



Dynamics of Institutions and Markets in Europe is a network of excellence of social scientists in Europe, working on the economic and social consequences of increasing globalization and the rise of the knowledge economy.
<http://www.dime-eu.org/>

DIME Working Papers on INTELLECTUAL PROPERTY RIGHTS



Sponsored by the
6th Framework Programme
of the European Union

<http://www.dime-eu.org/working-papers/wp14>

Emerging out of DIME Working Pack:
'The Rules, Norms and Standards on Knowledge Exchange'

Further information on the DIME IPR research and activities:
<http://www.dime-eu.org/wp14>

This working paper is submitted by:

Grid Thoma

University of Camerino

A survey of sources and databases of IP information

**This is Working Paper
No 69 (May 2008)**

The Intellectual Property Rights (IPR) elements of the DIME Network currently focus on research in the area of patents, copyrights and related rights. DIME's IPR research is at the forefront as it addresses and debates current political and controversial IPR issues that affect businesses, nations and societies today. These issues challenge state of the art thinking and the existing analytical frameworks that dominate theoretical IPR literature in the fields of economics, management, politics, law and regulation- theory.



WP 1.4 New Component

The Performance of IPR Systems and Differences in Potentials for Growth: Complementary Data Base Development Activities

A survey of sources and databases of IP information (D 1.4.8)

Grid Thoma

University of Camerino, Italy

April 2008

Index

1. Motivations	3
2. Survey Based Datasets	4
Online databases	5
US Patent Granted Full Text Database (USPFT).....	5
US Published Application Full Text Database (USAFT)	6
USPTO Trademark Electronic Search System (TESS)	6
ESPACE on-line database.....	6
CTM – on line	7
Delphion.....	7
3. Off-line databases	11
The NBER patent database	11
PATSTAT	12
Patent data integration with other indicators.....	12
References.....	14

1. Motivations

Until recently empirical studies on the economics and management of innovation have suffered from a paucity of data at the firm level. Scholars of technical change have addressed the lack of data by following two directions. A first approach has tried to collect firm-level information through surveys based on representative samples of the population of innovators.

Regarding the US context two widely cited surveys are the Yale survey (Levin et al 1987) administrated in the early 1980s and its subsequent version conducted by scholars at the Carnegie Mellon University in the 1990s (Cohen et al 2000) . These two surveys provide an useful source of detailed information on the nature and strategies of innovation and the means used to appropriate the economic returns generated innovative activities.

Similarly, in the European context the Community Innovation Survey (CIS) collects detailed data on innovation and other firm characteristics such sales, employment, exports/imports, etc. Unlike Yale and Carnegie Mellon surveys, which have been administrated by academic researchers, the CIS is conducted by National Statistical Offices with the aim of achieving a large coverage of industries and types of innovators (large and small firms etc.) (Arundel, 2003). Unfortunately, integration of CIS data with other information, like patents and accounting data is made difficult by the limitations to the use of CIS data imposed by privacy laws in countries like Italy. These shortcomings of the CIS dataset limit its use for the purposes of research in economics, management and public policy. More recently, scholars have conducted new innovation surveys providing very detailed information on the factors driving innovation at the level of individual inventors (Harhoff et al. 1999; Gambardella et al., 2000; Giuri, Mariani et al., 2007).

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

A third line of exploration is centred on innovation counts and R&D. R&D expenditures are a measure of input and do not tell much about the ‘success’ of innovative activities. Moreover, especially in the case of European firms, data on R&D expenditures are often missing because reporting these expenditures is not required by accounting and fiscal regulations in some countries.

An increasing number of studies have used patent counts and patent-related indicators to measure the quantity and the ‘quality’ of inventive output. Patents as a measure of inventive success have their own drawbacks too but they are the most direct, detailed and objective measure of innovation (Griliches, 1981 and 1990; Pavitt, 1988).¹

¹ Typically, patent offices and examiners rely on standard classification to classify patented inventions. The most widely used patent classification is the International Patent Classification (IPC) which has been established by the Strasbourg Agreement (1971) and provides a common classification for patents, utility models and utility certificates. The IPC is organized hierarchically in terms sections, classes, subclasses and groups and it is updated and revised regularly to keep track of emerging new technological areas. The current version - the Eighth Edition of the IPC system, has been in use since January 1, 2006 (see <http://www.wipo.int/classifications/ipc/en/>).

Patent analysis has been pioneered by Zvi Griliches and colleagues (Griliches, 1981 and 1990; Griliches, Hall and Pakes, 1991) at the National Bureau of Economic Research (NBER) and by Keith Pavitt and colleagues at the Science Policy Research Unit (SPRU) -University of Sussex (Pavitt, 1985 and 1988; Patel and Pavitt, 1991). The NBER patent dataset on US data has represented a path-breaking effort in this field providing new data that are useful to account for differences in the ‘value’ of patents (Hall, Jaffe and Trajtenberg (2001 and 2005). Bronwyn Hall and colleagues have made public the NBER patent citation database. They have also disclosed to the research community the links between the names of USPTO patent assignees with the names of US companies listed in the Compustat dataset.

Following this stream of research in this work we analyze the available sources of microdata on patents and other IP information. In particular, we will consider their content, time coverage, mode access, complementary search and management tools and potential integration with other sources. Moreover the analysis will focus on the potential updatetability and scalability of the databases to other sources of company information, as discussed by the paper of Thoma and Torrisi (2007).

2. Survey Based Datasets

The so called “Yale Survey” administrated in the early 1980s – see Levin et al (1987) – have constituted the first large inquire regarding the use of patents in the manufacturing US sector and the effectiveness of different mechanisms through which firms appropriate the returns to their product and process innovations, including patents, secrecy, lead time, complementary sales and service and complementary manufacturing facilities and know-how. The main finding of the Yale Survey has been that of highlighting clear differences across industries and between product and process innovations in the effectiveness of the appropriability mechanisms used by firms. Moreover, it documented that for a large number of respondents more than one of the mechanisms was judged effective to appropriate the returns of the innovation. In most industries, including the most R&D intensive ones but not Pharmaceuticals respondent firms did not consider patents – but other mechanisms – as one of the crucial ways in which they profited from their R&D investments.

The Yale Survey has been followed to the so called Carnegie Mellon University (CMU) Survey administrated by Cohen, Nelson and Walsh (1994). The CMU survey builds in substantially and tries to improve some of the Yale survey. On the hand, it improves the Yale survey with respect to question wording, definition of response scales, and sampling strategy. It starts from the population of all the R&D labs or units located in the U.S. conducting R&D in manufacturing industries and links it with Compustat directory allowing researchers to conduct more complex and complete analysis on the economic evaluation of the IPR strategies pursued by manufacturing firms in US. On the other hand, the CMU survey aimed to built a comparative view of the Yale survey to see if pro-patent reform in the US legislations have been associated with changes in the effectiveness of patents and the other mechanisms in appropriating the return of innovation. They documented that increased importance of the use of patents following the legislative changes have been limited only to large firms and that at the same time the mechanism of industrial secrecy to have increased significantly since the Yale survey.

In the European context during the 1990s there has been different survey tentative to documenting the use of appropriability mechanisms by innovators. Arundel and Steinmueller (1998) used the Community Innovation Survey to look at patents as information channels in Europe. Meyer (2000) interviewed a group of European inventors of nanotechnology patents to understand the connection between their invention and the scientific research that they cite. Tijssen (2002) performed a mail survey amongst Dutch inventors to understand the contribution of science to successful technical inventions, and to test the validity of patent citations to scientific literature as indicators of science dependency. While these surveys provide new data, they have limited European coverage and are mostly biased towards large companies. More recently in 2003, in order to overcome some of the weaknesses implicit in earlier studies, an enormous and substantial effort has been advanced with the PatVal survey administrated by a pan-European team of research (see Giuri, Mariani et al 2007). PatVal is a large-scale survey designed to be representative of the universe of patents in our EU6 countries. It covers all technological fields, deals with both for-profit and non-profit applicants, and collects information on small, medium and large business companies. In 2003, patents with the first inventor located in one of our EU6 represented 42.2% of all EPO patents, and 88% of the EPO patents whose first inventor was in one of the EU-15 countries. PatVal's main objective is to collect information about patents and the underlying invention process on issues that had not previously been explored in depth because of lack of information in the patent documents. It also provides new proxies for variables like knowledge flows or patent value for which the present measures are subject to the discussions noted earlier.

Another interesting attempt has been conducted by the JPO with the launched of the Survey on Intellectual Property Related Activities (SIPA). JPO started this survey in 2002 for collecting data on various IP related activities including application, licensing and litigation of patent, utility, design and trademark. The survey is conducted for all applicants with over a certain threshold number applications in the previous year and randomly sampled ones for the rest of group. The sample size of 2004 survey is about 12,300 applicants, including firms, individual inventors and research organizations, and JPO collected 5,300 responses (response rate: 43.1%). SIPA covers a broad range of survey items. The survey consists of four parts, (1) applications of IPR, (2) usage of IPR, (3) information on IPR section at firm and (4) IP related infringements. A follow-up survey has been launched for improving and completing the SIPA survey (see Motohashi, 2007).

Online databases

US Patent Granted Full Text Database (USPFT)

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

Patents from January 1976 to the present offer the full searchable text, including all bibliographic data, such as the inventor's name, the patent's title, and the assignee's name; the abstract; the full description of the invention; and the claims. The display of each patent's full-text includes a hyperlink to obtain full-page images of each page of the patent.

Patents from 1790 to December 1975 offer only the patent number, issue date, and current US patent classification in the text display, and can be searched only within those fields. However, this limited text display also includes a hyperlink to obtain full-page images of all pages of the patent.

US Published Application Full Text Database (USAFT)

As in the case of USPFT, the database is available from the USPTO and consists of the full text of US published applications (including new utility and plant). The full text of a published application includes all bibliographic data, such as the inventor's name, the published application's title, and the assignee's name, as well as the abstract, the full description of the invention, and the claims. All of the textual words in the publication are searchable.

USPTO Trademark Electronic Search System (TESS)

The USPTO website at <http://www.uspto.gov/main/trademarks.htm> provides freely a wide variety of information about trademarks. The TESS contains more than 4 million pending, registered and dead federal trademarks and it provides access to the same text and image database of trademarks. The date of the last update to TESS is displayed on the search screens. Updates are performed prior to the start of business for the date indicated. Updates are scheduled daily on Tuesdays through Saturdays. Marks registered and published in the Official Gazette -- Trademarks preceding the date of the last update should be available in this search database.

SPACE on-line database

The database enables to search freely for information about published patent applications from over 80 different countries and regions. It is based on the PCT minimum documentation, which is defined by WIPO as the minimum requirement for patent collections used to search for prior-art documents for the purpose of assessing novelty and inventiveness. The managing institution – the EPO - has expanded the coverage of its internal database far beyond the PCT minimum documentation to include data from other countries and other time periods.

In March 2007, esp@cenet® held data on 60 million patents from 81 countries. A total of 30.5 million of these patents have a title, while 29.5 million have an ECLA class (the European Patent Classification) and 19.5 million an abstract in English. The following table gives an overview of the availability of the PCT minimum documentation in the worldwide database.

Table 1 Time coverage of the esp@cenet®: starting year of availability for the main patent offices

Patent Office	Facsimiles	Full Text	ECLA
CH	1888	1970	1888
DE	1877	1970	1877
EP	1978	1978	1978
FR	1900	1970	1902
GB	1859	1893	1859
US	1836	1970	1836
WO	1978	1978	1978

CTM – on line

CTM –ONLINE provides free access to information on EU Community trade mark applications and Community trademarks, updated on a daily basis. Initial searches are performed from either the basic or the advanced search screen. Further searches may be performed on the basis of the results of a prior search using the refine search option. Twenty results are displayed per screen, with a summary given for each case.

The advanced search screen offers the following input options: Trade mark number, Trade mark basis, Trade mark name, Trade mark type, Vienna Codes (this field is active only when Figurative, 3D or Colour is selected as Trade mark type), Owner name, Owner ID (this field is visible only when CTM is selected as Trade mark basis), Representative name, Representative ID (this field is visible only when CTM is selected as Trade mark basis), Nice class number, Status (this field is active only when a basis is selected in Trade mark basis), Filing Date, Registration Date, Date of International Registration (this field is active only when IR designating the EC is selected as Trade mark basis), Publication Date (the corresponding part of the CTM bulletin should be selected), Expiry Date, Acquired distinctiveness

Delphion

Delphion is proprietary database of Thomson and offers the patent collections to search contemporaneously inside the most important patent databases at the worldwide level. Moreover, Delphion analytical tools include some value added services: Citation Link creates graphical maps of forward and backward references; Snapshot allows quick online analysis of the results using bar charts; PtentLab-II supports offline analysis of results with 3D graphs and charts; Clustering performs keyword-based linguistic analysis; Corporate Tree facilitates targeted Assignee name searching. The patent collections of Delphion regard:

- i) The US patents application collection contains complete text and images of all patent applications published by the USPTO. This collection has bibliographic text, full text and images of published US patent applications from March 2001 to present.
- ii) The US granted patents provides complete text and images of all patents issued by the USPTO since 1974, and bibliographic text and some images since 1971. This collection also includes full images for backfile patents dated 1790-1971.
- iii) Derwent World Patents Index (DWPI) is a value-added patent information database covering over 13 million separate inventions from 41 different international patent-issuing authorities. The main advantage of the database consists in the English translation in the natural language of the its patent documents.
- iv) European Patents Applications contains European patent applications published by the EPO, including text and full document images of applications for European patents issued from 1979 to present, and full text from 1987 to present.
- v) European Granted Patents contains European patents published by the EPO, including bibliographic text and full document images of patents for European patents issued from 1980 to present, and full text from 1991 to present.

- vi) German Patent Applications This collection has biblio, first claims, and full document images of German patent applications published by the German Patent and Trademark Office combined with German data provided by INPADOC from 1968 to present, and full text from 1987 to present.
- vii) German Granted Patents collection contains German granted patents and utility models published by the German Patent and Trademark Office combined with German data provided by INPADOC. The collection has biblio, first claims, and full document images of German granted patents and utility models issued from 1968 to present, and full text from 1987 to present. It also includes some biblio and image coverage from 1967.
- viii) INPADOC is one of the most comprehensive patent collections in the world. It contains patent family documents from 71 world patent signatories and legal status information from 42 patent offices. Delphion brings together both of these pieces of information into a single searchable database and makes available 30 million patent family documents and 45 million legal status actions. Family and legal status data are displayed in a patent's Integrated View.
- ix) Patent Abstracts of Japan database includes Japanese unexamined patent applications in English for both Japanese and non-Japanese priorities from JAPIO, combined with Japanese data provided by INPADOC. At present, representative first pages are available for viewing online since 1976.
- x) WIPO Patent Cooperation Treaty (PCT) Publications are abstracts, full document images, and full text from over a hundred member countries since 1978.

The following table reports the searchable textual fields in Delphion for documents from the main patent offices. In bold there are depicted the information that can directly read in the patent document. As it can be noticed Delphion includes more detailed information for USPTO with respect the WIPO, EPO, and JPO.

Table 2. Search Fields in Delphion for the documents from the main largest patent offices

Id	USPTO	EPO	JPO	WIPO	Information	Description
1	X	X	X	X	Publication Number	The unique number assigned to the patent publication.
2	X	X	X	X	Publication Date	The date a patent was officially published.
3	X	X	X	X	Publication Country	The Country in which this patent was published or the patent issuing authority for this patent
4	X	X	X	X	Title	The title shown on the patent document.
5	X	X	X	X	Abstract	A brief summary or description of the invention.
6	X	X			First or Exemplary Claim	
7	X				Independent Claims	This field contains the claims made in the patent about the subject matter and scope of the invention.
8	X	X			Number of Claims	
9	X	X	X	X	Assignee/Applicant Name	The person(s) or entity(ies) to whom ownership of a patent was assigned at the time the patent was issued (the original assignee). In the case of application, the Assignee field shows the applicant.
10	X			X	Assignee/Applicant City/State	The address of the assignee
11	X			X	Assignee/Applicant Country	The address of the assignee
12	X				USPTO Assignee Code	The assigned special code given to assignees by the USPTO.
13	X				USPTO Assignee Name	
14	X	X	X	X	Application Number	Unique number assigned to applications when they are filed. This is not the same as the Publication Number, which is assigned when the application is published.
15	X	X	X	X	Application Date	The date when the patent or application was filed.
16	X	X	X	X	Application Country	The country where the patent or application was filed.
17	X	X		X	Attorney Name	The name of the legal representative for the patent applicant.
18	X				Domestic References	US patents and applications cited as references by this patent or application.
19	X				Number of Domestic References	
20	X				Forward References	US patents or applications that cite this patent as a reference.
21	X				Number of Forward References	
22	X				Foreign References	Non-US patents and applications cited as references by this patent or application.
23	X				Other References	Non-patent prior art that this patent references.
24	X	X		X	Designated States National	A designated country is one in which an invention is protected in addition to the country in which the patent for the invention is filed.
25	X	X		X	Designated States Regional	A designated regional state is one in which an invention is protected in addition to the region in which the patent for the invention is filed.
26	X	X			ECLA Codes	The European Patent Office Classification code to which this patent is assigned. The classification system is administered by the European Patent Office.

27	X				Examiner - Primary	The assistant patent examiner who examined the patent.
28	X				Examiner - Assistant	The primary patent examiner who examined the patent.
29	X	X	X	X	Family Patent Numbers	Set of patents filed with different patenting authorities that refer to the same invention. Delphion extends the family data received from INPADOC to create the family unit.
30	X	X	X		Inventor Name	The name of the person(s) registered on a patent or application as the inventor(s).
31	X				Inventor City/State	
32	X				Inventor Country	
33	X	X	X	X	IPC Codes	The International Patent Classification code to which this patent is assigned. The classification system is administered by the World Intellectual Property Organization (WIPO).
34	X	X	X	X	Main IPC	
35	X	X	X	X	Main IPC (1st 4 Digits)	
36	X				National Class	The National Class code to which this patent is assigned. The National Class code is taken from the INPADOC record and is not standardized. There is no list of definitions available for these National Class codes.
37	X				Original National Class	
38	X				Main National Class	
39	X				Field Of Search	The US Class codes that represent the fields (by US Class) that were examined prior to the granting of the patent. This data is developed/entered by the patent examiner.
40	X				Maintenance Status Code	The status regarding fee maintenance for this patent.
41	X	X	X		Number of Pages	
42	X	X	X	X	Priority Number	A number assigned to a patent application when it is first filed.
43	X	X	X	X	Priority Date	The date a patent application is first filed, important in establishing novelty of an invention.
44	X	X	X	X	Priority Country	The country in which the priority application was filed.

3. Off-line databases

While the on-line databases provide real time and very updated information, researchers are more often interested to off-line databases in spite of higher costs and difficulties of having updated information. Off-line databases can allow an easy generation and manipulation of indicators. Moreover the ex-post scalability and integrability with other sources of information is significantly higher, as demonstrated by the interesting study of Thoma and Torrisi (2008). The most important sources in terms of easy and economical access are given by the NBER patent and EPO Patstat database.

The NBER patent database

The NBER patent dataset on US data has represented a path-breaking effort in this field providing new data that are useful to account for differences in the 'value' of patents (Hall, Jaffe and Trajtenberg (2001 and 2005). Bronwyn Hall and colleagues have made freely available the NBER patent and citation database through the NBER website. These data comprise detail information on almost 3 million U.S. patents granted between January 1963 and December 1999, all citations made to these patents between 1975 and 1999 (over 16 million).

Moreover, they have also disclosed to the research community the links between the names of USPTO patent assignees with the names of US companies listed in the Compustat dataset and a reasonably broad match of patents to Compustat (the data set of all firms traded in the U.S. stock market). The CUSIP match is based on the 1989 universe of companies.²

These data are described in detail in the following working paper Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498. All users of these data should read this paper, and should cite it as the source of the data

The data are freely available in two compressed (".zip") formats: SAS transport (.tpt) files and ASCII comma-separated variable (.csv) files. The link is <http://www.nber.org/patents/>. The NBER is working on a major NSF-funded update and extension of this data. A new release of these files, bringing existing data up to date through December 2004, is anticipated for spring 2008. A variety of additional fields and indexes will also be provided. These are anticipated to include "link-out" tables connecting patent numbers to geographic entities (e.g. SMSAs), and a codification of inventor names. Some experimental updates up to 2002 have been published on the Prof. Bronwyn H. Hall's website.

² Further documentation on uses of the patent citation data, including the methodology paper and a CD containing the complete dataset itself, is available in the book *Patents, Citations and Innovations: A Window on the Knowledge Economy* by Adam Jaffe and Manuel Trajtenberg, MIT Press, Cambridge (2002). The book may be ordered from MIT Press. ISBN 0-262-10095-9.

PATSTAT

The EPO Worldwide Patent Statistical Database (PATSTAT), which is available under license from OECD-EPO Task Force on Patent Statistics, includes not only data on patent and utility models indicators such as citations and IPCs codes, but also on patent families based on priority links. The coverage of the database regards documents from more than 80 patent offices worldwide since the 1970s. The elements of PATSTAT are: The title and abstract of application; filing, priority and publications dates of the application; Applicants and Inventors and detailed addresses, IPC classification symbol, priority applications. Moreover, PATSTAT provides also complementary information on the citation links such as category of the citation, Citation identification, Origin of the citation, Non-Patent Literature bibliography, etc.

Around 50 institutions have subscribed PATSTAT and it is reasonable to expect a vast its deployment of this database in innovation studies.

Patent data integration with other indicators

In innovation studies a major obstacle to the integration of patent data with other indicators of firm performance in large samples is represented by the difficulty of univocally matching the names of patent assignees with the corresponding legal entity in business directories such as Compustat or Who Owns Whom.

Previous studies have addressed this issue by trying automatic matching procedures to reduce the cost of data standardization and integration. The first step in this setting is represented by name standardization. To our knowledge, the most important attempts at standardizing patentee names are the Thomson Scientific's Derwent World Patent Index (2002) and the USPTO's CONAME standardization files. More recently, another standardization method has been developed by a group of researchers from the K.U. Leuven for the Eurostat (Magerman, Van Looy and Song, 2006).

The Derwent Index is constructed by assigning a code to 21,000 patentees. This index accounts for legal links between parent companies and subsidiaries thus achieving a legal entity standardization. This is made possible by the use of information on corporate structure collected from secondary business sources. This includes also information on M&As, changes of names and reorganization (e.g., new subsidiaries). Legal entity standardization requires substantial manual, labor-intensive work and some loss of accuracy in name matching thus giving rise to a potentially large number of false positives. Moreover, the process leading to standard names and in case of M&As and name changes the criteria adopted for name standardization are case specific (Magerman, Van Looy and Song, 2006).

The CONAME file compiled by the US Patents and Trademarks Office is a semi-automatic standardization procedure which focuses on the first-named assignee reported in the patent document. For patents granted after July 1992 the assignee name is standardized and matched automatically with other standardized names in the same dataset. New assignees that are not matched automatically with standardized names in the dataset are matched manually. For instance, the entry of a new assignee whose standardized name does not match any previously standardized

names is examined by looking and the names of inventors. The CONAME file accounts for changes in assignee names but does not account for legal links between assignee names. Moreover, similar names with a different legal form or from different countries are not matched.

The K.U.Leuven (KUL) methodology consists in the standardization of patentee names and perfect matching of names. The advantage of this method is a high level of accuracy at the cost of some loss of completeness. This is a conservative, fully automatic methodology which, like the CONAME file, does not try to establish links between similar names neither it seeks to find legal links among assignees. The main advantage of this procedure is high precision, i.e., a limited number of false matches. Inevitably, this method does not fare well in terms of completeness since a high number of good matches may remain unmatched. The KUL methodology has been used to standardize and match assignee names of EPO patent applications published between 1978 and 2004 and USPTO granted patents published between 1992 and 2003 (Magerman, Van Looy and Song, 2006).

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

An interesting advancement in this direction has been done by the study of Thoma and Torrisi (2007), who have conducted a matching test by comparing assignee names in the PATSTAT dataset with company names in the Amadeus dataset for a sample of around 2,197 European publicly listed firms and their 146,728 subsidiaries. These firms have disclosed information on their R&D expenditures.

The names found in the two datasets are standardized using a variant of the KUL methodology and then matched by the Jaccard similarity string function (Jaccard 1901). Their experiment shows that approximate string matching (ASM) yields a substantial gain over perfect matching in terms of number of patent assignees found in the Amadeus dataset. However, these gains are obtained at the cost of a loss of accuracy. Depending on the level of precision which one aims to achieve, matching similar names implies a higher risk of false matches as compared with perfect matching. Moreover they estimated the number of false positives and false negatives at different levels of the Jaccard similarity (J) score, find a low incidence of these for levels of J score higher than 70%. These results suggest that using the approximate matching methodology yields significant improvements in terms of completeness at the price of a relatively small cost in terms of loss of precision.

References

- Arundel, A. (2003), Patents in the Knowledge-Based Economy, Report of the KNOW Survey, MERIT, University of Maastricht.
- Arora, A., Fosfuri, A. and Gambardella, A., (2003), “The Division of Inventive Labor: Functioning and Policy Implications”, Paper presented at the CREST conference in honour of Zvi Griliches, Paris August 25-27, 2003
- Cohen, W. M., R. R. Nelson, et al. (2000), “Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)”, NBER Working Paper No. 7552. Washington, DC, NBER.
- Derwent (2000), WORLD PATENTS INDEX - Derwent Patentee Codes, Revised Edition 8 ISBN: 0 901157 38 4, Thomson Publishers.Leuven Manual.
- Fosfuri, A., and Giarratana, M.S. (2007), “Product Strategies and Survival in Schumpeterian Environments: Evidence from the US Security Software Industry”, *Organization Studies* 28 (6): 909-929.
- Gambardella, A., D. Harhoff, and B. Verspagen (2005), “The Value of Patents.” *Universita Bocconi, Ludwig-Maximilians Universitaet, and Eindhoven University, Working Paper*: http://www.creiweb.org/activities/sc_conferences/23/papers/gambardella.pdf
- Giarratana. M. and Torrisci, S. (2004.), “Entry and Survival in Foreign Markes: Technology, Brand Building and International Linkages”, *Social Science Research Network - Electronic Paper Collection, SSRN_ID577401_code386435.pdf* (<http://papers.ssrn.com>).
- Giuri, P., Mariani, M. et al. (2005) “Everything you Always Wanted to Know about Inventors (but Never Asked): Evidence from the PatVal-EU Survey”. *LEM Papers Series 2005/20, Sant'Anna School of Advanced Studies, Pisa, Italy.*
- Griliches, Z. (1981), “Market Value, R&D and Patents.” *Economic Letters* 7: 183-87.
- Griliches, Z. (1990), “Patent Statistics as Economic Indicators: A Survey”, *Journal of Economic Literature*, XXVIII (Dec.): 1661-1707.
- Griliches, Z., Hall, H. B. and Pakes, A. (1991), “R&D, Patents. And Market Value Revisited: Is There a Second (Technological Opportunity) Factor?.” *Economics of Innovation and New Technology* 1: 183-202.
- Hall, B. H., A. Jaffe, and M. Trajtenberg (2001), “The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools.” In A. Jaffe and M. Trajtenberg (eds.), *Patents, Citations and Innovations*, Cambridge, MA: The MIT Press. Also Cambridge, Mass.: National Bureau of Economic Research Working Paper 8498 (October).
- Hall B. H., A. Jaffe, and M. Trajtenberg (2005), “Market Value and Patent Citations,” *Rand Journal of Economics* 36: 16-38.
- Hall H. B., Thoma G. and Torrisci S. (2007), “The market value of patents and R&D: Evidence from European firms”, Working paper 13426, National Bureau of Economic Research, Cambridge, Mass. (<http://www.nber.org/papers/w13426>).

- Harhoff, D., F. Narin, and K. Vopel (1999), "Citation Frequency and the Value of Patented Inventions." *Review of Economics and Statistics* 81(3): 511-15.
- Jaccard, P. (1901), "Bulletin del la Société Vaudoisedes", *Sciences Naturelles* 37, 241-272.
- Lanjouw, J. O., and M. Schankerman (2004), "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators", *Economic Journal* 114: 441-465.
- Levenshtein, V.I. (1966), "Binary codes capable of correcting deletions, insertions, and reversal", *Soviet Physics Doklady*, 10(8) S. 707-710.
- Levin, R. C., A. K. Klevorick, et al. (1987), "Appropriating the Returns from Industrial Research and Development." *Brooking Papers on Economic Activity* 3: 783-831.
- Magerman, T. Van Looy B., and Song X. (2006), "Data production methods for harmonized patent statistics: Patentee name standardization", Technical report, K.U. Leuven FETEW, MSI.
- Moser, P. (2005), "How Do Patent Laws Influence Innovation? Evidence from Ninetheenth-Century World Fairs", *The American Economic Review*, vol. 95 (4), September, pp. 1215-1236
- Navarro, G. (2001), "A guided tour to approximate string matching". *ACM Computing Surveys* 33 (1): 31--88.
- Griliches, Z. (1990), "Patent Statistics as Economic Indicators: A Survey, *Journal of Economic Literature*", Vol. XXVIII, December 1990, 1661-1707.
- Patel, P. and K. Pavitt (1991), "Large firms in the production of the world's technology: an important case of 'non-globalisation'." *Journal of International Business Studies* 22 (1), 1-21.
- Pavitt, K. (1985), "Patent Statistics as an Indicator of Innovative Activities: Possibilities and Problems", *Scientometrics*, 7 (1-2): 77-99.
- Pavitt, K. (1988) "Uses and abuses of patent statistics," in van Raan, A. (ed.) *Handbook of Quantitative Studies of Science Policy*, Amsterdam: North Holland.
- Pavitt, K., Robson, M. and Townsend, J. (1987), "The Size Distribution of Innovating Firms in the UK: 1945–1983", *Journal of Industrial Economics*, March, 35, 291–316.
- Powell, W. W., D. R. White, et al. (2005), "Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences", *American Journal of Sociology* 110(4): 1132-1205.
- Salton, G. and Buckley, C. (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing & Management* 24(5): 513–523.
- Schmookler J. (1966), "Invention and Economic Growth", Harvard University Press, Cambridge, MA
- Smith, T. F. and Waterman, M.S. (1981), "Identification of common molecular subsequences", *Journal of Molecular Biology* 147: 195-197.

Thoma G. and Torrisi S. (2007), “Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases”, CESPRI Working Paper 2007, Bocconi University.