EPO ARP Project

From patents to trademarks: towards a concordance map

Project leader: Prof. Carolina Castaldi

Project team: dr. Milad Abbasiharofteh and dr. Sergio Petralia

Timeline: May 2020-May 2022

Final report

DISCLAIMER

This report was prepared as final report for the European Patent Office: please contact the project leader before citing and/or quoting, to make sure you refer to the latest published version of the research findings.

Title	From patents to trademarks: towards a concordance map
Research theme	**Primary theme:** *7. Advanced use of PATSTAT, patent searching, and analytics (e.g. classification, potential of IP linked open data)* **Secondary themes:** *1.Measuring the impact of patents on innovation* *6. Patents and climate change mitigation technologies*
Summary	An emerging research strand is revealing the complementarity between patents and trademarks. Patents play a role in protecting the inventive output of R&D, while trademarks can flag actual innovations to the market. Trademarks have the potential to inform us about the commercialization of patents, but patent and trademark data are hard to link. The project has two main objectives. First, it aims at developing and validating a novel concordance map linking patents to trademarks through linking the underlying classifications. This concordance will be shared in an open data setting to allow researchers to exploit it. Second, it plans to use the concordance to provide answers to research questions relevant to the study of cleantech patents.

Project summary

Motivation:

An important mechanism through which patents generate economic value is the way in which they lead to the the development of actual new goods and services that firms can sell in the markets and users can adopt. In this downstream phase of the innovation process companies often rely on trademarks to signal the introduction of their innovations. As such, combining patent and trademark data can allow a richer understanding of how patent technologies reach the market and generate economic value.

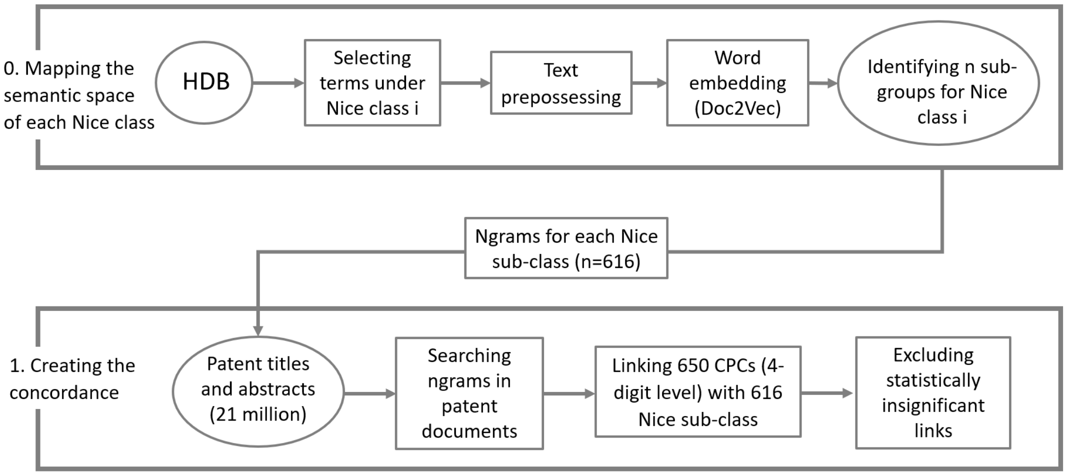
Unfortunately, linking these two sources of data has only been done on ad-hoc basis so far, frustrating efforts of researchers and policymakers to leverage both innovation data. This project aimed at filling this gap by systematically mapping relations between patented technologies and trademarked goods and services.

Key aim: to link patent to trademark data by mapping patent classes to trademark classes (Nice codes and the keywords in the detailed goods and services descriptors).

The main **scientific objective** of this project was threefold:
(1) to develop a concordance map between patent and trademark classes,
(2) to validate it extensively using complementary data sources and alternative techniques
(3) to illustrate its use for cleantech patents.

Key analytical steps and results

0.**Preliminary step**: given the coarse nature of trademark classification, a preliminary step involved identifying subclasses within the 45 broad Nice classes => this resulted in a list of **616 subclasses**, obtained from semantic analysis of harmonized EUIPO keywords (HDB) in Goods and Service (GS) descriptions.


0. Mapping the semantic space of each Nice class
HDB
Selecting terms under Nice class i
Text prepossessing
Word embedding (Doc2Vec)
Identifying n sub-groups for Nice class i
Ngrams for each Nice sub-class (n=616)
1. Creating the concordance
Patent titles and abstracts (21 million)
Searching ngrams in patent documents
Linking 650 CPCs (4-digit level) with 616 Nice sub-class
Excluding statistically insignificant links

1. **Concordance**: we were then able to link 662 CPC codes to the 616 subclasses. This concordance will be shared in an online data platform and described in a methodological paper (targeted to Nature, Scientific Data).

2. **Validations**: we checked our concordance approach in three different validation sets:

a. a firm-level dataset, based on the IPR bundles of EUIPO (2020);
b. a dataset on green goods and services, using the EUIPO green trademarks of EUIPO (2021);
c. a product-level dataset, linking to IProduct database (former ARP by Gaetan de Rassenfosse)
The last validation is leading to a collaboration on linking patents, products, and trademarks (with de Rassenfosse's team).

3. **Illustrations:** the project resulted in two lines of research related to cleantech patents and their link to green trademarks at the regional level:

Castaldi, C., Abbasiharofteh, M. and Petralia, S.G., *Greening EU regions: patterns of development between technological opportunities and product applications*, presented at the GEOINNO22 conference, July 2022.
Castaldi, C., Abbasiharofteh, M. and Petralia, S.G., *Greening at the periphery: an analysis of geography, technology and markets across EU regions,* presented at the Global Conference in Economic Geography, June 2022.

Table of Contents

1. Motivation and project objectives

Patents are an incredible source of data for economic and management research on innovation. Economists and management scholars have exploited patents to capture several aspects of inventive activity, from the measurement of firm capabilities to develop innovative technologies to the measurement of the very sources of economic growth. Counts of patents are linked to indicators of economic performance at different levels of analysis under the assumption of a link between technology generation and creation of wealth. Yet, any analysis assuming a direct or linear relation between technology and economic value leaves open the question of how exactly this economic value is created. This question is highly relevant to EPO as well since it relates to its core mission of raising awareness about the importance of patents (EPO-EUIPO, 2016). An important mechanism through which patents generate economic value is the way in which patents enable the development of actual new products and services that firms can sell in the markets and users can adopt. Here lies the fundamental difference between invention and innovation (Greenhalgh and Rogers, 2010).

An emerging research strand is exploiting trademarks to capture actual innovations being introduced in the market. A main theme is the complementarity between patents and trademarks, in IPR strategies, but also in generating economic value for companies (Castaldi, 2019).

1.1 On the complementarity between patent and trademarks

The rich literature on IPR strategies has generated theoretical insights and empirical evidence on how innovative firms appropriate the returns from their R&D and other innovation efforts (Hall et al., 2014). These studies show how patents play a role in protecting the inventive output of R&D, while trademarks are used in a later stage, when actual innovations are being brought to the market. Especially in complex industries, multiple patents are typically needed to allow the market introduction of new products. An essential part of the success of new products is an effective branding strategy, linked to new or existing trademarks (Castaldi, 2019). In this sense, investments in R&D and branding are complementary in generating value for innovating firms. An additional complementarity comes from the fact that while patents tend to act as information signals to competitors and investors, trademarks do so for consumers and users, next to competitors (Castaldi et al., 2020). In line with their primary economic function, trademarks are symbols that lower consumers' search costs by differentiating goods and services in markets (WIPO, 2013).

Empirical evidence on the complementarity between patents and trademarks comes from several recent firm-level studies (Schautschick and Greenhalgh, 2015). Patent and trademark stocks are found complementary in increasing market value (Sandner and Block, 2009), in securing venture capital funding for start-ups (Zhou et al., 2007) and in generating sales growth for R&D investors (Castaldi and Dosso, 2018).

A last insight is that combining information on patents and trademarks provides a stronger measure of innovation (Flikkema et al., 2019). While not all patents lead to actual innovations, it is also the case that not all trademarks proxy innovation. Based on the studies so far, we expect that combining patent and trademark data will strengthen our understanding and measurement of innovation processes.

1.2 Research gap

The studies reviewed above typically link patents and trademarks at the firm level, but they do not attempt to reconstruct the deeper connections between specific patents and specific trademarks.

Mapping patents to trademarks is in its infancy and deserves targeted efforts. By focusing on mapping patent and trademark classification systems, this project aimed at contributing novel ways of capturing the qualities of technological and market specializations and how they map.

There are **several research questions that researchers and policymakers will be able to tackle with such a mapping.** Let us illustrate three key questions, at three different levels of analysis.

> **RQ1. Technology level: how can we track the diffusion of new patents in actual products and services, as captured by trademarks?**

Scientific and societal relevance: Mapping patents into trademarks provides a novel take on studying diffusion of new technology. While prior research has mostly mapped patents into broad industries of use, a mapping to trademarks has the potential to provide a richer narrative of the market applications of new technology. This endeavour is also expected to increase societal awareness of actual applications of patents. Concerns around patenting for sake of patenting can be addressed with an informed analysis of which market offerings can be mapped to specific patents.

RQ2. Firm level: how can trademarks reveal the commercialization capabilities of patenting firms?

Scientific and managerial relevance: The project contributes to innovation management research aiming to investigate the missing link between investment in technology and generation of economic value. This has also strategic relevance for managers: they could use the concordance as a map of commercialization opportunities and analyse their position of the one of competitors in this map.

RQ3: Regional/National level: how can trademarks capture place-based advantages in patent commercialization?

Scientific and policy relevance: Studies on regional specialization in economic geography are mostly focused on patents. A capability approach relying on trademarks can reveal place-based advantages in market capabilities (Castaldi and Mendonça, 2022). While several innovation scoreboards, like the European Innovation Union index, do include the amount of trademark applications as an indicator of innovation output, no information on the qualitative properties of these trademark applications is included.

The above questions become even more relevant if we ask them for specific emerging technologies (Negro et al., 2012). In this project we considered the case of cleantech as a meaningful application. Cleantech refers to all technologies or applications for mitigation or adaptation against climate change. The CPC Y02 classification captures these technologies and prompts novel empirical research.

In addition to its policy relevance, cleantech is an interesting application of our concordance as the link between invention and successful market application is often theorized to be more problematic for cleantech, due to the double externality problem. Furthermore, indicators of eco-innovation are diverse but imperfect. Linking cleantech patents to trademarks would allow us to identify ‘eco-trademarks’ linked to eco-patents at different levels. At the firm level, climate change is also acting as a major driver of innovation. Companies are motivated by regulation and normative pressures to invest in greener alternatives. At the regional level, cleantech can offer opportunities for regions to develop new place-based advantages. More specifically it allows us to identify those firms and regions that are more successful in bringing cleantech to the market

1.3 State of the art of research on mapping and concordance approaches using patent or trademark data

For patent data, a number of concordances exists linking patent classes to industrial classes. Good examples are the MERIT concordance table (Verspagen et al., 1994) between IPC and ISIC rev2 or the one-to-one (thus not fractional) concordance provided in the latest PATSTAT editions which map 4-digit IPC codes to 84 NACE2 codes.

These initiatives have been motivated by the wish to uncover the 'industry of use' of different technologies and date back to the early 1980s (Scherer, 1982). Economic analysis would then look for economic effects, in terms of productivity growth or other, in the linked sectors.

More recently, trademarks have started to emerge as a valuable new source of data for economic analysis of innovation (see the reviews in Schautschick and Greenhalgh, 2016 and Castaldi et al., 2020). By now, commercial databases like Bureau van Dijk's ORBIS include both patent and trademark portfolios of firms. The top R&D investors scoreboard compiled by the European Commission's Joint Research Center also analyses both patents and trademarks (Dernis et al., 2015).

Zolas et al. (2017) have proposed a concordance between trademark and industrial classes. Since both patents and trademarks can linked to industry codes, one can in principle link patents to trademarks through their industry equivalent (like Grazzi et al., 2019 do). Yet, this is a rather indirect way of linking patents to trademarks. A more direct way is to construct patent-trademark pairs, but this approach is very labour intensive and also focuses on a one-to-one mapping that is hardly the norm in several high-tech complex industries. An alternative approach is constructing patent-product pairs through information on virtual patent marking (Rassenfosse, 2018). This is a very promising technique, with the limitation of being focused on tangible products, hence not covering services.

1.4 The contribution of this project

In this project, we focused on linking patent to trademark data by mapping patent classes (IPC codes) to trademark classes (Nice codes and the keywords in the detailed goods and services descriptors). By focusing on classification systems, we aimed at capturing the qualities of technological and market specialization patterns. In this respect, the envisioned concordance map would allow us to tackle the three types of research questions outlined above.

Given the exploratory nature of this project, two complementary contributions are: (1) that we validated the concordance map by using multiple approaches and multiple complementary data and (2) that we applied the concordance map to a specific set of technologies (i.e., Cleantech) to illustrate its relevance for economic and innovation research.

The main scientific objective of this project was threefold: (1) to develop a concordance map between patent and trademark classes, (2) to validate it extensively and (2) to illustrate its use for Cleantech.

1.5 Key scientific objectives and policy relevance

These scientific objectives are challenging, given specific properties of trademark classifications that make mapping to patents a non-trivial task. As we will detail later in the methods, three key challenges are: 1) the non-hierarchical structure of the Nice classification, 2) the multiple mapping and 3) the common use of user-generated keywords in trademark descriptions.

The scientific output of this project bears relevance for several stakeholders.

Societal relevance: by mapping patents into trademarks, our project contributes to the public perception of patents. We show how to map the application of patented knowledge into actual solutions commercialized to the benefit of users in society (either individuals, households or organizations). This is in line with one of the key objectives of both EPO and EUIPO (see the joint report EU-EUIPO, 2016, including a whole chapter on cleantech, which is also our key application in this study).

Managerial relevance: by mapping patents into trademarks, we provide a tool for companies to visualize their performance in terms of realized commercialization opportunities. Especially, SMEs and starting firms that might be mostly active on the technology development side of the innovation processes, are likely to benefit from benchmarking the opportunities for commercialization of their technology portfolios.

Policy relevance: by mapping patents into trademarks, regional/national market capabilities can be assessed next to technological capabilities. Capability development is increasingly seen as a promising policy rationale, also in the case of green development (Rodrik, 2014). Our application to cleantech provides novel insights to inform the current policy debates on sustainability transitions.

2. Research approach

2.1 Properties of the EUIPO trademark data

We investigated the trademark data provided by the European Union Intellectual Property Office (EUIPO) including trademarks filed between 1996 and 2020. The starting date corresponds to the time that EUIPO started to accept and examine trademark applications (EUIPO, 2021). First, we extracted the raw data (available from the Open Data resources of EUIPO) from a set of chronologically saved hierarchical xml files into a tabular format. The result includes 1,913,468 unique trademark applications. We only included applications that reached registrations. Since there is a lag between filing and registration (see Figure 2), we have to deal with truncation issues for the last year.

The applications were filed by 735,297 applicants. Applicants should specify one or several Nice codes for their product or service to provide information on the purposes of the filing. The Nice Classification is a system of classifying goods and services that range between 1 and 45. Classes from 1 to 34 represent coarse categories for goods, and classes from 35 to 45 include service categories. Figure 1 shows that the number of trademark filings distinguished by the type of trademark (i.e., products or services). This figure shows that trademark filings in both goods and services categories increased linearly on a log-scale, hence grew exponentially, and they followed a similar growth pattern, whereas the number of applications for goods was about 1.6 times higher than the ones for services. It is important to note that the drop in the number of filings in 2020 relates to the truncation problem discussed above.

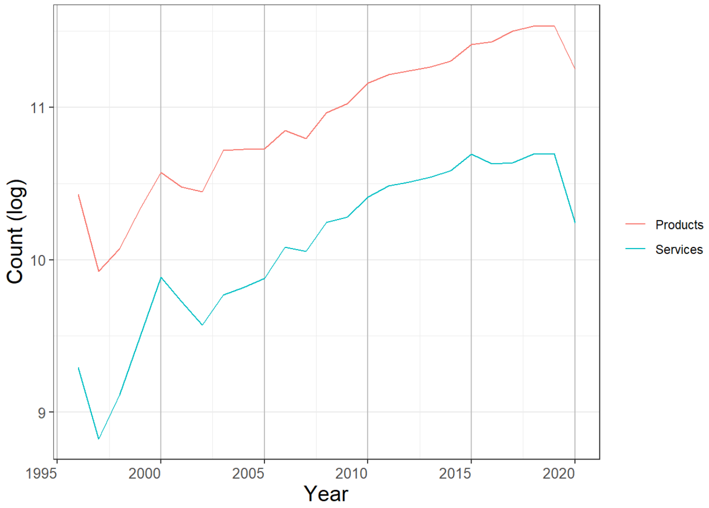


Figure 1. Number of trademark filings (log-scale) over time.

Figure 2 shows the time difference between trademark filing and registration. Interestingly, the time difference has significantly decreased since the mid-1996 from more than two years to less than a year. One change at EUIPO that can explain this pattern is the introduction and integration of the Harmonised Database (HDB) in the application process, in which applicants can select predefined terms under each Nice class and benefit from a fast-track application option (we will further discuss the HDB in the next sub-section).

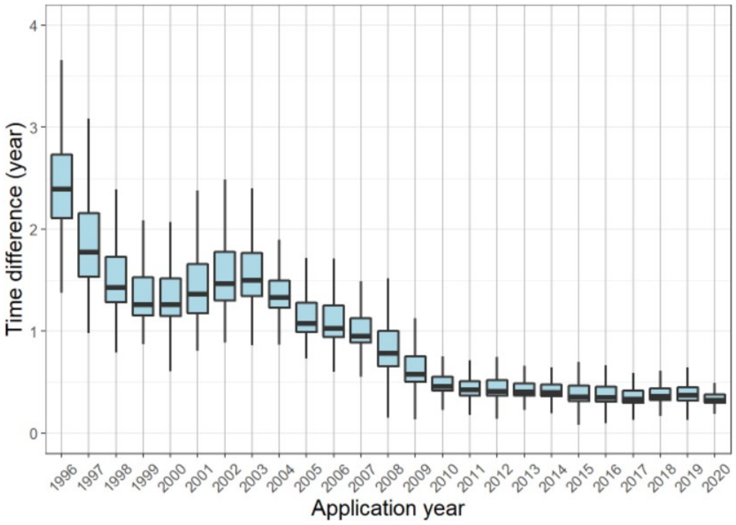


Figure 2. Time difference between trademark filing and registration.

As noted earlier, applicants have to identify one or several Nice codes that capture the markets where they will actually use the trademark. Figure 3 shows the median and mean of the number of Nice classes per application between 1996 and 2020. The shift in the central tendency of the number of Nice classes is driven by a change in the application process. Applicants were allowed to indicate up to three Nice classes (before 2005, two Nice classes) without any extra application costs.

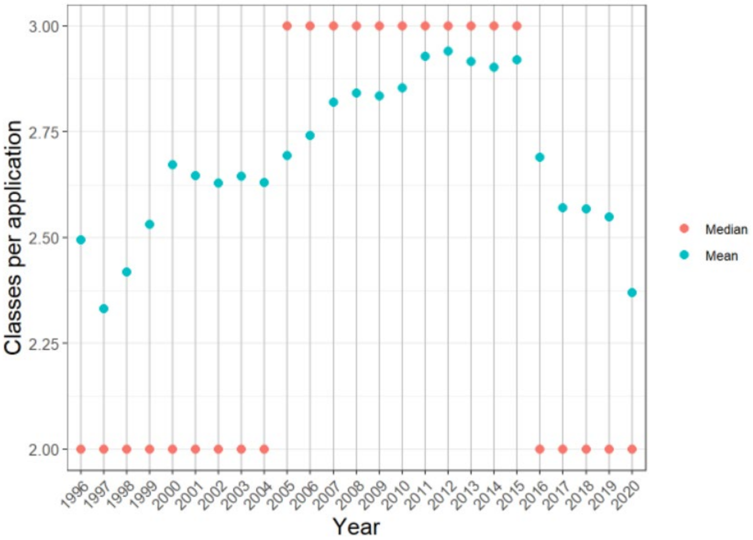


Figure 3. Median and mean of the number of Nice classes per application.

While the technological classifications (CPC and IPC) for patent data are hierarchical ones, with a high degree of detail, the Nice classification entails 45 non-hierarchical coarse categories. Figure 4 illustrates the distribution across the 45 Nice classes. Most trademarks in the goods category are filed in class 9 that includes a wide range of products in electronic components and devices. Similarly, most filings in the service category are related to Nice classes 35, 41, and 42 covering various services ranging from administrative to financial and to design consulting services. Conversely, the number of trademarks related to firearms, ammunition, and weapons (the Nice class: 13) is relatively low.

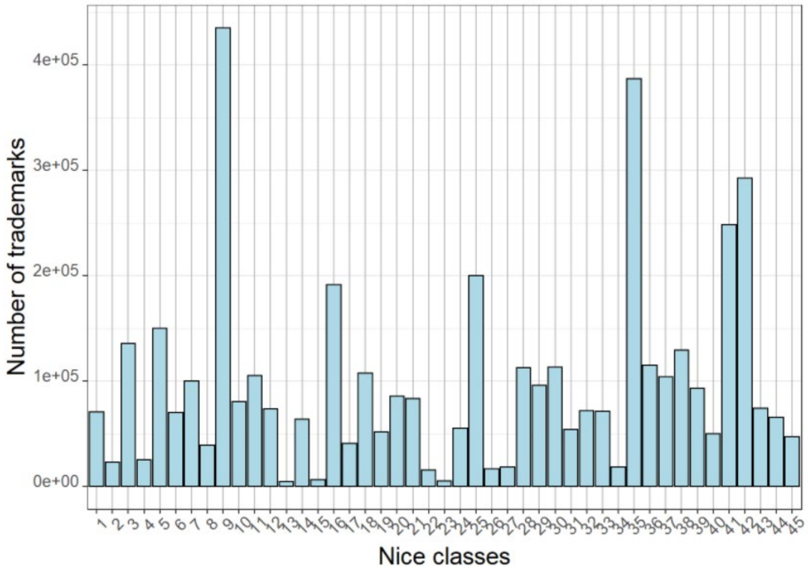


Figure 4. Number of trademarks across 45 Nice classes.

Beside specifying Nice classes, trademark applicants can add multiple goods and services descriptors to describe their good or service more clearly. For instance, while the Nice class 9 include a wide range of products in electronics, an applicant may add the descriptor ‘Artificial intelligence software’ to describe the product in more details. Figure 5 shows that most trademarks have between two and eight descriptors (except the Nice classes 23 and 33). The figure demonstrates a relatively consistent pattern across 45 classes.

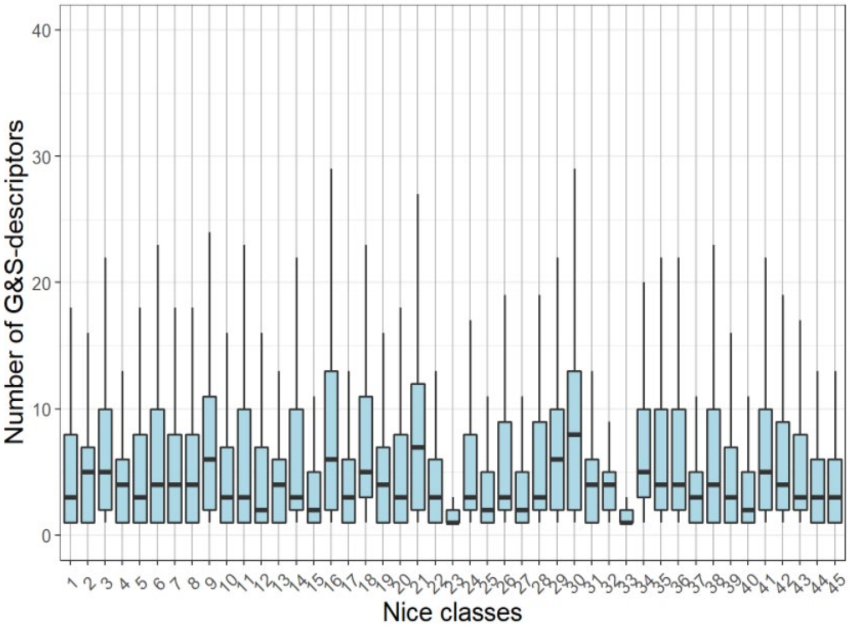


Figure 5. Number of goods and services descriptors across 45 Nice classes.

Also, the application process includes a step in which applicants provide information about themselves including, among others, their (company) names and addresses. This information can be easily linked to trademarks by merging applicant and trademark IDs. By utilising information on the geographic location of the applicants, we provided descriptive statistics on the geographic distribution of applicants. Most applicants (70%) are located in one of European countries. Barcelona, Madrid, Milan, Paris, and Berlin are cities with the highest number of trademark applicants. Shenzhen followed by New York and Guangzhou has the highest share of applicants outside Europe. Figures 6 and 7 show the geographic distribution of applicants and the kernel density estimation of applicant numbers across European NUTS2 regions. These figures show that a positively skewed geographic distribution of applicants with long right tails that resemble a log-normal distribution characterized by a low mean and high variance. This implies many regions have a few applicants, whereas few regions have more than 5000 companies or individuals that filed at least one trademark application. Regions in Germany, Spain, Northern Italy, Southern France seem to be home to businesses active in filing trademarks.

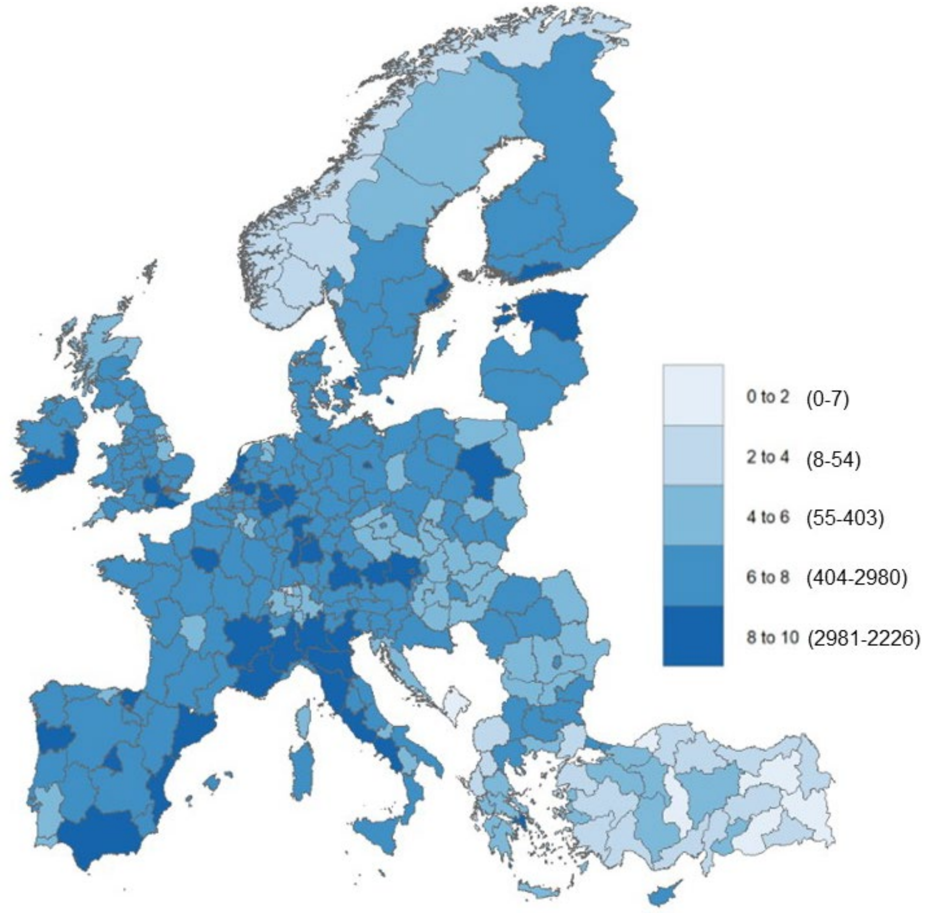


Figure 6. Number of applicants (log-transformed) across European NUTS2 regions. Number of applicants on a linear scale provided in parentheses.

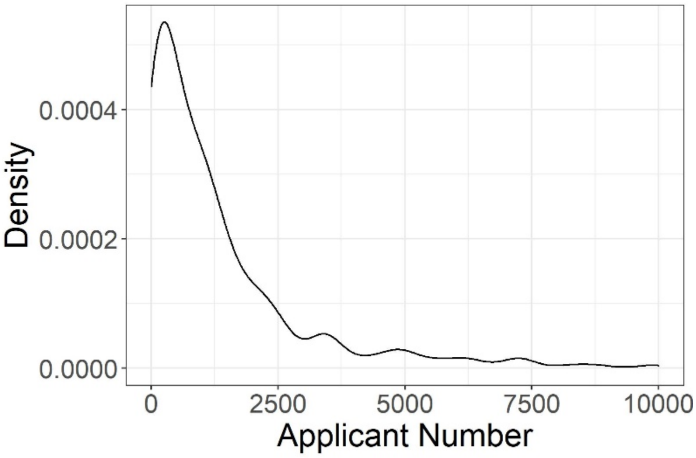
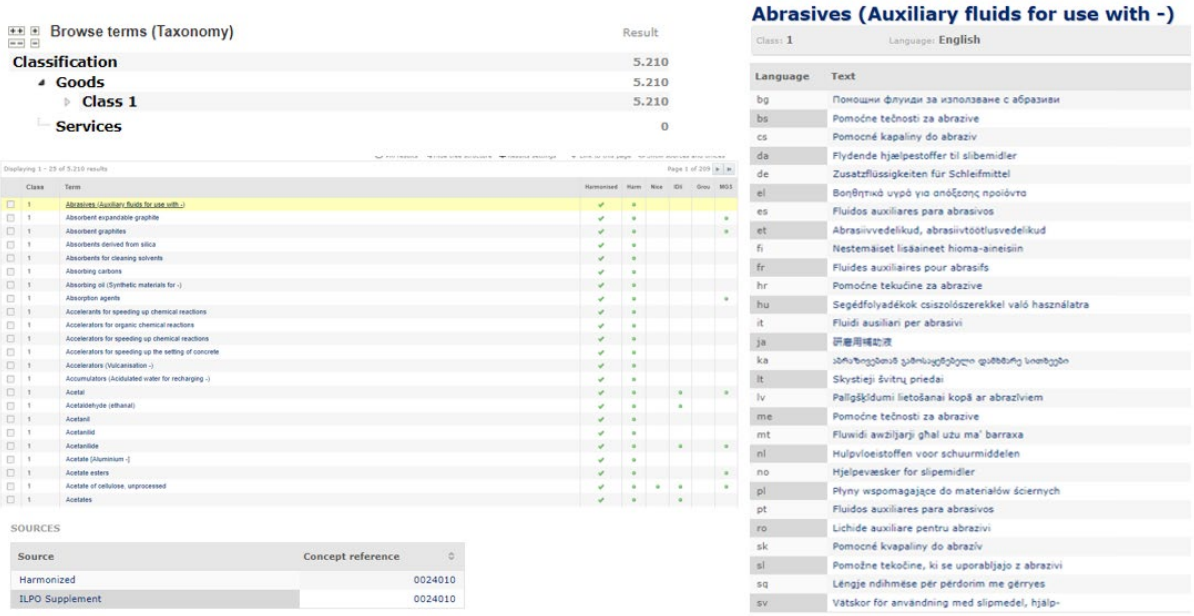


Figure 7. The kernel density estimation of applicant numbers across European NUTS2 regions.

2.2 The EUIPO Harmonised dataset

EUIPO developed and maintains the Harmonised Database (HDB) to classify goods and services protected by trademark filings[1]. The HDB includes fine-grained descriptions (>70,000 terms) of goods and services that are acknowledged by all national intellectual property offices in the European Union. HDB terms are defined based on common descriptors acknowledged by trademark application examiners to facilitate the application process. The terms reflect market realities in general and new terms reveal the current business trends. Based on the actual state of the market, new terms (obsolete terms) can be added (removed) from the database. Also, terms are based on the Nice classification system, publicly available, and translated to all EU languages (except Irish). HDB terms are short and precise (the average number of words per term: 12.18) and focus on the key features of products and services. Figure 8 demonstrates an example of multiple HDB terms listed on EUIPO's website[2]. As shown, terms are organized under each Nice class, and each has a unique Harmonized identifier (HDB ID). For instance, the highlighted term 'Abrasives (Auxiliary fluids for use with -)' corresponds to a HDB ID (0024010) and linked to its official translations.



Browse terms (Taxonomy)

	Result
Classification	5.210
Goods	5.210
Class 1	5.210
Services	0

Displaying 1 - 25 of 5.210 results

Page 1 of 209

	Class	Term	Harmonised	Harm	Nice	IDI	Grou	MGS
□	1	Abrasives (Auxiliary fluids for use with -)	✓	●				
□	1	Absorbent expandable graphite	✓	●				●
□	1	Absorbent graphites	✓	●				●
□	1	Absorbents derived from silica	✓	●				
□	1	Absorbents for cleaning solvents	✓	●				
□	1	Absorbing carbons	✓	●				
□	1	Absorbing oil (Synthetic materials for -)	✓	●				
□	1	Absorption agents	✓	●				●
□	1	Accelerants for speeding up chemical reactions	✓	●				
□	1	Accelerators for organic chemical reactions	✓	●				
□	1	Accelerators for speeding up chemical reactions	✓	●				
□	1	Accelerators for speeding up the setting of concrete	✓	●				
□	1	Accelerators (Vulcanisation -)	✓	●				
□	1	Accumulators (Acidulated water for recharging -)	✓	●				
□	1	Acetal	✓	●		●		●
□	1	Acetaldehyde (ethanal)	✓	●		●		
□	1	Acetanil	✓	●				
□	1	Acetanilid	✓	●				
□	1	Acetanilide	✓	●		●		●
□	1	Acetate (Aluminium -]	✓	●				
□	1	Acetate esters	✓	●				●
□	1	Acetate of cellulose, unprocessed	✓	●	●	●		●
□	1	Acetates	✓	●		●		

SOURCES

Source	Concept reference
Harmonized	0024010
ILPO Supplement	0024010

Abrasives (Auxiliary fluids for use with -)

Class: 1 Language: English

Language	Text
bg	Помощни флуиди за използване с абразиви
bs	Pomoćne tečnosti za abrazive
cs	Pomocné kapaliny do abraziv
da	Flydende hjælpestoffer til slibemidler
de	Zusatzflüssigkeiten für Schleifmittel
el	Βοηθητικά υγρά για απόξεσης προϊόντα
es	Fluidos auxiliares para abrasivos
et	Abrasiivvedelikud, abrasiivtöötlusvedelikud
fi	Nestemäiset lisäaineet hioma-aineisiin
fr	Fluides auxiliaires pour abrasifs
hr	Pomoćne tekućine za abrazive
hu	Segédfolyadékok csiszolószerekkel való használatra
it	Fluidi ausiliari per abrasivi
ja	研磨用補助液
ka	აბრაზივებთან გამოსაყენებელი დამხმარე სითხეები
lt	Skystieji švitrų priedai
lv	Palīgšķidumi lietošanai kopā ar abrazīviem
me	Pomoćne tečnosti za abrazive
mt	Fluwidi awżiljarji għal użu ma' barraxa
nl	Hulpvloeistoffen voor schuurmiddelen
no	Hjelpevæsker for slipemidler
pl	Płyny wspomagające do materiałów ściernych
pt	Fluidos auxiliares para abrasivos
ro	Lichide auxiliare pentru abrazivi
sk	Pomocné kvapaliny do abrazív
sl	Pomožne tekočine, ki se uporabljajo z abrazivi
sq	Lëngje ndihmëse për përdorim me gërryes
sv	Vätskor för användning med slipmedel, hjälp-

Figure 8. HDB terms listed on EUIPO's website.
Source: http://euipo.europa.eu/ec2/?lang=en

[1] See: https://euipo.europa.eu/ohimportal/en/harmonised-database
[2] http://euipo.europa.eu/ec2/?lang=en

2.3 Building the concordance

No concordance between patent and trademark classes exists yet, but we can take stock of the state-of-the-art methods by Zolas et al. (2017) on linking trademark and industrial classes. Moreover, thanks to recent progress in NLP (Natural Language Processing) algorithms and software, methodologies for matching textual data have advanced substantially. Accordingly, we exploit ideas from other mapping efforts. As explained by Zolas et al. (2017), there are several challenges when one aims at building concordances that involves trademark classification.

A first challenge is that the Nice classification systems lacks a hierarchical structure similar to the one of the IPC/CPC patent classification systems. Using the 45, very broad, Nice classes (1-34 for goods, 35-45 for services) would deliver a concordance where many patent classes link to most trademark classes, without much use for the envisioned applications.

Yet, trademark records also include short text descriptions compiled by trademark applicants, commonly referred to as goods and service descriptors. This text can be used to better characterize specific markets within each broad Nice class. One caveat is that at most offices trademark applicants can add user-generated goods and services descriptors, not following the standardized goods and service indicators that come with the official Nice classification. Also, applicants may strategically or even randomly add extra Nice classes and goods and services descriptors that introduce bias in developing the patent-trademark concordance. Our solution to this problem is to utilise HDB terms instead of goods and services descriptors to decrease random noise introduced by the myopic behaviour of applicants and self-defined goods and services descriptors. These terms can be used as standardized keywords in text analysis. The fined-grained information and hierarchical structure of the HDB database are helpful because each term can be analysed in the context of a given Nice class. Figure 9 provides a schematic overview of each step we took to develop our concordance map. The remainder of this section discusses each step as well as how we tackled the above-mentioned problems in the trademark data.

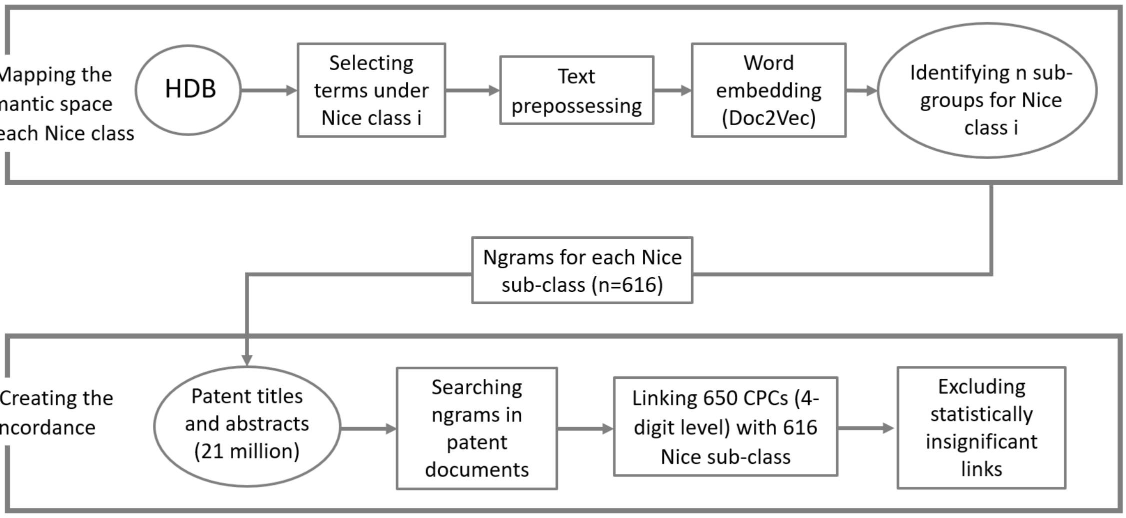

0. Mapping the semantic space of each Nice class
HDB
Selecting terms under Nice class i
Text prepossessing
Word embedding (Doc2Vec)
Identifying n sub-groups for Nice class i
Ngrams for each Nice sub-class (n=616)
1. Creating the concordance
Patent titles and abstracts (21 million)
Searching ngrams in patent documents
Linking 650 CPCs (4-digit level) with 616 Nice sub-class
Excluding statistically insignificant links

Figure 9. Scheme of the research steps from semantic analysis of HDB terms in Nice classes to creating the concordance

Mapping the semantic space of each Nice class (the preliminary step)

Analysing textual data and using text-based indices have become popular thanks to enhanced computational capacity, high-performance machine learning and natural language processing (NLP) techniques (Abbasiharofteh et al., 2021; Kinne and Lenz, 2021; Krüger et al., 2020; Ozgun and Broekel, 2021). Scholars extensively use text-based measures to analyse patent similarity (Abbas et al., 2014; Brachtendorf et al., 2020). Also, scholars use vector space models to map each patent document in the space and approximate their similarity with other patent documents based on their position in the numeric space (Magerman et al., 2010). Other studies benefited from text analysis to map general purpose technologies with patent data (Petralia, 2020) and Natural Language Processing (NLP) techniques to identify patents that are more likely to be linked to prestigious awards (Arts et al., 2021). Although analysing trademark documents has attracted far less attention compared to patent data, there are some recent examples of studies leveraging text analysis of trademarks. For instance, von Graevenitz et al. (2022) used goods and services descriptions of trademarks to identify radically new products and services and investigate the spatial diffusion of these innovations.

We build on the above-mentioned empirical works and use NLP techniques to map a semantic space for each Nice class. This preliminary phase consists of four steps. First, we filtered out all terms listed under each Nice class. Second, we pre-processed the text by 1) converting all terms to lowercase, 2) omitting terms that include negative words (e.g., not), 3) removing English stop-words (e.g., the), 4) removing general words that do not carry a specific meaning and can be misleading (e.g., contain), 5) removing numbers and punctuations, and 6) removing the suffix of each term's word and bringing them to corresponding base words (also known as stemming). Since HDB terms are relatively short and specific, the remaining words clearly describe the purpose and application field of each term. Table 1 shows the number of terms (after text pre-processing) for each Nice class. The frequency of terms corresponds to the number of trademarks in each Nice class. For instance, the Nice classes 15 and 23 have less than 300 terms, whereas the Nice class 9 entails more than 5700, and the classes 41 and 42 include more than 2000 terms. The table also demonstrates the most frequent stems for each Nice class. Interestingly, the information provided by the most frequent stems aligns with Schmoch's (2003) work, specifying on technologies related to several Nice classes. Our approach, however, goes beyond the work of Schmoch and provides first insights into related technologies across all 45 Nice classes.

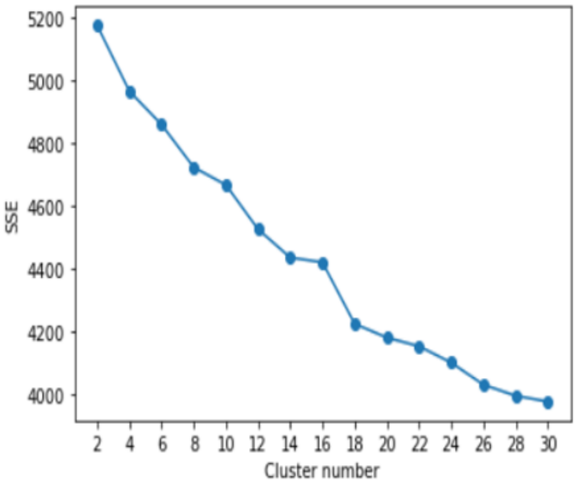
Nice	Number of terms	Top five keywords (stems)	Technologies*
1	3791	chemic, composit, manufactur, food, water	Chemistry
2	820	paint, natur, color, protect, metal	Chemistry
3	1309	cosmet, skin, lotion, cream, impregn	Chemistry
4	485	metal, briquett, oil, addit, concret	Chemistry
5	1967	medic, agent, pharmaceut, cream, care	Pharmaceuticals
6	1798	metal, rail, hook, plug, roof	Metals
7	3714	machin, cut, part, print, electr	Machines
8	852	hand, drill, saw, blade, bit	Machines
9	5761	electr, apparatus, applic, comput, download	Electronics
10	2145	medic, apparatus, mask, protect, devic	Medical technologies
11	2291	apparatus, electr, water, instal, sanitari	Electronic devices
12	1621	vehicl, land, adapt, air, bodi	Vehicles
13	387	firearm, cartridg, ammunit, weapon, pistol	Chemistry
14	535	metal, precious, jewelleri, jewelri, clock	
15	261	electron, apparatus, instrument, music, string	
16	1629	paper, write, decor, label, tissu	
17	1065	rubber, insul, pipe, build, articl	
18	404	case, umbrella, bag, travel, luggag	
19	964	metal, build, structur, glass, floor	
20	1520	furnitur, metal, door, box, plastic	
21	1470	glass, clean, cloth, bucket, hand	
22	415	rope, plastic, stuf, cover, wool	
23	116	yarn, thread, textil, fiber, twist	
24	604	fabric, textil, towel, cloth, manufactur	
25	963	pant, short, sock, tror, wear	
26	351	textil, cloth, hair, button, ornament	
27	122	floor, carpet, cover, tile, wall	
28	1460	set, toy, apparatus, game, play	
29	1198	chees, pickl, potato, preserv, sausag	
30	1722	food, flavour, confectioneri, flour, cake	
31	867	fresh, fruit, edibl, anim, plant	
32	188	drink, beverag, beer, water, alcohol	
33	145	alcohol, beverag, liqueur, drink, whiski	
34	183	pipe, tobacco, cigarett, electron, cigar	

35	1972	busi, advertis, manag, commerci, equip	
36	1682	financi, manag, administr, inform, brokerag	
37	2070	mainten, repair, build, instal, apparatus	
38	795	communic, data, transmiss, network, comput	
39	1214	transport, good, rental, passeng, travel	
40	946	fur, food, gas, mould, oil	
41	2510	educ, cours, organ, event, show	
42	2170	design, comput, develop, softwar, program	
43	352	hire, rental, bar, cater, drink	
44	877	inform, medic, field, advisori, breed	
45	464	inform, advisori, date, social, support	

Table 1. The counts of terms, top five keywords (stems), and related technologies across 45 Nice classes.
*Technologies related to Nice classes are identified by Schmoch (2003)

The third step is to cluster terms together which are semantically similar. By doing so, we create a meso-level Nice sub-classes between 45 Nice classes and 70,000 HDB terms. Nice sub-classes provide a detailed information about corresponding terms and become more comparable to patent technological classifications (e.g., 662 CPC technology codes). Neuhäusler et al. (2021) used a Levenshtein-based measure to match trademarks. This measure is based on the Levenshtein distance between two words or terms corresponding to the minimum number of characters edit to change one word into the other. This measure is normalized based on the length of the matched terms. Although this measure resonated well with patterns of data in fields as diverse as information theory, linguistics, and computer science, the application in the context of trademarks appears less clear. The measure cannot find association between words that come together in the context of one Nice code but have a low text similarity. For instance, in the Nice class 33 (drink and beverage) alcohol and whisky, and coffee and milk are semantically similar whereas their text similarity is low. To remedy, we opted for an alternative method that more efficiently captures the semantic similarity of pre-processed terms (hereafter terms or HDB terms). We mapped HDB terms using the Doc2Vec method which is a generalization of the word2vec method (Le and Mikolov, 2014). This method provides numeric vectors that represent terms and can be used to capture their relation in a high-dimensional vector space (e.g., [vector king]-[vector queen]~[vector man]) (Mikolov et al., 2013). Doc2Vec transforms and maps HDB terms to numeric vectors. The text corpus used in Doc2Vec is based upon all HDB terms under each Nice class. Doing this, we ensure that the semantic space is Nice-specific and we do not create false associations between terms in different contexts, e.g. cream and milk; cream and lotion; and NOT milk and lotion. Then, we used k-means clustering to clusters semantically similar HDB terms in the vector space. We used the common elbow plot to specify more efficiently the number of cluster (i.e.,

k) for each Nice class. This method provides 616 Nice sub-classes for 45 Nice classes. Of course, large Nice classes with a wide range of terms include more sub-classes compared to smaller ones including rather homogenous terms. For instance, the k-means algorithm found 8 Nice sub-classes for the Nice class 13 (387 terms) and 19 Nice sub-classes for the Nice class 19 (2145 terms). We created unique identifiers for each of 616 Nice sub-classes. Figure 10 gives an example of an elbow plot and the number of Nice sub-clusters for the Nice class 5. The elbow plot provides a value of within-clusters Sum of Squared errors (SSE) for each clustering with a certain number of clusters. The so-called elbow point corresponding to the optimal number of k is the value after which SSE values start decreasing in a linear manner. In the provided example, the value 16 seems to be the reasonable number of clusters for the Nice class 5. Schmoch (2003) identified pharmaceuticals as the dominant technology underlying the Nice class 5. While top keywords provided for each Nice sub-class align with Schmoch's claim, Nice sub-class keywords provide more information on the application fields within pharmaceuticals. By creating Nice sub-classes we tackled the problem of the non-hierarchical structure of the Nice classification.



Dental Alloy Metal Teeth Ceramic
Repellent Insect Impregnated Animal Camphor
Deodorant Hormonal Destroyed Agricultural Air
Gel Cream Skin Medical Care
Medical Oil Powder Babies Cream
Medical Reagent Chemical Diagnostic
Dress Bandage Medical Spray Adhesive
Dip Iodine Muscle Ophthalmic Paracetamol
Pharmaceutical Treat Capsule Agent Alginate
Medicinal Extract Herb Herbal Bark
Supplement Vitamin Animal Food
Medical Adaptation Dietetic Food Beverage
Diaper Babies Disposable Nappies Paper
Acaricides Antitussives Germicidal Larvae Pen
Medical Cream Agent Care Skin
Pharmaceutical Disease Prevent System Disorders
Tissue Surgical Implant Live Biological
Agent Drug Pharmaceutical Deliveries Form

Figure 10. The elbow plot and top keywords for each Nice sub-class.

Finally, we created bigrams for HDB terms that belong to each Nice sub-class. Ngrams are widely used in statistical natural language processing and data comparison. For instance, scholars used ngrams to map general purpose technologies in patent documents (Petralia, 2020) and identify breakthrough patents (Kelly et al., 2021). More specifically, we opted for using bigrams because (compared to unigrams) they provide more contextual information, and (compared to trigram) they provide a relatively higher number of matches. For instance, '*chemic seal grout construct build*' is a HDB term (before the text pre-processing: *Chemical sealing grout for use in the construction industry; Chemical sealing grout for use in the building industry*) belonging to the Nice class 1 (most frequent stems: chemic, composit, manufactur, food, water). As a result, bigrams representing this HDB terms are *chemic seal*,

seal grout, *grout construct*, and *construct build*. We utilised bigrams in the next phase to create the patent-trademark concordance to which we turn in the next sub-section.

Creating the concordance

In this sub-section, we report on the procedure to develop our patent-trademark concordance. While we argue that HDB terms present the application fields of new products and services being introduced in the market, patents capture inventive activities that represent technological change or a new technological solution. Scholars have used patent documents extensively to approximate innovative activities, knowledge exchange, and technological change (Abbasiharofteh et al., 2020; Abbasiharofteh and Broekel, 2020; Arts et al., 2021; Breschi and Lissoni, 2009; Castaldi et al., 2015; Fleming and Sorenson, 2001). Patent documents include information on, among others, the location of inventors, technologies, and the description of a given invention. We utilised this information to identify relations between HDB terms and technologies. We used PATSTAT Global (version: 2020) provided by the EPO that includes more than 100 million patent documents filed in more than 40 patent offices worldwide. We filtered out patents filed between 2000 and 2020 and excluded redundant patents that were filed in multiple patent offices. The latter was done to avoid searching through the same patent that are filed in different patent offices. This led to more than 21 million patents. It is important to note that we pre-process the text of patents' title and abstracts as explained in the previous sub-section. Then, we iteratively searched the bigrams of each HDB term in the title and abstract of patent documents. This search strategy aligns with the works of Bergeaud et al. (2017) and Brachtendorf et al. (2020) who rely on searching though abstracts rather than the full texts because it includes the most important terms associated with the invention and less legal terms.

We assume that there is a relation between a Nice sub-class and a CPC technology code if at least one of the bigrams of the given Nice sub-class is included in a patent. We sum the number of relations across all 407,792 pairs (616 Nice sub-class × 662 CPC codes). Next, we used a statistical method to create a null expectation to ignore the identified relations that are not statistically significant. Zolas et al. (2017) matched the generated keywords for trademarks and industries to create a trademark-industry concordance. They used a probabilistic approach by using Bayes Rules to take into consideration trademarks and industries have different propensity to be linked due to the frequency, and broad or narrow definition of trademarks and industries. Similar to Zolas et al., we used the z-score to identify statistically significant relations by accounting for the frequency of each Nice sub-class and CPC code as well as the count of their co-occurrence. Scholars used the z-score to identify novel combination of technologies and interdisciplinary inventions (Abbasiharofteh et al., 2020; Fontana et al., 2020; Mewes, 2019). The z-score is defined as follows:

$$Z_{i,j} = \frac{O_{i,j} - E_{i,j}}{\sigma_{i,j}} \tag{1}$$

where $O_{i,j}$ is the number of the co-occurrence of Nice sub-class i and CPC technology code j. $E_{i,j}$ is the statistical expectation of Nice sub-class i and CPC technology code j co-occurring randomly. Teece et al. (1994) argue that the co-occurrences of i and j is more likely to be random if the relative number of occurrence of two units is relatively high. The expected co-occurrence ($E_{i,j}$) is given by:

$$E_{i,j} = \frac{n_i n_j}{N} \tag{2}$$

where n_i and n_j are the overall number of Nice sub-class i and CPC technology code j, and N is the number all Nice sub-classes and CPC technology codes. The standard deviation is defined as:

$$\sigma^2{}_{i,j} = E_{i,j}\left(1 - \frac{n_i}{N}\right)\left(\frac{N - n_j}{N - 1}\right) \tag{3}$$

Intuitively, the positive value of the z-score indicates that the number of random co-occurrences is lower than the number of observed ones, and therefore a positive value reflects common co-occurrences of Nice sub-classes and CPC technology codes. In other words, Nice sub-class and CPC technology pairs are linked in the concordance that have a z-score value that is greater than zero. In the next section, we present the concordance and discuss its attributes.

3. Results

Estimated z-score values reveal whether there is a relation between each Nice sub-class and CPC technology pairs. As expected, Figure 11 shows that only 16% of pairs are statistically significant (i.e., 65,928 out of 407,792 pairs). The z-score values range between -4.81 and 673.65 (median: -0.7 and mean: 0.133).

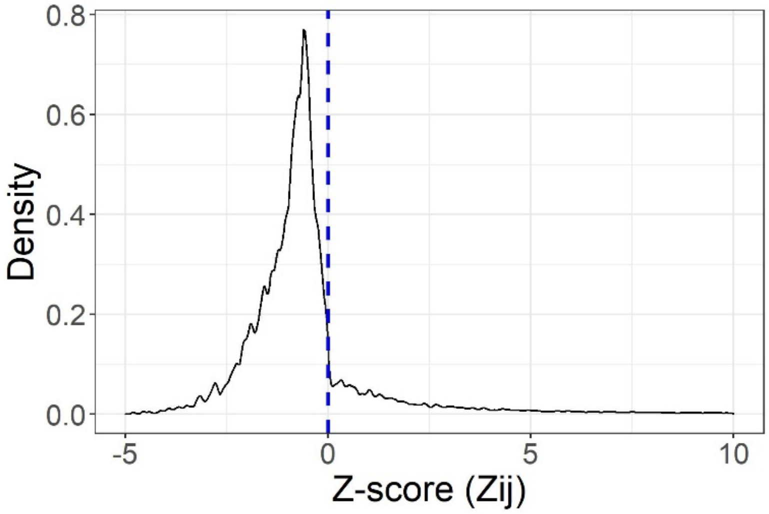


Figure 11. The kernel density of z-score values.

Figure 12 demonstrates the visual representations of concordance based on Nice sub-class and CPC technology pairs having the z-score values greater than zero. 644 (out of 662) CPC technologies and 611 (out of 616) Nice sub-classes have at least one relation. The median Nice sub-class has more relations to CPC code (Nice sub-classes: 110, CPC codes: 60). However, greater standard deviation on the technology side (Nice sub-classes: 47.24, CPC codes: 111.46) suggests that it is plausible that some new technologies find a lot of application in the market at the expense of perhaps obsolete or not directly applicable technologies. It seems that the technology code Y02E[3] (i.e., reduction of greenhouse gas emissions, related to energy generation, transmission, or distribution) finds a lot of applications in the market as this technology has the highest number of relations with Nice sub-classes. These are clearly all those products and services that help companies and consumers to change their energy production and consumption patterns, as part of the ongoing sustainable energy transition. Conversely, 16 technologies have only one relation to a Nice sub-class. F15C (fluid-circuit elements predominantly used for computing or control purposes) is an example of these technologies with one relation to a Nice sub-classes. Moreover, 18 technologies have no relations to Nice sub-classes. These technologies seem to be either obsolete ones (B01B: boiling apparatus) or the ones that have no direct implication in the market (G21J: nuclear explosives and applications thereof).

On the Nice sub-class side, a Nice sub-class associated with printing and cutting machines (top stems: machin, print, cut, chemic, devic) has the highest number of relations to technologies, whereas 5 Nice sub-class (e.g., a Nice sub-class related to alcoholic beverages with top stems such as cider, absinth, alcopop, brandi, calvado) have no link to technology codes.

[3] For the description of CPC codes, see: https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html

As shown in Figure 12, we take the example of the CPC code A01G. As described by the EPO, this code revolves around technologies utilised in horticulture; cultivation of vegetables, flowers, rice, fruit, vines, hops or seaweed; forestry; watering. The concordance suggests that this technology code is related, among others, to Nice sub-classes that can be described by mild, beverage, substitute, yogurt, and coconut; and drink, soft, flavour, water, and carbon.

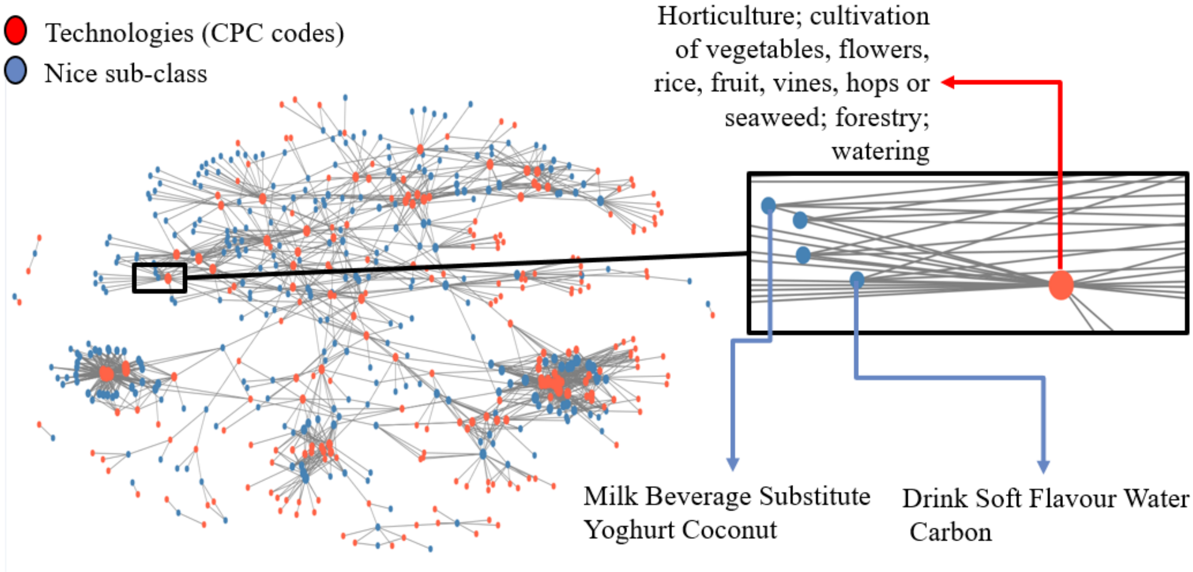

Technologies (CPC codes)
Nice sub-class
Horticulture; cultivation of vegetables, flowers, rice, fruit, vines, hops or seaweed; forestry; watering
Milk Beverage Substitute Yoghurt Coconut
Drink Soft Flavour Water Carbon

Figure 12. The visualization of the concordance.

It is worth mentioning that the provided descriptive statistics on the concordance is based on z-score values greater than zero, and greater values imply that a given Nice sub-class and CPC technology pair co-occurred more often. Therefore, in the developed concordance we list all Nice sub-class and CPC technology pairs with the z-score value greater than zero. By providing the z-score values, one can exclude some relations by setting a threshold on greater values representing the extent to which Nice sub-class and CPC technology pairs co-occur. For instance, if one set the threshold on 2.5 (instead of zero) the new concordance includes 6.8% of pairs (instead of 16%). While multiple manual checks are in line with our expectations, we conduct three validation exercises to ensure that the concordance built based on semantic similarities is supported by market realities.

4. Validations

As discussed earlier, this is of crucial importance to ensure that identified links between technologies and HDB terms reflect real-world patterns and not only potential links. The linkages we draw are in a sense ‘abstract’ and we wish to validate them with ‘real’ linkages. We turn to databases where patents and trademark data are linked to each other in an independent way. We found three different validation sets, at the firm (patent-trademark bundles), trademark (green EU trademark), and product level (IProduct). It is important to note that as these datasets are much smaller than the overall patent and trademark databases, a validation based on comparing predicted and actual linkages would give a biased picture of the validity of our concordance. Instead, the validation will revolve around validating our methodology for building the concordance. We will keep track of all four possible cases resulting from using the concordance and comparing it to the actual linkages (see Figure 13).

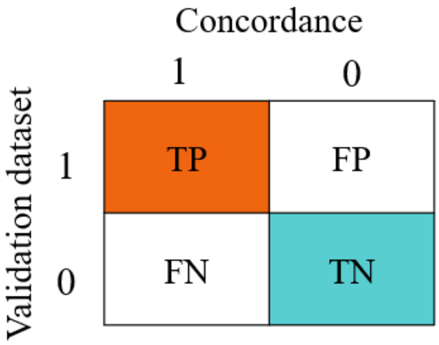


Figure 13. Classification of cases in the validity tests: TP=true positives, TN=true negatives, FP=false positives, FN= false negatives

4.1 Firm level patent-trademark bundles

Firms often file for both patents and trademarks to protect different features of the same product and with different motives in mind (Helmers and Schautschick, 2013, Castaldi et al., 2020). EUIPO (2020) relied on identifying IPR bundles of European firms in the period of 2014-2015, where such bundles were defined as different IPRs filed by the same company within a short period of time. Hence, such bundles represent a good proxy of specific innovation projects developed by companies. They are highly relevant for our validation purposes, since they indicate actual patent-trademark linkages within the same innovation project. The sample in EUIPO (2020) consists of 63,286 firms holding 76,202 European patents and 98,257 European trademarks, and 21,676 registered community designs. To ensure that the technologies included in patent documents are related to Nice sub-classes, we take a rather conservative approach and investigate those patent-trademark bundles by firms that file for only one patent and one trademark in a six-month time period. This approach is in line with the findings of Helmers and Schautschick (2013) who provide evidence that most companies file for related patents and trademarks in a short time-window. This enables us to avoid biases (false positives) through linking multiple unrelated patents and trademarks, whereas this may increase the number of false negatives. This selection results in 501 trademark-patent pairs.

4.2 Green EU trademarks

“Green EU trademarks” is a recent EUIPO report (EUIPO, 2021) that proposed a methodology to identify green goods and services descriptors with trademark data. More specifically, this study investigates more than 85,000 terms that are acknowledged and accepted by all European and several non-European IP offices. EUIPO (2021) utilised a so-called semi-supervised technique to develop Green Term Classifiers and assign goods and services terms to the appropriate green category. Green categories consist of energy production, transportation, energy conservation, reuse/recycling, pollution control, agriculture, climate change, and environmental awareness.

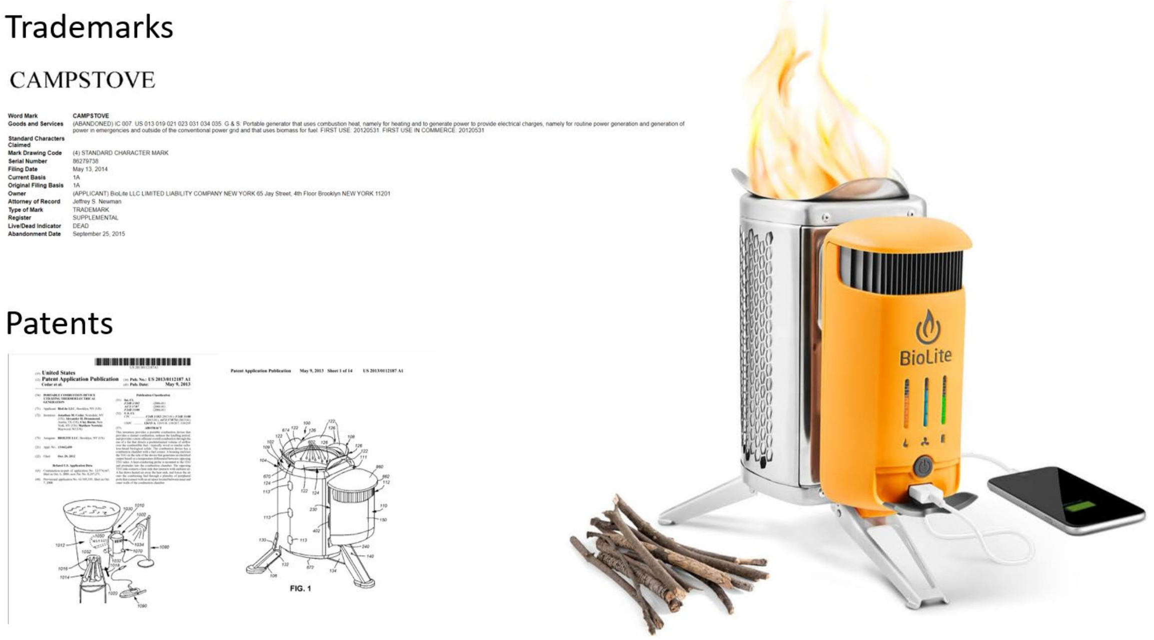
We utilised terms across all categories (except the ones of environmental awareness) and used the text pre-processing technique described in Section 2.3. We assume that these green terms are related to Cleantech CPC technology codes (Table 2). For this validation, we used these 501 green terms coupled with the same number of randomly selected terms and expect that our semantic method can classify terms as green and non-green far better than a random assignment (i.e., true positive rate and true negative rate: 50%).

CPC technology code	Digit-level	Description
Y02	3	Technologies or applications for mitigation or adaptation against climate change
Y02B	4	Climate change mitigation technologies related to buildings
Y02C	4	Capture, storage, sequestration or disposal of greenhouse gases
Y02D	4	Climate change mitigation technologies in information and communication technologies. That is, information and communication technologies aiming at the reduction of their own energy use
Y02E	4	Reduction of greenhouse gas emissions, related to energy generation, transmission or distribution
Y02P	4	Climate change mitigation technologies in the production or processing of goods
Y02T	4	Climate change mitigation technologies related to transportation
Y02W	4	Climate change mitigation technologies related to wastewater treatment or waste management

Table 2. List of cleantech CPC codes.

4.3 IProduct data

One of the earlier EPO academic grants has funded a highly original project linking patents to products (also known as IProduct[4]). Gaétan de Rassenfosse, the principal investigator of that project, allowed us access to the data to construct *patent-product-trademark triads*. The triads were based on matching product names to trademarks and/or patent and trademark applicant names. First, we used algorithmic and manual string matching between the trademark and patent applicants. Second, we matched the name of products in the IProduct and in the EUIPO trademark data. This method gave us 1973 patent-product-trademark links, whereby we derived links between HDB terms and CPC technology codes. Figure 13 provides an example of a patent-product-trademark. This figure shows that a product (e.g., BioLite) is linked to a USPTO patent in the IProduct dataset. Through string matching, we found a trademark for the same product in the EUIPO dataset.



Trademarks

CAMPSTOVE

Word Mark	CAMPSTOVE
Goods and Services	(ABANDONED) IC 007. US 013 019 021 023 031 034 035. G & S: Portable generator that uses combustion heat, namely for heating and to generate power to provide electrical charges, namely for routine power generation and generation of power in emergencies and outside of the conventional power grid and that uses biomass for fuel. FIRST USE: 20120531. FIRST USE IN COMMERCE: 20120531
Standard Characters Claimed	
Mark Drawing Code	(4) STANDARD CHARACTER MARK
Serial Number	86279738
Filing Date	May 13, 2014
Current Basis	1A
Original Filing Basis	1A
Owner	(APPLICANT) BioLite LLC LIMITED LIABILITY COMPANY NEW YORK 65 Jay Street, 4th Floor Brooklyn NEW YORK 11201
Attorney of Record	Jeffrey S. Newman
Type of Mark	TRADEMARK
Register	SUPPLEMENTAL
Live/Dead Indicator	DEAD
Abandonment Date	September 25, 2015

Patents

United States
Patent Application Publication
Pub. No.: US 2013/0112187 A1
Cedar et al.
Pub. Date: May 9, 2013
Patent Application Publication   May 9, 2013   Sheet 1 of 14   US 2013/0112187 A1
FIG. 1
BioLite

Figure 13. An example of linking patents and trademarks through the IProduct dataset.

[4] http://www.iproduct.io/data

4.4 Validation results

We created a concordance for each validation dataset and tested the quality of the concordance based on each percentile of the z-score value. Theoretically, we expect that the thresholds near zero provide the best result (i.e., high values of both true positive rates and true negative rates), whereas concordances based on z-scores smaller or greater than zero score high on one measure at the expense of the other measure. Figure 14 reveals that this expectation is supported by the validation results. Table 3 shows the results of the validation exercises based on the Zij threshold corresponding to zero.

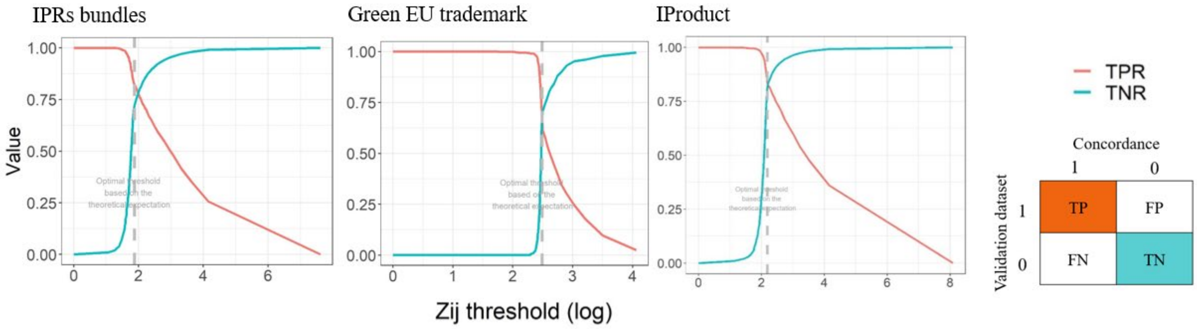


Table 14. The results of the validation exercises for varying Z_{ij} thresholds.
Note: the x-axis is log-transformed. That is, log(Zij + (the minimum value of Zij×-1) + 1). The grey dashed line corresponds to Z_{ij} threshold=0.

Validation datasets	Observations	TPR (%)	TNR(%)
EPO-EUIPO bundles	501 trademark-patent pairs linked through bundles	71	83
EUIPO green terms	1208 EUIPO green and non-green terms	67	73
IProduct	1973 trademark-patent pairs linked through IProduct	87	76

Table 3. The results of the validation exercises based on the Z_{ij} threshold corresponding to zero.

To conclude, all three validation results provide evidence that the links between patent classes and trademark classes identified by our concordance are significantly related to actual linkages between patents and trademarks as filed by innovative companies.

5. Conclusion and deliverables

We started this project with the wish of better connecting patent and trademark data. We managed to develop a concordance between patent and trademark classes that we will make available to all who aim to better understand how patented inventions connect to trademarked products. As discussed in the introduction, the potential applications are many and they range from technology-level analysis to firm-level and to regional-level ones.

The project resulted in the following deliverables

1. The definition of trademark sub-classes and the concordance between those sub-classes and CPC classes: both deliverables will be shared through an open data platform in the next months. The sharing will also come with an extensive manual on how to understand and use the two deliverables.

2. A methodological paper on the concordance, detailing the potential applications of the concordance, its limitations and specificities, in preparation for submission to an international journal

3. Two scientific papers on regional capabilities in cleantech, in preparation for submission to key journals in regional economics.

4. One paper in the making, focused on the link between patents, products and trademarks, co-authored with Gaetan de Rassenfosse and his team.

5. A podcast about the project, recorded in May 2022 and available on the EPO ARP website.

Acknowledgements

We are grateful for the financial support from the EPO ARP program. We benefited greatly from conversations and advice on the trademark data from Michal Kazimierczak and Francisco Garcia-Valero at EUIPO and also from their willingness to share the data underlying the recent EUIPO reports on IP bundling and green trademarks. We also wish to thank Georg von Graevenitz for his feedback on our intermediate report and Gaetan de Rassenfosse for sharing the IProduct database.

All errors are our own.

References

Abbas, A., Zhang, L., Khan, S.U., 2014. A literature review on the state-of-the-art in patent analysis. World Patent Information 37 (1), 3–13. doi:10.1016/j.wpi.2013.12.006.

Abbasiharofteh, M., Broekel, T., 2020. Still in the shadow of the wall? The case of the Berlin biotechnology cluster. Environment and Planning A: Economy and Space 46 (3). doi:10.1177/0308518X20933904.

Abbasiharofteh, M., Kinne, J., Krüger, M., 2021. The Strength of Weak and Strong Ties in Bridging Geographic and Cognitive Distances. ZEW Discussion Paper No. 21-049, Mannheim.

Abbasiharofteh, M., Kogler, D.F., Lengyel, B., 2020. Atypical Combination of Technologies in Regional Co-inventor Networks. Papers in Evolutionary Economic Geography (PEEG) 20.55.

Arts, S., Hou, J., Gomez, J.C., 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. Research Policy 50 (2), 104144. doi:10.1016/j.respol.2020.104144.

Bergeaud, A., Potiron Yoann, Raimbault, J., 2017. Classification Data for "Classifying Patents Based on their Semantic Content". Harvard Dataverse.

Brachtendorf, L., Gaessler, F., Harhoff, D., 2020. Approximating the Standard Essentiality of Patents – A Semantics-Based Analysis. EPO.

Breschi, S., Lissoni, F., 2009. Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. Journal of Economic Geography 9 (4), 439–468. doi:10.1093/jeg/lbp008.

Castaldi, C., 2019. All the great things you can do with trademark data: Taking stock and looking ahead. Strategic Organization 18 (3), 472–484. doi:10.1177/1476127019847835.

Castaldi, C., Dosso, M., 2018. From R&D to market: using trademarks to capture the market capability of top R&D investors.

Castaldi, C., Frenken, K., Los, B., 2015. Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting. Regional Studies 49 (5), 767–781. doi:10.1080/00343404.2014.940305.

Castaldi, C., Block, J., & Flikkema, M. J., 2020. why and when do firms trademark? Bridging perspectives from industrial organisation, innovation and entrepreneurship. Industry and Innovation, 27(1-2), 1-10.

Castaldi, C., Mendonça, S., 2022. Regions and trademarks: Research opportunities and policy insights from leveraging trademarks in regional innovation studies. Regional Studies 56 (2), 177–189. doi:10.1080/00343404.2021.2003767.

Dernis, H., Dosso, M., Hervás, F., Millot, V., Squicciarini, M., Vezzani, A., 2015. World Corporate Top R&D Investors: Innovation and IP bundles. Publications Office of the European Union, Luxembourg.

EPO-EUIPO, 2016. Intellectual property rights intensive industries and economic performance in the European Union, Industry-Level Analysis Report (Second edition). EUIPO and EPO.

EUIPO, 2021. Green EU trade marks: Analysis of goods and services specifications, 1996-2020. European Union Intellectual Property Office, Alicante.

EUIPO, 2020. Use of IPR bundles by EU firms 2014-2015. European Union Intellectual Property Office, Alicante.

Fleming, L., Sorenson, O., 2001. Technology as a complex adaptive system: Evidence from patent data. Research Policy 30 (7), 1019–1039. doi:10.1016/S0048-7333(00)00135-9.

Flikkema, M., Castaldi, C., Man, A.-P. de, Seip, M., 2019. Trademarks' relatedness to product and service innovation: A branding strategy approach. Research Policy 48 (6), 1340–1353. doi:10.1016/j.respol.2019.01.018.

Fontana, M., Iori, M., Montobbio, F., Sinatra, R., 2020. New and atypical combinations: An assessment of novelty and interdisciplinarity. Research Policy 49 (7), 104063. doi:10.1016/j.respol.2020.104063.

Graevenitz, G. von, Graham, S.J.H., Myers, A.F., 2022. Distance (still) hampers diffusion of innovations. Regional Studies 56 (2), 227–241. doi:10.1080/00343404.2021.1918334.

Grazzi, M., Piccardo, C., Vergari, C., 2019. Concordance and Complementarity in Intellectual Property Instruments. SSRN Electronic Journal 50 (4), 8. doi:10.2139/ssrn.3316020.

Greenhalgh, C., Rogers, M., 2010. Innovation, intellectual property and economic growth. Princeton University Press, Princeton, N.J., Woodstock.

Hall, B., Helmers, C., Rogers, M., Sena, V., 2014. The Choice between Formal and Informal Intellectual Property: A Review. Journal of Economic Literature 52 (2), 375–423. doi:10.1257/jel.52.2.375.

Helmers, C., Schautschick, P., 2013. The use of intellectual property right bundles by firms in the UK. Intellectual Property Office, Newport.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2021. Measuring Technological Innovation over the Long Run. American Economic Review: Insights 3 (3), 303–320. doi:10.1257/aeri.20190499.

Kinne, J., Lenz, D., 2021. Predicting innovative firms using web mining and deep learning. PloS one 16 (4), e0249071. doi:10.1371/journal.pone.0249071.

Krüger, M., Kinne, J., Lenz, D., Resch, B., 2020. The Digital Layer: How Innovative Firms Relate on the Web (20 ZEW Discussion Paper). ZEW, Mannheim.

Le, Q.V., Mikolov, T., 2014. Distributed Representations of Sentences and Documents. arXiv:1405.4053.

Magerman, T., van Looy, B., Song, X., 2010. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics 82 (2), 289–306. doi:10.1007/s11192-009-0046-6.

Mewes, L., 2019. Scaling of Atypical Knowledge Combinations in American Metropolitan Areas from 1836 to 2010. Economic Geography 95 (4), 341–361. doi:10.1080/00130095.2019.1567261.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Negro, S.O., Alkemade, F., Hekkert, M.P., 2012. Why does renewable energy diffuse so slowly?: A review of innovation system problems. Renewable and Sustainable Energy Reviews 16 (6), 3836–3846. doi:10.1016/j.rser.2012.03.043.

Neuhäusler, P., Feidenheimer, A., Frietsch, R., Kroll, H., 2021. Generating a classification for EUIPO trademark filings: A string matching approach. Discussion Papers "Innovation Systems and Policy Analysis" 69, Karlsruhe.

Ozgun, B., Broekel, T., 2021. The geography of innovation and technology news - An empirical study of the German news media. Technological Forecasting and Social Change 167 (6), 120692. doi:10.1016/j.techfore.2021.120692.

Petralia, S., 2020. Mapping general purpose technologies with patent data. Research Policy 49 (7), 104013. doi:10.1016/j.respol.2020.104013.

Rassenfosse, G. de, 2018. Notice failure revisited: Evidence on the use of virtual patent marking: . . National Bureau of Economic Research.

Rodrik, D., 2014. Green industrial policy. Oxford Review of Economic Policy 30 (3), 469–491.

Schautschick, P., Greenhalgh, C., 2015. Empirical studies of trade marks – The existing economic literature. Economics of Innovation and New Technology 25 (4), 358–390. doi:10.1080/10438599.2015.1064598.

Scherer, F.M., 1982. The office of technology assessment and forecast industry concordance as a means of identifying industry technology origins. World Patent Information 4 (1), 12–17. doi:10.1016/0172-2190(82)90086-2.

Schmoch, U., 2003. Service marks as novel innovation indicator. Research Evaluation 12 (2), 149–156. doi:10.3152/147154403781776708.

Teece, D.J., Rumelt, R., Dosi, G., Winter, S., 1994. Understanding corporate coherence. Journal of Economic Behavior & Organization 23 (1), 1–30. doi:10.1016/0167-2681(94)90094-9.

WIPO, 2013. Nice classification: The International Classification of Goods and Services for the Purposes of the Registration of Marks. World Intellectual Property Organization.

Zhou, T., Ren, J., Medo, M., Zhang, Y.-C., 2007. Bipartite network projection and personal recommendation. Physical review. E, Statistical, nonlinear, and soft matter physics 76 (4 Pt 2), 46115. doi:10.1103/PhysRevE.76.046115.

Zolas, N., Lybbert, T.J., Bhattacharyya, P., 2017. An 'Algorithmic Links with Probabilities' Concordance for Trademarks with an Application Towards Bilateral IP Flows. The World Economy 40 (6), 1184–1213. doi:10.1111/twec.12382.