Fraunhofer ISI Discussion Papers *Innovation Systems and Policy Analysis* No. 32

ISSN 1612-1430

Karlsruhe, July 2012

Identification of university-based patents: a new large-scale approach

Friedrich Dornbusch

Ulrich Schmoch

Nicole Schulze

Nadine Bethke

Fraunhofer Institute for Systems and Innovation Research ISI

Contents Page

Tables and Figures

1 Introduction

Measuring academic patent activity remains a complex task, especially in the case of Germany. The major problem here is the lack of a solid and comprehensive approach to identify academic patents that have not been applied for by the university themselves, but by other organizations (enterprises, research institutes etc.) or the inventors themselves. This paper briefly introduces a recently developed approach to identify university-invented patents based on the idea to match author names from scientific publications with inventor names from patent filings. By applying this new method to Germany, we attempt to address the lack of systemic evidence on patenting which involves German universities. This contributes to the methodological discussion and promises to be a new solution to the need for a systematic measurement method which can be compared across countries. Since the end of the 1990s, most European countries have been moving away from the individual ownership of academic patents towards systems of institutional ownership by the universities (Geuna/Rossi 2011). This trend was initiated based on the assumption that the levels of university patenting in Europe were low compared to the US. The Bayh-Dole Act, introduced in 1980 in the US, was seen as the main driver behind the growing patent portfolios of US universities. It acted as a prototype and role model for many European countries, even though the conclusions about its effect on knowledge and technology transfer were far from definite or conclusive (Kenney/Patton 2009; Mowery/Sampat 2004). Germany was one of the countries which introduced rules similar to Bayh-Dole and abolished the traditional professor's privilege (*Hochschullehrerprivileg*) in 2002.[1] Since then, employee inventions are owned by the employing university and no longer by the inventors themselves. If, however, research is financed fully or partly by external contractors like private companies, it remains possible for parties to negotiate the allocation of patent rights between the university, the company and the individual inventor (Geuna/Rossi 2011).

Thus, one of the fundamental aspects relevant for our research is the typology of university-based patents. Following Meyer (2003), we consider "university-owned patents" as those patents, in which universities or their technology transfer offices (TTOs) are listed as applicants. "University-invented patents", in contrast, are those patents, which list university affiliated authors as inventors. In this case, the applicant is a third party, e.g. an individual inventor or a private firm. Both groups taken together will be referred to as "academic patents" (Lissoni et al. 2008). The university-invented patents, in particular, remain a blind spot in analyses of academic patenting due to missing or poor

1 In the former East Germany the professors' privilege did not exist until 1989.

quality data. The lack of reliable methods to accurately measure the number of academic patents has become particularly valid recently, due to changing intellectual property rights (IPR) regimes and government models for technology transfer and the patenting activities of universities (Crespi et al. 2011; Geuna/Rossi 2011; Lissoni et al. 2008; Lissoni et al. 2009).

In the last few years, there have been growing concerns that the policy decisions in favor of Bayh-Dole-like IPR-regimes were made on the basis of missing or partially erroneous empirical data, leading to inadequate policy proposals. Structural differences between innovation systems in the US and those in Europe are obvious and raise questions about the transferability of Bayh-Dole-like regimes (Mowery/Sampat 2004). Recently conducted European studies have shown that the majority of patents in Europe are assigned outside the university, while the opposite is true in the US (over 62 percent of patents are university-owned) (Thursby et al. 2009). Hence, a large share of university-invented patents is not being filed by the universities themselves in Europe. This is not really surprising. It is important to recognize that, until recently, the professors privilege (faculty ownership) was common in Europe and inventors typically assigned patent rights to their research sponsors rather than to their universities. This results in different structures to academic patenting behavior and calls for methods to account for patents that have not been filed by the university itself. In this context, recent literature has highlighted the scarcity of statistical information (Geuna/Nesta 2006). Many attempts have since been made to improve the situation but there are still no systematic ways to measure academic patenting in Europe, which enable frequent and up-to-date cross-country comparisons. This is due to two reasons:

- A large share of university-invented patents is applied for by companies and is the result of co-operations with firms and
- in some cases the inventors apply for a patent on their own and appear as the applicant on the patent.

Recent studies have made significant contributions to identifying the share of university-invented patents in Europe but many gaps still remain (Geuna/Rossi 2011). The correct identification of academic patenting is essential in order to enhance our understanding of its organizational and individual implications, allow reliable estimations of its effects on social welfare and enable more suitable policy suggestions.

2 Previous work and the need for a comprehensive approach

The most significant problem with identifying structures and trends of patent applications from universities is to find those applications which have not been applied for by the university itself. Since the name of the applicant is listed on each patent application, patents filed by universities (university-owned) can easily be identified by keyword searches. But cursory searches for university names or surveys submitted to university TTOs are unlikely to map the reality in academic patenting. They do not capture the shares of university-invented patents and consequently tend to substantially underestimate the true number of patent filings with the participation of academic inventors. Hence, investigations aiming to draw a complete picture of academic patenting have to concentrate on finding the names of university scientists who are also registered as inventors in patent databases. In recent years, two different types of approach have been taken to identify university-based patents. These are:

1. Searching for keywords: The basic idea here is that it is common in Germany to give the academic titles on official documents such as patent applications, even though this is, legally speaking, not part of the name. Academic patents can then be found by searching for these titles (PROF/UNIV PROF/PROFESSOR etc.). This approach has been used several times for German patent applications by, e.g. Schmoch (2007), Czarnitzki et al. (2007; 2011), von Ledebur (2009); von Proff et al. (2011).
2. Matching lists: The basic idea behind this approach is to match existing staff lists of universities or other types of documents listing the names of professors with inventor data on patent specifications. This has been done for the US by Thursby et al. (2009), and by Lissoni et al. (2008; 2009) for France, Italy and Sweden in the so-called KEINS project (Knowledge-based Entrepreneurship: Innovation, Networks and Systems).

Both approaches offer useful and interesting insights into the research field of academic patenting. But they also have some limitations and weaknesses:

- The first approach is limited to Germany and Austria since only these countries commonly indicate the "Professor" title. Additionally, the title of professor is usually not denoted at the European Patent Office (EPO). Hence this procedure can only be used for the Austrian Patent Office and the German Patent and Trade Mark Office. Another, probably greater weakness of this approach is that it only finds those inventions which list a professor among the inventors of a patent. It fails to identify patents filed by another university staff member (e.g. assistant or PhD student) as inventor. Another shortcoming is – of course – that the number of professors not citing their title on the patents is also unknown. This number could be estimated and there are reasonable assumptions which can be made about this share, but it remains unsatisfactory. Furthermore, anecdotal evidence at least for Germany seems to indicate that the tendency to document the professor's title in patent applications is decreasing.

- The main weakness of the second approach is that generating the staff lists needed is costly and time-consuming because this data is usually difficult to find. Most countries do not keep comprehensive and up-to-date lists of university staff so these would have to be created and updated.

  Another problem is that if such staff lists do exist, they are usually limited to a certain group of people, e.g. those with an official function at the university like tenured professors. Thus, there is a similar problem as in the first approach: There is the risk of missing certain groups of inventors. Inventions often stem from persons other than those listed on official staff registers, like PhD students, assistants and lecturers without an official affiliation. This share is, as already stated, difficult to determine.

We are therefore proposing a new approach that might resolve some of these issues. To the best of our knowledge, there are no approaches to date which can capture all the patent applications from universities with a reasonable effort for handling and updating the databases. Our approach is based on the idea of checking for the same names of authors of scientific publications and inventors on patents. The main difference to previous attempts is that the names of university employees and associates are not derived from official staff lists of universities. We aim to capture every individual active in research who publishes in scientific journals[2].

One of the main advantages of this new approach is that we are able to update the author lists regularly and with reasonable effort.[3] The main precondition is to ensure that a reasonable matching algorithm is applied, which matches inventor and author data as precisely as possible. In this context, one of the biggest challenges for a project like this handling and matching large quantities of authors' and inventors' names is the occurrence of different persons with identical names, called homonyms. This can lead to erroneous assignments of patent applications to universities. Thus, our work aims to test the general feasibility of identifying patent applications from universities by matching the names of authors and inventors. In doing so, we have to manage the trade-off between tagging the greatest possible number of university patents and keeping the rate of incorrect assignments as low as possible.

---

2 We codified nearly all the organizational affiliations of the authors listed in Scopus. This allows us to state that 70 percent of all German articles stem from university authors (65000 of 78000 German articles in 2009).

3 An additional advantage is that the methodology is extendable to other countries even if personnel data from universities is not available on a larger scale.

3 Identification of university-based patents: Methodological approach

In this paper we apply our method to analyses of German applications at the DPMA (Deutsches Patent- und Markenamt; engl.: German Patent and Trademark Office) and EPO (European Patent Office) as an example. While other approaches have focused on other European countries, particularly the analyses done for Germany up to now have failed to provide data of sufficient quality to be used on the institutional level, particularly for the EPO. Previous attempts have relied strongly on estimations (see e.g. Schmoch 2007). One of the first attempts to use keyword searches (c.f. chapter 2) was made by Schmoch (2007) for German applications at the DPMA[4]. Despite the named methodological problems, this approach can be used as a helpful reference for the magnitude of expected patent applications. In the following, we will use this reference as a benchmark.

The following chapter describes our approach in more detail. Basically, we used a two-step process. The first step includes the construction of the appropriate databases by parsing the information required into specially designed tables in a relational database. In order to obtain all the information and provide usable datasets, we had to clean, harmonize and supplement missing data. This step is referred to as the "parsing stage" (Raffo/Lhuillery 2009). The second step involves the actual matching process. Here, we match the names of inventors and authors and use the filtering criteria applied in the parsing stage to increase the matching accuracy. We refer to this step as the "matching and filtering stage" (Raffo/Lhuillery 2009).

To judge the quality of our matching, we use precision and recall rates. These are commonly used parameters in information retrieval procedures. Precision is the proportion of correctly identified documents among identified ones; recall is the proportion of all correctly identified documents to the total number of all relevant documents, including the relevant documents not found. These parameters are associated with two different types of errors. When a Type I error (or false negative) occurs, it decreases the recall rate, whereas a Type II error (or false positive) decreases the precision rate.[5] It is usually assumed in the literature that the higher both precision and recall rates are, the better the match is (Baeza-Yates/Ribeiro-Neto 2011; Raffo/Lhuillery 2009).

---

4 For a detailed description, see Schmoch (2007).

5 Recall rate is defined as CR/(CR + CM), where CR is Correct Recall and CM is Correct Missing (error type I or false negative) and Precision Rate is given as CR/(CR + IR), where IR is Incorrect Recall (errors type II or false positive).

3.1 Matching author and inventor data: Parsing and choosing the selection criteria

Our aim is to provide a methodology that enables us to identify all the patents generated with the participation of university inventors. The basic idea is, as described above, to match patent information with information from scientific articles. In order to do this, we used two databases containing large datasets of information on the individual level. The patent data for the study were extracted from the "EPO Worldwide Patent Statistical Database" (PATSTAT), which provides information about published patents collected from 81 patent authorities worldwide. For the publications, we chose Scopus, provided by the company Elsevier, which encompasses information on articles in about 18,500 peer-reviewed journals and another 1,000 titles from trade publications, book series and conference proceedings. The next sections describe which information we drew on for both databases and how we applied it.

Both databases were designed as server-based relational databases and are exclusively used for statistical analyses. PATSTAT is updated every six months; Scopus every year. This meant we could use up-to-date data. Besides the standard information delivered by the providers EPO and Elsevier, both databases contain additional information from other sources. In order to target our research question, it was most important to get the inventors' addresses information for applications at the DPMA[6] (those of European patents are delivered by the EPO) and manually fill data gaps in Scopus. Data retrieval is done solely in the query language SQL. The following section describes the implementation, the data generation and the cleaning of the used datasets.

To give a complete picture of the data used, we briefly describe how the actual parsing took place. The parsing stage is a data preparation strategy to reduce noise in the name field (e.g. address, institution, title) without removing any information which might be useful in subsequent stages. We applied different parsing strategies in order to clean our data. Multiple parsing strategies have been shown to be the most successful in creating synergies and optimizing the matching results with regard to precision and recall (Raffo/Lhuillery 2009). We drew a random sample and conducted manual case studies to check the dataset in order to find the factors which could produce noise during the matching stage. Those factors identified as potential noise were removed. In doing so, we had to keep the balance between high recall and not restraining the precision of our matching. We cleaned all dots, symbols, blanks, hyphens, apostrophes etc. Furthermore, we separated double-barreled names and the initials of middle-names

6 This data, which was directly bought from the DPMA, was added to PATSTAT.

from first names unless they were connected by a hyphen. Finally, we had to deal with a specific characteristic of the German language – the "umlaut-problem". We decided to remove each umlaut and replace it with one character (e.g. ä > ae; ü>ue; ö > oe).

We created different tables and integrated them into an existing relational database. Each of these tables contains only the data we need for the project. They are mainly derived from the two databases described above. The first table contains all the relevant information from PATSTAT regarding German inventors. The second table contains all the relevant information from Scopus regarding German authors:

1. Country of origin: On the patent side, usually the country of the inventor not the applicant is used (Hinze/Schmoch 2004). This should ensure that, when a multinational company is the applicant, the country of the parent company does not appear as the applicant country if the invention was filed by a subsidiary in another country. On the publication side, we use the location of the organization to which the author is affiliated. We restrict our dataset on both sides to authors from German organizations and to inventors with a residence in Germany.

2. The organization: The applicant of patents invented with the involvement of a university researcher is only shown as the university itself in some cases. In most cases, these are other organizations or, in some cases, private persons. This is the unknown aspect and our research question. Our objective is to define this share. In doing so, those patents which are known to be university-owned are used to create a control set to estimate the share of patents we identify correctly. In the table containing the information from Scopus, we confined our dataset to those working at universities. Hence, the indication of the author's institution is important for our approach, because it significantly reduces the amount of data we have to deal with. The codification of Higher Education Institutions (HEIs) was conducted manually and supported by a keyword search. In Germany, HEIs can be captured relatively easily by keywords.

3. Name: In the majority of cases both databases include the full first name and the surnames. The inventors are complete, because patents are legal documents and the names have to be indicated. The large majority of authors' names in Scopus are almost completely mapped. From 1996 onwards, we can work with at least 88 percent of the documents, since they contain the author's first name. For current documents, this share rises to ca 97 percent. To keep precision high, we left out names where only an initial was available. Tests showed that the error rate would be too high with initials only. This means that we would match many wrong names, but those which could be additionally and correctly identified are too few in relation to the mistakes.

4. Location/region: Here we use postal codes, which are available in patent as well as publication databases. PATSTAT provides the address of the inventor's resi-

dence in the majority of cases, or, in a few cases, that of the individual's employer listed on a patent[7]. In the Scopus records, address information is only available for the organization where the academic is employed. The home addresses of authors are not usually documented. To identify the author's location, we use the postal codes of their employing organization. In doing so, we assume that the academic inventor lives in geographical proximity to the university. Some studies found that a number of academics live a long way from their university. We checked the inventor and applicant postal codes for this issue and found that, for 96.5 percent, the first digit and, for 85.9 percent, the first two digits are identical. Thus most inventors live near his/her university and the loss of excluding those living far away is limited. The postal codes were extracted from both databases automatically and cross-checked manually. Missing data were completed manually where we could be certain of assigning the correct postal code. In most cases this was possible up to the two-digit level.

5. Time window: The publication of an article and the registration of an invention should be timely contiguous. As a time reference, we refer to the priority year of a patent, which shows the year the invention was first filed and to the publication year of each article in Scopus. Here the time delay between submission and publication should be kept in mind. For a patent we know the precise point in time it was filed for the first time (the so called priority year).[8] One – from our perspective reasonable – assumption is that at least one of the inventors of an academic patent is or was also active in scientific publishing. However, we are fully aware that a publication prior to a patent harms the application process as it might be considered prior art. Hence, we can assume that the research activities leading to the invention took place immediately before the patent application. For publications we use the publication year. For our purpose, since we are focusing on the time of the research activity, it might be more appropriate to use information on the point in time when the article was filed, but this kind of information is not systematically captured by Scopus. Additionally, case studies showed that university employees moving to a company can be a methodological problem. While they have been publishing as a university employee, they apply for a patent as a company employee. For these two reasons, we have to estimate a time-lag between the filing and publication of an article. Since the journals covered by Scopus are predominantly higher ranking journals and have a comparatively intense review process, we assume the average time-lag to be between one and two years. It is generally assumed that extending the time window increases the number of authors, who rarely publish in journals but are registered in Scopus. This increases the possibility of additional homonyms and therefore reduces the

---

7 While the EPO provides these addresses as standard information, those from the DPMA were complemented by us later on.

8 However, we have to keep in mind that the patents are published 18 months after their initial registration.

precision. Furthermore, we wanted to keep the possibility to conduct up-to-date analyses and decided to keep the two year window. This is why we match data of a 2-year-publication period to each priority year of patent filings, considering a time-lag of one year that is needed for the review of scientific publications. To give an example: If an invention was filed in 2005, the related publication should be published in 2006 or 2007. Another reason for restricting the publication window to only two years is the fact that academics do change their affiliation over time – especially by entering industry– so that a longer publication window would increase the probability of incorrect affiliation assignments even in the case of correct author matches.

6. Subject: Each patent application is given a specific code describing its technological assignment and field, following the so-called International Patent Classification (IPC). The IPC is constantly being revised, reclassifying patent applications if new codes are introduced. This reclassification is also done in a backwards direction, meaning that older patent applications are adapted to the revised classification as well as new ones. A specific advantage for this analysis is that all the applications are indexed by patent examiners who are experts in their fields, assuring a very high quality compared to other classifications, for instance, of journal publications (Frietsch et al. 2010a). The IPC is built up hierarchically and consists - in descending order - of sections, classes, subclasses, main groups and subgroups. The respective hierarchical level corresponds to the number of digits of the encoding, where four digits correspond to the subclasses (World Intellectual Property Organization (WIPO) 2006). Thus, main classes can be divided into subclasses and the single IPC-groups allow a very precise classification of patents with ca 75000 single positions. The classification of articles in Scopus is more general. Single articles are classified according to the journal where they were published. Each journal is classified in a hierarchical order by scientific field. The classification system is also built up hierarchically and consists – in descending order - of main subject areas and subclasses. 343 positions with 4-digits are available for the classification of scientific fields. Thus, this classification is not very precise, but allows the general classification of a publication.

7. The challenge here is that the classifications of technology and science follow different criteria and a perfect concordance can not be established. However, we find at least a more general coherence between them. In order to connect these diverging classifications, we chose an existing technology classification with 34 technology fields (defined by IPC classes) (Schmoch 2008) and linked each Scopus code to one of these technology fields. In doing so we developed two different types of classifications. The first version was rather coarse-grained and assumes that a technological invention can stem from a broad spectrum of scientific disciplines. The Scopus classes were mostly assigned to one technology field on a two-digit level and are truncated. This means they are openly classified on the right string-side and cover all subclasses with the same two first digits. Hence, a larger spectrum of subclasses and a broader range of scientific fields are as-

signed to one technology field. The second version was much finer-grained. The Scopus classes were only assigned to technology fields on a four-digit level. This means we were much more precise in the assignment of scientific and technological fields. In this second variant, accurate mapping between the author and inventor is more likely, but it is also more probable that appropriate allocations are excluded because of a too narrow definition. To give an example: SCOPUS codes consist of four digits. If only the first two digits are given, all classes beginning with these numbers are targeted. Thus, code 17 describes all fields in computer sciences, while 1701 describes "computer science, miscellaneous", 1702 "artificial intelligence" and 1703 "computational theory and mathematics" and so forth.

Table 1 summarizes the remarks made above and shows the information available to match inventor and author data and to develop an appropriate and precise matching algorithm.

Table 1: Matching criteria: What do we have?

	Criteria	Patent information (PATSTA)	Article information (Scopus)
1	Country	Inventor country	Organization country
2	Organization	Applicant type	Author organization
3	Name	Inventor last and first name	Author last and first name
4	Location	Inventor address: Postcodes	Institution address: Postcodes
5	Time window	Priority year	Publication year
6	Subject	Technology classification	Scopus classification

Source: Own compilation.

Figure 1 refers to these matching criteria and shows schematically how they are exploited in order to construct a matching algorithm, which is as precise as possible without forfeiting too much recall. The missing variable x represents our research question: the patent invented by a university. This information is missing in PATSTAT. Information with regard to patenting activity is only available when patents are applied for by the university. But these patents remain a subset of all the patents stemming from university-related research. The information 1 (country) and 2 (organization) are utilized in order to reduce the amount of data we have to handle and to create a control sample for the recall rate. We will come back to this aspect later. The other four matching criteria form the actual basis of our matching algorithm. To optimize our results, we tested for different combinations of these criteria. Furthermore, we are able to apply more or less restrictive versions of these criteria, e.g. in postal codes we can use only the first digit as a matching criterion or we can use the two-digit level, which is much more restrictive. The same is possible with the subject matching, where we can use the finer-grained, more restrictive, or the coarser-grained, less restrictive concordance.

Figure 1: The basic matching model

$$\overset{\text{2) Organization}}{x_{\text{uni-inv}} = 1} \text{ if } (\overset{\text{3) Names}}{a_{\text{names match}}} + \overset{\text{4) Time}}{b_{\text{time match}}} + \overset{\text{5) Location}}{c_{\text{location match}}} + \overset{\text{6) Subject}}{d_{\text{subject match}}})$$

	Organization matching	Name matching	Time window matching	Location matching	Classification matching
PATSTAT	X	Full strings of last- and first name	Priority year	First two digits of the postcode	IPC classification = WIPO 34
SCOPUS	Author affiliation = university	Full strings of last- and first name	Publication year: One year time-lag and time-window	First two digits of the postcode	Scopus classification: fine- / coarse-grained

Source: Own compilation.

The following chapter shows how these different selection criteria work and presents the results for Germany. In addition, we compiled a sample of identified university inventors and questioned them via an online-survey to discover whether the identified patent was theirs. These results are also presented as a validation of our methodology.

3.2 Results: Matching and filtering stage

So far, the matching has been done for Germany only, since the semi-automated approach requires considerable data cleaning and a broad understanding of the research landscape and structures of a research system. Thus, we focus our empirical analyses on the German research system only. In addition, the abolishment of the university teachers' privilege in 2002 is also a motivation for us to analyze the impact on the structures and behavior of the actors involved.

We apply our approach to German applications at the DPMA and at the EPO. Both offices are of interest to innovation policy. On the one hand, applications at the DPMA give the broadest view of patent activities of German universities, since they represent domestic patent applications. On the other hand, applications at the EPO might be interpreted as those with higher economic prospects for two reasons. The first is the fact that an EPO application, which is often downstream to a DPMA filing, has significantly higher fees. The second is that we can assume that the EPO filing is only made if several European markets are to be covered (otherwise the EPO application is not necessary) and it is expected that the additional costs will pay off. Hence, a smaller share of patents is applied for internationally and we assume higher market expectations for these.

Because of the large number of author and inventor names, a simple test of conformity runs the risk of many false-positive errors. Hence, it is necessary to apply the additional selection criteria we described in Section 3.1. We show the results obtained with different combinations of these criteria. Furthermore, we discuss the validity of our method in the light of a survey we conducted among university inventors to estimate the precision of the algorithm and benchmark our results with those of a previous approach by Schmoch et al. (2007). Both benchmarks refer to DPMA patents. These findings are used to assign the approach to EPO patents. This means, that we verify our algorithm for DPMA applications and adapt it to the estimation of EPO filings.

The matching algorithm

Choosing the string matching algorithm was the first step in defining the actual matching procedure. We are aware that different types of matching algorithms exist (e.g. Soundex, Metaphone, N-gram, Token, etc. For a useful review with regard to the matching of patent data and name lists, see Raffo/Lhuilliery (2009)). Each of these algorithms has different advantages in different contexts depending on the used datasets and the research question. Since all string match algorithms standardize and therefore reduce the information (to different degrees) contained in the data, they can lead to higher recall rates. At the same time, however, they also reduce the matching precision (i.e. they cause higher type II error rates); this has been empirically tested and confirmed by Raffo/Lhuilliery (2009). Our dataset contains large numbers of names. In the year 2006, for example, 43,000 inventor names and 160,000 author names affiliated to a German university need to be matched. We crosschecked our dataset for homonyms and found that even relatively rare names can have identical counterparts. Hence, a single matching without further restrictions harbors the danger of identifying a huge number of matches, but is also most likely to return type II errors. We preferred to keep the level of precision as high as possible and decided, like several other authors, to apply a simple-string-match algorithm (Kim et al. 2005; Thoma/Torrisi 2007; Traijtenberg et al. 2006). This is combined with a multi-filter approach comprising the criteria described in Chapter 3.1. It was difficult to predict the impact of the different criteria on our matching approach. Therefore, we had to test the different selection options and compare them with the benchmark set. Since the name matching, the time window, the country and the organization are fixed, we had to employ and test the remaining two criteria in order to optimize our results:

- The fine-grained (F-conc) and the coarse-grained concordance (C-conc),
- the 1-digit (1digit pc) and 2-digit postcodes (2digit pc),
- different combinations of variations in postcodes and concordance are described below.

The different concordance lists (F-conc/ C-conc) enable us to test for the influence on the matching of a narrower and a broader overlap between technological and scientific fields. By using 1- and 2-digit postal codes, we can switch between a wider and a closer distance for inventor residence and author's institution.

Approach to verification

Figure 2 shows the results for each criterion. It can be observed that different selection criteria lead to similar trends, but significantly differing results in terms of absolute numbers. Although we compare these results with our benchmark set in order to obtain basic reference dimensions, we do not have a reliable method to specify the selection criteria which come closest to the actual proportions.

We need some way of verifying the most useful selection criteria. In information retrieval methodology, this is usually measured by recall and precision. One precondition for these types of quality measures is to have exact reference datasets available. We had to create these ourselves:

- Recall: Defining the control set for the recall rates was quite simple. We identified the number of patents on which the universities themselves appear as applicants. This number can be identified by quite simple keyword searches. Based on this number, we can calculate the share of patents we have identified with the help of the different selection criteria with high reliability and benchmark it against our university-applicant-only dataset.
- Precision: Since precision measures the share of correctly identified elements in all identified elements, we needed a sample in which all the identified elements can be correctly assigned as positive or negative. In order to create such a control group, we conducted a short online survey among all the authors in Scopus identified by the criteria "1-digit pc" or "F-conc", for which a valid email address could be found in the dataset.[9] We selected deliberately broad criteria in order to be able to test as many combinations of more restrictive criteria as possible. We addressed 1681 persons with 2782 patent applications. We received 435 exploitable answers which equals a response rate of 26 percent. A total of 678 patents could be tested whether they were correctly identified.[10]

---

9 We found e-mail addresses for the authors identified as inventors for 2799 of the total 4852 identified academic patents. This corresponds to 58 percent of all identified patents.

10 For a more detailed description of our survey, see the project report Schmoch et al. (2011).

Due to the large datasets, we cannot expect to reach one hundred percent for both recall and precision. We have to find the best fit between the two. In information retrieval this can be calculated with the F-score. The F-score represents the harmonized mean between recall and precision.[11] It can be weighted. By varying the ß-coefficient, it is possible to emphasize recall or precision or to give both equal weights.

In the following we discuss the matching results in light of the recall and precision scores and the benchmark set.

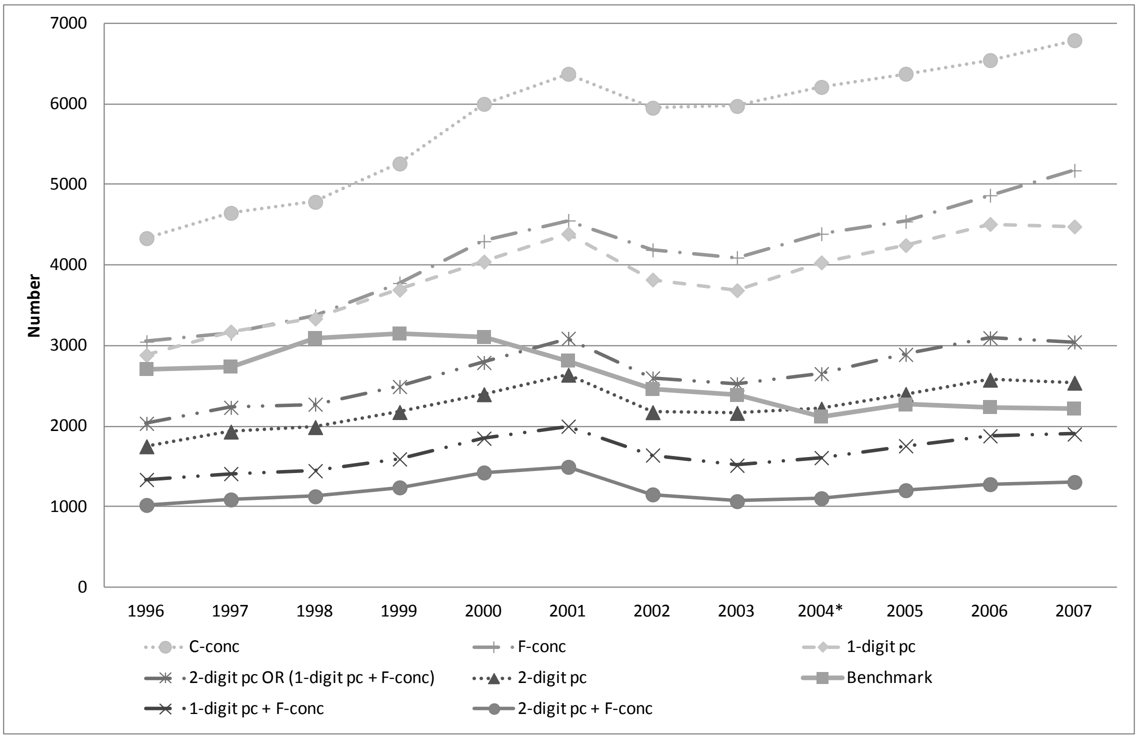
Academic patents at the DPMA: Discussing results and verification

Figure 2 presents the time series from 1996 to 2007 for university patents of German origin at the DPMA with the single application of each selection criterion. Basically, we find that all lines show a similar trend in reference to the benchmark line, with a maximum around the year 2000, a decline until around 2004 and then stabilization or a slight recovery. There are three obvious differences between the benchmark and the new approach. Firstly, all new lines show a rising trend and a peak in 2001. This cannot be observed for the benchmark set, which is more or less stable from 1996 onwards. This is certainly due to the changing degree of coverage of first names in Scopus, which is significantly higher after 2001 (cf. Chapter 3.1). Secondly, the main events using our approach seem to occur two years later than in the benchmark line. While for the latter, the maximum occurs in 1999, the maxima for the new approach are in 2001. From 2004 onwards all multi-filter lines show a similar development, but with different volumes. It is noticeable that the numbers rise significantly from 2004 onwards. Here one can assume that this is due to a larger dataset and the increased general coverage of publications in Scopus in more recent years.[12]

11 It is calculated using $F_ß = (1+ß^2)\ (p*r)/(ß^2*p*r)$; p = precision = $t_p/(t_p+f_n)$ and r = recall = $t_p/(t_p+f_p)$ where t_p means true positive, f_n false negative and f_p false positive (Baeza-Yates/Ribeiro-Neto 2011).

12 In fact, the number of registered publications in Scopus increased by around 30% between 2003 and 2007, which also applies to the publications of German origin.

Figure 2: Academic patents at the DPMA using different selection criteria[13]



Source: Own calculations and compilation.

We see that the four single selection criteria lead to a clear graduation of the obtained numbers of academic patents. The introduction of more and stronger selection criteria reduces the dataset in clear steps. Table 2 displays the recall, the precision and the F-scores for the application of different selection criteria. As expected, the stronger the selection criterion, the lower the recall rate and the higher the precision, implying a reciprocal relation between the two. The following conclusions can be drawn for the different criteria:

- The coarse-grained concordance (C-conc) leads to a significant reduction to almost half of the matched elements. However, since it is the weakest criterion in terms of selection, we decided to exclude it for the following analyses. We assume that it does not provide us with enough precision and that the addition of the concordance does not provide a reasonable number of retrieved patents. It delivers a high recall rate, but does not help us improve our algorithm as it is likely to cause many type II errors.
- This latter assumption is further supported by calculations made to define the precision of the fine-grained concordance (see Table 2). The application of the finer concordance (F-conc) leads to a reduction of a further 20-25 percent. While it has the

13 * Due to missing data (25 percent) in the postal codes of the DPMA data in 2004, we added a ratio of 1.25 to the data in 2004 in order to avoid distortions in the trend curve.

same recall rate as the two-digit postcode criterion, it shows significantly lower precision rates. This is also shown in the F-scores, where the concordance criterion alone shows a weak performance. Thus, we conclude that the concordance criteria alone are insufficient to improve the matching results and the only one which might be conducive in combination with the postcode criteria is the fine-grained concordance.

- The one-digit postal code criterion (1-digit pc) is a few percent points lower than the fine concordance, but stays significantly below the scores for the two-digit criterion in terms of the F-scores.

Table 2: Recall, precision and F-score

Selection criteria	Recall	Precision	F-scores		
			R=P (F_1)	P>R ($F_{0.5}$)	R>P (F_2)
1-digit pc	0.76	0.63	0.69	0.65	0.73
2-digit pc	0.71	0.77	0.74	0.76	0.72
F-conc	0.71	0.52	0.60	0.55	0.66
1-digit pc, F-conc	0.64	0.82	0.72	0.78	0.67
2-digit pc, F-conc	0.59	0.93	0.72	0.83	0.64
2-digit OR (1-digit pc + F-conc)	0.74	0.72	0.73	0.72	0.74

Source: Own calculations and compilation.

- The two-digit postal code shows similar results to our benchmark set from 2001 onwards (see Figure 2) and seems to be the selection criterion which achieves the best balance when weighting recall and precision equally (F_1).
- Combing the one-digit postcode with the fine-grained concordance (1-digit pc + F-conc) and the two-digit postcode with the fine-grained concordance (2-digit pc + F-conc) follow the same patterns as the other lines, but show a gradual reduction of matched patents in total numbers. They represent the lower border of Figure 2. The combination of the two-digit postcode with the fine-grained concordance represents the most restrictive criterion available to us. Thus, it is not surprising that it scores the highest with regard to precision and reaches the highest overall F-score if precision is given a higher weight than recall ($F_{0.5}$).
- Since the two-digit postcode excludes around 14 percent of the inventors (see Chapter 3.1), we decided to combine it with the one-digit postcode and the fine-grained concordance (1-digit pc OR (1-digit pc + F-conc)). Both are connected with an "or" constraint, meaning that either the two-digit postcode or the one-digit postcode is matched with the fine-grained concordance. In doing so, we tried to complement the good fit of the two-digit postcode criterion and improve the recall rate by searching in the missed inventor set, while keeping acceptable precision. Thus, we added the one-digit postcode to include the missed inventors and restricted the dataset to those which match the technological profile. This led to an increase in the numbers of about 20 percent. If we give recall a higher weight than precision, this combined criterion obtains the highest F-score.

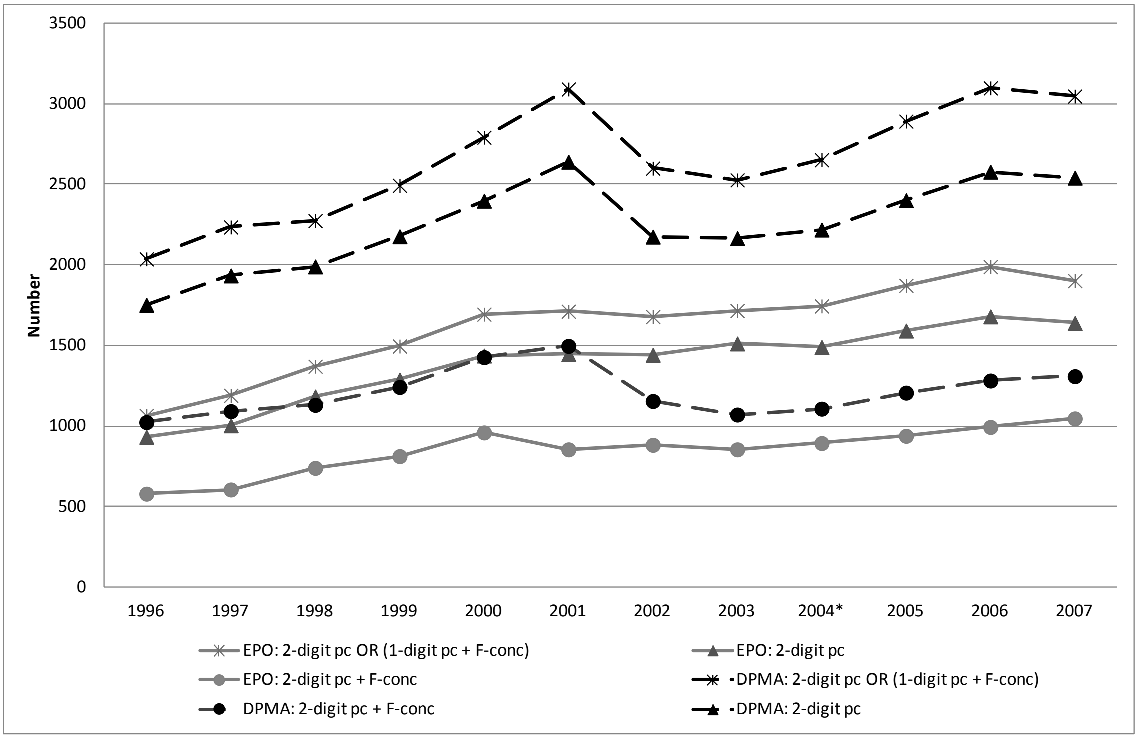
Our analyses reveal that the proper selection algorithm might depend on the intention underlying the matching. The F-scores indicate the choice made between a higher valuation of recall or of precision. The first is most likely to be met by combining the two-digit postcode with the one-digit postcode and the fine-grained concordance. The balance between the two is provided by the two-digit postcode criterion. The highest F-scores are reached by combining the two-digit postcode with the fine-grained concordance if precision is assigned greater weight than recall.

To sum up, we propose three different versions for future analyses:

- Conservative: Since the two-digit criterion delivers reasonable results in comparison with the benchmark dataset and both recall and precision rates are acceptably balanced, we suggest using this as the standard criterion for estimating and analyzing the structures and trends in German academic patenting. This criterion is still conservative with regard to the overall estimation of the number of academic patents. It also remains relatively restrictive in the sense of precision vs. recall. Thus, we suggest this criterion should be used to conduct conservative estimates of the total amount of academic patenting.
- High recall: Combining the two-digit postcode with the one-digit postcode and the fine-grained concordance will deliver a high recall at the expense of precision. This is likely to be useful if the analysis aims at broadly identifying academic patents with acceptable precision. This might be of particular interest, e.g. when the identified patents are going to be merged with clearly demarcated data (e.g. small datasets with information on a firm level).
- High precision: If the intention is to conduct analyses on an institutional or patent level, we suggest accepting lower recall rates in return for higher precision. In doing so, we can reduce statistical noise. This should make such analyses more solid and conclusions or policy suggestions more profound, but will reduce the number of cases.

The last step in the development of our method is to transfer our approach to the EPO. Therefore, we applied the three different criteria (described above) to the German applications at the EPO. The results are shown in Figure 3.

Figure 3: Comparing German academic patenting at the DPMA and the EPO



Source: Own calculations and compilation.

As expected, there are similar trends for the EPO and the DPMA. The EPO patents exhibit a stronger rise in application numbers from 2003 onwards. The different selection criteria show a comparable impact on the volume of the query outputs. The main difference lies in the lower total number of applications at the EPO. Academic EPO patents are significantly below the DPMA patents. However, this is not surprising since the total number of EPO patents with German origin is in the range of 60 percent of the patents filed at the DPMA. This share seems to be even higher in academic patenting, around 70 percent (with deviations depending on the year and criteria). One explanation might be that academic patents are usually research-intensive and are connected to higher economic expectations (Czarnitzki et al. 2011; Frietsch et al. 2010b). Thus, they are more likely to be filed at the EPO.

Main limitations and critical reflections of the approach

Even though we managed to make some advances with regard to the problems found in previous approaches to identifying academic patents, we still have to deal with some limitations. These are mainly connected to the availability and quality of the databases. One of our problems was identified by an analysis of our recall-dataset. We manually checked the university-owned patents that had not been identified by our algorithm in order to determine the main errors behind the missed matches, apart from the postcodes. As a result, we can state that, in 18 percent of the cases, the concordance (F-conc) proved to be not fitting and in seven percent the attributed organization was incorrect. Therefore our algorithm was not able to identify these persons as academics. In the majority of the missed cases (75 percent), the main reason was that the searched-for name was missing, i.e. not available in Scopus. Thus recall errors are primarily due to missing publications in the selected time window or to different spellings in the two databases:

- Dealing with the different spellings seems to be difficult, considering the size of the used datasets and the number of possible homonyms. This makes it difficult to apply string matching algorithms without causing type II errors (reducing precision). Furthermore, we have to acknowledge that research-active individuals, even if they generate inventions, do not necessarily publish in scientific journals. Hence we have to expect some omissions in the lists which are (in the short term) difficult to resolve.
- It is also possible that the publication databases do not cover all the journals in which university staff and associates publish. A bias could be observed for the missing names, with missing patents mainly in technology-intensive fields, especially mechanical and civil engineering. These subject areas are not satisfactorily covered by our publication database.[14] We have to admit that this might be of particular relevance for Germany. Many German academics in the engineering sciences are more prone to publish in German language journals, which remain underrepresented in Scopus. However, we did observe a rising coverage of these scientific areas, which might lead to a significant reduction of this bias and an improvement of our approach in the future.

These observations might explain one remarkable difference that appears when comparing the recall rates between the EPO and the DPMA (see Table 3). The recall at the EPO is significantly higher than at the DPMA. One reason might be that especially research-intensive inventions are registered at the EPO, while more applied, less research-intensive areas tend to apply to the DPMA. Academic research is generally closer to research-intensive technology, so that a higher number of publications can be observed here.

Another area for further improvements is the regionalization of the data. We used postal codes which proved to be quite useful. But we are missing at least around 15 percent of potential matches due to our assumption that the academic inventor lives close to his university. Although we decided that this is an acceptable limitation, we suggest searching for a more fine-grained solution to the location issue as a selection criterion.

---

14 The situation is even worse in Web of Science. The main reason we decided to work with Scopus is that it covers more of the technological fields which are relevant for academic patenting, even though the coverage remains (in parts) unsatisfactory.

Table 3: Comparing the recall at the DPMA and the EPO for 2006

Selection criteria	Recall DPMA	Recall EPO
2-digit pc	0.71	0.75
2-digit pc, F-conc	0.59	0.67
2-digit OR (1-digit pc + F-conc)	0.74	0.78

Source: Own calculations and compilation.

This will also be necessary when applying our approach to other countries. Thus, we plan to implement an additional geographical concordance, which will enable us to translate the postal codes into NUTS codes (nomenclature des unités territoriales statistiques). Another option is to use a geographical concordance based on geographical distances and not related to administrative boundaries that do not satisfactorily map the functional relations between entities like the working and living places of inventors. In addition, the different granularity of postal codes in different countries will limit the comparability of the findings. The implementation of NUTS codes in combination with geographical distances as a geographic concordance could be particularly helpful to improve the matching at the borders between two 1-digit areas where a neighboring 2-digit area may belong to a different 1-digit area than the central 2-digit one.

4 Further research: Identifying universities as centers of research excellence – the patent dimension

When looking at the total number of academic patent applications, universities' contribution to technology seems to be relatively moderate at first sight. However, a more detailed analysis of especially research-intensive fields shows that their input to new technology is substantial (Schmoch 2004). Therefore, patent applications should be considered a relevant output of universities in addition to other ones such as publications.

However, identifying all university-based patents in a consistent, replicable and automated manner remains a difficult task. A major restriction of the approach commonly used so far based on the title "Professor" is that the share of inventions by academic staff not bearing the title "Professor" had to be estimated and it was impossible to assess the number of patent applications by single universities (Schmoch 2007). Using the suggested match of authors and inventors, it is now easier to introduce a further filter for publications in order to restrict the search to selected universities. This may be useful for identifying "entrepreneurial universities" (see, e.g. Ranga et al. 2003), or for the assessment of universities in general. Furthermore, the method can be used to supplement large datasets of higher education institutions (HEIs) such as the EUMIDA[15] dataset for all European HEIs with output indicators for each individual HEI.

Additionally, our approach enables us to further our understanding of the interaction between universities and their local environment in terms of technology transfer. Some studies hint at the significant contributions of universities to their local environment, even in technology transfer (see, e.g. Youtie/Shapira 2008). By regionalizing our data and complementing it with data on regionalized structural indicators from official databases, it will be possible to assess not only the output of academic patents by regions, but also to explain which regional characteristics might influence academic patenting behavior. In addition, it can be analyzed which local branches benefit from academic output in terms of patents and university-based innovations.

Another direction for further research is the analysis of the determining factors for collaborations of universities and other types of patent applicants regarding patent filings. Small and medium-sized enterprises (SMEs) and large multinational companies differ significantly in their affinity towards and activity in academic patenting (see, e.g. Fontana et al. 2006; Lissoni et al. 2008). Matching data on the level of single institutions between universities and firms enables us to gain further insights into co-patenting and collaboration patterns between firms and universities.

5 Summary and conclusions

This paper described a method to consistently and automatically identify all university-based patents by matching the names of inventors on patent filings with authors of scientific publications. This methodological procedure can be adjusted for different purposes. Handling selection criteria rigorously allows us to generate a dataset of university-based patents with high precision. A less restrictive setting of selection criteria yields a higher coverage in terms of the overall picture and total numbers of academic patents. Using the authors of scientific publications allows us to identify research-active staff at universities without having to continuously compile and update staff lists for comparison with inventor lists. In particular, the author lists also cover research-active staff without official teaching functions who are often not covered by the official staff lists.

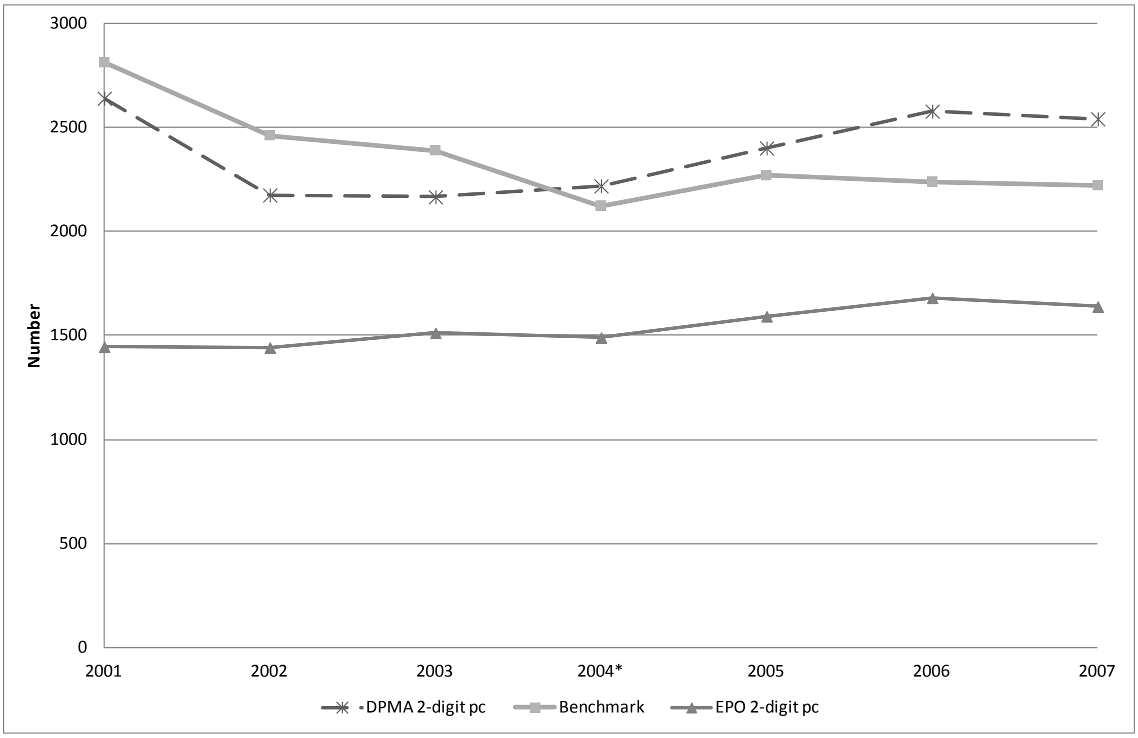
When considering the results of the recall and precision analysis and its good congruence with other previous approaches, especially in comparison to the benchmark data-

---

[15] See http://thedatahub.org/dataset/eumida

set, the 2-digit postal code criterion proved to be the most suitable for a balanced dataset in terms of recall and precision (see Chapter 3.2). Hence, we suggest using this as a standard criterion for the estimation and further analysis of the structures and trends in German academic patenting.

Figure 4: Academic patenting in Germany by application authority and by standard criterion



Source: Own calculation and compilation

Figure 4 displays our results for the overall estimation of academic patenting in Germany. We can state that, in general, the results of the new method are in line with the results of previous approaches for the number of academic patents at the DPMA. Although we are not able to provide benchmark data for German EPO filings, we obtain reasonable results from applying our algorithm to filings at the EPO. We observe a slight increase in the numbers between 2001 and 2007. This shows that the academic EPO patents are not affected by the decline in patenting activity, which appears to be logical when considering the arguments about patent quality and value. Patents with higher economic expectations are usually filed at the EPO. It is more likely that potential applicants decide not to patent inventions for which they have lower expectations.

The major difference between the old and new approach is that the new one no longer relies on the professor's title being stated in the patent application. This has two advantages. First, the number of professors who do not list their title and the share of academic staff in academic patenting without a professor's title are unknown. Thus, up to now, a large part of the identification procedure relied on estimations. This is not only error-prone and unsatisfactory, even though reasonable assumptions can be made about this share, but it also hinders analyses on an institutional level. Second, we no longer have to rely on the professor's willingness to indicate the title on the patent declaration. There is some anecdotal evidence, at least for Germany, that there is a decreasing tendency to document the professor's title in patent applications. Our newly developed approach is a promising solution to circumvent this problem and provide a reliable tool for the identification of academic patenting in Germany. This will enable us to conduct analyses on the level of individual universities and to combine our data with other information like that contained in the EUMIDA-dataset.

In summary, we can state that our approach has several advantages compared to previous approaches. Even though we were able to prove the applicability and reliability of this kind of large-scale approach, there is still some room for improvements. Using NUTS codes in combination with geographical matrices seems to be beneficial. Furthermore, one could imagine experimenting with string matching algorithms and additional matching criteria like, e.g. the appearance of co-inventors on more than one patent filing. In combination with a regression-based weighting of selection criteria, these approaches could be used for a more fine-grained determination of the optimal matching.

A substantial advantage of the present approach is that university-based patents can also be analyzed for countries where the use of the title "Professor" in patent applications is less frequent. In the research project underlying this paper we already analyzed the academic EPO applications of France and Switzerland with encouraging results (Schmoch et al. 2012). In any case, it will be important to expand the analysis to other countries to achieve suitable comparisons. However, this would involve conducting a recall-precision analysis for each country, as the selection criteria, especially the regional ones, may have different effects in different countries.

One remaining limitation is our reliance on the quality of the bibliometric data. We are expecting some improvements with regard to this issue, since rising recall rates from 2000 onwards suggest an improved coverage of the relevant authors in the SCOPUS database over time.

In conclusion, our new approach can make a significant contribution to identifying academic patenting in Germany and harbors potential for future analyses. Although we have shown that this approach could still be improved, we were able to provide a new and reliable way of constructing a database, which enables promising future research, not only in Germany but also in other countries.

References

Baeza-Yates, R./Ribeiro-Neto, B. (2011): *Modern Information Retrieval - The concepts and technology behind*, Second Edition. Pearson Education Limited.

Crespi, G./D'Este, P./Fontana, R./Geuna, A. (2011): The impact of academic patenting on university research and its transfer, *Research Policy*, 40, 55-68.

Czarnitzki, D./Glänzel, W./Hussinger, K. (2007): Patent and publication activities of German professors: An empirical assessment of their co-activity, *Research Evaluation*, 16, 311-319.

Czarnitzki, D./Hussinger, K./Schneider, C. (2011): Commercializing academic research: the quality of faculty patenting, *Industrial and Corporate Change*, 20, 1403-1437.

Fontana, R./Geuna, A./Matt, M. (2006): Factors affecting university-industry R and D projects: The importance of searching, screening and signalling, *Research Policy*, 35, 309-323.

Frietsch, R./Schmoch, U./Neuhäusler, P./Rothengatter, O. (2010a): *Patent Applications – Structures, Trends and Recent Developments* (=Studien zum deutschen Innovationssystem 10-2010). Berlin: Expertenkommission Forschung und Innovation (EFI).

Frietsch, R./Schmoch, U./van Looy, B./Walsh, J.P./Devroede, R./Du Plessis, M./Jung, T./ Meng, Y./Neuhäusler, P./Peeters, B./Schubert, T. (2010b): *The Value and Indicator Function of Patents* (=Studien zum deutschen Innovationssystem, 15-2010). Berlin: Expertenkommision Forschung und Innovation (EFI).

Geuna, A./Nesta, L.J.J. (2006): University patenting and its effects on academic research: The emerging European evidence, *Research Policy*, 35, 790-807.

Geuna, A./Rossi, F. (2011): Changes to university IPR regulations in Europe and the impact on academic patenting, *Research Policy*, 40, 1068-1076.

Hinze, T./Schmoch, U. (2004): Opening the black box: analytical approaches and their impact on the outcome of statistical patent analyses. In: Moed, H.F./Glänzel, W./Schmoch, U. (eds.): *Handbook of Quantitative Science and Technology Research, The Use of Publication and Patent Statistics in Studies on S&T Systems*. Dordrecht: Kluwer Academic Publishers, 215-235.

Kenney, M./Patton, D. (2009): Reconsidering the Bayh-Dole Act and the Current University Invention Ownership Model, *Research Policy*, 38, 1407-1422.

Kim, J./Lee, S.L./Marschke, G. (2005): The influence of university research on industrial innovation (=NBER Working Papers No. 11447). National Bureau of Economic Research, Inc.

Lissoni, F./Llerena, P./McKelvey, M./Sanditov, B. (2008): Academic patenting in Europe: New evidence from the KEINS database, *Research Evaluation*, 17, 87-102.

Lissoni, F./Lotz, P./Schovsbo, J./Treccani, A. (2009): Academic patenting and the professor's privilege: Evidence on Denmark from the KEINS database, *Science and Public Policy*, 36, 595-607.

Meyer, M. (2003): Academic patents as an indicator of useful research? A new approach to measure academic inventiveness, *Research Evaluation*, 12, 17-27.

Mowery, D.C./Sampat, B.N. (2004): The Bayh-Dole Act of 1980 and university-industry technology transfer: A model for other OECD governments?, *Journal of Technology Transfer*, 30, 115-127.

Raffo, J./Lhuillery, S. (2009): How to play the "Names Game": Patent retrieval comparing different heuristics, *Research Policy*, 38, 1617-1627.

Ranga, L.M./Debackere, K./Von Tunzelmann, N. (2003): Entrepreneurial universities and the dynamics of academic knowledge production: A case study of basic vs. applied research in Belgium, *Scientometrics*, 58, 301-320.

Schmoch, U. (2004): *The technological output of scientific institutions: Handbook of quantitative science and technology research.* Dordrecht: Kluwer Academic Publishers, 717-731.

Schmoch, U. (2007): *Patentanmeldungen aus deutschen Hochschulen* (=Studien zum deutschen Innovationssystem 10-2007), Studien zum deutschen Innovationssystem, Karlsruhe.

Schmoch, U. (2008): *Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectial Property Office (WIPO*), Karlsruhe: Fraunhofer ISI.

Schmoch, U./Dornbusch, F./Mallig, N./Michels, C./Schulze, N./Bethke, N. (2012): *Vollständige Erfassung von Patentanmeldungen aus Universitäten. Bericht an das Bundesministerium für Bildung und Forschung (BMBF)*. Revidierte Fassung., Karlsruhe: Fraunhofer ISI.

Thoma, G./Torrisi, S. (2007): Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases (=CESPRI Working paper No. 211). Centre for Research on Innovation and Internationalisation, Universita' Bocconi.

Thursby, J./Fuller, A.W./Thursby, M. (2009): US faculty patenting: Inside and outside the university, *Research Policy*, 38, 14-25.

Traijtenberg, M./Shiff, G./Melamed, R. (2006): The "names game": Harnessing Inventors' Patent Data for Economic Research (=NBER Working Papers No. 11447). National Bureau of Economic Research, Inc.

Von Ledebur, S. (2009): University-owned patents in West and East Germany and the abolition of the professors' privilege (=Working Papers on Innovation and Space). Marburg: Philipp Universität Marburg.

Von Proff, S. (2011): Patent and publication productivity of German professors - a life cycle view, *International journal of technology transfer and commercialisation*, 10, 392-410.

Youtie, J./Shapira, P. (2008): Building an innovation hub: A case study of the transformation of university roles in regional technological and economic development, *Research Policy*, 37, 1188-1204.