited to EPO).
↓
Step 3: Matching reference files and parsed PATSTAT (EPO) address fields in order to assign NUTS codes
↓
Step 4: Classifying unassigned addresses Manual search/lookup efforts based on the underlying volume of unassigned address fields (coverage objective of 99%+).
↓
Step 5: Assessing accuracy (QC) Controlling parsing efforts. Assessing 1:1 relationship between PC/CN combinations and NUTS levels assigned. Assessing 1:1 relationship between PC and NUTS levels assigned. • Assessing 1:1 relationship between CN combinations and NUTS levels assigned. • Independent assessment of address samples (at least 100/country). • Re-adjust applied rules/allocations (steps 2, 3, 4) and/or extend reference lists (step 1) where needed.

2.2. Methodology

2.2.1. Country cases and the creation of reference lists

The first step is to describe the current administrative subdivisions and the hierarchical NUTS classification of a country, as well as historical changes. Table 2.4 gives an overview of the current NUTS classification for each country. Similarly, the current postal code system is described and historical changes are identified.

Examples of historical changes
- Reorganisation of the postcode system in Belgium in 1990
- Introduction of a new postcode system in Germany in 1993
- Reorganisation of the administrative subdivisions in Denmark in 2007
- Several changes in the province subdivisions of Italy over the years

Data sources used in this respect are numerous and include websites describing countries’ administrative subdivisions and postcode systems as well as national post office directories and databases containing lists of a country’s current and past postal codes, cities and regions (provinces, counties, etc.), and lists made available by Eurostat containing Local Administrative Units (LAU) and NUTS 3 codes.

In a next step, reference files were created country by country by compiling different databases with postcode, city, region and NUTS information in each country. Only in a few countries could postcodes be directly related to the country’s administrative subdivision (e.g. in France the first two digits of the postcode refers to the *Département*, or NUTS 3 level). In most countries, however, the correspondence between postal groupings and administrative regions is not always “logical”, requiring further information (name of city, county, province, department, etc.) in order to allocate postcodes and city names to NUTS regions.

For each country, this leads to the creation of a list of possible combinations of postcodes and city names, and their respective NUTS 3 code[5]. It is important to note that none of these lists turns out to be exhaustive in terms of covering all variants observed within patent databases: small villages may be missing, as well as some combinations of postcodes and city for larger cities.

In addition, city homonyms may exist and may or may not appear in the list. If these city homonyms are located in different NUTS regions, it is not possible to allocate an address to its respective NUTS code based on the city name only. Similarly, one postcode area may be located in two or more NUTS regions as postal districts and post town do not necessarily follow administrative subdivisions. Such features of the postal code and administrative system of a country are included in the reference list and are taken into account in order to correctly allocate addresses to NUTS codes.

Once reference lists for all countries were created, additional efforts were made to harmonise and extend the lists as to maximise the number of matches with the parsed inventor and patentee addresses in a later stage. Such efforts include the replacement of special characters in city names by generic characters, the inclusion of name variants in different languages and the harmonisation of abbreviations.

[5] NUTS 2 code for the United Kingdom, Denmark, Malta and Poland.

Examples
- Achbrücke, Austria (original name)/Achbruecke, Achbrucke (name variants): special characters have been replaced by the generic characters
- Villabuena de Álava/Eskuernaga, Spain (original name)/Villabuena de alava, Eskuernaga (name variants): name variants in different languages are included in the reference list
- Saint-Didier-sur-Rochefort, France (original name)/St Didier sur Rochefort: harmonisation of notation in the reference list by using official abbreviations and removing the hyphens; the same was done for the parsed city names

2.2.2. Parsing of postcodes, city and region names

Before the matching can begin, patent address information needs to be parsed: postcode, city name, and for some countries the region name (province, department, etc.). While for the majority of address fields this is a rather straightforward exercise, mistakes can be made at this level, resulting in fewer or incorrect allocations later on. Therefore, parsing rules have been developed and validated after inspecting considerable samples of address data for each country separately. In addition, quality control procedures were introduced later on to detect potential mistakes in the parsing process (cf. infra).

2.2.3. Matching parsed addresses

Matching takes place in different steps. As cities may have homonyms in different region(s) of the country and as postcode areas may extend over more than one NUTS area, both the postcode and the city name of the patent address should correspond with the postcode and city name in the reference list in order to allocate a NUTS code to the patent address.

The first step thus comprises the allocation of addresses to NUTS regions when the postcode and city name in the address match with a postcode and city name in the reference list. However, not all addresses can be matched in this way. For a considerable number of addresses, the parsed postcode and city name in the address field do not correspond with a combination of postcode and city name in our reference list. In addition, for other address fields, only a postcode or city name is available. In the next two steps, matching address information with NUTS codes is performed by using either the postcode or the city name. The use of this rule is made conditional on the presence of a unique combination of postcode (or city name) and NUTS code in the reference list.

In some countries, address fields also contain information on the region (province/county) in which the city is located. The region in combination with postcode or city name, or the region in itself, is also used to allocate unassigned addresses to corresponding NUTS codes.

Examples
- City homonyms:
  81829 München (Bayern) DE212 München, Kreisfreie Stadt
  99438 München (Thüringen) DEG0G Weimarer Land
- Postcode homonyms:
  99438 München (Thüringen) DEG0G Weimarer Land
  99438 Possendorf (Thüringen) DEG05 Weimar, Kreisfreie Stadt

2.2.4. Classifying unassigned addresses

After the matching procedure, a substantial number of addresses still remain unassigned because the postcodes and city names available in these addresses cannot be uniquely assigned to a NUTS code according to our reference list. In addition, many addresses do not contain complete address information or some parts of the address information may be incorrect (e.g. typing error in postcode and/or city name).

Hence, the remaining unassigned address fields must be classified manually by conducting web searches on the available address information. The manual search procedures allow the identification of city homonyms and the correct assignment of addresses with missing or partially incorrect address information. Efforts are guided by the underlying patent volume, until a coverage of 99 % is reached and the majority of unassigned addresses turns out to contain no or very incomplete address information, e.g. city homonym without postcode or street name.

2.2.5. Quality controls

As indicated above, cities may have homonyms within a country and postcode areas may spread over more than one NUTS region. As the lists of postal codes and city names and their corresponding NUTS codes are not exhaustive, not all addresses can be matched based on the simultaneous presence of postcode and city name. Allocation based on postcode or city only may be subject to misallocation when a city has a homonym or when a postcode area spreads over more than one NUTS region, and when either case was not identified through the reference list. Typing errors in the postcode or city name may also lead to misallocation of the address. Likewise, changes in the postcode system, or wrong or partial parsing of the postcode or city name, may lead to allocation errors.

Different quality control procedures have been implemented to assess the accuracy of the matching procedures and to correct for misallocations when the address contains typing errors, unidentified city homonyms or when the city was not correctly parsed out. In a limited number of cases, this resulted in the adjustment of the NUTS allocation. Quality control procedures focused on the presence of multiple allocations for city names and postcodes; all such cases have been verified and, where needed, adjusted. Finally, for each country, an individual assessment of the NUTS allocation was done for a sample of 100 addresses that could not be matched with an allocation based on the simultaneous presence of both postcode and city. Throughout this validation exercise, no methodological mistakes could be identified.

As the regional allocation was carried out in sequence and on a country-by-country basis, the evaluation of the accuracy and efficacy of the parsing and matching rules used in the preceding countries provided feedback regarding their usefulness in the subsequent countries. Some rules were taken out because they only marginally increased the allocation performance (volume of addresses matched) and other rules were refined.

When all the countries were completed, the parsing and allocation rules were harmonised for the different countries, thus allowing country-specific rules where needed. The parsing, allocation and quality-control procedures were integrated in a common toolbox. In addition, the reference lists were extended with the manual search results.

2.3. Results

The following tables present the outcomes of the regionalisation (NUTS2 level) applied to all EU-27 countries (EPO patents, 1978–2008). The tables present the number of allocated addresses, the number of addresses that could not be assigned due to incomplete or missing address information, and the number of unallocated addresses. In the EU-27 Member States, the methodology for regionalising patent data was used successfully to allocate the bulk (99 % of all patents) of the addresses to their respective NUTS2 region. A similar methodology yielding the same level of quality for the allocation of regions at NUTS3 level will be made available in 2011, allowing the dissemination of the whole set of regional patent data in Eurostat's dissemination environment (Eurobase — Statistics Explained).

Table 2.2: EPO Unique Address Count (1978–2008), based on inventor addresses from EU-27 countries

	Allocated		Unassignable		Unallocated		Total
	Count	Per cent	Count	Per cent	Count	Per cent	
EU-27	1 083 240	99.15 %	693	0.06 %	8 639	0.79 %	1 092 572
Belgium	29 326	99.91 %	20	0.07 %	6	0.02 %	29 352
Bulgaria	1 024	99.32 %	6	0.58 %	1	0.10 %	1 031
Czech Republic	1 976	99.00 %	11	0.55 %	9	0.45 %	1 996
Denmark	20 246	99.87 %	13	0.06 %	14	0.07 %	20 273
Germany	483 764	98.86 %	141	0.03 %	5 446	1.11 %	489 351
Estonia	232	99.57 %	1	0.43 %	0	0.00 %	233
Ireland	5 594	99.43 %	5	0.09 %	27	0.48 %	5 626
Greece	1 514	97.99 %	18	1.17 %	13	0.84 %	1 545
Spain	20 193	99.18 %	15	0.07 %	152	0.75 %	20 360
France	202 778	99.86 %	207	0.10 %	75	0.04 %	203 060
Italy	89 037	98.64 %	29	0.03 %	1 195	1.32 %	90 261
Cyprus	121	100.00 %	0	0.00 %	0	0.00 %	121
Latvia	179	100.00 %	0	0.00 %	0	0.00 %	179
Lithuania	195	100.00 %	0	0.00 %	0	0.00 %	195
Luxembourg	1 899	99.89 %	2	0.11 %	0	0.00 %	1 901
Hungary	8 886	99.42 %	7	0.08 %	45	0.50 %	8 938
Malta	87	98.86 %	1	1.14 %	0	0.00 %	88
Netherlands	46 456	98.95 %	15	0.03 %	477	1.02 %	46 948
Austria	25 795	99.74 %	44	0.17 %	24	0.09 %	25 863
Poland	2 736	97.99 %	10	0.36 %	46	1.65 %	2 792
Portugal	1 220	97.13 %	4	0.32 %	32	2.55 %	1 256
Romania	459	98.71 %	2	0.43 %	4	0.86 %	465
Slovenia	1 248	98.58 %	3	0.24 %	15	1.18 %	1 266
Slovakia	475	98.34 %	7	1.45 %	1	0.21 %	483
Finland	26 447	99.68 %	22	0.08 %	62	0.23 %	26 531
Sweden	48 171	99.39 %	54	0.11 %	242	0.50 %	48 467
United Kingdom*	63 182	98.74 %	56	0.09 %	753	1.18 %	63 991

* United Kingdom figures from 2000.

Table 2.3: EPO Patent Count (1978–2008), based on inventor addresses from EU-27 countries

	Allocated		Unassignable		Unallocated		Total
	Count	Per cent	Count	Per cent	Count	Per cent	
EU-27	1 919 253	99.15 %	5 663	0.29 %	10 702	0.55 %	1 935 618
Belgium	51 792	99.77 %	114	0.22 %	6	0.01 %	51 912
Bulgaria	1 081	99.17 %	8	0.73 %	1	0.09 %	1 090
Czech Republic	2 493	99.01 %	16	0.64 %	9	0.36 %	2 518
Denmark	30 371	99.74 %	63	0.21 %	17	0.06 %	30 451
Germany	950 135	98.90 %	3 623	0.38 %	6 914	0.72 %	960 672
Estonia	260	99.62 %	1	0.38 %	0	0.00 %	261
Ireland	7 192	99.43 %	14	0.19 %	27	0.37 %	7 233
Greece	1 852	98.09 %	23	1.22 %	13	0.69 %	1 888
Spain	27 896	99.31 %	32	0.11 %	162	0.58 %	28 090
France	313 590	99.77 %	634	0.20 %	77	0.02 %	314 301
Italy	135 602	98.79 %	328	0.24 %	1 338	0.97 %	137 268
Cyprus	135	100.00 %	0	0.00 %	0	0.00 %	135
Latvia	319	100.00 %	0	0.00 %	0	0.00 %	319
Lithuania	281	100.00 %	0	0.00 %	0	0.00 %	281
Luxembourg	2 960	99.73 %	8	0.27 %	0	0.00 %	2 968
Hungary	11 685	99.35 %	29	0.25 %	47	0.40 %	11 761
Malta	99	95.19 %	5	4.81 %	0	0.00 %	104
Netherlands	115 339	99.43 %	120	0.10 %	547	0.47 %	116 006
Austria	43 983	99.39 %	243	0.55 %	25	0.06 %	44 251
Poland	3 308	98.22 %	13	0.39 %	47	1.40 %	3 368
Portugal	1 446	97.57 %	4	0.27 %	32	2.16 %	1 482
Romania	493	98.80 %	2	0.40 %	4	0.80 %	499
Slovenia	2 045	98.98 %	6	0.29 %	15	0.73 %	2 066
Slovakia	582	98.31 %	8	1.35 %	2	0.34 %	592
Finland	41 225	99.69 %	53	0.13 %	75	0.18 %	41 353
Sweden	75 164	99.43 %	167	0.22 %	266	0.35 %	75 597
United Kingdom*	97 925	98.76 %	149	0.15 %	1 078	1.09 %	99 152

* United Kingdom figures from 2000.

Table 2.4: Nomenclature of Territorial Units for Statistics

Countries		NUTS 1		NUTS 2		NUTS 3	
EU-27			97		271		1 303
Belgium	BE	Regions	3	Provinces (+ Brussels Capital Region)	11	Arrondissements	44
Bulgaria	BG	Regions	2	Planning regions	6	Oblasts	28
Czech Republic	CZ	-	1	Oblasts	8	Regions	14
Denmark	DK	-	1	Regions	5	Lands	11
Germany	DE	States (Länder or Bundesländer)	16	Government regions (Regierungsbezirke)	39	Districts (Kreise)	429
Estonia	EE	-	1	-	1	Groups of counties	5
Ireland	IE	-	1	Regions	2	Regional Authority Regions	8
Greece	EL	Groups of development regions	4	Peripheries	13	Prefectures	51
Spain	ES	Groups of autonomous communities	7	Autonomous communities and cities	17	Provinces + Islands	57
	ES			Ceuta and Melilla	2	Ceuta and Melilla	2
France	FR	ZEAT	8	Régions	22	Départements	96
	FR	Overseas departments (DOM)	1	Overseas departments (DOM)	4	Overseas departments (DOM)	4
Italy	IT	Groups of regions	5	Regions	21	Provinces	107
Cyprus	CY	-	1	-	1	-	1
Latvia	LV	-	1	-	1	Regions (+ Riga)	6
Lithuania	LT	-	1	-	1	Counties	10
Luxembourg	LU	-	1	-	1	-	1
Hungary	HU	Statistical large regions	3	Planning and statistical regions	7	Counties + Budapest	20
Malta	MT	-	1	-	1	Islands	2
Netherlands	NL	Lands	4	Provinces	12	COROP regions	40
Austria	AT	Groups of states	3	States	9	Groups of districts	35
Poland	PL	Regions	6	Voivodeships	16	Subregions	66
Portugal	PT	Continent	1	Regional Coordination Commissions	5	Groups of municipalities	28
	PT	Azores and Madeira	2	Autonomous regions	2	-	2
Romania	RO	Macroregions	4	Regions	8	Counties + Bucharest	42
Slovenia	SI	-	1	Macroregions	2	Statistical regions	12
Slovakia	SK	-	1	Oblasts	4	Regions	8
Finland	FI	Mainland Finland	1	Large areas	4	Regions	19
	FI	Åland	1	-	1	-	1
Sweden	SE	Regions	3	National areas	8	Counties	21
United Kingdom	UK	Government Office Regions (England)	9	(Groups of) Counties; Inner and Outer London	30	Unitary authorities or groups of districts	93
	UK	Wales	1	Groups of unitary authorities	2	Groups of unitary authorities	12
	UK	Scotland	1	Groups of unitary authorities	4	Groups of council areas	23
	UK	Northern Ireland	1	Groups of unitary authorities	1	Groups of districts	5

3. SECTOR ALLOCATION

(M. Du Plessis; B. Van Looy; X. Song; T. Magerman)

3.1. Introduction

From the mid-1980s onwards, a broader conception of the dynamics underlying innovative performance, synthesised by the concept of the 'innovation system', has emerged (e.g. Freeman, 1987; Lundvall, 1992; Nelson, 1993, Nelson and Rosenberg, 1993). This concept sees innovative performance at the level of regions, nations or industries as driven by industrial innovative activity and the pursuit of scientific excellence, both of which are influenced and shaped by institutional frameworks. Moreover, interaction among different institutional actors is advanced as a further explanation for differences in technological and innovative performance. These interactions are seen as critical in the process of knowledge generation and diffusion on a national, regional and industrial level.

A corollary of this conception of innovation dynamics is the need for refinements in patent indicators. Sector assignment — i.e. identifying whether patentees are companies (private business enterprise), universities and higher education institutions, or governmental agencies — becomes a necessary condition for further analysis of the dynamics underlying technological performance.

This section outlines an updated version of the sector allocation methodology that was developed in 2006 (Van Looy, du Plessis & Magerman, 2006). It starts with an overview of previous efforts in sector assignment of patentees, indicating the relevance of additional development efforts. After that, the currently developed methodology and its outcomes are outlined. Conclusions are drawn on the performance of the current sector allocation methodology, and future avenues for further improvement are delineated.

3.2. Existing sector typologies

The objective of the sector allocation methodology is to allocate each patentee to one of the following sectors: (a) individual (private) patentee (b) private business enterprise (c) government (agency) (d) university/higher education (e) private non-profit. This classification shows similarities with the existing sector classification developed by OECD in the context of conducting surveys on research and development, as outlined in the Frascati Manual (2002).

The Frascati Manual builds on the classification of the System of National Accounts (SNA). This system distinguishes between the following sectors: non-financial corporations, financial corporations, general government and non-profit institutions serving households, and households. In the OECD Frascati Manual (2002), largely based on the SNA 1993, higher education has been designated as a separate sector, and households are considered part of the private non-profit sector. Five sectors are identified in the Frascati Manual:

(1) Business enterprise
Includes: (a) all firms, organisations and institutions with the primary activity of the production of goods or services for sale to the general public, (b) the private non-profit institutions mainly serving them. The core of this sector is made up of private enterprises. Additionally, this sector includes public enterprises and non-profit institutions that are market producers of goods and services other than higher education. Examples of these non-profit institutions include: research institutes, clinics, hospitals, private medical practitioners, chambers of commerce, and agricultural, manufacturing or trade associations.

(2) Government
The government sector is composed of all departments, offices and other administrative bodies which do not normally sell to the community, as well as those that administer the state and the economic and social policy of the community. Non-profit organisations controlled and mainly financed by government but not administered by the higher education sector are also included in this sector. Furthermore, units associated with the higher education sector but mainly serving the government sector should also be included in the government sector.

(3) Private non-profit
This sector includes private non-profit institutions serving the general public and private individuals or households.

The following types of private non-profit institution should not be included in this sector:

- *Those mainly rendering services to enterprises,*
- *Those primarily serving government,*
- *Those entirely or mainly financed and controlled by government,*
- *Those offering higher education services or those controlled by higher education institutions.*

(4) Higher education
The higher education sector includes all universities, colleges of technology and other institutions providing post-secondary education, irrespective of their source of finance or legal status. Research institutes, laboratories and clinics operating under the direct control of, administered by, or associated with higher education institutions should also be included in this sector.

(5) Abroad
This sector consists of all institutions and individuals located outside the political borders of a country and all international organisations including facilities and operations within the country's borders.

It should be noted that individual (private) patentees do not show up as a separate category in the Frascati classification; in addition, the ‘Abroad’ category carries little relevance when classifying patentee names. Finally, whilst the definition of categories is generally clear and precise, the matching of name characteristics to the different categories is not clear-cut for certain types of organisation. For instance, hospitals could be classified as either‘business enterprise’, ‘private non-profit’ or ‘higher education’ depending on the governance mode under which they operate. As demonstrated later in this paper, the sector in which a given organisation should be classified is not always clear from looking solely at name field information found in the patent system. There is also the problem of a given institution being allocated to two sectors, e.g. when different objectives are being pursued by one and the same organisation.

Overview of approaches for sector allocation

Broadly speaking, one can make a distinction between two approaches for assigning sector codes. The first option involves building further on existing efforts and classification schemes that already make a distinction between different types of actors, and refining them so that they correspond to the targeted classification. The second option consists in developing ‘bottom-up’ methods to assign patentees to different categories. Given the amount of effort required to assign all patentees to categories from scratch, the first option is clearly preferable.

The most exhaustive effort to allocate patentees to different sectors has been undertaken within the framework of the USPTO system. As the USPTO patent system already allocates patentees to different categories, this classification provides the obvious starting point to further develop a sector classification. It should be noted that a similar codification does not exist in the EPO database. Nevertheless, if the USPTO classification proves to be relevant and accurate, the sector information available in the USPTO system could be related to the EPO database using harmonised names. The NBER patent citation data file (Hall et al, 2001, Jaffe and Trajtenberg, 2002) also classifies USPTO patentees into sectors. A closer inspection of the NBER sectors reveals that the same classification as the USPTO database system is used.

Hence, the first exercise conducted to develop an appropriate sector assignment method is related to assessing the accuracy and relevancy of the existing USPTO sector classification. We used a sample of patentees from USPTO to validate the sector classification of the USPTO. The USPTO patentee table provides information on all the patentees for each of the granted USPTO patents in the USPTO dataset. For each patentee, the USPTO has provided an organisational type code: namely, US company (2 or 12[6]), foreign company (3 or 13), US individual (4 or 14), foreign individual (5 or 15), US government (6 or 16), foreign government (7 or 17), county government (8 or 18), and state government (9 or 19). It should be noted that this classification does not coincide with the target categories: Universities and private non-profit sector categories are missing.

To validate whether the organisational types allocated to the patentees by the USPTO are correct, we assessed a sample of 500 patentees for each organisational type. As the total number of patentees with sector codes 8 and 9 did not exceed 500, all patentees in these two categories have been validated. Table 3.1 provides a summary of the findings.

Table 3.1: Validation of patentee types given in the USPTO patent database

Patentee types	Number of patentees incorrectly assigned to patentee type*	Number of patents incorrectly assigned*
2. US Company	65/500 (13 %)	7 419 (4.5 %)
3. Foreign Company	70/500 (14 %)	6 948 (4.6 %)
4. US Individual	0/500 (0 %)	0 (0 %)
5. Foreign Individual	21/500 (4 %)	72 (7.2 %)
6. US Government	39/500 (8 %)	60 (0.4 %)
7. Foreign Government	48/500 (10 %)	96 (6.4 %)
8. County Government	5/9 (56 %)	5 (56 %)
9. State Government	30/42 (71 %)	56 (68 %)

* The percentage for the sample analysed is given in parenthesis.

As Table 3.1 demonstrates, the existing sector allocation has certain shortcomings. Although the ‘individual (private) patentee’, ‘private business enterprise’, and ‘government’ sectors are present, the ‘university/higher education’ and ‘private non-profit’ sectors are not included. In addition, the existing allocation of patentees to organisational types includes a considerable level of error, except in the case of ‘US individuals’. Moreover, the following issues merit our attention:

- In the existing USPTO classification, organisations such as hospitals, higher education, and private non-profit organisations do not have a unique code to identify them. In the sample, universities and hospitals are usually given the types 2 or 3 to identify US and foreign universities/hospitals respectively. It should be noted that a separate list for US universities, developed independently from this categorisation, is available at USPTO. A similar list is not, however, available for foreign universities.

- Institutes (public/non-profit) are mostly assigned types 2 and 3 for US and foreign institutes respectively but are also found in categories 6 and 7; the criteria used to arrive at these classifications remain unclear. (Battelle Memorial Institute — type 2; Florida Institute of Phosphate Research — type 2; Institut National De La Recherche Agronomique — type 3; Fruit Tree Research Station, Ministry Of Agriculture, Forestry And Fisheries — type 3; Institut National De La Sante Et De La Recherche Medicale (INSERM) — type 6; Commissariat A L'energie Atomique — type 6; Stichting Rega Vzw — type 7; Hadasitmedical Research Serv. & Devel. LTD. — type 7).

[6] The number one in front of the code identifies part interest.

Having observed that several sectors are in need of refinement and that some categories need to be developed in their entirety, it was decided to adopt a different approach. In this approach, a set of rules will be developed that relates relevant information from the name field of patentees to specific sector categories. In applying this logic to the full patentee list as identified in the USPTO and EPO patent system, it is evident that different types of rules are needed; besides more generic rules that relate several patentees to one sector, a set of rules will be required targeted at specific organisations. In addition, conditionality will be introduced to minimise the number of multiple sector assignments. Without case-based allocation criteria and conditionality, accuracy as well as completeness will be negatively affected. 'Completeness' refers to the extent to which the sector allocation methodology is able to assign all patentees to a discrete category. 'Accuracy' refers to the extent to which the sector allocation correctly identifies the actual status of the patentee.

3.3. Methodology

Developing a methodology with a comprehensive set of rules is a highly iterative process in which it is eminently desirable to work on the full set of patentee names in order to adequately assess the impact of discrete rules. Accordingly, development and production efforts tend to coincide. It should be kept in mind that any methodological development reflects the particularities of the underlying database. This sector allocation work was based on the patentee list extracted from EPO PATSTAT (September 2009 version).

Whilst the overall logic strives for a maximum number of rules that follow logically from information found in the name fields of the patent database, concerns about completeness and accuracy point to the need for assessment and a certain level of expert involvement. In some cases, the category to which an organisation belongs is not clear from the patentee information alone because the name gives no real indication. In addition, some categories where the governance mode is crucial for sector allocation pose specific challenges, as in the case of hospitals, which can be private business sector, university/higher education, government or private non-profit. Equally, additional information would be required on whether certain research organisations funded by government are administered by the Ministry of Education, in which case they would fall within the university/higher education sector. Finally, there are cases where clues found in the name fields result in multiple sector allocations. Such cases will require a specific assessment resulting in case-based decision criteria. Depending on the desired levels of accuracy and completeness, additional data verification efforts could become considerable.

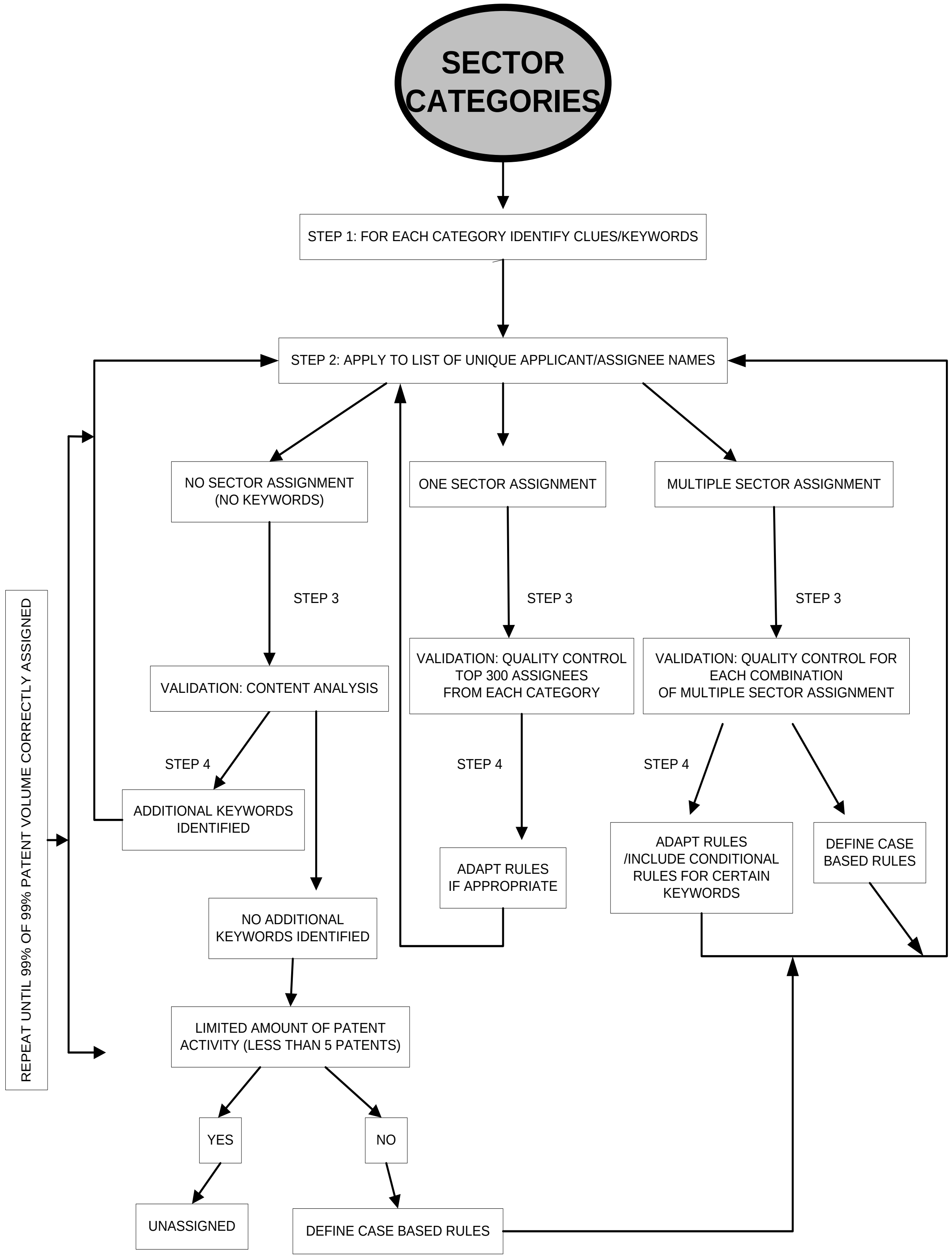
Within the framework of the development of this methodology, levels of completeness and accuracy of 99 % were targeted. This means that in applying all rule-based and case-based criteria to the patentee list, 99 % of all patents[7] must be assigned to just one sector, with a degree of error of less than 1 %.

The first principle underlying the methodology is straightforward: maximising the number of generic rules that can translate clues found in patentee names into the proper sector code. This rule-based logic works on the assumption that information found in the patentee names can provide clues to 'sector' membership. Such clues can be parts of names, specific words (e.g. government) and/or terms signalling specific legal forms (Inc.). If such clues can be identified in a systematic manner, they can be integrated into one script, which in itself allows for an automated allocation of sector codes. From an efficiency point of view, such an approach is clearly preferable, but it implies several assumptions. First of all, a sufficient number of patentee names should include such clues. Secondly, one-to-one relationships between clues and specific sector codes are preferable. Finally, a single name should only contain clues pertaining to one specific sector code. As the following sections demonstrate, several cases do not meet these ideal criteria. In order to remedy this situation, additional principles have been introduced. For patentees characterised by larger patent portfolios and for which generic rules do not result in an assignment, sectors are allocated on the basis of case-by-case decisions. Moreover, validation efforts —

[7] Levels of accuracy and completeness have been assessed on patent volume coinciding with allocated patentees. As the majority of patentees hold only one patent, striving for accuracy and completeness at patentee level would involve considerable additional resources, mainly for verification purposes.

applied throughout the process — reveal that generic rules generate occasional errors and assign certain patentees to the wrong sector (e.g. GMBH is often found in association with companies, but not always). Validation efforts have been undertaken for patentees with more than five patents, resulting in the development of an extensive set of additional case-oriented rules. A final principle has been introduced in order to address the occurrence of multiple sector assignment. Again, for patentees with more than five patents, conditional rules have been developed that result in a proper allocation of specific names (e.g. a patentee name has the words University, Foundation and a company legal form, e.g. Ltd. The sector codes 2, 4 and 6 are allocated to the name (Georgia State University Research Foundation, INC.). This is corrected by the conditional rule: if University and Foundation are both in the patentee name, then the sector code 4 should be given. *City Of Hope Research Institute* received sector codes 3 and 6. As the correct code is 4, a conditional rule was added to correct for this incorrect double sector code assignment.

Figure 3.1: Diagram of the methodology used to assign sector codes to patentees


SECTOR CATEGORIES
STEP 1: FOR EACH CATEGORY IDENTIFY CLUES/KEYWORDS
STEP 2: APPLY TO LIST OF UNIQUE APPLICANT/ASSIGNEE NAMES
NO SECTOR ASSIGNMENT (NO KEYWORDS)
ONE SECTOR ASSIGNMENT
MULTIPLE SECTOR ASSIGNMENT
STEP 3
STEP 3
STEP 3
VALIDATION: CONTENT ANALYSIS
VALIDATION: QUALITY CONTROL TOP 300 ASSIGNEES FROM EACH CATEGORY
VALIDATION: QUALITY CONTROL FOR EACH COMBINATION OF MULTIPLE SECTOR ASSIGNMENT
STEP 4
STEP 4
STEP 4
ADDITIONAL KEYWORDS IDENTIFIED
ADAPT RULES IF APPROPRIATE
ADAPT RULES /INCLUDE CONDITIONAL RULES FOR CERTAIN KEYWORDS
DEFINE CASE BASED RULES
NO ADDITIONAL KEYWORDS IDENTIFIED
LIMITED AMOUNT OF PATENT ACTIVITY (LESS THAN 5 PATENTS)
YES
NO
UNASSIGNED
DEFINE CASE BASED RULES
REPEAT UNTIL 99% OF 99% PATENT VOLUME CORRECTLY ASSIGNED

An exercise of this nature is time-consuming since it involves an assessment based on secondary sources (mostly web searches). However, it is feasible within a limited time frame, if an exhaustive allocation result (100 %) is not required. Figure 3.1 presents a diagrammatic summary of the approach followed in allocating sector codes to patentee names. The starting point is an initial list of keywords that are considered indicative of a certain category in the sector classification. Table 3.2 provides a sample of the keywords used for each sector.

Table 3.2: Examples of keywords/clues used to identify patentee sectors

Sector	Keywords
(1) Individual	"*DIPL.-ING.*"; "*PROF.*"; "*DR.*"; "*DECÉDÉ*"; "*DECEASED*"; "*DIPL. ING.*"; "*PH.D*"; "*DIPL.-GEOGR.*"; "*ING.*"; "*ÉPOUSE *"
(2) Private Enterprise	"* SA *"; "*S.R.L*"; "*HANDELSBOLAGET*"; "*INC."; "*LTD."; "*S.A.R.L"; "* BVBA *"; "*S.P.R.L.*"; "*NAAMLOZE VENNOOTSCHAP*"; "*AKTIEBOLAG*"
(3) Public and Private Non-Profit	"*GOUVERNMENT*"; "* MINISTRO*"; "*INSTIT*"; "*INSTYTUT*"; "*FONDATION*"; "*FOUNDATION*"; "*CHURCH*"; "*TRUST*"; "*KENKYUSHO*"; "*STIFTUNG*"
(4) University	"*UNIVERSI*"; "*UNIV.*"; "*COLLEGE*"; "*SCHOOL*"; "*REGENTS*"; "*ECOLE*"; "*FACULTE*"; "*SCHULE*"; "*UNIVERISTY*"; "*UNIVERSTIY*"
(5) Hospital	"*HOSPITAL*"; "*MEDICAL CENTER*"; "*MEDICAL CENTRE*"; "*ZIEKENHUIS*"; "*CLINIQUE*"; "*NOSOCOMIO*"; "*CLINICA*"; "*POLICLINICA*"; "*HOPITAL*"; "*HOPITAUX*"

These keywords/clues were applied to the full list of unique patentee names, extracted from the EPO PATSTAT September 2009 version. This involved a total dataset of 9 310 595 (11 100 882 based on person IDs that are not necessarily unique) unique patentee names, of which 349 765 are derived from EPO, 1 560 738 from WIPO and 1 250 384 from USPTO. The total number of patent documents related to these patentees amounts to 44 383 534 for EPO PATSTAT as a whole; 2 089 060 unique patents for EPO, 1 607 554 for WIPO and 6 032 306 for USPTO. In the update of the methodology an extra field was included in the table, which gave the patentee name a code 1 if it ever appeared as inventor. This field will be used in a final step to identify individuals which cannot be identified with one of the existing rules.

Patentees were previously assigned to the following six sectors: (1) individuals; (2) private enterprises; (3) government; (4) university; (5) hospital; (6) research institutes & non-profit. This — more disaggregate — classification is better suited to the rule-based methodology adopted here. In a subsequent step, these different categories can then be aggregated in accordance with the reporting objectives sought. However, as will become clear, even this approach will not produce a straightforward sector assignment as found in the Frascati Manual. Problems increase when considering whether hospitals and research organisations (although clearly identifiable as such, in most cases) should be included in the government, private business enterprise, or private non-profit sectors depending on the organisation's funding and governance. Sufficient levels of accuracy within each category will only be achieved by engaging in extended validation efforts based on secondary sources. In other words, the name alone does not reveal sufficient clues to arrive at a sector allocation, with sufficient levels of accuracy, for the majority of patentees in these categories. For this reason, analysing information on establishment, funding and governance is needed in order to assign patentees correctly, as the following section demonstrates. In this update of the methodology, we opted for a combination of the previous public and private non-profit institution sectors into a single ‘public and private non-profit’ sector. Hence, we recommend that patentees be allocated to 5 sectors: namely, (1) individuals; (2) private enterprises; (3) public and private non-profit organisations (4) universities and (5) hospitals.

On a technical level, it should be noted that the order of execution — i.e. which category is analysed first — is also a point of attention, since it significantly affects accuracy levels. In the methodology developed here, the rules are applied to patentees in the following sequence: private enterprises; universities; government & private non-profit sector; hospitals; and finally individuals.

After application, validation efforts are geared to all possible outcomes. If only one sector code is allocated, the accuracy of the findings obtained is assessed. Case-based adaptations are introduced as needed, i.e. when too many false hits are generated by a particular rule. If names do not obtain a code, a search for additional generic rules is initiated. Moreover, if patent volume for such patentees is considerable, case-based decision criteria will be considered. Finally, the occurrence of multiple codes is assessed and additional, conditional, rules to remedy the situation are introduced. In the revised methodology, we incorporated the field 'inventor' to identify more individuals that were placed in the category unknown. If the patentee name had the code 1 in this field and was given the code 6 (unknown), then this name was identified as an individual (e.g. "ITO TAKESHI"; "SATO KOJI" and "KOIKE NORIO").

Harmonised patentee names, created by the methodology of Magerman *et al.* (2009; cf. also infra) were used in a final validation step. These harmonised names were also used to identify the sectors for patentee names that were not yet identified (unknown). For these cases, the following rules were applied. If several patentee names received the same harmonised name, but were identified once as the sector government & non-profit and once as unknown, then the 'unknown' records became identified as belonging to the sector government & non-profit. Parallel rules were applied to the sectors of university and hospital. If several patentee names received the same harmonised name and they were once identified as company and once as unknown, then the unknown names were given the sector 'company' if their length exceeded 5 characters. The latter condition was included because for short names, there are many abbreviations and which may cause cases to be incorrectly assigned to companies. Table 3.3 provides further examples of these additional rules.

A next step implies the application of refined rules, which results in a more complete and accurate sector allocation, again requiring additional validation efforts. These steps are repeated until 99 % of all patent volume is assigned to a sector and 99 % of patents are correctly allocated to the sector.

Table 3.3: An example of patentee names with the same harmonised name but allocated once to a sector and once unknown

HARMONISED NAME	UNKNOWN	COMPANY	GOV NON-PROFIT	HOSPITAL	UNIVERSITY
INFOBLOX	INFOBLOX	INFOBLOX, INC.			
AQUA AIR	AQUA AIR	AQUA-AIR, INC.			
ALUGLACE	ALUGLACE	ALUGLACE S.A.			
EUROTUBE	EUROTUBE	EUROTUBE AB			
EUROFILM	EUROFILM	EUROFILM SRL			
STADT WIEN	* STADT WIEN		STADT WIEN		
SOUTHWEST RES INST	SOUTHWEST RES INST		SOUTHWEST RES INST.		
NG DER WISSENSCHAFTEN	NG DER WISSENSCHAFTEN		NG DER WISSENSCHAFTEN E.V.		
NEDERLANDSE ORGANISATIE VOOR TOEGEPAST-NATUURWETENSCHAPPELIJK ONDERZOEK TNO	NEDERLANDSE ORGANISATIEVOOR TOEGEPAST-NATUURWETENSCHAPPELIJK ONDERZOEK TNO		NEDERLANDSE ORGANISATIE VOOR TOEGEPAST-NATUURWETENSCHAPPELIJK ONDERZOEK-TNO		
NATAL SHARKS BOARD	NATAL SHARKS BOARD.		NATAL SHARKS BOARD		
L'ASSISTANCE PUBLIQUE - HOPITAUX DE PARIS	L'ASSISTANCE PUBLIQUE - HÔPITAUX DE PARIS			LASSISTANCE PUBLIQUE - HOPITAUX DE PARIS	
FONDAZIONE CENTRO SAN RAFFAELE DEL MONTE TABOR	FONDAZIONE CENTRO SAN RAFFAELE DEL MONTE TABOR.			FONDAZIONE CENTRO SAN RAFFAELE DEL MONTE TABOR	
CATHOLIC HEALTHCARE WEST, D.B.A. ST. MARY'S MEDICALCENTER	CATHOLIC HEALTHCARE WEST, D.B.A. ST. MARY'S MEDICALCENTER			CATHOLIC HEALTHCARE WEST, DBA ST. MARY'S MEDICAL CENTER	
UNIV BRITISH COLUMBIA	* UNIV BRITISH COLUMBIA				UNIV BRITISH COLUMBIA
STC.UNM	STCUNM				STC.UNM
UNIVERSITY OF WASHINGTON	UNIVER SITY OF WASHINGTON				UNIVERSITY OF WASHINGTON.
LUDWIG-MAXIMILIANS-UNIVERSITAET	LUDWIG-MAXIMILIANS-UNI VERSITAT				LUDWIG-MAXIMILIANS-UNIVERSITÄT

* A few examples where the unknown cases were not changed to the sector company due to the risk of incorrect sector allocation are: IT (IT:KK); CA (C A:KK); KA (YUGEN KAISHA K & A); HP (HP, SPOL. S R.O.); CIB (C-I-B-, INC.); AL (OBSHCHESTVO S OGRANICHENNOJ OTVETSTVENNOST'JU AL')-patentee name assigned to company in parenthesis.

3.4. Results

Table 3.4 further clarifies the principles and logic adopted. It contains a sample of 10 patentees for each category and the sector(s) to which they are assigned after application of an initial set of rules. Table 3.4 also includes 10 patentees that are not yet allocated to a sector. In addition, some patentees are allocated incorrectly to a sector. For instance, VITO, a public research organisation in Flanders (BE), has been given a company code since its name includes a legal form (N.V.) that is mostly associated with companies. Likewise, Andreas Stihl is a company and not an individual inventor. Finally, it should be noted that multiple codes occur frequently as one and the same name might include clues that suggest different sector allocations. For example, Virginia Foundation for Independent Colleges obtains a sector code for universities (Colleges) and one for private non-profit (Foundation).

The observations in Table 3.4 justify further efforts, focusing on: 1) creating additional rules based on a content analysis of non-assigned patentee names; 2) refining rules in order to avoid multiple codes; and 3) verifying whether assigned codes are accurate and, if not, introduce more refined rules. These refinements imply that certain rules are made conditional.

After several iterations, a total level of completeness of over 96.02 % was obtained (see Table 3.5) for EPO PATSTAT patentees as a whole. For the EPO, USPTO and WIPO data respectively 99.65 %, 98.74 % and 99.57 % of the patent volume has obtained a code. The proportion of patentee names with a double count is limited. For the total EPO PATSTAT list, the number of patentees with more than one sector amounted to 0.29 %; for EPO, USPTO and WIPO patentees with a double sector accounted for respectively 0.23 %, 0.28 % and 0.25 %.

Table 3.4: An example of 10 patentees in each of the sectors and 10 patentees with multiple sector codes

	Sector	Patentees
Single code	Individual	“DELAFON; JACOB”; “STIHL; ANDREAS”**; “KOIKE; YASUHIRO”; “PREGENZER, BRUNO”; “STOBBE, ANATOLI”; “TRAWÖGER, WERNER”; “FREI, SIEGFRIED”; “UEGAKI; TATEO”; “IKEDA, TAKESHI”; “NILL, WERNER”
	Private enterprise	“SIEMENS AKTIENGESELLSCHAFT”; “INTERNATIONAL BUSINESS MACHINES CORPORATION”; “BASF AKTIENGESELLSCHAFT”; “MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.”; “ALCATEL*”; “L'OREAL*”; “TEXAS INSTRUMENTS INCORPORATED”; “MOTOROLA, INC.”; “CIBA-GEIGY AG”; “VLAAMSE INSTELLING VOOR TECHNOLOGISCH ONDERZOEK, AFGEKORT V.I.T.O., ONDERNEMING VAN OPENBAAR NUT ONDER DE VORM VAN EEN N.V.”**
	University	“UNIVERSITE DE MONTREAL”; “DREXEL UNIVERSITY”; “CHINESE ACADEMY OF SCIENCES”**; “THE GOVERNORS OF THE UNIVERSITY OF ALBERTA”; “PRESIDENT & FELLOWS OF HARVARD COLLEGE”; “GEORGIA TECH RESEARCH CORP.”*; “TECHNISCHE UNIVERSITEIT DELFT”; “YALE UNIVERSITY”; “THE UAB RESEARCH FOUNDATION”*; “AUBURN UNIVERSITY”
	Hospital	“CEDARS-SINAI MEDICAL CENTER”; “BRIGHAM AND WOMEN'S HOSPITAL”; “BETH ISRAEL DEACONESS MEDICAL CENTER”; “ASSISTANCE PUBLIQUE HOPITAUX DE PARIS”; “MOUNT SINAI HOSPITAL”; “SHRINERS HOSPITALS FOR CRIPPLED CHILDREN”; “ORTHO-CLINICAL DIAGNOSTICS”**; “DANA FARBER CANCER INSTITUTE”*; “RHODE ISLAND HOSPITAL”; “THE HITCHCOCK CLINIC”
	Public and private non-profit	“UNITED KINGDOM ATOMIC ENERGY AUTHORITY”; “COMMISSARIAT A L'ENERGIE ATOMIQUE”; “UNITED STATES DEPARTMENT OF ENERGY”; “NATIONAL INSTITUTE OF AGROBIOLOGICAL SCIENCES”; “BATTELLE-INSTITUT E.V.”; “INSTITUT FRANÇAIS DU PÉTROLE”; “SAGAMI CHEMICAL RESEARCH CENTER”; “FOX CHASE CANCER CENTER”; “CSIR”*; “CORNELL RESEARCH FOUNDATION”**
	Unknown	“F.A.S.”; “DEVICES FOR VASCULAR INTERVENTION”; “STRATOS LIGHTWAVE”; “COLAS”; “FIOR DE VENEZUELA”; “INTERAG”; “KENZO”; “WEST POINT PEPPERELL”; “FRED”; “SIARC”
Multiple codes***	University and public and private non-profit	“CTRC RESEARCH FOUNDATION BOARD OF REGENTS”; “THE KANAGAWA ACADEMY OF SCIENCE AND TECHNOLOGY FOUNDATION”; “ACADEMY OF APPLIED SCIENCE (DIVISION OF ALLOR FOUNDATION)”; “WISCONSIN ALUMNI RESEARCH FOUNDATION”; “TRUSTEES OF TUFTS COLLEGE”; STICHTING HOGESCHOOL VAN UTRECHT”; “DEUTSCHES WOLLFORSCHUNGSINSTITUT AN DER RHEINISCH-WESTFÄLISCHEN TECHNISCHEN HOCHSCHULE AACHEN E.V.”; “KIRKWOOD COMMUNITY COLLEGE FACILITIES FOUNDATION”; “VIRGINIA FOUNDATION FOR INDEPENDENT COLLEGES”; “TRUSTEES OF BOSTON COLLEGE”
	Private enterprise and public and private non-profit	“INSTITUTE FOR INFORMATION INDUSTRY”; “KOREA INSTITUTE OF SCIENCE AND TECHNOLOGY”; “INSTITUT FÜR MIKROTECHNIK MAINZ GMBH”; “INSTITUT FÜR NEUE MATERIALIEN GEM. GMBH”; “GENETICS INSTITUTE, LLC”; “INSTITUT STRAUMANN AG”; “ISTITUTO DI RICERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S.P.A.”; “INSTITUT CERAC S.A.”; “NEC RESEARCH INSTITUTE INC.”; “DANA-FARBER CANCER INSTITUTE, INC.”
	Private enterprise and university	“ABC SCHOOL SUPPLY, INC.”; “IT'S ACADEMIC OF ILLINOIS, INC.”; “COLLEGE PARK INDUSTRIES, INC.”; “ACADEMIC PHARMACEUTICALS, INC.”; “F.H. SCHULE MÜHLENBAU GMBH”; “CAMBRIDGE UNIVERSITY TECHNICAL SERVICES LIMITED”; “UNIVERSITY PATENTS, INC.”; “IDAHO RESEARCH FOUNDATION, INC.”; “TOKYO INSTITUTE OF TECHNOLOGY”; “HELSINKI UNIVERSITY LICENSING LTD. OY”

* Case-based allocation.

** Incorrectly assigned; corrected by case-based rule.

*** Corrected by case-based conditional rules.

Table 3.5: Number of patents assigned after the final round of sector assignments

	Total PATSTAT			EPO			USPTO			WIPO		
	Number of unique patentees	Number of unique patents	% patent volume	Number of unique patentees	Number of unique patents	% patent volume	Number of unique patentees	Number of unique patents	% patent volume	Number of unique patentees	Number of unique patents	% patent volume
INDIVIDUAL	5 011 211	17 199 381	31.27 %	132 484	166 273	7.80 %	585 590	855 718	13.77 %	1 308 911	1 289 590	47.17 %
COMPANY	2 995 830	33 827 475	61.51 %	202 671	1 858 775	87.20 %	575 764	5 042 010	81.11 %	227 023	1 294 044	47.33 %
GOV NON-PROFIT	117 349	891 571	1.62 %	4 045	47 225	2.22 %	12 314	12 1354	1.95 %	6 635	48 708	1.78 %
UNIVERSITY	50 879	694 094	1.26 %	3 434	43 233	2.03 %	8 024	95 109	1.53 %	6 091	77 083	2.82 %
HOSPITAL	5 335	35 877	0.07 %	348	3 795	0.18 %	786	6 565	0.11 %	598	6 083	0.22 %
TOTAL SINGLE CODE	8 180 604	52 648 398	95.73 %	342 982	2 119 301	99.42 %	1 182 478	6 120 756	98.46 %	1 549 258	2 715 508	99.32 %
COMPANY GOV NON-PROFIT	39 506	148 821	0.27 %	1 450	4 564	0.21 %	5 015	16 606	0.27 %	2 002	6 089	0.22 %
COMPANY HOSPITAL	1 867	4 515	0.01 %	75	145	0.01 %	200	504	0.01 %	140	275	0.01 %
COMPANY UNIVERSITY	1 464	3 751	0.01 %	58	92	0.00 %	126	232	0.00 %	89	208	0.01 %
GOV NON-PROFIT UNIVERSITY	489	2779	0.01 %	19	38	0.00 %	69	188	0.00 %	44	158	0.01 %
COMPANY GOV NON-PROFIT UNIVERSITY	167	721	0.00 %	12	47	0.00 %	16	74	0.00 %	13	33	0.00 %
TOTAL MULTI-CODES	43 493	160 587	0.29 %	1 614	4 886	0.23 %	5 426	17 604	0.28 %	2 288	6 763	0.25 %
UNKNOWN	1 087 808	2 189 578	3.98 %	5 170	7 537	0.35 %	62 511	78 074	1.26 %	9 204	11 732	0.43 %
GRAND TOTAL	9 311 905	54 998 563	100 %	349 766	2 131 724	100 %	1 250 415	6 216 434	100 %	1 560 750	2 734 003	100 %

To assess accuracy, a random sample of patentees (500 for the smaller sectors and 1000 for the larger sectors) in each category was checked[8], whereby the accuracy levels for EPO, WO and USPTO were assessed separately. Table 3.6 provides an overview of the results of this accuracy assessment. In several categories, the level of accuracy is sufficient, with less than 1 % of the patent volume incorrectly assigned to the sector.

Table 3.6 : Results on the number of incorrectly assigned organisations from the final validation of sector assignment for a random sample of 500 patentees (1000 patentees for the individuals and private enterprises) in each sector

Sector	Total PATSTAT		EPO*		USPTO*		WIPO*	
	Number of patentees	% Patent volume	Number of patentees	% Patent volume	Number of patentees	% Patent volume	Number of patentees	% Patent volume
Individual (n=1000)	5	0.8 %	1	0.13 %	4	0.57 %	0	0 %
Company (n=1000)	3	0.01 %	3	0.05 %	2	0.26 %	6	0.85 %
Government (n=500)	17	1.27 %	1	0.72 %	1	0.17 %	4	1.21 %
University (n=500)	3	0.60 %	0	0 %	0	0 %	0	0 %
Hospital (n=500)**	5	0.23 %	1	0.17 %	0	0 %	1	0.09 %
Total sample	33	0.10 %	6	0.06 %	7	0.25 %	11	0.52 %

* For the patent systems a sample of 500 patentee names in the sectors company and individual were evaluated and 100 patentee names for the sectors government & non-profit, university and hospital.

** In the sector hospital, there may still be hospitals affiliated to universities, but these cannot be identified based on their name e.g. if they do not include terms like: “Teaching hospital”; “academic hospital”; or the like.

Similar conclusions can be drawn for unassigned patentees and for multiply allocated patentees. While the first category accounts for just below 4 % of the total patent volume, a complete removal of this category entails manual verification of 1 087 808 patentee names. The total number of patentees with multiple allocations is 43 493. Although this is a small proportion, the absolute numbers mean considerable resources will need to be deployed if a diagnosis based solely on manual verification efforts is to bear fruit. Also here, the analyst can opt for further refining cases which receive multiple codes, or consider these cases as ‘unknown’.

3.5. Conclusion

This section outlined a methodology designed to assign patentees to different sectors, based on a combination of a rule-based and a case-based logic. The rule-based logic works on the assumption that information incorporated in patentee names can provide clues on ‘sector’ membership, which can then be translated into a set of rules for the automated allocation of sector codes. In practice, such a rule-based approach proves to be insufficiently complete and accurate. The absence of clues, as well as the simultaneous presence of several clues that suggest different sectors, is a common feature. In order to remedy this situation, a second case-based layer is introduced. When applied in an iterative and sometimes conditional manner, quality levels — both in terms of completeness and accuracy — of 99 % are obtained.

[8] This assessment implied for each name a verification of the actual status, based mainly on analysing information found on the websites of the organisations involved.

3.6. References

System of National Accounts, 1993, Commission of the European Communities (CEC), International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank (1994).

Frascati Manual (2002). Proposed Standard Practice for Surveys on Research and Experimental Development. OECD Publications, Paris Cedex, France.

Freeman, C., (1987). Technology policy and economic performance. Pinter, London.

Hall, B. H., A. B. Jaffe, and M. Tratjenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Jaffe, A. B., and M. Trajtenberg, (2002). Patents, Citations, and Innovations: A Window on the Knowledge Economy. MIT Press, Cambridge.

Lundvall, B.-Å., (1992). National Systems of Innovation. Pinter, London.

Magerman, T., Grouwels, J., Song, X., & Van Looy, B. (2009), “Data production methods for harmonized patent statistics: Patentee name harmonization,” Eurostat Working Paper and Studies, Luxembourg.

Nelson, R.R., (1993). National Innovation Systems. Oxford University Press, New York.

Nelson, R.R., Rosenberg, N., (1993). Technical innovation and national systems, in: Nelson, R.R. (Ed.), National Innovation Systems. Oxford University Press, Oxford, pp. 3–21.

Van Looy, B., du Plessis, M. &Magerman, T. (2006) Data Production Methods for Harmonized Patent Indicators: Patentee sector allocation. Eurostat Working Paper and Studies, Luxembourg

.

4. NAME HARMONISATION

(T. Magerman; B. Peeters; X. Song; J.Grouwels; J. Callaert; B. Van Looy)

4.1. Introduction

The development of patent indicators on the micro-level of specific entities like companies, universities and individual inventors is faced with specific concerns stemming from the heterogeneity of patentee names that appear in patent documents within and across patent systems. Whereas this poses no challenge to the functioning of the patent system itself, it does complicate analyses at patentee level: the analyst is confronted with inconsistencies such as spelling mistakes, typographical errors and name variants, which often also reflect idiosyncrasies in the organisation of R&D and/or IPR activities within a single organisation.

These discrepancies in the naming of identical patentees in current patent databases justify efforts to achieve name harmonisation 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-harmonisation procedure is able to capture all name variants of the same patentee. ‘Accuracy’ relates to the extent to which the name-harmonisation procedure correctly allocates name variants to a single, harmonised patentee name. Unfortunately, completeness and accuracy do not go hand in hand. Efforts directed to maximising the number of identified name variants will ultimately lead to decreasing accuracy, while maximising accuracy inevitably leads to an increase in missed or unidentified name variants, or to a decrease in completeness.

With the objective of reconciling completeness and accuracy, a comprehensive methodology was developed to obtain harmonised patentee names in an automated way. The methodology consists of several harmonisation layers. In a first layer, which emphasised accuracy or ‘precision’, the number of unique patentee names was reduced by approximately 20 % and the average number of patents per patentee increased from 5.5 before to 6.8 after harmonisation. In a second layer, emphasis was placed on ‘recall’ (a high coverage in terms of patent volumes). This layer covers the top 500 most active patentees, as well as university patentees. For the top 500 patentees, this additional harmonisation layer resulted in allocating over 30 000 patentee names to the top organisations, raising their aggregated patent volume by almost 70 %.

Before presenting the developed layered methodology, focus should be placed on the difference between patentee name harmonisation on the one hand and legal entity harmonisation on the other hand. Legal entity harmonisation is concerned with the identification of all patents owned by one and the same legal entity. In this respect, legal entity harmonisation is not only concerned with name inconsistencies but takes into account:

- 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

For instance, when aiming at legal entity harmonisation, 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 harmonised to “Accenture” (name change).

In other words, the harmonisation of legal entities requires that, for every patentee name, historical information be checked on (changes in) naming practices and ownership structures. This type of information is not available in current patent databases. External information is needed — on ownership, changes of ownership, and organisational practices with regard to names — to arrive at a comprehensive methodology for legal entity harmonisation. Given the absence of databases providing exhaustive coverage of information needed to achieve legal entity harmonisation[9], such efforts are not included in the name-harmonisation methodology outlined here.

Before discussing the methods and their impact in detail, we will briefly discuss patentee name harmonisation efforts that have been undertaken in the past, notably by USPTO and by Derwent (Thomson Scientific).

USPTO co-name patentee name harmonisation

As part of the USPTO TAF database, first-named patentee names of organisational entities are harmonised for utility patents granted since 1969.

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

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

- Extract name of first-named patentee;
- Condense patentee name by removing spaces and non-alphanumeric characters;
- Convert to uppercase characters;
- Match condensed name with existing list of condensed and harmonised names;
- Manual review of all new patentee names not yet matched to an existing name in previous step (e.g. by looking at patentees of other patents granted to the same inventor or inventors);
- Annual large-scale manual review to verify integrity of the entire patentee 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 harmonised names, allows for a complete harmonisation solution.
  On the other hand, the USPTO harmonisation has several shortcomings:
- The partial manual approach implies significant resources every time new patentee names appear in the database;
- Only the first patentee is processed;
- Names reflecting different legal forms or associated with different countries are not combined[10];
- The manual review process is not transparent and might cause rule variation since harmonisation is performed by different persons, jeopardising the reproduction on a broader set of names (e.g. EPO patentee names, second patentee)[11]

[9] 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.

[10] For example, in the USPTO harmonisation, the following name variations of “BURR-BROWN” can be found in the list of harmonised names: “BURR-BROWN CORPORATION”, “BURR-BROWN INC.” and “BURR-BROWN LIMITED”.

[11] For instance, this can be observed in the list of original patentee names harmonised 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.

DERWENT WPI company name harmonisation

The DERWENT WORLD PATENT INDEX provides patentee codes for all patentees. One can summarise the DERWENT WPI method to produce these patentee codes as follows[12]

- Take the name and replace commonly occurring words with a standardised 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[13].
- Next, a selection of these patentees is analysed in depth to obtain unique standard patentee codes. The emphasis in this phase shifts towards legal entity harmonisation. 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 harmonisation: mergers and acquisitions, name changes and subsidiaries.

The index of standard patentee codes provided by WPI contains 21 000 entities and can be considered the most comprehensive harmonised index currently available, as it includes legal entity harmonisation. 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, or a new code is created). The precise rules that have been applied in each case are only evident after analysis of the names associated to a certain standard patentee code (information which is not publicly available)[14.]

For companies for which a standard code is not available (because of a limited number of patents), or for companies not recognisable as a subsidiary of a company that already has a standard code, the automatically generated non-standard code cannot be considered appropriate to achieve harmonisation 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[15.]

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

[13] 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".

[14] 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".

[15] 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.

4.2. Methodology layer 1

(T. Magerman; J. Grouwels; X. Song; B. Van Looy)

The first layer in the name harmonisation methodology emphasises accuracy or 'precision'. It is based on a previously developed comprehensive method to achieve harmonisation of patentee names in an automated way (Magerman et al., 2006). The 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 one or several of 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";
- 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 have been analysed in a systematic manner in order to develop an appropriate methodology. Whereas spelling variations, typographical errors and the additions of legal forms can be addressed in an automated manner; errors, acronyms and business unit or department extensions require additional validation efforts for assuring accuracy.

The developed methodology, which is an update of the 2006 version, builds on the complete person table provided by EPO PATSTAT, i.e. more than 11 000 000 names. It results from significant steps that have been undertaken to improve the 2006 version of the name harmonisation methodology. First, completeness was improved by adding extra rules (cf. infra) — mainly addressing legal forms and country indications — while remaining true to the philosophy of emphasising accuracy. Second, to cope with the considerable increase of treated data (from 450 000 to 11 000 000 + names), 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. Besides these major improvements, some smaller modifications have been introduced (e.g. the possibility of restricting rules to country codes).

4.2.1. Approach

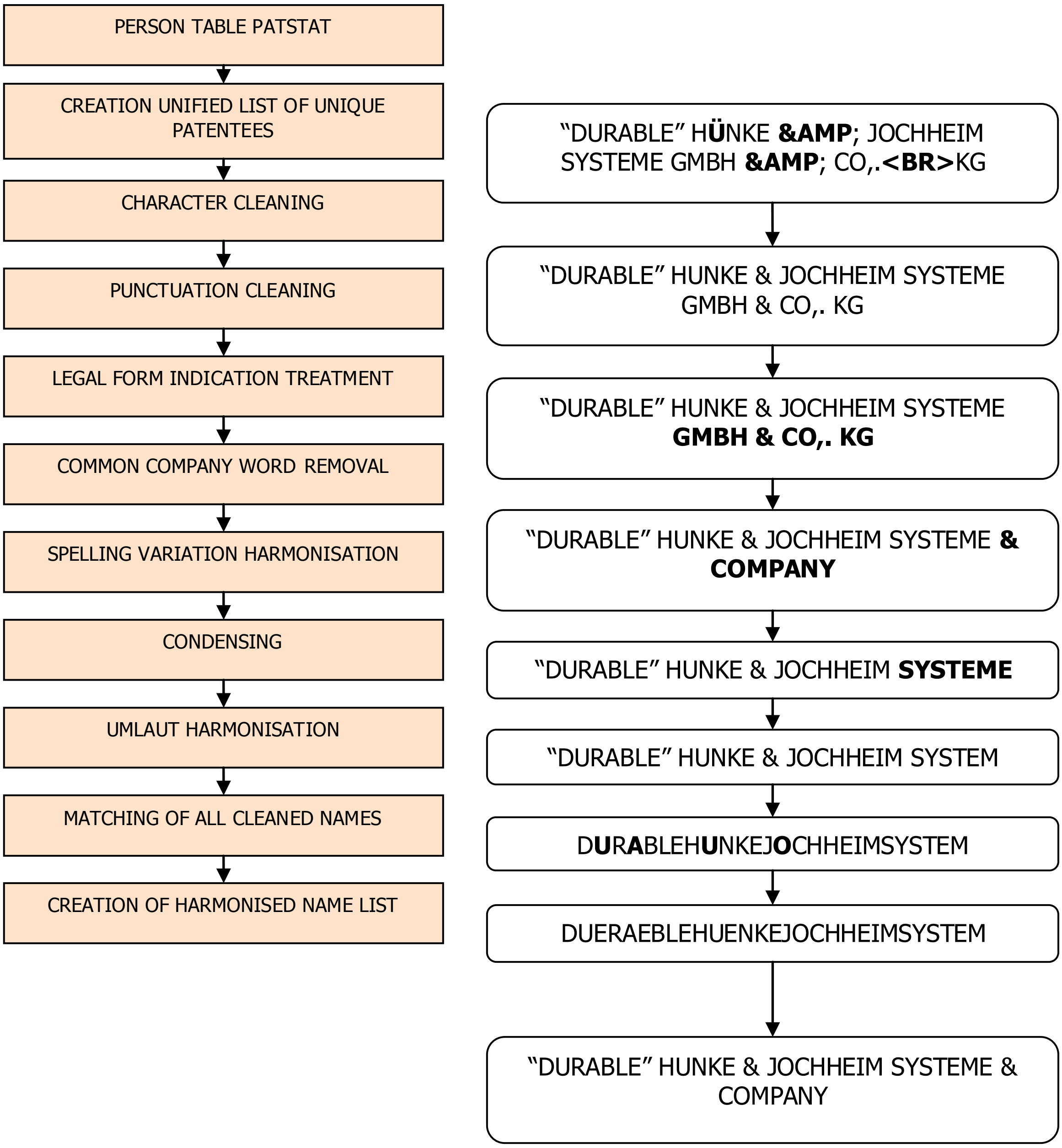
As indicated in the introduction, name harmonisation involves a trade-off between completeness and accuracy. It has been a deliberate choice in the methodology outlined here to favour accuracy over completeness for reasons of transparency, as it is easier to combine additional names than to separate combined names. An accurate but somewhat incomplete set of harmonised names provides users with ample opportunities to extend the methodology and its results to a broad range of applications. Given an accurate set of harmonised 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 endeavour than combining disaggregated names. Hence, this methodology, conceived as a transparent and accurate set of harmonised 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 harmonisation process must increase completeness without decreasing accuracy. Every action that jeopardises 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 labour-intensive, development process.

The final procedure can be completely automated in a modular approach to allow further refinements and improvements. The entire procedure is organised 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-harmonisation applications, some procedures are highly content-specific and additional analysis and refinements might be needed to apply the methodology to a different set of organisation names.

Figure 4.1 provides an overview of the developed methodology, consisting of a sequence of steps that include both data pre-processing and name-harmonising 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).

Figure 4.1: Overview schema name cleaning and harmonisation


PERSON TABLE PATSTAT
CREATION UNIFIED LIST OF UNIQUE PATENTEES
CHARACTER CLEANING
PUNCTUATION CLEANING
LEGAL FORM INDICATION TREATMENT
COMMON COMPANY WORD REMOVAL
SPELLING VARIATION HARMONISATION
CONDENSING
UMLAUT HARMONISATION
MATCHING OF ALL CLEANED NAMES
CREATION OF HARMONISED NAME LIST
"DURABLE" HÜNKE &AMP; JOCHHEIM SYSTEME GMBH &AMP; CO,.<BR>KG
"DURABLE" HUNKE & JOCHHEIM SYSTEME GMBH & CO,. KG
"DURABLE" HUNKE & JOCHHEIM SYSTEME GMBH & CO,. KG
"DURABLE" HUNKE & JOCHHEIM SYSTEME & COMPANY
"DURABLE" HUNKE & JOCHHEIM SYSTEME
"DURABLE" HUNKE & JOCHHEIM SYSTEM
DURABLEHUNKEJOCHHEIMSYSTEM
DUERAEBLEHUENKEJOCHHEIMSYSTEM
"DURABLE" HUNKE & JOCHHEIM SYSTEME & COMPANY

Data pre-processing

In the pre-processing steps, data are prepared for processing to facilitate actual name cleaning and harmonisation. 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.

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 “<BR>” 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.

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 harmonise 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 that 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 harmonised, 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.

Name cleaning

In the name-cleaning steps, the actual name cleaning and harmonisation is performed. As mentioned above, 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.

• 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 harmonised and moved to a separate field, thereby considerably reducing name variations.

- **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 organisation. Examples include “3COM” and “3COM CORPORATION”, “AMIC” and “AMIC COMPANY”, “BAUR SPEZIALTIEFBAU” and “BAUR SPEZIALTIEFBAU GESELLSCHAFT”, and “SOCIETE NOVATEC” and “NOVATEC”.

- **Spelling variation harmonisation**

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, harmonised effortlessly. For example, “SYSTEM”, “SYSTEMS”, “SYSTEMEN”, and “SYSTEMES” can all be harmonised to “SYSTEM” or “SYSTEMS”. Spelling variation harmonisation replaces all variants of common words with one harmonised variant that will be used to match name variants.

- **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 harmonised variant can be used to match names.

- **Umlaut harmonisation**

Although accented characters have already been replaced, 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 harmonisation identifies and matches different variants of words including “ä”, “ö” and “ü”.

Improvements 2009

• 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 favour 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. 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 for the last words and their respective names in a hierarchical way (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 jeopardised.

The country codes that were visually inspected are: US, JP, DE, UK, 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 minimise 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, AT, 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 variations of legal forms and/or common words.

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

Patent data from EPO PATSTAT are full of spelling variations. This is also the case for legal forms in patentee 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[16].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 harmonisation algorithm. A similar procedure was followed for common words. The total amount of new rules for legal forms and common words together was 2 227.

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

- **Country code correction**

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

- **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. In order to apply 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 made it possible to process large amounts of records and at the same time to program the application in a generic way, allowing new rule mechanisms that can be used in future releases (e.g. full support for regular expressions.).

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

4.2.2. Results

The complete name cleaning and harmonisation procedure has been applied to all 11 100 882 patentee records present in the PERSON table of EPO PATSTAT (October 2009 edition). These 11 100 882 patentee records contain 9 310 595 distinct names (ignoring uppercase/lowercase variations in names). The cleaning and harmonisation 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 EPO PATSTAT patentees.

- **EPO PATSTAT data pre-processing**

A particular observation for the EPO 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 harmonisation method. Before processing EPO 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 cannot 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, or a reduction of 1.1 %.

It should be noted that we only removed country code suffixes in the name field of patentee records in EPO 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.

- **Impact of intermediate steps**

The more than 4 000 rules of the name harmonisation method were executed on all patentee names. Tab 4.1 contains the impact of intermediate name harmonisation steps (numbers in the column 'DISCTINCT NAMES' represent the number of distinct names after the name harmonisation step mentioned in column 'STEP').

Table 4.1: Impact of intermediate name harmonisation 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 harmonisation	8 493 979	10 299	0.1 %
Condensing	7 548 399	945 580	10.2 %
Umlaut harmonisation	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 harmonisation), followed by legal form indication removal (about 25 % of total impact).

- **Validation**

Before discussing final results and impact, focus should be placed 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 harmonised correctly. In the recall validation, we checked for the presence of missed names, i.e. names that should have been harmonised, 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.

- **Precision validation**

In the precision validation, we verified to what extent the method correctly harmonises 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 harmonised names, and a big set of 1 000 harmonised names. The harmonised names were randomly selected from the full population of harmonised names being linked to at least two different original patentee names. For all harmonised names in the sample sets, all original patentee names were retrieved, and for each original name–harmonised name pair, a validation score was given by the human raters (Y, the original name is correctly linked to the harmonised name; N, the original name is incorrectly linked to the harmonised name, or there is doubt whether the original name is correctly linked to the harmonised name).

Table 4.2 contains the results of the precision validation. The number of errors and error rates in this table are presented at the level of harmonised names, i.e., they indicate the number and rate of harmonised names that have at least one original name that is incorrectly linked to that original name.

Table 4.2: 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 harmonised names have at least one original name that is incorrectly linked, with small variations between individual sample sets. Regarding the patent volume, the number is slightly lower: the harmonised names having at least one original name incorrectly linked to them represent about 0.5 % of the patent volume of all harmonised 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 harmonised 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 harmonised 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 harmonised 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 harmonised 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 harmonised name ‘SCHNELL’ with original names “SCHNELL & CO.”, “SCHNELL S.P.A.” and “SCHNELL S.R.L.”. All names are harmonised 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 harmonised name were to be used to identify unique entities, “SCHNELL & CO.” would be incorrectly taken together with the others. If the combination of harmonised name and removed legal form were to 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.

- **Recall validation**

The objective of the recall validation is to estimate how many names were missed by the name harmonisation method, i.e., how many original names have not been harmonised when they ideally should have been. The recall rate is calculated by counting the number of harmonised names over the number of names that should have been harmonised. The recall rate was calculated based on a sample set of harmonised names. In practice, this exercise involved three activities. First, a list of relevant keywords was constructed for every harmonised 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 were automatically retrieved using an approximate string search algorithm based on the Levenshtein distance (using the TRE-AGREP tool [17]). Finally, all retrieved original names were verified to confirm whether or not they should be linked to the harmonised name.

The 500 top patentees (after name harmonisation) 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 (see section below).

Table 4.3 contains the results of the recall validation. Results are expressed at the level of names (how many name variants are captured by the harmonisation 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 harmonisation 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 EPO PATSTAT are also broken down by three major patent authorities/offices (WIPO, EPO and USPTO).

Table 4.3: 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 EPO 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.

4.2.2.1. Final results and impact

Overall, harmonisation 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 harmonisation. 13.4 % of the harmonised names are related to more than one original name, ranging from 2 to 418 original names. Table 4.4 displays the harmonisation impact on the number of patentee names, overall and broken down by three major patent authorities/offices (WIPO, EPO and USPTO) [18].

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

[18] 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 the overall total patent count as present in the table is higher than the total number of patents present in PATSTAT.

Table 4.4: Harmonisation 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
Harmonised 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 harmonised name	6.8	7.0	6.9	14.3
(rate)	23 %	6 %	7.8 %	16.2 %
Harmonised 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 harmonisation	2 785 935	178 210	43 068	278 451
(rate)	30 %	11 %	12 %	22 %
Patent volume affected by harmonisation	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 harmonised 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 at patentee level 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 4.5 provides an overview of the most extreme cases overall and for EPO, USPTO and WIPO separately. It becomes clear that for a number of organisations, name harmonising is an essential requirement to create a more accurate view of the relevant patent portfolio.

Table 4.5: Highest harmonisation impact (patentee names)[19]

	Harmonised name	Number of names	Number of patents	Maximum number of patents	Additional patents	Additional share
Max additional patents	CANON	96	317 663	202 820	114 843	36 %
Max. additional patents EPO	UNILEVER	43	62 314	21 738	40 576	65 %
Max. additional patents USPTO	E.I. DU PONT DE NEMOURS & COMPANY	274	103 134	40 225	62 909	61 %
Max. additional patents WO	UNILEVER	43	62 314	21 738	40 576	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	36 874	14 833	22 041	60 %
Max. matched names EPO	ABB	54	6 690	3 787	2 903	43 %
Max. matched names USPTO	SAMSUNG ELECTRONICS COMPANY	195	265 656	201 932	63 724	24 %
Max. matched names WO	SIEMENS	195	204 387	104 848	99 539	49 %

[19] Figures in table refer to overall EPO PATSTAT data (number of names, patents, etc.).

4.3. Methodology layer 2

(B. Peeters; X. Song; J. Callaert; J. Grouwels; B. Van Looy)

The first layer of the name harmonisation methodology offers a transparent and comprehensive approach with emphasis on maximising the accuracy of procedures that can be implemented automatically. The method avoids the need for additional and time-consuming validation efforts that require secondary information sources. For example, name variations are not combined if there is any doubt that the names relate to different legal entities. The outcome of this method is a considerable reduction in the volume of unique patentee names. In spite of its merits, the first layer of the name harmonisation methodology — as any other automated method — has its limitations in terms of coverage, or the number of names retrieved and harmonised. This has different reasons. First, spelling corrections and grammatical or language variations are limited to plain English words, and do not consider proper names. Second, the layer 1 method does not cover organisations occurring under their full name and their abbreviated name (e.g. “International Business Machines” and “IBM”) and organisations 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 such name variants can partly be addressed by introducing more complex algorithms. But efforts like this typically require a close inspection of harmonised names and decisions made on a case-by-case base. Therefore, an additional harmonisation layer was introduced where the feasibility and impact of such closer inspection is studied. The emphasis in this layer is on high coverage in terms of patent volumes, on high accuracy (‘conservative’ rules) and on completeness.

4.3.1. Approach

This second harmonisation layer builds further on the first layer, and emphasises completeness. To keep the required work manageable, two subsets of the harmonised patentee list from layer 1 were considered.

On the one hand, a manual harmonisation layer was applied to all university patentees (identified through the sector allocation) with a patent volume above a threshold of 5. This additional effort was undertaken because of the growing demand for indicators of technology development at universities and knowledge-generating institutes. It accommodates current needs for mapping and monitoring innovation in today’s knowledge-based systems, where knowledge-generating institutes — and more specifically entrepreneurial universities — are cast in a central role. This further cleaning of university names was done manually, by searching for name variants of each university through keywords. To the maximum extent possible — and to our best knowledge — keywords were also included that depict institutes which at first sight seem unrelated, but which are actually part of a university (e.g. “Isis Innovation” is the technology transfer office of Oxford University).

On the other hand, the additional harmonisation effort in layer 2 was aimed at the most active patenting companies. For this purpose, the top 500 patentees in terms of cumulative patent counts for EPO, USPTO and WIPO patent documents were considered. Several organisations occurred multiple times in this top 500 (e.g. “IBM” and “International Business Machines”, “Celanese Corporation” and “Celanese Corporation of America”), which already signals the importance of this second layer. As such, 453 top organisations remained in the target list for this second layer. Adopting the sector allocation methodology developed by Du Plessis et al. (2009, see section 3 of this compendium) reveals that these 453 organisations consist of 427 companies, 15 governmental non-profit organisations, 10 universities and 1 hospital. The following paragraphs present the approach used for the top 500 institutes.

This approach is based purely on name similarity. To search for all possible name variants of an organisation, 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). Condensed names were used to calculate these distances, as condensing already eliminates some 'noise' (cf. layer 1 methodology). For example: the distance between the harmonised names "AgfaGevaert" and "Agfa-Gevaert" is 1; while their condensed counterparts in uppercase both equal "AGFAGEVAERT", hence distance zero.

a) Defining search keys and selection of new harmonised name

For the 453 organisations retained for the additional harmonisation efforts, search keys were developed by removing all common words from the condensed names (e.g. search key 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 organisation names were identified by an online search. Consequently, search keys were developed for both the former and the current company names. When a geographical suffix occurs in a patentee name, this can refer to the company's address — country, city, street or combination (e.g. "IBM Armonk"="IBM") — or it can indicate another legal entity (e.g. "Bayer Antwerpen" <> "Bayer"). While in the automated procedures from layer 1, country codes were removed for harmonisation purposes, other geographical references remained. For layer 2, these variants were visually inspected and, if considered appropriate, additional harmonisation was performed. A distinction between different entities can still be made by using the address field present in EPO PATSTAT 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 were present in conjunction with a name, no harmonisation was done as this might signal co-patenting or a different legal entity (including joint ventures). So for instance "Bayer Cropscience" has not been harmonised to "Bayer".

The names of companies that have changed their name over time, or that have taken over another company, have been harmonised in cases that could be identified after a brief online search (e.g. "Minnesota Mining and Manufacturing"="3M"). Note that this approach excludes mergers and acquisitions followed by a name change (e.g. "GlaxoSmithKline"='GSK' <> "GlaxoWellcome" <> "SmithKline Beecham"). When an organisation applied for patents both under its full and abbreviated name, these names were harmonised (e.g. "BASF"="Badische Anilin- und Soda-Fabrik"). Finally, for Japanese companies, often both the Japanese name and the English translation occur. This was 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').

b) Approximate string searching

Before applying approximate string searching, a crucial decision is to be made with respect to the Levenshtein distances to include for consideration. For longer search keys, the allowed Levenshtein distance between the search keys and the matching part in the harmonised names can and should be higher. To illustrate this: "International Busines Machines" is the same as "International Business Machines", but "Imtech" is different from "Amtech", although both pairs have a Levenshtein distance of 1. But working without an upper boundary on the Levenshtein distance would result in an explosion of the number of potential names requiring inspection. And working with a too small Levenshtein distance might on the other hand result in less coverage. The appropriate balance in this trade-off 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 4.6.

Table 4.6: 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 4.7. 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", in contrast, higher Levenshtein distances do reveal patentee names which are relevant for harmonising purposes (e.g. the Levenshtein distance for "INTERNATIOANL BUSINESS MACHINES CORPORATION" is 3).

Table 4.7: Examples of number of hits per Levenshtein distance per search key for 3 harmonised names

Harmonised 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.

c) Validation and quality control

A validation table was constructed by combining the output of approximate string searching with the observed number of related patent documents. Table 4.8 shows an example of the distribution of the patent counts associated with the number of retrieved names for "Deutsche Thomson-Brandt".

Table 4.8: Number of harmonised 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 was observed also in the other cases. Further analysis of this distribution for a sample of firms (n=50) revealed 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.

All retrieved harmonised names above this threshold were inspected manually and, if needed, were additionally harmonised. In case of doubts about the validity of harmonising two names, a brief online search was performed.

Several quality controls were performed after the manual validation, including verification of multiple or conflicting allocations. Most importantly, an analysis of inter-rater reliability was performed. For 22 harmonised names (i.e. 6 % of the total number of names), two persons independently engaged in the harmonisation exercise. Their inter-rater correlation was calculated by a kappa-score. The results in Table 4.9 show a very satisfying kappa score of 95 %, signalling consistent scoring.

Table 4.9: Kappa scores

		Value	Approx. Sig.
Measure of Agreement	Kappa	.952	.000
N of Valid Cases		2 915	

For this sample of 22 firms, recall and precision data were calculated. Table 4.10 shows the obtained results: a precision[20] rate of 99.5 % (proportion of correct validations by initial rating) and a recall rate of 99.8 % (number of hits correctly identified by the initial rating).

Table 4.10: Initial Rating * Validated Rating Cross tabulation (harmonised names)

		Validated Rating		Total
		0	1	
Initial Rating	0	2 638	16	2 654
	1	7	254	261
Total		2 645	270	2 915

When the same statistics are based on patent volumes, a precision rate of 99.9 % and a recall rate of 99.7 % are obtained (see Table 4.11).

Table 4.11: 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		910 857	300 909	1 211 766

4.3.2. Results

- **University patentees**

The second harmonisation layer consisted, on the one hand, of a manual harmonisation round for all university patentees with a patent volume above threshold 5. The improvement that is due to this additional layer is depicted in Table 4.12.

Table 4.12: Reduction in the number of distinct patentee names for university patentees throughout harmonisation layers

Patent system	Origin	After layer 1	Improvement layer 1 versus origin	After layer 2	Improvement layer 2 versus origin	Improvement layer 2 versus layer 1
All	50 252	40 561	19.28 %	33 602	33.13 %	17.16 %
EPO	3 434	3 138	8.62 %	1 739	49.36 %	44.58 %
USPTO	7 947	6 780	14.68 %	3 050	61.62 %	55.01 %
WIPO	6 091	5 201	14.61 %	2 596	57.38 %	50.09 %

The results clearly show the usefulness of this manual harmonisation layer for university patentees. Within patent systems, additional reductions of 45 %, 55 % and 50 % in the number of unique university names were achieved for EPO, USPTO and WIPO respectively. The lower improvement for all patent systems (17 %) is due to the fact that name variants are often scattered over patent systems, rather than

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

occurring within one system. In other words, this reflects that naming within patent systems is more consistent than naming across patent systems.

- **Top patentees**

In addition, the top patentees were subjected to this second harmonisation layer, based on an approach of matching similar names by using Levenshtein distances (complemented with exhaustive manual checks of the results). Again, the impact of this step proves to be considerable. This is clearly shown in Table 4.13, which reports on the number of identified unique name variants that were harmonised for the top ten companies. The example for “F. Hoffmann-La Roche” shows that 1431 unique name variants (present in the EPO PATSTAT person table) were harmonised after layer 2. The automated procedure from layer 1 resulted in 132 harmonised names. Besides the relevance of harmonisation efforts, this shows the usefulness of complementing the automated first layer with a manual second layer. On average, for the 453 organisations that were involved in this second-layer methodology, the number of unique name variants that were harmonised per organisation equals 106.

Table 4.13: Top 10 organisations in terms of underlying unique person names after harmonisation

Rank	Harmonised name (after second round of harmonising)	# Person names (after first layer of harmonising)	# Person names (after second layer of harmonising)
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

- **Improvement within Top 500**

Selecting the top players for this manual harmonisation already showed that several among the top 500 patenting organisations occur multiple times under a different name within the top 500. This in itself illustrates the importance of harmonisation. For the 453 unique organisations, the first harmonisation layer 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, representing an increase of 31.5 % (Table 4.14). The harmonisation efforts from layer 2 result in allocating yet an extra 30 960 names to these 453 organisations. This raises the aggregated patent volume from 10 328 128 to 13 251 949 patents: an increase of 28.31 %. Overall, an increase of 68.73 % in terms of patent volume is achieved.

Table 4.14: Impact of harmonisation for the 453 organisations 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 the same figures are calculated for the EPO, USPTO and WO patent documents separately, the overall increase in patent volume for the top 500 patentees amounts to 13.72 %, 21.98 % and 18.06 % respectively (see Tables 4.15, 4.16 and 4.17 for detailed results). Results for the publication authorities separately are lower, because name variants associated with high patent volumes are mostly scattered over publication authorities, rather than appearing within one publication authority.

Table 4.15: Impact of harmonisation for the 453 organisations 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 4.16: Impact of harmonisation for the 453 organisations 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 4.17: Impact of harmonisation for the 453 organisations 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 organisations are presented in Table 4.18. The table shows the often non-trivial changes in their ranking before and after harmonisation. “NEC Corporation” for example occupies the 7th place before harmonisation and the 2nd place after harmonisation. “Canon”, in contrast, drops from the 4th place before harmonisation to the 5th place after harmonisation.

Table 4.18: Top 10 patenting organisations with patent count and ranking before and after harmonisation

Harmonised name (after second round of harmonising)	After harmonisation		Before harmonisation		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 organisations before and after harmonisation is 0.92 (rank order correlation: 0.97).

Improvement of Top 500 harmonisation for EPO PATSTAT as a whole

The impact of the second-layer harmonisation for the top 500 (~453) organisations in terms of patent volume is considerable. As mentioned above, these 453 organisations have a total patent volume of 13 251 949 patents. This represents 26 % of the total patent volume in the EPO PATSTAT database (October 2009). Respective shares for EPO, USPTO and WO are 36 %, 35 % and 11 % (see Table 4.19).

Table 4.19: Patent volume of the 453 organisations overall, for the EPO, USPTO and WO

	Overall	EPO	USPTO	WO
Total patent count of the 453 organisations	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 mentioned above, 427 of the 453 organisations are companies. They hold over 98 % of the patent volume of the 453 organisations (13 004 136 patents). The total patent volume of all companies in EPO PATSTAT 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 4.20.

Table 4.20: 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 harmonisation effort has additionally reduced the number of unique patentee names by 0.16 % (from 7 536 191 to 7 523 564 unique names).

4.4. Conclusion

When creating patent statistics at patentee level, it is of prevailing importance to identify all the different name variants under which an organisation applies for a patent. Automated harmonisation methods result in 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 harmonise harmonised names. This second layer starts from the results of the first layer of automated harmonisation, developed by Magerman et al. (2009; cf. supra).

First, additional cleaning of university patentees (> 5 patents) further reduced the amount of unique name variants by approximately 50 % (averaged over EPO, USPTO and WIPO). Second, by additionally harmonising patentee names of 453 top patenting organisations, approximately 99.6 % of the total patent volume of these organisations has been allocated with a precision rate of 99.9 % and a recall rate of 99.7 %. For the 453 unique organisations, the first harmonisation layer had succeeded in allocating 16 670 extra name variants to these companies. This raised their aggregated patent volume by 31.5 %. The additional harmonisation efforts from layer 2 resulted in allocating an extra 30 960 names to these 453 organisations, which raised their aggregated patent volume by a further 28.31 %. Overall, an increase of 68.73 % in terms of patent volume is reached. The considerable overall improvement illustrates the benefits of combining automated procedures with case-based rules.

4.5. 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, Luxembourg.

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

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

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

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.

Schmoch U., Laville F., Patel P., Frietsch R. (2003). “Linking Technology Areas to Industrial Sectors” Final Report to the European Commission, DG Research.

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, Luxembourg.

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.

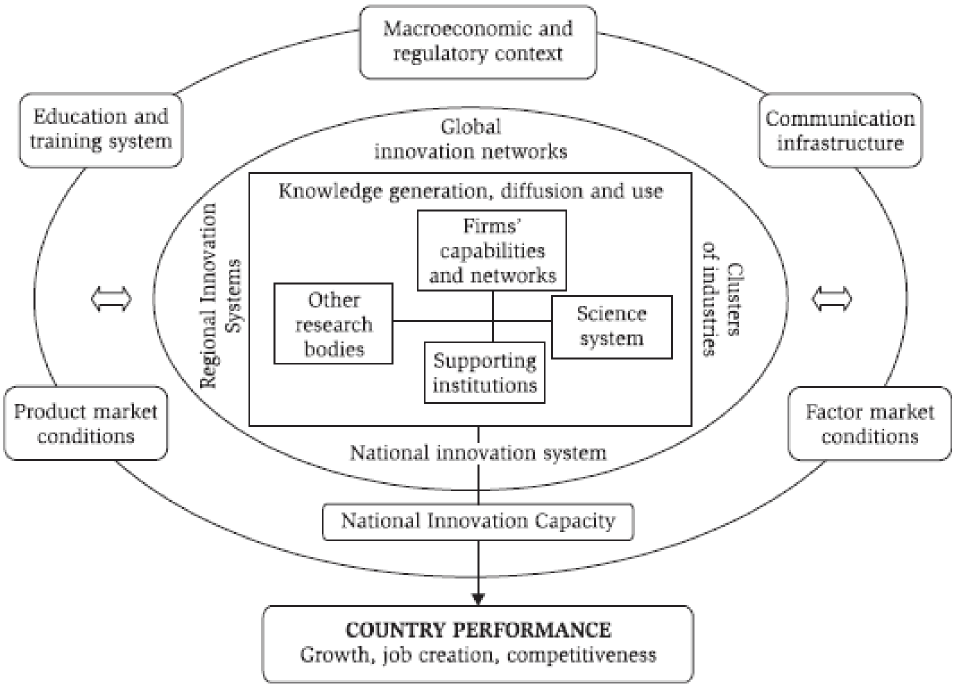
5. APPLYING THE METHODOLOGIES: AN ANALYTICAL STUDY ON THE EVOLUTION OF INNOVATION ACTORS AND THE INFLUENCE OF LEGISLATION

(B. Van Looy; M. Du Plessis; J. Callaert)

5.1. Introduction

The notion of innovation systems emphasises interactions between multiple actors, encompassing not only firms, but also knowledge-generating institutes like universities and research laboratories. In addition, governmental agencies can play a pivotal role in terms of supporting innovative efforts by means of providing funding and creating appropriate legislative framework conditions. Pioneered by scholars like Freeman (1987), Lundvall (1992), and Nelson and Rosenberg (1993), the concept of national innovation system (see Figure 5.1) has gained wide acceptance[21] as a guiding framework to understand and analyse innovation dynamics on a more aggregated level (OECD, 1999; European Innovation Scoreboard, 2002).

Figure 5.1: The European Research Area: New Perspectives (2007). Green paper published by the Commission of the European Communities, 4.4.2007


Macroeconomic and regulatory context
Education and training system
Communication infrastructure
Global innovation networks
Knowledge generation, diffusion and use
Firms' capabilities and networks
Other research bodies
Science system
Supporting institutions
Regional Innovation Systems
Clusters of industries
Product market conditions
Factor market conditions
National innovation system
National Innovation Capacity
COUNTRY PERFORMANCE
Growth, job creation, competitiveness

[21] See in this respect the triple helix framework which models mutual interdependencies between government, industry and academia (Etzkowitz & Leydesdorff, 1997).

Studies on innovation dynamics adopting this multiple-actor perspective have long faced the challenge of collecting and analyzing data whereby different types of institutional actors were identified. Whereas certain statistical data already reflected this perspective of institutional types (e.g. R&D expenditures broken down by firms (BERD) and Higher Education (HERD), an encompassing and accurate classification and methodology for the identification of institutional types for patent data was lacking. The methodologies outlined in the previous sections of this compendium provide a way of filling this gap.

This chapter highlights the relevance of such methodologies by illustrating one of their many potential applications. The reported study uses enriched patent data (mostly based on sector allocation, but also on name harmonisation) to help our understanding of innovation dynamics on a national innovation system level. More specifically, the impact is examined of legislative framework conditions — pertaining to ownership rights of publicly funded research — on the behaviour of universities institutions in terms of patenting activity.

5.2. A closer look at the patenting behaviour of universities (hei)

Based on the sector allocation methodology (see section 3 in this compendium), Table 5.1 shows the distribution of the proportion of patent applications per sector over time (1995–2005). A gradual increase (almost 1 %) is visible for the higher education sector. The proportion of patent applications for applications with individuals as patentee show a gradual decrease over time (almost 1 %), especially after 2001. All the other sectors display a more constant pattern over time[22] .

Table 5.1: Total Distribution of the five economic sectors over the application years 1995 to 2005

SECTOR	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
INDIVIDUAL	9.37 %	9.24 %	9.15 %	9.31 %	9.30 %	9.29 %	9.16 %	8.98 %	8.45 %	8.12 %	8.21 %
COMPANY	86.66 %	86.46 %	86.47 %	86.08 %	86.05 %	86.06 %	86.19 %	86.12 %	86.48 %	87.03 %	86.66 %
GOVENRMENT & NON-PROFIT	1.99 %	1.98 %	1.97 %	2.13 %	2.12 %	2.17 %	2.11 %	2.33 %	2.39 %	2.30 %	2.32 %
HIGHER EDUCATION	1.75 %	2.10 %	2.20 %	2.24 %	2.33 %	2.26 %	2.30 %	2.36 %	2.49 %	2.37 %	2.63 %
HOSPITAL	0.23 %	0.22 %	0.21 %	0.24 %	0.19 %	0.21 %	0.23 %	0.20 %	0.19 %	0.18 %	0.19 %

Table 5.2 presents the top 100 higher education institutions active within the EPO patent system. Together, these 100 institutions account for more than 70 % of all patents applied for by universities (the total number of harmonised higher education institutions is 1 125). From this table, it can be seen that the United States accounts for the majority of institutions. But the top 100 also includes higher education institutions in Belgium, Canada, Israel, the United Kingdom, Australia, Germany, France, South Korea, the Netherlands, Switzerland, Spain, Japan, Singapore, Denmark and China. Overall, this table shows that patenting activity by universities (and higher education institutions) differs across countries. A more formal test of whether country differences exist in the patenting activity of universities and other higher education institutions reveals high levels of significance ($p< 0.0001$)[23] . While in some countries universities and higher education institutions only account for a fraction of patenting activity (< 1 % in Japan, Germany and Sweden), this share is more substantial in other countries (> 4 % in the United States, Canada, the United Kingdom and Belgium[24]). This raises the question: what explains such country differences in terms of patenting behaviour?

[22] In absolute numbers, this means an overall increase over a ten-year period.

[23] ANCOVA results with amount of patenting activity acting as a dependent variable, GERD and year acting as control variables.

[24] The high share for Belgium is to a large extent due to the technological activities of IMEC, a research institute in the field of micro-and nano-electronics, which originates out — and is governed by — Flemish universities. IMEC has developed a pro-active stance towards the creation of IP, enabling the formation of R&D partnerships on a global scale.

Table 5.2: Top 100 higher education institutions (harmonised names) with > 50 patent applications within the EPO patent system (time period 1995 to 2006)

Country	Higher education institution	Grand total	% 1995–2006	Cumulative %	Rank
US	REGENTS OF THE UNIVERSITY OF CALIFORNIA	1538	5.37 %	5.37 %	1
US	BOARD OF REGENTS, THE UNIVERSITY OF TEXAS SYSTEM	587	2.05 %	7.42 %	2
BE	INTERUNIVERSITAIR MICROELEKTRONICA CENTRUM VZW	490	1.71 %	9.13 %	3
US	JOHNS HOPKINS UNIVERSITY	467	1.63 %	10.76 %	4
US	MASSACHUSETTS INSTITUTE OF TECHNOLOGY	452	1.58 %	12.34 %	5
US	CALIFORNIA INSTITUTE OF TECHNOLOGY	397	1.39 %	13.72 %	6
US	WISCONSIN ALUMNI RESEARCH FOUNDATION	390	1.36 %	15.09 %	7
US	BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY	377	1.32 %	16.40 %	8
US	UNIVERSITY OF WASHINGTON	376	1.31 %	17.71 %	9
IL	WEIZMANN INSTITUTE OF SCIENCE	349	1.22 %	18.93 %	10
IL	YISSUM RESEARCH DEVELOPMENT COMPANY OF THE HEBREW UNIVERSITY OF JERUSALEM	339	1.18 %	20.12 %	11
UK	CHANCELLOR, MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD	299	1.04 %	21.16 %	12
US	REGENTS OF THE UNIVERSITY OF MICHIGAN	297	1.04 %	22.20 %	13
US	TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK	284	0.99 %	23.19 %	14
US	UNIVERSITY OF FLORIDA	274	0.96 %	24.14 %	15
US	TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA	272	0.95 %	25.09 %	16
US	PRESIDENT AND FELLOWS OF HARVARD COLLEGE	262	0.91 %	26.01 %	17
US	CORNELL RESEARCH FOUNDATION, INC.	253	0.88 %	26.89 %	18
US	DUKE UNIVERSITY	249	0.87 %	27.76 %	19
US	UNIVERSITY OF NORTH CAROLINA	248	0.87 %	28.63 %	20
US	YALE UNIVERSITY	235	0.82 %	29.45 %	21
CA	UNIVERSITY OF BRITISH COLUMBIA	231	0.81 %	30.25 %	22
US	UNIVERSITY OF UTAH RESEARCH FOUNDATION	228	0.80 %	31.05 %	23
US	BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS	220	0.77 %	31.82 %	24
US	UNIVERSITY OF SOUTHERN CALIFORNIA	218	0.76 %	32.58 %	25
UK	IMPERIAL COLLEGE INNOVATIONS LIMITED	199	0.69 %	33.27 %	26
US	REGENTS OF THE UNIVERSITY OF MINNESOTA	192	0.67 %	33.94 %	27
UK	CAMBRIDGE UNIVERSITY TECHNICAL SERVICES LIMITED	190	0.66 %	34.61 %	28
CH	EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE ZÜRICH	189	0.66 %	35.27 %	29
US	RESEARCH FOUNDATION OF STATE UNIVERSITY OF NEW YORK	184	0.64 %	35.91 %	30
US	UNIVERSITY OF MARYLAND	179	0.62 %	36.53 %	31
US	EMORY UNIVERSITY	177	0.62 %	37.15 %	32
US	NORTH CAROLINA STATE UNIVERSITY	172	0.60 %	37.75 %	33
CH	ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE	171	0.60 %	38.35 %	34
US	UNIVERSITY OF ALABAMA	171	0.60 %	38.95 %	34
CA	MCGILL UNIVERSITY	168	0.59 %	39.53 %	35

Country	Higher education institution	Grand total	% 1995–2006	Cumulative %	Rank
US	PURDUE UNIVERSITY	167	0.58 %	40.12 %	36
US	UNIVERSITY OF ROCHESTER	158	0.55 %	40.67 %	36
BE	K.U. LEUVEN RESEARCH & DEVELOPMENT	154	0.54 %	41.21 %	38
US	UNIVERSITY OF VIRGINIA	149	0.52 %	41.73 %	39
US	TRUSTEES OF PRINCETON UNIVERSITY	144	0.50 %	42.23 %	40
US	UNIVERSITY OF MASSACHUSETTS	139	0.49 %	42.71 %	41
US	RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY	137	0.48 %	43.19 %	42
US	TEXAS A&M UNIVERSITY SYSTEM	136	0.47 %	43.67 %	43
NL	UNIVERSITEIT LEIDEN	136	0.47 %	44.14 %	43
US	PENN STATE RESEARCH FOUNDATION	133	0.46 %	44.61 %	44
US	BOARD OF TRUSTEES OPERATING MICHIGAN STATE UNIVERSITY	131	0.46 %	45.06 %	45
AU	UNIVERSITY OF QUEENSLAND	131	0.46 %	45.52 %	45
UK	UNIVERSITY COLLEGE LONDON	128	0.45 %	45.97 %	46
UK	UNIVERSITY OF SOUTHAMPTON	128	0.45 %	46.41 %	46
BE	UNIVERSITEIT GENT	123	0.43 %	46.84 %	47
NL	TECHNISCHE UNIVERSITEIT DELFT	121	0.42 %	47.27 %	48
US	NEW YORK UNIVERSITY	119	0.42 %	47.68 %	49
NL	UNIVERSITEIT UTRECHT	118	0.41 %	48.09 %	50
CH	UNIVERSITY OF ZURICH	117	0.41 %	48.50 %	51
US	THOMAS JEFFERSON UNIVERSITY	115	0.40 %	48.90 %	52
US	BAYLOR COLLEGE OF MEDICINE	111	0.39 %	49.29 %	53
US	GEORGIA TECH RESEARCH CORPORATION	109	0.38 %	49.67 %	54
AU	UNIVERSITY OF SYDNEY	106	0.37 %	50.04 %	55
US	GEORGETOWN UNIVERSITY	106	0.37 %	50.41 %	55
US	UNIVERSITY OF CONNECTICUT	106	0.37 %	50.78 %	55
US	VANDERBILT UNIVERSITY	105	0.37 %	51.15 %	56
FR	UNIVERSITE PIERRE ET MARIE CURIE (PARIS VI)	103	0.36 %	51.51 %	57
IL	HADASIT MEDICAL RESEARCH SERVICES AND DEVELOPMENT LTD.	102	0.36 %	51.86 %	58
US	TRUSTEES OF BOSTON UNIVERSITY	102	0.36 %	52.22 %	58
US	UNIVERSITY OF CHICAGO	102	0.36 %	52.58 %	58
US	OHIO STATE UNIVERSITY	99	0.35 %	52.92 %	59
US	ROCKEFELLER UNIVERSITY	98	0.34 %	53.26 %	60
BE	VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE VZW	97	0.34 %	53.60 %	61
US	UNIVERSITY OF IOWA RESEARCH FOUNDATION	97	0.34 %	53.94 %	61
US	UNIVERSITY OF PITTSBURGH	97	0.34 %	54.28 %	61
US	UNIVERSITY OF SOUTH FLORIDA	95	0.33 %	54.61 %	62
US	CASE WESTERN RESERVE UNIVERSITY	94	0.33 %	54.94 %	63
US	CITY UNIVERSITY OF NEW YORK	93	0.32 %	55.26 %	64
BE	UNIVERSITÉ DE LIÈGE	92	0.32 %	55.59 %	65

Country	Higher education institution	Grand total	% 1995–2006	Cumulative %	Rank
UK	UNIVERSITY OF BRISTOL	91	0.32 %	55.90 %	66
UK	UNIVERSITY OF MANCHESTER	90	0.31 %	56.22 %	67
UK	UNIVERSITY COURT OF THE UNIVERSITY OF GLASGOW	89	0.31 %	56.53 %	68
BE	VRIJE UNIVERSITEIT BRUSSEL	88	0.31 %	56.84 %	69
BE	UNIVERSITE LIBRE DE BRUXELLES	88	0.31 %	57.14 %	69
IL	TECHNION RESEARCH & DEVELOPMENT FOUNDATION LTD.	88	0.31 %	57.45 %	69
US	REGENTS OF THE UNIVERSITY OF COLORADO	87	0.30 %	57.75 %	70
US	TRUSTEES OF TUFTS COLLEGE	87	0.30 %	58.06 %	70
BE	UNIVERSITE CATHOLIQUE DE LOUVAIN	86	0.30 %	58.36 %	71
UK	UNIVERSITY OF STRATHCLYDE	86	0.30 %	58.66 %	71
US	UNIVERSITY OF GEORGIA RESEARCH FOUNDATION, INC.	86	0.30 %	58.96 %	71
DE	ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG	85	0.30 %	59.25 %	72
UK	UNIVERSITY OF SHEFFIELD	85	0.30 %	59.55 %	72
DE	EBERHARD-KARLS-UNIVERSITÄT TÜBINGEN	84	0.29 %	59.84 %	73
US	CARNEGIE MELLON UNIVERSITY	84	0.29 %	60.14 %	73
US	NORTHWESTERN UNIVERSITY	84	0.29 %	60.43 %	73
AU	UNIVERSITY OF MELBOURNE	83	0.29 %	60.72 %	74
CA	UNIVERSITE DE MONTREAL	82	0.29 %	61.01 %	75
US	UNIVERSITY OF TENNESSEE RESEARCH CORPORATION	82	0.29 %	61.29 %	75
US	WILLIAM MARSH RICE UNIVERSITY	81	0.28 %	61.58 %	76
DE	HUMBOLDT-UNIVERSITÄT ZU BERLIN	79	0.28 %	61.85 %	77
CA	UNIVERSITE LAVAL	79	0.28 %	62.13 %	77
IL	RAMOT AT TEL AVIV UNIVERSITY LTD.	79	0.28 %	62.40 %	77
US	OREGON HEALTH & SCIENCE UNIVERSITY	79	0.28 %	62.68 %	77
US	UNIVERSITY OF ARIZONA	78	0.27 %	62.95 %	78
ES	UNIVERSIDAD POLITECNICA DE VALENCIA	76	0.27 %	63.22 %	79
KR	KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY	75	0.26 %	63.48 %	80
FR	UNIVERSITE LOUIS PASTEUR	72	0.25 %	63.73 %	81
JP	UNIVERSITY OF TOKYO	71	0.25 %	63.98 %	82
UK	UNIVERSITY COURT OF THE UNIVERSITY OF EDINBURGH	69	0.24 %	64.22 %	83
KR	POSTECH FOUNDATION	68	0.24 %	64.46 %	84
AU	MONASH UNIVERSITY	67	0.23 %	64.69 %	85
UK	UNIVERSITY COLLEGE CARDIFF CONSULTANTS LIMITED	67	0.23 %	64.92 %	85
SG	NATIONAL UNIVERSITY OF SINGAPORE	67	0.23 %	65.16 %	85
US	WAYNE STATE UNIVERSITY	67	0.23 %	65.39 %	85
JP	KYOTO UNIVERSITY	66	0.23 %	65.62 %	86
NL	UNIVERSITEIT VAN AMSTERDAM	66	0.23 %	65.85 %	86
US	UNIVERSITY OF MEDICINE AND DENTISTRY OF NEW JERSEY	66	0.23 %	66.08 %	86
UK	UNIVERSITY OF DUNDEE	65	0.23 %	66.31 %	87
DK	DANMARKS TEKNISKE UNIVERSITET	62	0.22 %	66.53 %	88

Country	Higher education institution	Grand total	% 1995–2006	Cumulative %	Rank
UK	UNIVERSITY OF NOTTINGHAM	62	0.22 %	66.74 %	88
US	UNIVERSITY OF KENTUCKY RESEARCH FOUNDATION	62	0.22 %	66.96 %	88
US	BOARD OF REGENTS OF THE UNIVERSITY OF NEBRASKA	60	0.21 %	67.17 %	89
US	UNIVERSITY OF MIAMI	60	0.21 %	67.38 %	89
DE	UNIVERSITÄT STUTTGART	59	0.21 %	67.58 %	90
UK	ABERDEEN UNIVERSITY	59	0.21 %	67.79 %	90
NL	ERASMUS UNIVERSITEIT ROTTERDAM	58	0.20 %	67.99 %	91
US	BROWN UNIVERSITY RESEARCH FOUNDATION	58	0.20 %	68.20 %	91
US	UNIVERSITY OF CINCINNATI	58	0.20 %	68.40 %	91
CA	QUEEN'S UNIVERSITY OF KINGSTON	57	0.20 %	68.60 %	92
UK	UNIVERSITY OF LEEDS	57	0.20 %	68.80 %	92
NL	RIJKSUNIVERSITEIT TE GRONINGEN	57	0.20 %	69.00 %	92
US	ARIZONA BOARD OF REGENTS	57	0.20 %	69.19 %	92
US	BOARD OF SUPERVISORS OF LOUISIANA STATE UNIVERSITY AND AGRICULTURAL AND MECHANICAL COLLEGE	57	0.20 %	69.39 %	92
US	CURATORS OF THE UNIVERSITY OF MISSOURI	57	0.20 %	69.59 %	92
CA	UNIVERSITY OF SASKATCHEWAN	56	0.20 %	69.79 %	93
CN	NATIONAL TSING HUA UNIVERSITY	56	0.20 %	69.98 %	93
FR	UNIVERSITE JOSEPH FOURIER	56	0.20 %	70.18 %	93
US	UNIVERSITY OF ARKANSAS	56	0.20 %	70.37 %	93
UK	KING'S COLLEGE LONDON	55	0.19 %	70.57 %	94
JP	KEIO UNIVERSITY	55	0.19 %	70.76 %	94
US	LOMA LINDA UNIVERSITY	55	0.19 %	70.95 %	95
US	YESHIVA UNIVERSITY	54	0.19 %	71.14 %	96
DE	TECHNISCHE UNIVERSITÄT DRESDEN	53	0.19 %	71.32 %	97
US	WASHINGTON STATE UNIVERSITY RESEARCH FOUNDATION	53	0.19 %	71.51 %	97
CH	UNIVERSITE DE GENEVE	52	0.18 %	71.69 %	98
FR	UNIVERSITE CLAUDE BERNARD - LYON 1	52	0.18 %	71.87 %	98
US	IOWA STATE UNIVERSITY RESEARCH FOUNDATION, INC.	52	0.18 %	72.05 %	98
JP	TOHOKU UNIVERSITY	51	0.18 %	72.23 %	99
CN	CHINESE ACADEMY OF SCIENCES	51	0.18 %	72.41 %	99
US	UNIVERSITY OF AKRON	51	0.18 %	72.59 %	99
CA	UNIVERSITY OF MANITOBA	50	0.17 %	72.76 %	100
US	ARIZONA STATE UNIVERSITY	50	0.17 %	72.94 %	100
US	AUBURN UNIVERSITY	50	0.17 %	73.11 %	100
US	RENSSELAER POLYTECHNIC INSTITUTE	50	0.17 %	73.29 %	100

5.3. A closer look at the role of legislative framework conditions

In terms of policy measures, the rise of the entrepreneurial university phenomenon is often associated with the Bayh-Dole Act (1980) and the Stevenson-Wydler Act (1980). These American legislative initiatives created transparency with respect to the ownership of intellectual property rights originating from publicly funded research; whether performed by universities or companies, the involved institutions obtain in principle the right of ownership (for a more technical account, see Colsaet, 2005). This new legislation was an important stimulus for adopting and/or further developing intellectual property-related procedures and regulations at universities and research institutes. Along with the rise of science-intensive fields of economic activity (like biotechnology), the introduction of the Bayh-Dole act undoubtedly contributed to the strong increase of patenting activity undertaken by American universities from the 1980s onwards (Branscomb et al., 1999; Mowery et al., 1999; 2001).

It is argued that Bayh-Dole-like legislative framework conditions could be an interesting option for European countries in order to further stimulate innovation activity. Economic theories on innovation provide additional arguments in this respect. The seminal work of Arrow (1962) has already pointed out that market failures occur frequently in innovation. When one scrutinises the nature of technology developed by academic scientists, it becomes apparent that these technologies are often of an embryonic nature, requiring additional investments to arrive at market applications (see Jensen, Thursby & Thursby, 2003 for a revealing account). The absence of ownership rights leads to incentive issues at the level of the academic inventor and of his/her principal (i.e. the university). In other words, if scientific inventors and/or their principals are not acknowledged as 'owners', incentives to engage in further development efforts are absent and follow-up efforts — towards market exploitation — will be driven by voluntarism only. On the other hand, granting IP rights results in the creation of entrepreneurial agency within the universities (and HEI).

The next question then relates to who should acquire such rights: individual inventors or their principal (the university)[25]? Situating these rights at the level of individual inventors might result in under-investment due to risk averseness and/or the lack of capabilities to further invest in the development of the technology. If, on top of that, universities acquire no rights, conflicts of commitment might arise between agent and principal; as academic inventors pursue technology development activities while universities limit their scope to education and research. Moreover, when situating these rights at the level of the university, it becomes feasible to address specific concerns that stem from the nature of scientific work (e.g. rules on disclosure, impact on science and education). In other words, such university-specific regulations seem justified to guarantee the co-presence of multiple academic missions (science, education & knowledge transfer) and to avoid potential conflicts (including secrecy and skewing). Finally, granting rights to universities creates a more transparent 'market' situation towards industrial partners. Being explicit on terms and conditions not only seems fair from a funding perspective, but it might also reduce transaction costs[26].

While conceptual arguments might be advanced in favour of granting IP rights to universities, an empirical assessment of their impact seems equally relevant (for a detailed account of the Flemish case see Du Plessis et al., 2005). This poses the question of whether or not different legislative framework conditions coincide with differences in the amount of technological activity undertaken by universities within a particular national innovation system. To answer this question, a distinction between three different 'regimes' in terms of legislative framework conditions is used. The first regime considers HEI-specific legislative framework conditions which broadly reflect the Bayh-Dole legislation (i.e. ownership rights are granted to the principal of the research team carrying out the research activities). The second regime ('Professor's privilege') grants the rights to the individual researcher. Finally, some countries (e.g. the Netherlands) opt for more general regimes, in which rights are granted to employers.

[25] One could also envisage a situation in which such rights are situated at levels above the principal of the inventors (e.g. a patent organisation for a region or country as a whole). This would only make sense if economies of scale are significant; these are however limited (and relate to IP procedures). Moreover, by aggregating relationships, new conflict situations (both within and between involved organisations) can and probably will arise like witnessed in the past in both the UK (BTG) and the US (NRC). For a revealing account on this issue, see Mowery & Sampat (2001).

[26] Whether it actually will depends on the behaviour of the negotiating partners.

Table 5.3 provides an overview of the selection of European countries under study as well as the relevant situation in terms of legislation. Table 5.4 reports the results obtained by applying a fixed effect econometric model (ANCOVA) in which different legislative framework conditions act as independent variables. Business expenditures on R&D (BERD) as well as expenditures on R&D by higher education institutions are included as control variables (HERD). The number of patent applications by universities acts as dependent variable.

Table 5.3: Overview of countries under study — Impact of legislative framework conditions

Belgium	The governance of universities has become a regional responsibility (state reform of 1991). In Flanders, all IP from university researchers belongs to the university. A similar logic has been adopted in 1998 by the French Community.
Germany	Private and public employers have the rights to patent service inventions; in parallel, university professors own the patent rights to university inventions (1994 law on employee inventions). The 2001 Reform of Employee Law has changed university inventions to "service inventions" which means they now belong to the university.
Denmark	The Act on Inventions at Public Research Institutions (2000) grants rights to Public Research Organisations (PRO), but allows right of first refusal to the inventor. Before 2000 the rights were owned by the researcher/professor.
Finland	Employer has the right to file for a patent, also in the case of PRO. University inventions are notable exceptions: the patent rights belong to the employee (1967). Finland is currently changing its legislation (towards granting rights to universities).
Sweden	Professor's privilege.
Netherlands, France and UK	Three countries in which legislation is general, i.e. universities are considered as employers and own the rights on inventions made by staff.

Table 5.4: Impact of different legislative framework conditions on universities' technological activity

IP Rights	Mean	Std. Deviation	N
Employee has right to file for a patent	0.4630	0.72297	47
Employer has right to file for a patent	4.9846	4.66603	17
General employer-oriented IP	1.7193	1.45154	45

Source	Type III sum of Squares	df	Mean square	F	Sig.
Corrected Model	564.789(a)	13	43.445	26.520	0.000
Intercept	15.105	1	15.105	9.221	0.003
HERD	4.406	1	4.406	2.690	0.104
GERD	0.174	1	0.174	0.106	0.745
Year	15.458	1	15.458	9.436	0.003
IP Rights	41.105	1	41.105	25.091	0.000
Country	174.400	6	29.067	17.743	0.000
IP Rights*Country	68.486	2	34.243	20.903	0.000
Error	155.630	95	1.638		
Total	1030.578	109			
Corrected Total	720.420	108			

R Squared=0.784 (Adjusted R Squared=0.754)

It is clear that specific HEI-tailored legislative framework conditions have a considerable impact on the amount of technological activity observed. Countries adopting such legislation display higher levels of technological activity, as compared to previous periods and as compared to countries in which legislation opts for the professor's privilege (i.e. situating the ownership rights at the level of the individual researcher). Not only does one observe a notable difference in the countries that opt for professor's privilege, but the difference with broader, employer-oriented, legislation is also significant. This raises the question of whether the observed differences stem from shifts in technological activity — from one type of actor towards another, e.g. from individuals toward universities — or whether they reveal an overall net gain in terms of technological activity within the innovation system. Previous findings reported by du Plessis et al. (2005) are unambiguous for Flanders: the observed impact can indeed be interpreted as a net gain. Likewise, for the European countries under study, no crowding-out effects were observed, neither in terms of patenting activity undertaken by individuals, nor in terms of patenting activity undertaken by firms.

Given the clear effect observed in terms of technological activity, adopting legislative framework conditions that provide incentives to universities and at the same time take into account the specific role of scientific actors seems highly appropriate. Introducing such 'best' practices on a larger European scale might be more beneficial for the innovative performance of Europe than preserving the diversity that is currently present within Europe.

5.4. References

Arrow, K.J. (1962). Economic Welfare and the Allocation of Resources for Invention. The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton NJ, Princeton University Press.

Branscomb, L.M., Kodama, F., & Florida, R. (1999). Industrializing Knowledge: University-Industry Linkages in Japan and the United States. London, MIT Press.

Du Plessis, M., Van Looy, B., Debackere, K., & Magerman, T. (2005). "Assessing academic patenting activity: The case of Flanders," Paper presented at the Triple Helix Conference, Turin, May 2005.

Etzkowitz, H. & Leydesdorff, L. (1997), "Introduction to special issue on science policy dimensions of the Triple Helix of university-industry-government relations," Science and Public Policy, 24, pp. 2-5.

European Innovation Scoreboard 2002, Cordis Focus, December 2002.

Frascati Manual (2002). Proposed Standard Practice for Surveys on Research and Experimental Development. OECD Publications, Paris Cedex, France.

Freeman, C. (1987). Technology Policy and Economic Performance. Pinter, London.

Jensen, R., Thursby, J., & Thursby, M. (2003). "Disclosure and licensing of university inventions: 'The best we can do with the s**t we get to work with'," International Journal of Industrial Organization, 21, 1271-1300.

Lundvall, B.A. (1992). National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. London: Pinter Publishers.

Mowery, D., Nelson, R., Sampat, N. & Ziedonis, A. (1999). "The effects of the Bayh-Doyle Act on US university research and technology transfer." In: Branscomb et al. (Eds.), Industrialising Knowledge. MIT Press.

Mowery, D., Nelson, R., Sampat B., & Ziedonis, A. (2001). "The growth of patenting and licensing by U.S. universities: An assessment of the effects of Bayh-Dole Act of 1980," Research Policy, 30, 99-119.

Mowery, D.C., & Sampat, B.N. (2001). "Patenting and licensing university inventions: Lessons from the history of the research corporation," Industrial and Corporate Change, pp. 317-355.

Mowery, D.C., & Nelson, R.R. (1999). Sources of Industrial Leadership. Cambridge University Press, Cambridge.

Nelson, R.R., & Rosenberg, N. (1993). "Technical Innovation and National Systems." In R.R. Nelson, (Ed.), National Innovation Systems. A Comparative Analysis. New York: Oxford University Press, Inc.

OECD (1999). University Research in Transition. OECD STI-Report. OECD Publications, France.

Van Looy, B., Du Plessis, M., & Magerman, T. (2007). "Data production methods for harmonised patent statistics: Patentee sector allocation", Eurostat Working Paper — KUL/MSI Working paper.

Van Looy, B., Meyer, M., du Plessis, M., & Debackere, K. (2007). "The impact of legislative framework conditions on the patenting behavior of universities: An empirical assessment," Paper presented at the Triple Helix Conference, Singapore.

Van Looy, B. (2009). "The Role of Entrepreneurial Universities within Innovation Systems: An Overview and Assessment", Review of Business and Economics, 54, pp. 62-81.

6. GLOSSARY

BERD	Expenditure on R&D in the business enterprise sector
EEE-PPAT	ECOOM-EUROSTAT-EPO PATSTAT Person Augmented Table
EPO	European Patent Office
Eurostat	Statistical Office of the European Union
HEI	Higher Education Institutes
HERD	Expenditure on R&D performed in the higher education sector
IP	Intellectual Property
IPR	Intellectual Property Right
LAU	Local Administrative Units
NBER	National Bureau of Economic Research (US)
NUTS	Nomenclature of Territorial Units for Statistics
OECD	Organisation for Economic Cooperation and Development
PATSTAT	EPO Worldwide Patent Statistical Database
PCT	Patent Cooperation Treaty
R&D	Research & Development
SNA	System of National Accounts
USPTO	United States Patent and Trademark Office
USPTO TAF	USPTO Technology Assessment and Forecast database
WIPO/WO	World Intellectual Property Organisation
WPI	Derwent World Patent Index

European Commission

Patent Statistics at Eurostat:
Methods for Regionalisation, Sector Allocation and Name Harmonisation

Luxembourg: Publications Office of the European Union

2011 — 65 pp. — 21 x 29.7 cm

Theme: Science and technology
Collection: Methodologies & Working papers

ISBN 978-92-79-20237-7
ISSN 1977-0375
doi:10.2785/13718
Cat. No KS-RA-11-008-EN-N

KS-RA-11-008-EN-N

ISBN 978-92-79-20237-7

9 789279 202377

Publications Office