Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2022

Technical Documentation

February 2022

Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2022

Technical Documentation

February 28, 2022

Submitted to:
SRI International

Authors
Guillaume Roberge
Alexandre Bédard-Vallée

Project Leader
Guillaume Roberge

By:

Science-Metrix
1.438.945.8000 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com

Contents

Tables

1 Introduction

Science-Metrix has been commissioned by SRI International, on behalf of the National Science Foundation, to develop measures and indicators of research and patent activity using bibliometrics and patent data for inclusion in the Science and Engineering Indicators (SEI) 2022. This technical document details the various steps taken to implement the databases, clean and standardize the data, and produce statistics on technometric data, including not only U.S. utility patents from the United States Patent and Trademark Office (USPTO), but also patent families with patents from dozens of patent authorities, and trademarks. The work done for the bibliometrics aspect is presented in a separate document. This documentation is accompanied by a collection of external files that are necessary complements to perform these tasks. The list of accompanying external files is as follows:

External File 1: IPC technology concordance table
External File 2: Patent number and uuid to Scopus ID
External File 3: Patent number and SEQ to countries and regions
External File 4: Patent number and SEQ to American states
External File 5: US applicant to sector

These external files are also introduced in the relevant sections of this documentation.

2 Patent indicators

USPTO

The patent indicators for the U.S. market in this report were produced using an in-house implementation of the PatentsView patent database, a platform derived from the USPTO bulk data files. To accomplish such tasks, an in-house version of the database was built in Databricks and was carefully conditioned for the production of large-scale comparative patent analyses based on utility patents.

Worldwide priority patents

While metrics based only on USPTO patent data inform innovation activities in the United States, they do not provide a global scope, because most inventions worldwide are not protected in the United States, even though it is one of the largest markets in the world—if not *the* largest. This may result in misleading inferences when it comes to comparing innovation around the globe, as innovation for countries closer to the U.S. market (e.g., the United States, Canada, Mexico) will tend to be overestimated compared to countries with lesser economic integration in it (e.g., European and Asian countries). Over the years, new metrics have been created to alleviate the effect of selecting a single patent authority when measuring innovation. For instance, the concept of triadic patent families—that is, patents being applied for in the United States,[1] Europe and Japan—was developed in the 1990s. It allowed for fairer comparisons across countries, measuring inventions of broader economic scope covering the three largest markets at the time. With the economic growth of other countries such as China and South Korea, the concept was expanded in recent years to five offices (IP5). Patent statistics based on the IP5 authorities are now becoming more mainstreamed and can be consulted online.

To provide a broader context to the patent analyses presented in the report, counts of patent families based on data indexed in the PATSTAT database were also provided. These metrics help alleviate the home advantage when measuring innovation within only a single market and alleviate differences in office practices as these can vary widely, resulting in very different patent counts when measuring similar innovations around the world. For instance, inventions protected with single utility patents in some countries could be split into multiple utility patents for a similar invention in another country.

Using PATSTAT data, which cover utility patents from close to 100 patent authorities, INPADOC patent families based on almost all data covered in PATSTAT were selected as the main unit of measure, following a methodology proposed by a team of researchers from academia and the OECD.[2] This method uses information within patent families to fill in gaps regarding inventorship for patent offices where data are not complete, looking at related patents in other offices when information is not available for a patent. When no information on inventorship can be retrieved from any office, the approach relies instead on assignees, using the assumption that in such cases inventors are frequently from the same country as the

[1] Originally, granted patents were used for the U.S. patents because data on patent applications were not available to the research community.

[2] De Rassenfosse et al. (2012). The worldwide count of priority patents: A new indicator of inventive activity, *Research Policy 42*(3), 720–737.

assignees who requested the patent, again using all patents within the family to fill remaining gaps. As a final step, for the remaining priority patents with missing information, the country of the patent authority is projected as the country of inventorship, because in most cases patents without any information and no related patent at the world level will be the fruit of local inventors. While this method is not perfect and can lead to errors when projecting inventorship at the level of individual patents, its level of precision is quite good overall when dealing with large-scale analyses such as the one prepared for this project.

However, one limitation of the methods described above is that it leaves out what are called "artificial patents" when dealing with computation of statistics. Artificial priority patents are created when a published patent mentions an earlier priority document not delivered at the EPO. Such a document may be missing if the office where it has been filed has not published it or if the priority document is not a patent of invention. Artificial priority patents contain only scarce information, including patent office, date, and type of applied document. Critical missing information includes names, addresses of applicants and inventors, and IPC codes. A large share of artificial patents come from provisional patents, which are patent applications often used in the U.S. and other markets to quickly protect an invention, at low cost, in the hope of later filing a patent application for a utility patent in the same market.

Normally, leaving these documents out of the equation should not drastically impact statistics based on PATSTAT. For instance, preparing counts of granted USPTO patents will result in counts mostly identical to those prepared according to other sources because the underlying data are the same. However, in the case of worldwide priority patents, because some patent families have such "artificial patents" as their priority patent for the whole family, about 15% of all patent families end up being left out of the analysis (i.e., about 3 million out of 35 million patent families). While this number, although non-negligible, may not seem too problematic, because most of these artificial patents are not evenly distributed across patent offices and are in fact overrepresented for the U.S. and some European markets, their exclusion leads to major underestimation of the real counts of innovations for multiple countries, including the United States. Furthermore, this is even more critical in some technological areas that have used provisional patents more and more frequently in recent years because these patent applications are not patents of invention and thus only appear as artificial patents in the database. In some areas, such as Biotechnology and Pharmaceuticals, the vast majority of recent patent families now start with a provisional patent application, resulting in drastically different results if artificial priority patents are not accounted for in the analysis. The impact of artificial patents is further detailed in a recent paper by Laurens et al.[3]

Regarding the limitation related to artificial priority patents, instead of dropping patent families that first came to life through these artificial priority patents, Science-Metrix selected the first utility patent in the family that was applied for after the artificial priority patent and used it as a replacement to act as the first priority patents in these cases. This made it possible to obtain the relevant information for IPC codes and inventorship in the same manner as for other priority patents, following the approach designed by De

[3] Laurens et al. (2018). The artificial patents in the PATSTAT database: How much do they matter when computing indicators of internationalization based on worldwide priority patents? *Scientometrics 114*, 91–112.

Rassenfosse. Implementing this correction results in more complete information for a large number of countries around the globe, with the biggest effect on those countries with a national patent office that allows provisional patent applications or where inventors often reach out to markets where these are available.

2.1 Data limitations

There is no notable limitation regarding the USPTO data because they provide complete coverage of U.S. patents, but the same cannot be said for some patent authorities available in PATSTAT data. Even though PATSTAT's coverage is quite expansive, data quality is unequal across patent offices, as the EPO relies on the offices to provide complete data of good quality. In the context of the SEI, the impact of lower-quality data for small patent offices is minimized because the level of output from these offices is extremely low. Science-Metrix performed data-quality checks for the largest patent authorities in the database to ensure that most problems could be identified and either corrected for or at least highlighted. Data gaps for specific patent authorities most often result in an underestimation of innovation for the host country as residents usually account for most inventions at their national office. Two cases stood out here: India and Italy.

India

In the case of India, coverage in PATSTAT is lacking, resulting in an underestimation of its output. However, according to WIPO statistics,[4] India is a special case because only about 37% of patent applications to its national office were made by residents, out of an annual number hovering at around 54,000 in 2019. This means that about 12,000 patent applications are made by the residents each year (8,800 in 2010, 19,500 in 2019), with foreign companies and inventors dominating in terms of patent applications at the office (e.g., Qualcomm, Samsung, Huawei, Microsoft, Philips, General Electrics, Ericsson, Mitsubishi, BASF).[5]

The choice of INPADOC patent families as the indicator reduces the extent of the undercount to substantially below the 12,000-patent number because each patent family may include multiple patents. Furthermore, some of these Indian patents are part of international patent families and should therefore be counted within the data using information provided by other patent offices for related patents, further reducing the actual undercount of India's contribution. The data presented in the *Indicators* are measuring granted patent families based on the year of the first grant in the patent family; however, the Indian patent authority is lagging behind in terms of processing of its patents, only granting about 11,000 patents to Indian residents over the last decade (less than 1,000 for most years). It takes an average of five years for the India office to evaluate each patent, and as a result the office has a backlog of hundreds of thousands of unprocessed patent applications.[6] While changes have been made to address this issue in recent years[7]

[4] https://www.wipo.int/ipstats/en/statistics/country_profile/profile.jsp?code=IN

[5] http://www.ipindia.nic.in/writereaddata/Portal/IPOAnnualReport/1_94_1_1_79_1_Annual_Report-2016-17_English.pdf

[6] https://www.hindustantimes.com/mumbai-news/india-takes-five-years-to-look-at-patent-applications-reveals-economic-survey/story-q1u11vKeg8lLtPqtdEtniM.html

[7] https://www.majumdarip.com/blog_post/indian-patent-office-shows-trends-of-speedy-grants/

and numbers of granted patents have started increasing, they are still relatively low in a global context (about 1,700 granted patents in 2017, 2,300 in 2018 and 3,700 in 2019). As a result, this does not substantially alter India's ranking among leading countries.[8]

Italy

The data quality is also limited for granted patents at the Italian patent offices in PATSTAT. To address this issue for *Indicators*, data on patent applications were used to correct the patterns observed; this suggests a small overestimation of invention for Italy and other countries that patent in this market.

2.2 Kind codes

Kind codes are a classification system used across patent offices to classify document types. Each patent office has its own classification system; although codes are often similar across offices, their implementation may differ across offices.

For the SEI 2022, USPTO kind codes were used to identify utility patents from the USPTO. For the patent family approach using PATSTAT data from multiple offices, standardized codes provided in PATSTAT were selected instead to limit the analysis to utility patents across these offices.

USPTO patents

The patent indicators for this study were produced using a set of kind codes[9] that select granted utility patents and applications, although the indicators were only computed on granted utility patents. Kind codes associated with utility patents at the USPTO were limited to three document types: A, B1 and B2. Kind code A applies to granted patents before 2001, while B1 and B2 replaced this kind code on 2 January 2001.

Patent families

Granted utility patents were selected using the keys *appln_kind*, *ipr_type* and *publn_first_grant* available in PATSTAT, which respectively identify document types (e.g., applications, grants), types of patents (utility, design, plant) and grant year.

2.3 Databases

PatentsView

Most of the patent analyses in *Indicators* were prepared using data from the USPTO indexed in PatentsView. The database provides details on patents such as full titles and abstracts, the country and state (when available) of the inventors and applicants, as well as names of the inventors and applicants. In most cases, applicants are organizations, although they are sometimes individuals when the patent is not assigned to any organization. The database also provides information on three classification schemes: the U.S. national classes (the USPC classes, although these are not available after 2015 as the system is no

[8] https://www.livemint.com/Politics/LkKhP62yJrhSRJZDoqDIiN/Indias-patent-problems.html

[9] http://www.uspto.gov/learning-and-resources/support-centers/electronic-business-center/kind-codes-included-uspto-patent

longer in use), the World Intellectual Property Organization's (WIPO) International Patent Classification (IPC), and the Cooperative Patent Classification (CPC). The CPC was produced in partnership between the USPTO and the EPO; it replaced the USPC classes after 2015, and the European Classification System (ECLA) after 2012. PatentsView is suitable for the production of technometric data dating from 1976, whereas patent data in the previous round of the SEI were largely prepared for the period 1996 to the present.

PatentsView tables were downloaded and uploaded into the Science-Metrix SQL server. The process is straightforward and does not require any treatment because the data are already parsed. Documentation[10] presenting the content of the tables is available on the PatentsView website.

PATSTAT

The European Patent Office Worldwide Patent Statistical database, better known as EPO PATSTAT or PATSTAT, is the database of reference in the field of international technometrics. Mainly developed for use by governmental organizations and academic institutions, it contains bibliographical and legal status patent data from most industrial and developing countries and covers major patent offices such as the EPO (Europe) and USPTO (United States). A conditioned in-house version of PATSTAT2021 Spring Edition, which consist of pre-defined tables with keys linking these together, was built on Databricks and used to prepare statistics on worldwide priority patents.

2.4 Data standardization

2.4.1 Mapping of patents by technical fields

In SEI 2016, patents were matched on a classification scheme of 35 technical fields[11] developed by the World Intellectual Property Organization (WIPO). The main objective behind the development of such a classification was to provide a tool for country comparisons.[12] The technical fields defined by this classification are listed in Table I.

[10] https://patentsview.org/download/data-download-dictionary

[11] Classification scheme from IPC8 codes to technical fields. Available at http://www.wipo.int/ipstats/en/statistics/technology_concordance.html

[12] Concept of a Technology Classification for Country Comparisons. Available at http://www.wipo.int/edocs/mdocs/classifications/en/ipc_ce_41/ipc_ce_41_5-annex1.pdf

Table I WIPO classification scheme for the production of SEI patent indicators

Technical Fields	
Analysis of biological materials	Macromolecular chemistry, polymers
Audio-visual technology	Materials, metallurgy
Basic communication processes	Measurement
Basic materials chemistry	Mechanical elements
Biotechnology	Medical technology
Chemical engineering	Micro-structural and nano-technology
Civil engineering	Optics
Computer technology	Organic fine chemistry
Control	Other consumer goods
Digital communication	Other special machines
Electrical machinery, apparatus, energy	Pharmaceuticals
Engines, pumps, turbines	Semiconductors
Environmental technology	Surface technology, coating
Food chemistry	Telecommunications
Furniture, games	Textile and paper machines
Handling	Thermal processes and apparatus
IT methods for management	Transport
Machine tools	

Source: IPC Technology Concordance Table

This classification scheme is based on the IPC classification. Since the most recent U.S. patents are natively classified using the CPC, which replaced the USPC classification scheme at the national level, using this scheme as a starting point is more practical. In order to classify the patents by technology fields, a concordance table between CPC and IPC codes prepared by the USPTO, in collaboration with the EPO, was used.[13]

The WIPO technical field classification scheme is mutually exclusive in that no IPC code is assigned to more than one technical field. In the rare cases that remained unmatched to a technical field after the code conversion process, the leftover IPC codes were assigned to an additional field entitled *Unclassified* so that the sum of patents across technical fields would add up to the total number of patents.

Patents can be assigned more than one IPC code and therefore potentially more than one technical field if multiple codes are not all assigned to the same field. To make sure that the sum of patents across technical fields added up to the total number of patents, it was necessary to fraction patent counts by technical field. Patents were fractioned according to the number of WIPO technical fields to which they were assigned, each technical field receiving an equal weight. For instance, a patent assigned to three different IPC codes pointing to two distinct technical fields would see each of these fields receive half of the patent count. The following example in Table II details this process for one patent.

[13] http://www.cooperativepatentclassification.org/cpcConcordances.html

Table II Example of a patent fractioned by technical fields according to IPC codes, following conversion from CPC codes

CPC Codes					IPC Codes (Concordance with CPC codes)					Technical Field
Section	Class	Subclass	Group	Main Group	Section	Class	Subclass	Main Group	Subgroup	
B	08	B	3	022	B	8	B	3	2	Chemical engineering
B	24	B	53	017	B	24	B	53	17	Machine tools
B	24	B	21	04	B	24	B	21	4	Machine tools
B	08	B	3	041	B	8	B	3	4	Chemical engineering
B	08	B	1	02	B	8	B	1	2	Chemical engineering
B	08	B	1	007	B	8	B	1	0	Chemical engineering
B	08	B	3	123	B	8	B	3	12	Chemical engineering

Total fraction of patent by technical field	
Chemical engineering	0.5
Machine tools	0.5

Source: Prepared by Science-Metrix using the IPC Technology Concordance Table (http://www.wipo.int/ipstats/en/statistics/technology_concordance.html)

The same approach was applied when counting worldwide patent families, accounting equally for all technical fields appearing at least once on any patent in the INPADOC patent family of the priority patents.

External File 1: IPC technology concordance table

or online at: http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls

2.4.2 Linking citations to non-patent literature to the bibliometric database

This section presents the various tasks that were performed in order to link USPTO utility patents with scientific publications by using the references made to scientific publications within patents.

Extracting references

All references from patents indexed in the USPTO that were tagged as "non-patent literature" were first extracted from the PatentsView patent database (i.e., in table "Otherreference"). This represented 37,113,970 reference strings, each tagged individually within the database using a unique identifier (uuid).

Although named "non-patent literature", the field contains many references to patent literature. It also contains numerous references to non-scientific literature such as handbooks, instruction manuals, Wikipedia pages, and so forth. Here are a few examples of reference strings to patent literature, incorrectly tagged as "non-patent literature" in the PatentsView database:

- International Searching Authority, International Search Report [PCT/ISA/210] issued in International Application No. PCT/JP2004/017961 on Feb. 1, 2005.
- Israeli Patent Office, Office Action issued in Israeli Application No. 187840; dated Mar. 10, 2010.
- New Zealand Patent Office, Office Action in NZ Application No. 563863; issued Jul. 1, 2010.
- Russian Patent Office, Office Action in Russian Application No. 2007148992; issued Jun. 23, 2010.
- European Patent Office, Supplementary European Search Report dated Feb. 12, 2010 in European Application No. 04819909.5.

And a few examples of reference strings leading to material that is neither peer-reviewed scientific nor patent literature:

- Webpage CLEAT from http://ezcleat.com/gallery.html dated Apr. 19, 2011.
- Automotive Handbook, 1996, Robert Bosch GmbH, 4th Edition, pp. 170-173.
- Periodic Table of the Elements, version published in the Handbook of Chemistry and Physics, 50thEdition, p. B-3, 1969-1970.
- Microsoft aggressive as lines between Internet, TV blur dated Jul. 29.

Here is an example of a proper reference string to peer-reviewed scientific literature with the various elements of bibliographic information indicated in different colors:

- Grinspoon, et al., Body Composition and Endocrine Function in Women with Acquired Immunodeficiency Syndrome Wasting, J. Clin Endocrinol Metab, May 1997, 82(5): 1332–7.

  Authors, Title, Journal, Date, Volume, Issue, Pages

Pre-processing: Removing references to patent literature and generic material

Identifying references to peer-reviewed scientific literature within this pool is an easy task if recall is not a concern. If, however, the goal is to identify all references to peer-reviewed scientific literature within the pool, the task becomes extremely arduous. It is easier and much more efficient to eliminate reference strings that are obviously patent related or that point to generic material and deem the remainder valid candidates for a match.

N-grams are contiguous sequences of *n* items from a given sequence. In this case, the items are words and sequences are reference strings. Studying high-frequency n-grams is a very efficient way of separating noise from useful data in a corpus. For example, the 10 most frequent 2-grams in the original pool of reference strings during data preparation for SEI 2014 are listed in Table III.

Table III Most frequent 2-grams in patent reference strings

Rank	2-grams	Frequency
1	ET AL	9,057,092
2	U S	2,385,810
3	APPL NO	2,036,765
4	S APPL	2,0246,20
5	OF THE	1,492,354
6	OFFICE ACTION	1,159,499
7	JOURNAL OF	954,351
8	APPLICATION NO	800,897
9	NO 11	794,935
10	SEARCH REPORT	760,949

Source: SEI 2014 technical documentation

In this small subset of 2-grams, there are six expressions that are obvious signifiers for patent literature (U S, APPL NO, S APPL, OFFICE ACTION, APPLICATION NO, SEARCH REPORT), two expressions very common to scientific literature (ET AL, JOURNAL OF) and two other expressions that are so generic as to be useless in this context (OF THE, NO 11).

Matching references to scientific literature

Advanced fuzzy matching algorithms that searched for hundreds of patterns used in bibliographic referencing were used to retrieve titles, pages, issues, volumes, publication years and journal names and their abbreviated forms appearing in the references. These extracted parameters were tested against article entries in the Scopus database in conjunction with similarity analyses between the references and publication titles and journal titles.

The matching algorithm was tuned to favor precision at the expense of recall because increasing recall above the current rate attained (i.e., 94%) would greatly increase the number of false positive matches, with minimal impact on recall. A total of 17,560,414 references were matched with high confidence to scientific literature in the Scopus database, going back to the 1800s.

External File 2: Patent number and uuid to Scopus ID

A large share of the remaining references are non-scientific references, references to scientific articles not indexed in the Scopus database, or references lacking information to confidently match them to a publication. Here are examples of unmatched references:

- Cohen et al. Microphone Array Post-Filtering for Non-Stationary Noise, source(s): IEEE, May 2002.
- Mizumachi, Mitsunori et al. Noise Reduction by Paired-Microphones Using Spectral Subtraction, source(s): 1998 IEEE. pp. 1001-1004.
- Demol, M. et al. Efficient Non-Uniform Time-Scaling of Speech With WSOLA for CALL Applications, Proceedings of InSTIL/ICALL2004 NLP and Speech Technologies in Advanced Language Learning Systems Venice Jun. 17-19, 2004.
- Laroche, Jean. Time and Pitch Scale Modification of Audio Signals, in Applications of Digital Signal Processing to Audio and Acoustics, The Kluwer International Series in Engineering and Computer Science, vol. 437, pp. 279-309, 2002.
- Tekkno Trading Project Brandnews, NSP, Jan. 2008, p. 59.
- Merriam-Webster Online Dictionary, Definition of Radial (Radially), accessed Oct. 27, 2010.
- Merriam-Webster Online Dictionary: definitions of uniform and regular, printed Jul. 8, 2006.
- Article: Mictrotechnology Opens Doors to the Universe of Small Space, Peter Zuska Medical Device & Diagnostic Industry, Jan. 1997.
- Article: For lab chips, the future is plastic. IVD Technology Magazine, May 1997.
- Affinity Siderails Photographs dated Dec. 2009, numbered 1-6.
- Information Disclosure Statement By Applicant dated Jan. 24, 2013.
- Merriam-Webster's Collegiate Dictionary, published 1998 by Merriam-Webster, Incorporated, p. 924.

At the end of the matching process, manual validations to estimate recall and precision were performed. Overall, the precision of the patent references matched to scientific publications stood at around 99%. Using a sample of 100 patent references that were not matched, recall within this sample was estimated at 95%—that is, only five of these references could be linked to scientific publications when searched for manually. This number is especially important because it makes it possible to estimate the number of references to scientific publications missed by the matching algorithms. In total, of the 44,676,474 references available in the "otherreference" table, 17,560,414 could be matched to a scientific publication indexed in Scopus. Since about 6.6 million references were filtered out in the pre-processing step (e.g., reference to patents, search reports), this left about 20.4 million references unmatched. Using the 95% recall estimated above on a sample of unmatched references, this means that approximately 5% of the 20.4 million references, or about 1.2 million results, could potentially be references to scientific publications that the algorithm could not match. Therefore, the expected total number of matched references should stand at about 18.7 million, meaning that recall for the current exercise stands at about 94%. While it is expected that further improvement to the matching algorithm could be performed in the future, it will become extremely difficult to increase recall without compromising on precision because the missed cases are all hard to catch and will not be easily retrieved as they mostly consist of exceptions and unstandardized ways of referencing literature.

2.4.3 Data standardization: country, country groups, regions

To provide comparisons across countries and regions, data are presented at the regional and national levels in the SEI. It is straightforward to identify publications at the national level in USPTO patents because the two-letter country codes for inventors and applicants are provided in PatentsView. Online documentation on the USPTO website includes a conversion table from country codes to country names.[14] Science-Metrix matched country groups and regions using the USPTO conversion table, which enables quick identification of all countries included under each country group or region. A few corrections to country codes were performed to reassign outdated country codes to new codes reflecting geopolitical changes (e.g., Yugoslavia used for addresses in Serbia, Serbia and Montenegro, Slovenia).

Similar corrections were applied for data on Puerto Rico and the U.S. Virgin Islands. These were included under "Central and South America" in the SEI 2016 edition, but in the following rounds they were included under "North America", with the U.S. Virgin Islands being included under the United States and Puerto Rico being presented separately from the United States. For this edition, Puerto Rico was moved to "Central America and Caribbean" to align with regional definitions used in the bibliometric analyses. To achieve this, country information had to be corrected for both of these countries because although they often appear under their proper country code in the database (i.e., PR and VI), in many cases the country code is instead set to "US", with "PR" and "VI" being instead displayed in the state information. As a result, all country codes set to "US" for which the state code was displayed as "PR"

[14] http://patft.uspto.gov/netahtml/PTO/help/helpctry.htm

were reassigned to "PR", and all country codes assigned to "VI" were replaced with "US", to provide the valid number of patents for both.

External File 3: Patent number and SEQ to countries and regions

2.4.4 Data standardization: U.S. states

Information regarding states for inventors and applicants on USPTO patents is provided in PatentsView; however, it is generally absent for most countries other than the United States. Science-Metrix matched the two-letter U.S. state codes provided in PatentsView to U.S. state names. The total for the United States is limited to one of the 50+1 states (including the District of Columbia), plus the Northern Mariana Islands (coded "MP") and the U.S. Virgin Islands (coded "VI") and the "unclassified" cases for those where state information was missing or invalid.

External File 4: Patent number and SEQ to American states

2.4.5 Data coding: U.S. sectors

Coding of U.S. sectors was prepared using information about applicants for which the country code is "US". U.S. applicants were assigned to five different sectors:

- Government
- Private
- Academic
- Individuals
- Others

Automated coding was used to assign non-ambiguous forms of applicant names (e.g., "Univ" in the academic sector, "inc." in private) to the corresponding sector. After this first matching step, manual coding was performed to assign the remaining applicants' names that could not be automatically assigned. Coding forms extracted from the SEI 2020 exercise were also used to help during the coding exercise. In the end, tests were performed to ensure that distinct forms appearing in the database were always coded under the same sector, ensuring the absence of any ambiguous decisions. Of all U.S. addresses, 99.7% could be assigned a sector, the remaining cases being listed under a sixth sector, "Unclassified".

The academic and government sectors have far lower patenting output than the private sector. Because it was important for the SEI report to have accurate output estimates for these two sectors, Science-Metrix prioritized the crediting of patents to the academic and government sectors in the rare cases of multiple matches. If these sectors had not been prioritized, it is believed that slightly inaccurate and lower estimates of patenting activity for these two sectors would have been obtained because these few cases, although almost unnoticeable at the level of output measured for the private sector (i.e., about 151,000 patents in 2020), still represent a sizable number of patents at the level of the government and academic sectors (i.e., about 1,200 and 7,800 patents in 2020, respectively). Also, because many applicants were assigned to both sectors because of university-affiliated companies, this guided the decision toward prioritizing the academic sector when dual assignments with the private sector were detected. Although

this decision resulted in a slight bias in favor of the academic and government sectors over the private sector, this bias is in the end negligible when considering the levels of output measured for the private sector (i.e., less than 0.05% difference for the private sector).

Manual validation of the sector coding was performed on a random sample of 100 U.S. addresses, resulting in a precision level of above 99%. Similar levels were observed with samples focusing on the five main categories individually, ensuring the precision of the results reported for each sector. A similar test was performed looking at the 0.3% of all addresses that could not be classified. Overall, most categories were represented in accordance with their expected frequency based on occurrences in coded addresses, the only notable difference being the small over-representation of the "Others" sector in unclassified addresses. The "Others" sector represents 0.41% of all addresses in the database, but around 4% of all unclassified addresses. Yet, because unclassified addresses account for such a small number of cases, correcting for this does not change the proportion of addresses coded under the "Others" sector in the United States, because correcting for this would only add about 120 publications to this sector (or 0.006% of all publications).

External File 5: US applicant to sector

2.5 Indicators related to utility patents

This section presents the patent indicators computed as part of this study. In the SEI 2020, patent counts were prepared based on utility, plant, and design patents; in contrast, only patent counts based on utility patents were prepared for the present edition.

2.5.1 Inventors versus applicants

Most of the indicators prepared for this project using utility patents are based on data pertaining to inventors. Science-Metrix assigned country and state affiliations to addresses on patents linked to the inventors (not the organization owning the rights on the patents, i.e., applicants/assignees). Statistics based on sectors were prepared using information on applicants because the coding of sectors of activity requires assigning organizations to their corresponding sector (e.g., a university to the academic sector, a company to the private sector), and there is no information available on inventors' affiliation. To avoid any potential confusion between both concepts, footnotes below the delivered statistics tables always clearly indicate whether the data presented are based on inventors or applicants.

In cases where information on applicants was not available, the information on inventors was used to assign patents to countries or regions, assuming that these individuals owned the patents.

2.5.2 Applications versus granted patents

All the statistics related to utility patents were based on granted patents. One important distinction between patent applications and patent grants is the considerable time lag between the two. While an application is made closer to the time of invention, the granted patent is closer to the commercial return of the invention. Useful and complementary statistics can be derived from both approaches. However,

several limitations in the quality of data on applications reduce their potential for the development of indicators. This is particularly true for U.S. applications, and Science-Metrix usually tries to avoid producing statistics for these. There are two main reasons for this:

- Applicants can ask that the application not be published.[15] Currently, only about 70% of patent applications are published. This proportion varies by type of industry, Patent Cooperation Treaty (PCT) versus non-PCT, size of company, country, and over time. Science-Metrix is not aware of any statistics on these variations. Importantly, once patents are granted, applications become public. So, this subsequently adds to the number of applications that were made public at the moment of application. Therefore, the exact number of applications for a given year is not known until at least 7–8 years later because of the time lapse between application and grant. These results have at least two implications: (1) statistics are always incomplete in more recent years, and (2) because of the variability in application-to-grant time, statistics for the most recent years are biased.
- The quality of data for applications is poor. Several applications do not have any information on the country and/or the state and/or the applicant name and/or the U.S. class. This information is sparse, and the quality varies from one provider to another. For instance, PatentsView appears to only have information regarding applications of granted patents.

2.5.3 Number of utility patents

Full and fractional counting are the two principal ways of counting the number of patents.

Full counting

In the full counting method, each patent is counted once for each entity listed in the address field (either for inventors or applicants depending on the statistic being prepared). For example, if two inventors from the United States and one from Canada were awarded a patent, the patent would be counted once for the United States and once for Canada. The same method applies for applicants. If a patent is assigned to Microsoft in the United States, IBM in the United States and Siemens in Germany, the patent will be counted once for Microsoft, once for IBM and once for Siemens. It will also be counted once for the United States and once for Germany. When it comes to groups of institutions (e.g., research consortia) or countries (e.g., the European Union), double counting is avoided. This means that if inventors from Croatia and France are co-awarded a patent, when counting patents for the European Union this patent will be credited only once, even though each country has been credited with one patent count at the country level.

[15] A few thousand patents cannot be accounted for because of the *Invention Secrecy Act* of 1951, which prevents disclosure of technologies presenting a possible threat to national security. However, given that both the granted patent and the application of these inventions are blocked from publication, this does not impact the decision related to the selection of applications or granted patents for the preparation of patent counts.

Fractional counting

Fractional counting is used to ensure that a single patent is not counted several times. This approach avoids the use of total numbers across entities (e.g., inventors, organizations, regions, countries) that add up to more than the total number of patents, as is the case with full counting. Ideally, each inventor/applicant on a patent should be attributed a fraction of the patent that corresponds to his or her level of participation in the invention process compared to the other inventors/applicants. Unfortunately, no reliable means exists for calculating the relative effort of inventors/applicants on a patent, and thus each is granted the same fraction of the patent.

For this study, fractions were calculated at the address level for the production of data based on inventors. In the example presented for full counting (two inventors with addresses in the United States, one inventor located in Canada), two thirds of the patent would be attributed to the United States and one third to Canada when the fractions are calculated at the level of addresses. Using the same approach for applicants in the other example (one address for Microsoft in the United States, one for IBM in the United States and one for Siemens in Germany), each organization would be attributed one third of the patent.

2.6 Indicators related to worldwide priority patents

Patent counts based on worldwide priority patents using patent families were also prepared, accompanied by specialization indexes based on these counts. The specialization index is computed using fractional counts indicating the level of involvement across categories, in this case technical fields. By definition, an entity cannot be specialized across all technical fields. Readers can find more details regarding the specialization index in the methodological report dedicated to bibliometrics.[16]

[16] https://www.science-metrix.com/bibliometrics-indicators-for-the-science-and-engineering-indicators-2022-technical-documentation/

3 Trademark indicators

In a spirit of broadening the scope of the SEI beyond traditional metrics based on patents, a decision to include statistics on trademarks in the SEI 2020 was reached by the NSF after consulting material prepared by Science-Metrix demonstrating the coverage of the data available. This decision was made possible by the recent addition of data sources covering trademark data, which were not available in the past. Science-Metrix prepared statistics using trademarks data from the USPTO for the SEI 2022.

3.1 Building databases

One database covering USPTO trademarks was built to prepare statistics on trademarks. XML files containing data are freely available online[17] and were downloaded by Science-Metrix. Science-Metrix built in-house versions of these databases covering a selection of fields essential to the preparation of the statistics:

- Addresses of trademark holders (to assign trademarks to countries, regions, and U.S. states)
- Names of holders (for sector analysis)
- Nice categories of goods and services (for comparison across categories)
- Registration year

3.2 International classification of goods and services

The international classification of goods and services, also known at the Nice classification, is a system used to register trademarks across categories of goods and services. It was adopted in 1957 following the Nice Agreement and comprises 45 classes. Classes 1 to 34 cover goods and 35 to 45 cover services.[18] The system operates in close to 90 countries as of 2020, with an additional 65 non-member countries using the classification.

3.3 Indicators related to trademarks

Around the end of 2018, the NSF requested a memo from Science-Metrix detailing which indicators could be prepared using trademark data. Below are the indicators that were selected for inclusion in the SEI 2022:

- Number of registered trademarks (USPTO), by region, country, or economy
- Number of registered trademarks (USPTO) for the U.S., per Nice categories of goods and services
- Number of registered trademarks (USPTO), by region, country, or economy, per industry sector (as defined by a mapping of Nice classes provided by Edital, a company specializing in trademark information)

[17] USPTO: https://bulkdata.uspto.gov/

[18] For details about the 45 categories: https://www.wipo.int/classifications/nice/nclpub/en/fr/

4 USPTO patents and trademarks at U.S. county level

Science-Metrix has been commissioned by SRI International, on behalf of the National Science Foundation (NSF), to develop measures and indicators of patent and trademark activity at the level of U.S. counties. The envisioned purpose of these data was to investigate regional patterns of innovation at more refined levels than previously available in the SEI reports. The SEI reports have been covering patent data for decades and introduced trademark data in the 2020 edition, but were limited to country and U.S. state data in its geographical scope. The provisioning of these data now enables the production of additional analyses. This section of the technical documentation details the various steps taken to implement the databases, sanitize and standardize the data, and produce statistics at the requested geographical level—according to both U.S. inventors and U.S. assignees across all USPTO utility patents covering patents grants between 1996 and 2020, and for all U.S. trademark owners on all registered USPTO trademarks over the same period.

4.1 Patent indicators

4.1.1 USPTO

The patent indicators for the U.S. market in this report were produced using an in-house implementation of the PatentsView patent database, a platform derived from the USPTO bulk data files that is further enriched by the PatentsView team with additional data treatment regarding names of inventors and assignees, geocoding of U.S. addresses and more. A version of the database was built in Databricks and was carefully conditioned for the production of large-scale comparative patent analyses. This database covers not only utility patents, which were the sole focus of the patent analyses in the SEI 2022, but also design patents and plant patents, which were covered in the SEI 2020.

4.1.2 Data limitations

There is no notable limitation regarding the USPTO data to be reported, because they provide complete coverage of U.S. patent grants. One note though is that small issues regarding some of the PatentsView tables have been detected in the past, which should not come as a surprise considering the complex endeavor of preparing such an exhaustive database. The PatentsView team has always been responsive when such issues have been identified and have corrected the data promptly.

One notable limitation is linked to the format of U.S. addresses as they appear on USPTO patents. The address format is mostly limited to U.S. states and U.S. cities, without more precise information that would be quite helpful for this geocoding exercise, such as zip codes, street names and street numbers. Although it is still possible to obtain a robust geocoding of U.S. counties using only state and city information, this adds a layer of uncertainty to the matching in cases where multiple cities share the same name in a given state, or for large cities encompassing multiple counties, or any cities overlapping multiple counties. Details about these limitations are addressed later in the report.

4.1.3 Kind codes

Kind codes are a classification system used across patent offices to classify document types. Each patent office has its own classification system; although codes are often similar across offices, their implementation may differ across offices.

For the SEI 2022, USPTO kind codes were used to identify granted utility patents from the USPTO. The same selection of kind codes was used for the regionalization at the level of U.S. counties.

4.1.4 Databases

PatentsView

Most of the patent analyses in *Indicators* were prepared using data from the USPTO indexed in PatentsView. The database provides details on patents, such as full titles and abstracts, the country and state (when available) of the inventors and applicants, as well as names of the inventors and applicants. In most cases, applicants are organizations, although they are sometimes individuals when the patent is not assigned to any organization. Federal Information Processing Standard (FIPS) codes for U.S. counties are also available in the data,[19] and at a relatively high frequency, with around 90% of all U.S. patents being assigned a county in the data. However, this high level is not spread equally across the U.S., as only a little more than 50% of the approximately 40,000 distinct U.S. addresses in the database are assigned a county FIPS code, a reflection of the high imbalances observed in the U.S. regarding patent output. Additionally, PatentsView does not allow for multiple county assignments per address, which is sometimes expected given that patent data only contain state and city information. This can become especially problematic in the case of large cities, which are assigned to a single county in the data but should theoretically be linked to multiple counties given the uncertainty regarding the assignment (e.g., New York City is always forced under county FIPS code 36061 of New York County).

The database also provides information on three classification schemes: the U.S. national classes (the USPC classes, although these are not available after 2015 as the system is no longer in use), the World Intellectual Property Organization's (WIPO) International Patent Classification (IPC), and the Cooperative Patent Classification (CPC). The CPC was produced in partnership between the USPTO and the EPO; it replaced the USPC classes after 2015, and the European Classification System (ECLA) after 2012. PatentsView is suitable for the production of technometric data dating from 1976, whereas patent data in the previous round of the SEI were largely prepared for the period 1996 to the present.

PatentsView tables were downloaded and uploaded into Science-Metrix's Databricks environment. The process is straightforward, does not require any treatment because the data are already parsed, and was

[19] Although Federal Information Processing Standards are not the norm anymore regarding geographic codes in the U.S., the American National Standards Institute (ANSI), which took over from the National Institute of Standards and Technology (NIST), still continues to issue the commonly used FIPS codes.

fully automated from online downloads to final tables. Documentation presenting the content of the tables is available on the PatentsView website.[20]

4.1.5 Data standardization

Although the main scope of this project is to prepare data at the level of U.S. counties, these might be further presented across multiple additional dimensions, including technical fields of the WIPO classification scheme and by U.S. sectors. More details about these dimensions can be found in the technical documentation of the SEI 2022 report.[21]

The focus of this technical documentation is to present the approach implemented to geocode addresses of U.S. inventors and assignees appearing on USPTO granted patents. The following sections detail the different steps of the process.

Mapping U.S. cities to U.S. counties using a mapping scheme between cities and counties

The main limitation regarding the geocoding of USPTO patent data at the level of U.S. counties is the limited completeness of U.S. addresses as they appear on patents. With only U.S. state and city information available, it is to be expected that some ambiguity in the geocoding process will arise. This ambiguity has already been reported by previous works reporting on the geocoding of U.S. addresses to U.S. counties—for instance, with the USPTO Patent Technology Monitoring Team (PTMT) managing to geocode U.S. addresses to U.S. counties.[22] Their results have been reproduced independently by Carlino et al.,[23] both efforts finding that it was possible to geocode more than 95% of all U.S. addresses to at least one U.S. county. Of these, only about 12% were assigned more than one U.S. county, and further work reaggregating these data at the level of MSAs further decreased the percentage of co-assigned addresses to only 2%.

For this project, an approach like those developed by the PTMT and Carlino was implemented. While the PTMT used a U.S. Post Office reference file to match cities and states of residence of inventors to U.S. regional components,[24] Science-Metrix identified a more recent reference file from the U.S. Census Bureau that linked place names with U.S. counties.[25] This list includes 41,414 entries with the following parameters:

- U.S. state
- U.S. state FIPS code
- Place name
- Place name FIPS code

[20] https://patentsview.org/download/data-download-dictionary

[21] https://www.science-metrix.com/bibliometrics-indicators-for-the-science-and-engineering-indicators-2022-technical-documentation/

[22] https://www.uspto.gov/web/offices/ac/ido/oeip/taf/countyall/explan_countyall.htm

[23] https://core.ac.uk/download/pdf/6887989.pdf

[24] The working paper by Carlino et al. does not details the source for the matching of U.S. cities to counties.

[25] https://www2.census.gov/geo/docs/reference/codes/PLACElist.txt

- Type of place (i.e., census designated place (DCP), incorporated place, county subdivision)
- County name

Contrary to the geocoding available in PatentsView, this list includes co-assignments, with place names sometimes being linked to even more than two counties. For instance, the entry for "New York City" is rightfully linked to its five constituting counties (i.e., Bronx county, Kings County, New York County, Queens County, Richmond County) as the information is not discriminant enough to select one of the five counties.

This mapping file from the U.S. Post Office acted as the main reference for the geocoding of U.S. cities in patents to U.S. counties. A simple, multi-step geocoding approach was implemented to assign U.S. addresses based on the state and city information available on both sides, starting with an exact match without any data treatment, and moving from this point to detect missed cases and develop the algorithm to further increase the coverage of the mapping process. Overall, about 15 steps were implemented to increase the rate of matched addresses, with the main corrections applied listed below:

- Place names in the reference file appear with their place types (e.g., city, town, township, charter township, village, CDP), whereas this is not often the case in the USPTO data. Most steps were dedicated to matching the data after removal of these place types and correcting for some specific cases identified by selecting combinations of state and city names not yet matched after each new step (e.g., Boise's namesake in the reference file appears under "Boise City city", which was not detected in the original steps).
- The final step of the process is a manual geocoding one for the remaining addresses based on the highest frequency counts using Google Maps and ArcGIS online maps.[26] Most of the place names not matched were smaller units of cities (e.g., neighborhoods) or unincorporated places, which are not covered in the reference file.

Overall, the initial matching steps before manual coding enabled the geocoding of about 94% of all U.S. addresses to at least one U.S. county. About 38,000 combinations of U.S. states and cities, accounting for about 6% of all patents, remained unassigned prior to the manual step, but geocoding a little more than 70 of these combinations increased the coverage of the geocoding to about 97% of all U.S. patent counts.

At the end of the matching process, 98.7% of all patents associated with U.S. applicants (98.2% for U.S. inventors) were assigned at least one U.S. county, and about 14% of these patents of U.S. applicants were assigned more than one county (12.8% for U.S. inventors), resulting in about 86% of all U.S. applicants' patents unambiguously assigned to a single county (87.2% for U.S. inventors' patents). These results are highly similar to those reported earlier in this documentation based on the works of the PTMT and Carlino.

One notable fact regarding the geocoding process is that, at first, a sequential mapping process was implemented, with the matched entries being removed from the pool so that the new steps only considered the remaining cases. However, because some cities share the same name (e.g., there is an

[26] https://hub.arcgis.com/datasets/esri::usa-counties/explore

Abbeville city in Alabama, Georgia, Louisiana and South Carolina), and some different places become identical when removing place types (e.g., Aberdeen town in North Carolina, Aberdeen township in New Jersey, Aberdeen village in Ohio), using a sequenced process could have led to biases for entries that were matched first when ambiguity remained for entries within the same state (e.g. there are five "Wilson town" in Wisconsin and one "Wilson village"; matching first on "town" would remove the opportunity to map to "Wilson village" in cases where only "Wilson, WI" is stated on patents). Therefore, the sequential mapping was replaced by the process described earlier, where each entry is tested at each step, and the result of all steps is considered at the end, allowing for multiple assignments when needed. To diminish co-assignment in cases where one matched county appeared much more probable than the others, manual checks were performed for the entries with co-assignment presenting the highest counts, and corrections were made accordingly. For instance, "Mountain View, CA", was first assigned to Santa Clara county and Contra Costa county, the namesake being held by a city of 75,000 inhabitants in Santa Clara county and a census designated place of about 2,500 inhabitants. It was deemed safe to assume that most of the output under this city tag would come from Santa Clara county, thus all patent output was given to Santa Clara county in that case. This also avoided drastically overestimating Contra Costa's output if the output was split equally across both counties.

Distribution of ambiguously assigned patents across counties and CBSAs

Although the proportion of ambiguously assigned U.S. patents is relatively low, at about 14%, this is nevertheless non-negligible. To account for this, a redistribution of the counts of the ambiguous cases was performed. At first, we envisioned redistributing the output following the proportions observed in the population that could be assigned unambiguously. However, it quickly became clear that doing so would lead to highly unreliable results. Indeed, for cities spread across more than one county, their output would in fact be redistributed based on the patent counts associated with these counties, based on mappings of other cities, which is not at all representative of the weight each county might have within these cities. For instance, if output for entries tagged "New York City" were to be redistributed across its five counties based on the level of output from each county, more declarative borough names, for which unambiguous assignment could be performed, would receive a larger share of the total output from the city.

Instead, it was decided that in cases of ambiguous assignment, each county would receive an equal share of the output from an entry, similarly to what is done by the PTMT team. This redistribution, although imperfect, should nevertheless be less biased than the other suggested approach. It is to be noted that co-assignment, when counties are reaggregated at the MSA level, drops to less than 4%, as most of the ambiguity in the mapping process comes from highly populated cities encompassing multiple counties.

Validation of U.S. county geocoding

To ensure that the analyses prepared during this project were of the highest possible quality considering the limitations associated with the data, manual and automatic validation was performed to check for the validity of the data obtained. A manual sampling approach checking for a sample of 200 U.S. addresses on U.S. patents was manually validated by analysts, looking for these addresses in Google to identify if

the county (or counties) assigned by the geocoding process were correct. This sampling approach enabled the computation of a global precision score for the process, which stood at 98%.

Alignment of the data with existing data sources

Two notable datasets with patent counts per U.S. counties were presented at the beginning of this report, the set from the USPTO PTMT and the data from Carlino's working paper. Since Carlino et al. mentioned in their paper that they compared their data with those from the PTMT and that they were highly similar, it was decided that data from the current exercise would only be compared with those from the PTMT as they are easily accessible online and in a more suitable format than those from Carlino's paper.

To make the comparison, since definitions of U.S. counties evolve over time with new censuses, it was important to ensure that the definition of U.S. counties used for the validation was the same as the one used by the PTMT. After inspection of the documentation associated with the PTMT data available online, it appeared that the definition was extracted from a file distributed to the public in March 2011 and based on U.S. Post Office information acquired from a private vendor. Because this date is after completion of the latest census in the U.S., a direct comparison was performed between the data prepared for this project and those from the PTMT available online for the 2000–2015 period. Overall, the comparison performed demonstrated that the findings of the current project were aligned with those from the PTMT, reinforcing the assessment of robustness of the data prepared. Some discrepancies were observed for a few counties, which is to be expected given that some cities overlap with multiple counties, and Science-Metrix has no way to identify exactly to which county all cities were mapped by the PTMT.

As a final step in the validation of the data, a triangulation with the geocoding available in PatentsView was also performed. Although PatentsView does not provide co-assignment of U.S. addresses to multiple counties, it was still possible to perform a comparison, checking that the non-ambiguously assigned cases from the match were linked to the same county in the PatentsView data, and that the cases that were assigned multiple counties had been assigned at least the single county available in PatentsView. Again, after performing this exercise, a high level of agreement between each set was detected, further confirming the quality of the match performed.

4.2 Trademark indicators

4.2.1 USPTO

The trademark indicators for the U.S. market in this report were produced using an in-house implementation of the USPTO bulk download trademark database.[27] The in-house version of the database was built on Databricks and was carefully conditioned for the production of large-scale comparative trademark analyses. This database covers USPTO trademarks, which have been covered in the SEI since the SEI 2020.

[27] https://developer.uspto.gov/data

4.2.2 Data limitations

Much like with patent data, there is no notable limitation regarding the USPTO data to be reported, because they provide complete coverage of U.S. trademarks. In fact, in comparison with patent data, USPTO trademark data are better suited to the geocoding exercise as addresses are much more complete, including not only U.S. states and U.S. cities, but zip codes, street names and street number as well. These more complete addresses are highly useful as they make it possible to differentiate cases that would be ambiguous if only cities were available, as is the case for patent data. Therefore, it was expected prior to the matching process that the percentage of U.S. patents assigned to more than one county would be drastically lower than the 14% measured for patents, and results presented later demonstrate that it is indeed the case.[28]

4.2.3 Databases

USPTO Bulk download files

The trademark analyses in *Indicators* were prepared using data from the USPTO bulk download files available online. The XML files, which were used to build an in-house production database, provide details on trademarks such as mark names and full addresses of the holders of the marks, in addition to their names (either of individuals or organizations owning the trademark). In most cases, holders are organizations, although about 10% of trademarks are owned by individuals. Contrary to PatentsView, which contained county geocoding through its enriched content, no geocoding at the level of U.S. counties is available for these files. Still, because of the more complete address data, the trademark data are much more suitable for a geocoding exercise.

To build the in-house version of the USPTO trademark database, Science-Metrix uploaded all the XML files from the USPTO website and reused a parser designed during the work performed for the SEI 2020 to extract the information needed and include these into Science-Metrix's Databricks environment. The process was straightforward and did not require any additional data treatment because the data parser is already complete.

4.2.4 Data standardization

Although the main scope of this project is to prepare data at the level of U.S. counties, these data could be further presented for multiple additional dimensions, including product and service categories from the Nice classification. Methodological details about these dimensions can be found in the technical documentation of the SEI 2022 report.[29]

[28] While there were existing sources against which to benchmark the geocoding of patent data, none were detected for trademark data, which added a layer of uncertainty regarding expectations for the process. However, because of the more complete address format available for trademarks, it was deemed safe to assume that the matching would be at least as good as the ones observed in the literature for patent data.

[29] https://www.science-metrix.com/bibliometrics-indicators-for-the-science-and-engineering-indicators-2022-technical-documentation/

The focus of this technical report is to present the approach implemented to geocode addresses of U.S. owners appearing on USPTO registered trademarks. These following section details the different steps of the process.

Mapping U.S. addresses to U.S. counties using a mapping scheme between zip codes, cities, and counties

As reported earlier, the main limitation regarding the geocoding of USPTO patent data at the level of U.S. counties was the limited scope of U.S. addresses as they appear on patents. With trademark data, this is not a problem anymore as most U.S. addresses are complete with state, city, zip code and even street information. This greatly reduces the number of trademarks co-assigned to more than one county since it is possible to precisely geolocate each address.

For this project, an approach similar to the one presented for the patent data was implemented. Science-Metrix used the same reference file from the U.S. Census Bureau, which linked place names with U.S. counties to geocode U.S. trademarks, allowing for co-assignments when city names were not discriminant enough to identify a single county using the same multi-step geocoding approach. Again, a manual step was performed to geocode the remaining addresses based on the highest frequency counts using Google Maps and ArcGIS online maps.

Overall, the initial matching steps before manual coding enabled the geocoding of about 94% of all U.S. addresses to at least one U.S. county. Geocoding a little more than 70 of these remaining state and city combinations increased the coverage of the geocoding to about 97% of all U.S. trademark counts. When dealing with patent data, this is where the matching process needed to be stopped, because all the available information had been used. However, zip codes are available for trademark data, so another round of matching was performed, this time using a crosswalk file between zip codes and U.S. counties, as defined in the 2010 Census, from the U.S. Office of Policy Development and Research of the Department of Housing and Urban Development.[30] This step provided an additional set of potential U.S. counties for each U.S. addresses, which could be tested against the mapping obtained at the city level. In cases where the city mapping was ambiguous, priority was given to non-ambiguous matches using the zip codes. Following the geocoding using zip codes, the level of matching reached a high of 98.8%, with no state presenting rates below 98%. In the end, only about 2.6% of all trademarks were assigned to more than one county. These results are highly similar to those reported for the patent data, except that co-assignment levels are much lower due to the more complete address format in the trademark data.

Distribution of ambiguously assigned trademarks across counties

Similar to the approach taken for patent data, an equal redistribution of the counts of the remaining ambiguous cases was performed. This redistribution, although imperfect, was performed on an extremely small proportion of all trademark addresses and should still provide robust data.

[30] https://www.huduser.gov/portal/datasets/usps_crosswalk.html

Validation of U.S. county geocoding

To ensure that the data prepared during this project were of the highest possible quality considering the limitations associated with them, manual and automatic validation was performed to check for the validity of the data obtained. A manual sampling approach checking for a sample of 200 U.S. addresses on U.S. trademarks was manually validated by analysts, looking for these addresses in Google to identify if the county or counties assigned by the geocoding process were correct. This sampling approach enabled the computation of a global precision score for the process, which stood at 98%.

Alignment of the data with existing data sources

No data source for USPTO trademark counts at the level of U.S. counties were detected during the literature review performed at the onset of this project. Therefore, it was not possible to compare the data prepared with an external source, as was performed for the patent data. Nevertheless, given the high level of agreement with other sources observed for the patent data, and the fact that trademark addresses are much more complete than those on patents, it is expected that the precision obtained for the mapping is high and that the results prepared are reliable and reproducible if other organizations tried to perform a similar exercise.