Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2026

Technical Documentation

April 2026

Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2026

Technical Documentation

April 30, 2026

Submitted to:
SRI International

Author
Guillaume Roberge

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 ... i
1 Introduction ... 1
2 Patent databases ... 2
2.1 Data limitations ... 4
Kind codes ... 7
2.2 Database ... 7
2.3 Data standardization ... 8
2.3.1 Mapping of patents by technical fields ... 8
2.3.2 Mapping of patents under Critical Technologies of the CHIPS and Science Act ... 10
2.3.3 Linking citations to non-patent literature to the bibliometric database ... 14
2.3.4 Data standardization: country, country groups, regions ... 16
2.3.5 Data standardization: U.S. states ... 17
2.3.6 Data coding: U.S. sectors ... 18
2.3.7 Data coding: U.S. counties ... 19
2.4 Indicators related to USPTO utility patents ... 23
2.4.1 Inventors versus applicants ... 23
2.4.2 Applications versus granted patents ... 23
2.4.3 Number of utility patents ... 24
2.4.4 Patent counts, publication output and patent citations related to Federally Funded Research and Development Centers (FFRDCs) ... 25
2.4.5 Top cited organizations patenting at the USPTO ... 25
2.5 Indicators related to worldwide priority patents ... 26
2.6 U.S. federal agencies mentioned in funding acknowledgments ... 28
3 Trademark indicators ... 30
3.1 Data limitations ... 30
3.2 Database ... 30
3.3 Data standardization ... 31
3.3.1 International classification of goods and services ... 31
3.3.2 Data coding: industry sectors ... 31
3.3.3 Data coding: U.S. counties ... 31
3.4 Indicators related to trademarks ... 33
4 Annex ... 34

Tables

Table I PATSTAT country coverage fact table of national patent authorities with known issues for counts of priority patent families 5
Table II WIPO classification scheme for the production of SEI patent indicators 9
Table III Example of a patent fractioned by technical fields according to IPC codes, following conversion from CPC codes 10
Table IV Most frequent 2-grams in patent reference strings 15
Table V Example of normalized patent family to patent family citation using a citing-side approach 27
Table VI Definition of industry sectors in trademark data 31

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) 2026. 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 trademarks from the same office, as well as granted patent families covering patents from dozens of patent authorities. The work done for the bibliometrics aspect is presented in a separate document.[1] 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

External Files 1, 3, 4 and 5 are available in an Elsevier Data Repository.[2] These external files are also introduced in the relevant sections of this documentation. In addition, Databricks notebooks created to download and prepare the patent and trademark databases for the project, and to generate the main indicators on patents and trademarks, are also accessible in an Elsevier Data Repository.

[1] https://www.science-metrix.com/bibliometrics-indicators-for-the-science-and-engineering-indicators-2026-technical-documentation/

[2] https://elsevier.digitalcommonsdata.com/datasets/xdg7w7vnf9

2 Patent databases

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. The technical team created an automated process to download data files from the PatentsView website and built from these files an in-house version of the database in Databricks carefully conditioned for the production of large-scale comparative patent analyses based on utility patents. The PatentsView snapshot released in March 2025 serves as the foundation for the USPTO patent indictors for SEI 2026.

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 itself, 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,[3] 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 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 mainstream 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 provided for the SEI 2026. These metrics help alleviate the home advantage when measuring innovation only within 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.[4] This method uses information within patent families to fill in gaps regarding inventorship for patent offices where data

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

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

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 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 work 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 this one 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 United States 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 the 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). Although this number may not seem problematic, because most of these artificial patents are not evenly distributed across patent offices and are overrepresented for the U.S. and some European markets, their exclusion leads to 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.[5]

To address the limitation related to artificial priority patents, instead of discarding patent families that initially originated through these artificial priority patents, Science-Metrix selected the first utility patent within the family that was applied for after the artificial priority patent and used it as a replacement to

[5] 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.

serve as the first priority patent 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 Rassenfosse. Implementing this correction provides more complete information for a large number of countries worldwide, with the biggest effect on countries whose national patent offices allow provisional patent applications or where inventors often seek to patent in markets where these are available.

2.1 Data limitations

There is no notable limitation regarding the PatentsView USPTO data because they provide complete coverage of U.S. patents. Science-Metrix performed data-quality checks after downloading the content to ensure that there were no issues with the content, either from the original source files or because of issues during data processing. Comparisons were made against official statistics provided by the USPTO[6] and by comparing with data prepared for the previous SEI.

One notable issue arises when trying to geocode addresses to U.S. counties, which is linked to the format of U.S. addresses as they appear natively on USPTO patents. The address format is mostly limited to U.S. states and U.S. cities, lacking more precise information that would be helpful for geocoding, such as zip codes, street names and street numbers. Although it is 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 in which 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 will be addressed later in the report.

While there are few limitations linked to the USPTO data, the same cannot be said for some patent authorities available in PATSTAT data. Even though PATSTAT's coverage is quite expansive, data quality varies across patent offices, since the EPO relies on the offices to provide complete and high-quality data. 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. Patent authorities with notable data issues are detailed in Table I. Two cases especially stand out and are further described here: India and Italy.

[6] https://www.uspto.gov/learning-and-resources/statistics

Table I PATSTAT country coverage fact table of national patent authorities with known issues for counts of priority patent families

Country	Type of explanation	Explanation
Algeria	Data gap	Data not shared fully with EPO since 2004
Austria	Fluctuations with valid explanation	A more than 50% year-on-year growth is observed in 2006, which is in line with other reported findings for that period (https://www.wipo.int/edocs/pubdocs/en/wipo-pub-rn2024-40-en-country-perspectives-austria-s-journey.pdf)
Belgium	Fluctuations with valid explanation	A more than 50% year-on-year growth is observed in 2016, which is in line with the reported increase of granted patents at the EPO for that period (https://report-archive.epo.org/about-us/annual-reports-statistics/annual-report/2016/statistics/granted-patents.html)
Brazil	Fluctuations with valid explanation	Brazil is experiencing major decreases in terms of granted patents for the last two years, in accordance with findings from the WIPO (https://www.wipo.int/web-publications/world-intellectual-property-indicators-2024-highlights/en/patents-highlights.html). This leads to decreases for the country's priority patent families. Notable drops in processed patents are reported in biotechnology in a recent report (https://phrma.org/blog/the-time-is-now-to-address-brazils-notorious-patent-backlog), in alignment with the observed decreases after 2021 in the SEI data.
Egypt	Data gap	Data not shared fully with EPO since 2013
Guatemala	Data gap	Data not shared fully with EPO since 2019
Honduras	Data gap	Data not shared fully with EPO since 2013
India	Fluctuations with valid explanation	Data from the Indian patent office are mostly incomplete. At the moment, given that the majority of patents granted in India are to foreign entities, this results in a limited impact on India's counts of priority patent families. However, as Indian individuals and corporation keep increasing their share of the national patenting activity, this could lead to an understimation of India's patenting activity in the future.
Indonesia	Data gap	No data
Italy	Data projection + data gap	Information on granted status of patents are incomplete, leading to applications being used instead. This leads to an overestimation of patent activity for Italy across years. For 2024 data, because of the delay to publish patent applications, data are incomplete
Malaysia	Data gap	Data not shared fully with EPO since 2020
Mexico	Fluctuations with valid explanation	The number of granted priority patent families between 2017 and 2017 is higher than usual trends for the country. This aligns with findings from recent report (https://www.lighthouseip.com/wp-content/uploads/2025/01/Trend_report_Patenting_trends_in_Mexico_LighthouseIP.pdf)
Panama	Data gap	Data not shared fully with EPO since 2010
Poland	Fluctuations with valid explanation	An above 60% year-on-year increase is observed in 2016, this aligns with increases in patent activity for Poland around the same period as reported in a separate report (https://www.lighthouseip.com/wp-content/uploads/2022/03/20220301_Patenting-trends-in-European-member-states.pdf)
Singapore	Data gap	Data not shared fully with EPO since 2020
Spain	Fluctuations with valid explanation	A drop of 30% year-on-year is observed from 2001 to 2002, this aligns with a decrease in patent activity for Spain around the same period as reported in a separate report (https://www.lighthouseip.com/wp-content/uploads/2022/03/20220301_Patenting-trends-in-European-member-states.pdf)

Source: Prepared by Science-Metrix

India

India's coverage in PATSTAT is limited, resulting in an underestimation of its output. However, according to WIPO statistics,[7] India is a special case, because historically a majority of applications to its

[7] https://www.wipo.int/edocs/statistics-country-profile/en/in.pdf

national office were made by non-residents. In 2019, only about 37% of patent applications to its national office were made by residents, or 19,500 out of an annual total of approximately 54,000 applications (just 8,800 applications were made by residents in 2010). Foreign companies and inventors have dominated in terms of patent applications at the office (e.g., Qualcomm, Samsung, Huawei, Microsoft, Philips, General Electrics, Ericsson, Mitsubishi, BASF).[8] However, in recent years, the share of applications attributable to India's residents has been increasing sharply, increasing year over year by almost 30% from 2022 to 2023, resulting in a share of applications from India's residents of 55% in 2023.

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 lags behind in processing 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.[9] While changes have been made to address this issue in recent years[10] and the number of granted patents has started increasing, numbers 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.[11] As of 2023, annual granted patents to Indian applicants still stood at a relatively low 27% (9,239 out of 34,124), and globally, of all patents in force, only 19% belonged to Indian patentees.[12]

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. However, because PATSTAT data are based on the Spring 2025 edition, and patent applications may take up to 18 months following initial filing or priority date before it is indexed, patent applications from 2024 are still incomplete, leading to a drop in Italy's output.

[8] https://www.ipindia.gov.in/writereaddata/Portal/IPOAnnualReport/1_94_1_1_79_1_Annual_Report-2016-17_English.pdf
[9] https://www.hindustantimes.com/mumbai-news/india-takes-five-years-to-look-at-patent-applications-reveals-economic-survey/story-q1u11vKeg8lLtPqtdEtniM.html
[10] https://www.majumdarip.com/blog_post/indian-patent-office-shows-trends-of-speedy-grants/
[11] https://www.livemint.com/Politics/LkKhP62yJrhSRJZDoqDIiN/Indias-patent-problems.html
[12] https://ipindia.gov.in/writereaddata/Portal/IPOAnnualReport/1_114_1_ANNUAL_REPORT_202223_English.pdf

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 2026, 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[13] linked to 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.

2.2 Database

PatentsView March 2025

All 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,[14] 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

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

[14] Although Federal Information Processing Standards are no longer the norm 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.

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 U.S. Patent Classification System (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 European Patent Office (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.

Disambiguated forms of assignees are provided in PatentsView in an attempt to unify patent portfolios. A new analysis looking into the most cited patent assignees at the USPTO relies on these curated profiles which can be accessed through table *g_assignee_disambiguated*.[15]

PatentsView tables were downloaded and uploaded into the Science-Metrix AWS S3 and Databricks environments. The process is straightforward and does not require any initial treatment because the data are already parsed. Documentation[16] presenting the content of the tables is available on the PatentsView website.

PATSTAT Spring 2025

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 PATSTAT 2025 Spring Edition, which consists of pre-defined tables with keys linking these together, was built on Databricks and used to prepare statistics on worldwide priority patents. Complete documentation of the database is openly accessible online.[17]

2.3 Data standardization

2.3.1 Mapping of patents by technical fields

In all SEI editions since 2016, patents have been matched with a classification scheme of 35 technical fields[18] developed by the World Intellectual Property Organization (WIPO). The main objective behind

[15] During preparation of the SEI 2026 indicators, PatentsView release a statement indicating data issues with this table. They replaced it with a corrected cut of the data, which was released as table *g_assignee_disambiguated_20241231*. Science-Metrix downloaded the updated table to replace the original one before preparing the indicators on top cited assignees.

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

[17] https://link.epo.org/web/searching-for-patents/business/patstat/data-catalog-patsat-global-spring-en.pdf

[18] Classification scheme from IPC8 codes to technical fields. Available at https://www.wipo.int/ipstats/en/docs/ipc_technology.xlsx

the development of such a classification is to provide a tool for country comparisons.[19] The technical fields defined by this classification are listed in Table II.

Table II 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, is used.[20]

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 of IPC codes 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 adds up to the total number of patents, it is necessary to fraction patent counts by technical

[19] 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

[20] https://www.cooperativepatentclassification.org/cpcConcordances

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 have each of these fields receive half of the patent count. The following example in Table III details this process for one patent.

Table III 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

External File 1: IPC technology concordance table

External File 1 is also available in the Elsevier Data Repository or online at: https://www.wipo.int/ipstats/en/docs/ipc_technology.xlsx

2.3.2 Mapping of patents under Critical Technologies of the CHIPS and Science Act

On August 9, 2022, the "CHIPS and Science Act of 2022" was signed into law, aiming to drive research and innovation in 10 key technology focus areas:

- Advanced computing and semiconductors
- Advanced materials
- Advanced communications
- Advanced energy and industrial efficiency
- Artificial intelligence
- Biotechnology
- Cyberinfrastructure and cybersecurity
- Disaster risk and resilience
- Robotics and advanced manufacturing
- Quantum information science and technology

To support the NSF in reporting on advances in these areas and considering that no broadly accepted and comprehensive mappings of these categories to patent activity existed at the time of date production, Science-Metrix developed a mapping of patents to create sets that could be used to generate patent activity indicators for these key technologies during the cycle for SEI 2024.

First version of Critical Technologies

The Act was not signed until the second half of 2022, so providing data for the key technology areas was not part of the original scope of the SEI 2024. Work on this was only requested a year later, in mid-August 2023, at the end of the originally planned timeline for producing the patent data. Under tight time constraints, Science-Metrix began defining the mapping, initially using seed keywords defined according to priorities related to each of the 10 key technologies, to identify relevant patents through titles and abstracts. While this approach yielded interesting and relevant results, the team was concerned with recall, as the approach missed a notable share of relevant patents. Furthermore, in recent SEI editions, the NSF has moved away from keyword-based mappings, as they can be quite challenging to maintain and update, especially as queries become more complex.

For these reasons, Science-Metrix moved to a second phase, focusing mostly on mappings of CPC codes. This had the advantage of aligning with other mappings used in the SEI and relied upon a well-established mapping system used at both the USPTO and EPO, ensuring its long-term sustainability. While using the CPC system proved to be advantageous, identifying the relevant CPC codes for each key technology area was still a challenge. With over 250,000 million codes, ensuring the capture of all relevant codes, at the correct hierarchical level, while maintaining a low level of false positive results, was not straightforward. As an additional challenge, while the key technologies are, in concept, quite distinguishable, developing an effective definition for each technology was not trivial. To assist with developing these definitions, SRI provided Science-Metrix with literature reviews for each category, filling in templated documents Science-Metrix had prepared to cover most of the relevant information needed to properly delineate the key technologies. Science-Metric and SRI also received internal feedback from other groups at NSF with a vested interest in the mapping, and some of this feedback was incorporated. Working from that basis, Science-Metrix used the original keyword queries, expanding these with additional keywords relevant to the areas identified in the literature reviews, to capture relevant CPC codes according to their definition. Analysts proceeded iteratively, identifying relevant CPC codes, excluding others, and adjusting the seed keyword sets to generate additional CPC candidates for inclusion. At the end of the process, more than 3,000 CPC codes had been identified as relevant to one or many of the key technologies.

To further validate the mapping, our team investigated how we could take advantage of newly emerging generative AI models. Ideally, Science-Metrix would have used generative AI to generate a mapping of all CPC codes, feeding in the information from the literature review and fully mapping the whole patent space, enabling a direct comparison with the human semi-automated approach. However, running the process for half a million candidates was not within the scope and budget of the project, and with limited time, the team had to compromise. Thus, instead of running generative AI on the full set of CPC codes, a decision was made to run it on a sample of about 100 main CPC codes identified by analysts spread across all 10 technologies, and to check whether generative AI models, based on SRI's literature reviews, would identify these codes as relevant. To do so, our analysts created a carefully drafted prompt for ChatGPT 4.0, asking the model to classify under one of the 10 key technologies each of the submitted CPC codes, accompanied by its label definition. The model was asked to return a score ranging from 1 to 10, with 10 being absolute certainty, and 1 being low certainty, and to also provide a short explanation as to why the code was relevant. The model was also asked to classify any non-relevant code under an 11th category to avoid forcing content that was not relevant. Our team was then able to compare ChatGPT 4.0 decisions with those made by analysts. Results demonstrated great alignment between analysts' decisions and the AI model, both mostly agreeing on cases that were unambiguously relevant, and both being similarly ambivalent on non-obvious codes. This proof of concept was enough to cement the existing mapping as the final one for SEI 2024.

New SEI 2026 version

The SEI 2024 approach paved the way for future improvements to the mapping. With more resources available to run the model fully on all CPC codes, it was considered feasible at the time to get a full mapping from generative AI models, potentially identifying additional codes that the current approach might have missed. Indeed, it would have been naïve to think that the SEI 2024 mapping, with more than a quarter of a million potential codes, captured every relevant code, and while analysts focused on CPC codes linked to high patent volumes during validation to ensure that most content was properly captured, there was certainly room for improvement. With more time, combined with quicky diminishing costs to run generative AI models, it seemed reasonable to revisit to further improve on it. Feedback from the key stakeholders was also anticipated, helping to converge on effective definitions as it is expected that even experts in the same disciplines might not agree on some of the inclusions and exclusions to be made.

Over the course of 2024, Science-Metrix worked in partnership with SRI, NCSES and TIP NSF to identify if alignment on the definitions of the Critical Technologies was possible, as TIP NSF had also been working in parallel to develop definitions for the Critical Technologies within the context of award grants. Over the course of multiple months, TIP NSF developed a set of 47 foci which would now define the 10 Critical Technologies, enriched with a set of keywords describing each of these foci. These 47 "definitions" acted as the new definitions to delineate more refined sets of patents.

Since the inception of SEI 2024, advances in GenAI models, in combination with reduced costs related to running these models, has made it possible to be much more ambitious with our use of these approaches to generate this new mapping. Using the "AI factory", a GenAI platform developed internally in the Analytics and Data Services team at Elsevier, we created prompts to label each of the nearly 250,000 CPC levels available for classification. For each focus area, a prompt was processed using the CPC label, which provides a general description of the content, to assess whether the CPC code was pertinent to the focus, based on the category at stake and the set of key keywords developed by TIP NSF. Across all 47 foci, this represented about 12.2 million combinations to treat through the GenAI tool. The model selected was the Mistral24b model, which ended up being the cheapest model deemed to provide robust findings based on manual validation from our team of experts. Other models, such as the llama 8b model, while cheaper, appeared to struggle with classifying CPC labels, often returning positive hits that were a little too broad in subject matter.

Processing a single focus with the Mistral24b model took on average 3 hours, resulting in nearly 150 hours of processing to generate mappings for each of the 47 foci. In addition, two analysts spent approximately 40 hours developing and refining the prompt, optimizing it through trial and error, running the models and checking the results. While this might seem a lot, this is in fact a lot less time than was needed to develop the original 10 set of Critical Technologies for SEI 2024. Having access to GenAI tools made it possible to scale up our methods to define sets of publications, not only for 10 broad categories as was done in 2024, but now for 47 sub-categories.

While assessing precision of the codes identified as relevant by the model was relatively straightforward, with analysts picking random selections of selected CPC codes and related patents and checking against NSF TIP definitions, checking for recall is a more difficult matter. Indeed, for most focus areas, the final selection of CPC codes consists of a few hundred elements at most, meaning that the non-selected codes more often than not covered almost the entirety of all 260,000 CPC codes in the database. To refine the mapping and minimize risk of missing a highly relevant CPC code through the prompt approach, we defined a second layer of testing, which relied on patent level identification. Starting from the principle that critical technologies are highly relevant, actual technologies, we used the 2024 USPTO patent data, the latest year available, and sampled all patents individually with a prompt similar to the one classifying CPC codes, but relying instead on patent titles and abstracts. Each of the more than 320,000 patents was thus prompted under each focus area to determine if it was relevant. The resulting mapping helped analysts identify, among the patents not already returned by the CPC definitions, additional CPC codes that were frequently captured through this patent level approach. While for most focuses, this step did not generate a lot of additional CPC codes for review, for a few this led to pertinent additions. The results thus served two purposes, first by enriching the mapping, and second by validating that the model did not miss much content in the first step.

Given that the costs of classifying individual patents using the AI factory have decreased dramatically in recent months, readers may wonder why the decision was not made to fully classify each patent individually instead of a CPC-to-focus approach. To promote reproducibility by external actors, a patent-level mapping was not considered the best option. Replicating the indicators prepared in the SEI would have required relying on a mapping file that links millions of patents to their respective foci, making this not only impractical, but also dependent on future updates of the mapping as new patents are added. With CPC-to-focus mapping, any team outside of our group can keep updating their data independently with a single, unique mapping file, facilitating updates in the coming years.

As a final validation step, analysts identified all patents from the 2024 sample that were not tagged to any focus area, and among these, computed counts of the most frequent CPC codes on these non-relevant patents. Analysts then went through the list in decreasing order of frequency, manually inspecting the CPC categories to validate that their content was not relevant to any of the 47 foci. Analysts stopped at a frequency of a few hundred cases, thus greatly diminishing the risk that any CPC codes linked to a sizable amount of patents would be missed by the final mapping.

The final mapping can be accessed as Supplemental Table STRN-18 of the Translation Report.

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

This section presents the various tasks performed to link USPTO utility patents with scientific publications, using the references made to scientific publications within the 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 about 40 million 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, and Wikipedia pages. 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. [Removed]; 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, 50th Edition, 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 IV.

Table IV 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 search for hundreds of patterns used in bibliographic referencing were used to retrieve titles, pages, issues, volumes, publication years and journal names and the 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 more than 24 million 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."

Although further improvement to the matching algorithm could be explored in the future, increasing recall without compromising precision will be extremely challenging. This is because the missed cases, which mostly consist of exceptions and unstandardized ways of referencing literature, are hard to catch and will not easily be retrieved.

2.3.4 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.[21] 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).

[21] https://www.uspto.gov/patents/apply/applying-online/country-codes-wipo-st3-table

In the past, 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 included in the United States and Puerto Rico presented separately from the United States. For the 2022 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" displayed in the state information. As a result, all country codes set to "US" for which the state codes were displayed as "PR" were reassigned to "PR", and all country codes assigned to "VI" were replaced with "US," to provide the valid number of patents for both. The newest 2026 edition is aligned with the 2024 edition, with the only notable difference being the list of countries covered in the data, and Puerto Rico moving back under the United States. In contrast with the previous SEI, all countries with data in the databases are included and align with the set of countries provided for the Bibliometrics component of the SEI. This ensures a more comprehensive view of patent output, which is especially relevant for the counts of granted international patent families.

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

External File 3 is also available in the Elsevier Data Repository.

2.3.5 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 U.S. Science-Metrix matched the two-letter U.S. state codes provided in PatentsView to U.S. state names. The total for the U.S. 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 where state information was missing or invalid.

External File 4: Patent number and SEQ to American states

External File 4 is also available in the Elsevier Data Repository.

2.3.6 Data coding: U.S. sectors

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

- Government
- Private
- Academic
- Individuals
- Others

Coverage of the academic and government categories is relatively straightforward, primarily covering publications from universities and governmental institutions respectively. Private is primarily defined as businesses. Individuals covers patents linked to individuals themselves who directly own rights to the inventions. Finally, the Others category covers the remaining cases, including foundations, trusts and other non-academic non-profit entities that do not fit into the other four categories.

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 initial matching step, manual coding was performed to assign the remaining applicants' names that could not be automatically assigned. Coding forms extracted from the SEI 2022 exercise were also used to help. In the end, tests were performed to ensure that distinct forms appearing in the database were consistently coded under the same sector, ensuring there were no ambiguous decisions. Of all U.S. addresses, 99.7% could be assigned a sector; the remaining cases were 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 129,000 patents in 2022), still represent a sizable number of patents at the level of the government and academic sectors (i.e., about 1,200 and 6,600 patents in 2022, 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 approximately 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 U.S., 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

External File 5 is also available in the Elsevier Data Repository.

2.3.7 Data coding: U.S. counties

Since 2022, SEI has been providing data at the level of U.S. counties. The provisioning of these data enables the production of additional analyses at a more granular geographic level. 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 granted between 1996 and 2024, and for all U.S. trademark owners on all registered USPTO trademarks over the same period.

Coding of U.S. counties was prepared using addresses of inventors for whom the country code is "US." U.S. addresses were geocoded to the 3,115 U.S. counties, county-equivalents and planning regions reflecting the latest changes made to definitions as of publication of this report. Details of the approach are presented below.

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, some ambiguity in the geocoding process is to be expected. This ambiguity has already been noted in previous work 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 that developed by the PTMT and Carlino was implemented. While the PTMT used a U.S. Postal Service 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

[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 detail the source for the matching of U.S. cities to counties.

Bureau that links 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
- Type of place (i.e., census designated place (CDP), incorporated place, county subdivision)
- County name

Unlike the geocoding available in PatentsView, this list includes co-assignments, with place names sometimes linked to more than two counties. For instance, the entry for "New York City" is correctly linked to its five constituent counties (i.e., Bronx County, Kings County, New York County, Queens County, Richmond County) as the information is not specific enough to discriminate between the five counties. A more recent file covering county definitions from the 2020 Census was retrieved online in March 2025.[26]

This mapping file from the U.S. Post Office was 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 of the remaining addresses based on the highest frequency counts using Google Maps and ArcGIS online maps.[27] 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.

For Connecticut's Planning Regions, crosswalk files for town names and zip codes were extracted from the GitHub repository "CT-Data-Collaborative."[28]

25 https://www2.census.gov/geo/docs/reference/codes/PLACElist.txt
26 https://www2.census.gov/geo/docs/reference/codes2020/cou/
27 https://hub.arcgis.com/datasets/esri::usa-counties/explore
28 https://github.com/CT-Data-Collaborative/2022-tract-crosswalk

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 just over 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 aspect of 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 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, in which each entry was tested at each step, and the result of all steps were considered at the end, allowing for multiple assignments when needed. To diminish co-assignment in cases in which 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, that name being held by a city of 75,000 inhabitants in Santa Clara county and a census designated place of about 2,500 inhabitants in Contra Costa County. 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 assigned 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, output would be redistributed based on the patent counts associated with those counties, based on mappings of other cities, which might not be at all representative of the weight each county has 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 precise 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 the remaining cases of ambiguous assignment, each county would receive an equal share of the output from an entry, similarly to what was done by the PTMT team. This redistribution, although imperfect, should nevertheless be less biased than the other suggested approach. It is notable 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 despite the limitations associated with the data, both manual and automated validation were performed. A sample of 200 U.S. addresses on U.S. patents was manually validated by analysts, who looked for the addresses in Google to verify if the county (or counties) assigned by the geocoding process was correct. This sampling approach enabled the computation of a global precision score for the process of 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 those data 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. Postal Service information acquired from a private vendor. Because that date was, at the time of performing this exercise, 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 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, we conducted a triangulation with the geocoding available in PatentsView. 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, a high level of agreement between each set was detected, further confirming the quality of the match performed.

Merging of counties to align with population data

Population data at the county level are used to normalize some indicators in the Translation Report. To ensure the population data accurately correspond to the measured patent levels, the definitions of U.S. counties in the patent and trademark data were aligned with modifications proposed by the U.S. Bureau of Economic Analysis (BEA). Most counties remain unchanged, with a few new counties created by merging two existing ones, following BEA guidance. These changes affect only a few dozen counties, spread across the states of Hawaii and Virginia.[29]

2.4 Indicators related to USPTO utility patents

This section presents the patent indicators computed as part of this study. As was the case in the SEI 2022, only patent counts based on utility patents were prepared for the present edition. Patent counts are provided across all USPTO patents, per WIPO technical fields and Critical Technologies, across countries, regions, and U.S. countries.

2.4.1 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.

2.4.2 Applications versus granted patents

All the statistics related to utility patents are 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 on 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.

[29] More details available in the Technical Notes at https://www.bea.gov/sites/default/files/methodologies/BEA-Local-Area-Personal-Income-and-Employment-Concepts-and-Methods.pdf

This is particularly true for U.S. applications, and as a rule Science-Metrix tries to avoid producing statistics for these. There are two main reasons:

- Applicants can ask that the application not be published.[30] 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 between providers.

2.4.3 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.

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

[30] 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 patent 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.

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.

2.4.4 Patent counts, publication output and patent citations related to Federally Funded Research and Development Centers (FFRDCs)

A new addition to the metrics prepared for the SEI are data related to FFRDCs. In recent years, Science-Metrix provided to the National Institute of Standards and Technology (NIST) a set of metrics related to utility patent and scientific publications linked to FFRDCs, in addition to patent citations made to scientific publications from these FFRDCs. For this year, these data were instead requested by the NCSES within the task order for SEI 2026 for inclusion into the data production process for the SEI. While these data had been aligned with the SEI in terms of data coverage and classification, their inclusion into the SEI 2026 process streamlined their production and should ensure broader access to these data.

To retrieve patent output from these FFRDCs, Science-Metrix built a dictionary of names covering not only the different FFRDCS, but also sub-entities that are part of them. Overall, more than 200 entities are included in this dictionary, which can be consulted in the Annex. Automated searches for all these variants were coded using regular expressions (regexes), which are searches of sequences of characters that can be programmed to retrieve occurrences in a text, and are run to retrieve new patents and publications each year. The automated set of rules was validated by comparing the amount of output retrieved with this approach with that retrieved using manual standardization. Overall, both recall and precision were extremely high, at more than 99% in each case, confirming the quality of the set of rules created to retrieve the output of these FFRDCs.

2.4.5 Top cited organizations patenting at the USPTO

Another new addition to the SEI 2026 is the inclusion of an analysis on the most cited organizations at the USPTO. To prepare this analysis, relative scores on highly cited USPTO patents had to be prepared. A normalized field and year weighted citation score was computed for each USPTO patent, relying on the WIPO technology areas for the field normalization and the grant year for the temporal normalization. Because patents may be classified under multiple WIPO technology areas, an approach similar to the computation of Elsevier's field-weighted citation impact (FWCI),[31] which accounts for expected citations in all scientific areas tagged to a scientific publication, was applied to the USPTO patents. After computation of the FWCI-equivalent for each USPTO patent, selection was made of the top 1% most cited per year, based on the highest FWCI. This enabled the computation of the share of patents in the top 1% most cited at organizational level. A share at 1% indicates that an organization is performing on par with expectations. Readers can find more details regarding the relative citation scores on highly cited documents in the methodological report dedicated to bibliometrics.[32]

2.5 Indicators related to worldwide priority patents

Patent counts based on worldwide priority patents using patent families were prepared, accompanied by relative scores on highly cited patent families (top 1%). Readers can find more details regarding the relative citation scores on highly cited documents at Section 2.4.5 and in the methodological report dedicated to bibliometrics.

Citation metrics on the 1% most cited patent families are also an addition to the SEI 2026. One notable difference when dealing with worldwide priority patents is that the whole set of documents in the family is accounted for, and thus to avoid double counting citations, citations are counted on a patent family to patent family basis, deduplicating so that each family gets only one citation from any given family, regardless of the number of elements citing that said patent family.

Additionally, while typical normalizations over year and disciplines are performed (based on first grant year in the family and combination of WIPO technology areas), one dimension that is difficult to control for is the difference in citation practices. While some families are domestic only and thus are only protected by a single patent authority, many encompass patents across multiple offices, each with its own practices regarding patent citations, leading to a wide distribution of average citation counts per patent around the world, with the USPTO presenting with a much higher number of references when dealing with acknowledging prior art. If preferential citations are made to patents from the same office, it could thus be expected that countries with a large share of their output protected in the U.S. market would present with higher citation scores. This effect is somewhat mitigated by the fact that patent references to prior art do not have to be limited to patents from the same patent authority. Still, it is expected that countries with a high proportion of their patenting activity in the U.S market will be strong performers on this indicator, in part due to differences in citation practices.[33] While it would at first appear that a simple additional normalization would solve the issue, such normalization would quickly become quite challenging, having to deal in some cases with multiple patent authorities on a single family, with different application and grant years.

---

[31] https://helpcenter.pure.elsevier.com/en_US/data-sources-and-integrations/field-weighted-citation-impact-fwci-metrics

[32] https://www.science-metrix.com/bibliometrics-indicators-for-the-science-and-engineering-indicators-2026-technical-documentation/

[33] https://www.sciencedirect.com/science/article/pii/S0172219024000577

To address the problem in defining a proper normalization for citation practices on the cited-side of patent families, we instead implemented a citing-side approach, drawing inspiration from existing works in bibliometrics.[34] Instead of only relying on the characteristics of the documents cited for normalization, the citing-side approach relies on the length of the reference list from the citing material, acknowledging that differences in citation patterns come mostly from different practices in the volume of content referenced in a given discipline. In the case at stake here, differences in patent office referencing practices are equivalent to differences in referencing practices across disciplines, and thus we can normalize citations from patent family to patent family by the length of the reference list on the citing side. In typical citing-source approaches, which rely on citing journals, the normalization would be measures as the average length of the reference list of each journal, and citation counts from the citing journal would be normalized by their respective average. In our case, the notion of journals does not exist, and the closest equivalent, the patent offices, would not be easy to implement because compared to journals, there is not a one-to-one mapping, with patent families often falling under multiple offices. Because of this difference, we instead normalize citations at the patent family level, each citation made by a patent family being normalized by the length of the reference list (i.e., distinct patent families) of that same family. Table V presents an example for three different citations

Table V Example of normalized patent family to patent family citation using a citing-side approach

Citing patent family	Cited patent family	Citation	Length reference list of citing patent family	Normalized citation
100000000	200000000	1	3	0.333
100000000	204000000	1	3	0.333
100000000	208000000	1	3	0.333
150000000	208000000	1	6	0.167
150000000	300000000	1	6	0.167
150000000	310000000	1	6	0.167
150000000	320000000	1	6	0.167
150000000	330000000	1	6	0.167
150000000	340000000	1	6	0.167

Source: Prepared by Science-Metrix

As can be observed, a citation is normalized against the number of referenced patent families, becoming the share of the citing patent family that is made to a patent family. For instance, patent family 1,000,000,000 presents with a reference list of three distinct patent families, while patent family 1,500,000,000 presents with 6. Citations from patent family 1,000,000,000 will thus contribute 0.33 citation, while citations from patent family 1,500,000,000, with a reference list of 6 documents, will contribute 0.167 citation. This normalization is critical to control for the drastically different length in reference list across offices, especially the USPTO where applicants and examiners as asked to report on all prior art. In comparison, the EPO only requires examiners to cite the most relevant prior art, leading

[34] https://direct.mit.edu/qss/article/1/4/1553/96120/How-can-citation-impact-in-bibliometrics-be

to 20-times less citation data at EPO compared to the USPTO (14 million versus 267 million citations, respectively).[35]

Following this citing-side normalization, citation scores are further normalized with the typical cited-side approach, against earliest year of patent families and the Critical Technologies, resulting in an average field-weighed citation impact (FWCI) of 1.00 at the world level. For each Critical Technology and year, the top 1% most cited patent families, according to this FWCI, are identified, and these scores are used to measure the share of patents of countries, regions, or economies falling in the 1% most cited. Again, readers can find more details regarding the relative citation scores on highly cited documents at Section 2.4.5 and in the methodological report dedicated to bibliometrics.

While the component on citing-side normalization greatly diminishes issues with country comparison when dealing with patent families across multiple offices, caution is still advised when interpreting scores on highly cited priority patent families as perfectly controlling for all effects at play is a challenging task and there are no set standards to do so currently in the literature, at least not to the level observed in traditional bibliometrics related to publications, which have been in place for decades, or for patent data, which often tend to remain within a single office to circumvent these issues.

2.6 U.S. federal agencies mentioned in funding acknowledgments

Since SEI 2024, data on U.S. papers acknowledging support from U.S. federal agencies has been included. For the purposes of these analyses a "U.S. paper" was defined as being any publication otherwise included in the SEI's database that listed at least one author as being affiliated with at least one U.S. institution.

This subset of papers was then analyzed for the presence of funding information, most of which can be found in the funding acknowledgments section. This section of a peer-reviewed paper is a brief statement at the end of a paper that acknowledges the support, financial or otherwise, for the research that was conducted. The funding acknowledgments section is generally expected to include information such as:

- The name of the funding agency or organization that provided the support.
- The grant number or other identifier that was assigned to the grant.
- A brief description of the type of support that was provided.

Scopus already has text-mining algorithms in place aimed at extracting relevant information from the funding acknowledgments sections of publications, which Science-Metrix knew to be especially good for large-scale funders, and in the U.S. context. As expected, coding precision of a random sample of U.S. papers having received federal funding was remarkably good, with 99% of the coded information being manually confirmed as correct. The sample included a variety of cases, ranging from complete mentions of the funder's name, a grant number, and a description of the support received, down to a single mention of the funder using only the acronym, all of which were correctly handled. It would be normal to expect that precision would be lower when only acronyms are used compared with the use of the full name and

[35] https://wipo-analytics.github.io/handbook/citations.html

a grant number, especially for agencies that share the same acronym. However, this precision benchmarking has also shown that these very partial occurrences are rare compared to complete mentions, which reduces the risk of misattribution.

This is possible because Scopus also sources funding information not only from the acknowledgments section, but also directly from funders through the existing reporting platforms of federal agencies. As a result, some publications with no funding acknowledgments text indexed still have funding information attached.

After preliminary checks, very little room for improvement was identified by Science-Metrix for tagging federally-funded publications. However, some improvements were made to the classification of funders into federal and state levels, and to the hierarchization of parent and children entities.

Indeed, funding data is hierarchical, which means that funders can be organized into parent-child relationships (e.g., the Directorate of Engineering is considered to be a "child" of NSF). In Scopus data, funding attribution is always done to the most precise level possible, and the attribution of funded publications to the parent organizations is not done automatically. To address this and to correctly report counts for parent institutions, Science-Metrix built and used an improved version of the hierarchical information available as part of Scopus funding data to roll up all funded publications for all relevant parents. This way, a publication that acknowledged support from the Directorate of Engineering would also be counted for NSF, and this extends to a parenting chain of any length. However, similar to the case for regional data, if two or more children of a funder are acknowledged by the same publication, this publication will still only be counted once at the level of any parents these entities have in common.

A final important element that must be kept in mind when analyzing the funding data provided as part of SEI 2026 is that Scopus' coverage of funding information has increased with time: 27.2% of all U.S. publications from 2003 have funding information indexed in Scopus while in 2022, this share rises to 67.7%. Therefore, any annual data showing raw volumes over a long period will be affected by this changing coverage rate. However, this improvement in coverage rates has slowed in recent years, and from 2018 onwards, it grew only 6%. It then stands to reason that more recent time trends (the last five years or so) are less likely to be strongly affected by coverage rates.

3 Trademark indicators

To broaden the scope of the SEI beyond patent-based metrics, NCSES decided to include statistics on trademarks in the SEI 2020, supported by Science-Metrix's analysis of available data coverage. This decision was made possible by the recent addition of data sources covering trademark data, which were not previously available. For this new cycle, Science-Metrix again prepared statistics using trademark data from the USPTO for the SEI 2026.

3.1 Data limitations

Much like with patent data, USPTO trademark data provide complete coverage of the U.S. market, with no notable limitations. In comparison with patent data, USPTO trademark addresses are better suited to geocoding as they are much more complete and include not only U.S. states and U.S. cities, but zip codes, street names and street number as well. This greater detail is highly useful as it makes it possible to differentiate cases that would be ambiguous if only cities were available, as is the case with patent data. Therefore, the percentage of U.S. trademarks assigned to more than one county is drastically lower than the 14% measured for patents, standing at only about 2.6%. More details are available at Section 3.3.3.

3.2 Database

One database covering USPTO trademarks was built to prepare statistics on trademarks. XML files containing data are freely available online[36] 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

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 contains county geocoding through its enriched content, no geocoding at the level of U.S. counties is available for these files. Still, due to the more complete address data, the trademark data are better suited to geocoding.

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 it in Science-Metrix's Databricks environment. The process

[36] USPTO: https://www.uspto.gov/learning-and-resources/bulk-data-products

was straightforward and did not require any additional data treatment, because the data parser was already complete.

3.3 Data standardization

3.3.1 International classification of goods and services

The international classification of goods and services, also known as 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.[37] The system operates in close to 90 countries as of 2023, with an additional 65 non-member countries using the classification.

3.3.2 Data coding: industry sectors

USPTO trademark can be further classified under industry sectors using a mapping of Nice classes by Edital. The mapping is a mutually exclusive alignment of each Nice class to a single industry sector. This mapping is presented below.

Table VI Definition of industry sectors in trademark data

Industry sector	Nice classes
Agriculture	[29, 30, 31, 32, 33, 34]
Business services	[35, 36]
Chemicals	[1, 2, 4]
Clothing	[14, 18, 22, 23, 24, 25, 26, 27, 34]
Construction	[6, 17, 19, 37, 40]
Health	[3, 5, 10, 44]
Household equipment	[8, 11, 20, 21]
Leisure and education	[13, 15, 16, 28, 41]
Research and technology	[9, 38, 42, 45]
Transportation	[7, 12, 39]

Source: Edital mapping of industry sectors

3.3.3 Data coding: U.S. counties

The following section details the steps to geocode U.S. addresses on USPTO trademarks.

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 no longer a problem, as most U.S. addresses are complete with state, city, zip code and even street

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

information. This greatly reduces the number of trademarks co-assigned to more than one county, because 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 distinguishing 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 just over 70 of these remaining state and city combinations increased the coverage of the geocoding to about 97% of all U.S. trademark counts. In the case of patent data, the matching process had to stop there, as all 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.[38] 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.

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 of 98%.

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

Alignment of the data with existing data sources

No data source for USPTO trademark counts at the level of U.S. counties was detected during the literature review at the start of this project. Therefore, it was not possible to compare the data prepared with an external source, as was done 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 attempt a similar exercise.

3.4 Indicators related to trademarks

The list of indicators on trademarks for the SEI 2026 remains the same as for SEI 2024. These included:

- 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 business sector (as defined by a mapping of Nice classes provided by Edital, a company specializing in trademark information)
- Number of registered trademarks (USPTO), by U.S. county (normalized against population data)

4 Annex

This annex presents all the entities included under each Federal agency. Name variants and acronyms for each of these entities were used as search terms and searched for in the addresses found in patents and publications, to ensure a high recall of all agency output. Cases in italic refer to sub-entities of specific organizations and are part of the parent organization listed just above in the table.

Department of Agriculture	
USDA	**Department of Agriculture**
ERS	Economic Research Service
FNS	Food and Nutrition Service
FSIS	Food Safety and Inspection Service
HNRCA	Jean Mayer Human Nutrition Research Center on Aging
NASS	National Agricultural Statistics Service
NIFA	National Institute of Food and Agriculture
NRCS	Natural Resources Conservation Service
ARS	**Agricultural Research Service**
CMAVE	Center for Medical, Agricultural and Veterinary Entomology
NADS	National Animal Disease Center
NCAUR	National Center for Agricultural Utilization Research
NCGR	National Clonal Germplasm Repository
NLAE	National Laboratory for Agriculture and the Environment
SHRS	Subtropical Horticultural Research Station
SRRC	Southern Regional Research Center
USHRL	Horticultural Research Laboratory
WRRC	Western Regional Research Center
RHRC	Robert W. Holley Center for Agriculture and Health

APHIS	**Animal and Plant Health Inspection Service**
CPHST	Center for Plant Health Science and Technology
NWRC	National Wildlife Research Center
PPQ	Plant Protection and Quarantine
VS	Veterinary Services
WS	Wildlife Services
USFS	**US Forest Service**
FPL	Forest Products Laboratory
RMRS	Rocky Mountain Research Station
Department of Defense	
DOD	**Department of Defense**
DTRA	Defense Threat Reduction Agency
USUHS	Uniformed Services University of the Health Sciences
USMA	Military Academy
MDA	Missile Defense Agency
WRNMMC	Walter Reed National Military Medical Center
	MIT Lincoln Laboratory
NDU	National Defense University
DARPA	**Defense Advanced Research Projects Agency**
DIA	**Defense Intelligence Agency**
Department of the Army	
APG	Aberdeen Proving Ground

AMRDED	Aviation and Missile Research Development and Engineering Center
AR	Army Reserve
ARDEC	Armament Research, Development and Engineering Center
ARL	Army Research Laboratory
ARL	Aeromedical Research Laboratory
ARO	Army Research Office
BAMC	Brooke Army Medical Center
CEHR	Center for Environmental Health Research
CID	Criminal Investigation Command
CoE	Corps of Engineers
CRREL	Cold Regions Research and Engineering Laboratory
DPG	Dugway Proving Ground
DTRD	Dental and Trauma Research Detachment
ECBC	Edgewood Chemical Biological Center
ERDC	Engineer Research and Development Center
ISR	Institute of Surgical Research
MEDCOM	Medical Command
MEDDC&S	Medical Department Center and School
MRICD	Medical Research Institute of Chemical Defense
MRIID	Medical Research Institute of Infectious Diseases
MRMC	Medical Research and Materiel Command
NSRDEC	Natick Soldier Research, Development and Engineering Center
PHC	Public Health Command
RDREOM	Research, Development and Engineering Command
RIBSS	Research Institute for the Behavioral and Social Sciences
RIEM	Research Institute of Environmental Medicine

SAMMC	San Antonio Military Medical Center
TARDEC	Tank Automotive Research Development and Engineering Center
USA	US Army
AWC	Army War College
WRAIR	Walter Reed Army Institute of Research
	All Army hospitals and medical centers
Department of the Air Force	
AFA	Air Force Academy
AFIT	Air Force Institute of Technology
AFRL	Air Force Research Laboratory
AFSAM	Air Force School of Aerospace Medicine
JBSA	Joint Base San Antonio
KAFB	Kirtland Air Force Base
KAFB	Keesler Air Force Base
LAFB	Lackland Air Force Base
MMD	Materials and Manufacturing Directorate
SVD	Space Vehicles Directorate
TAFB	Travis Air Force Base
USAF	US Air Force
WHASC	Wilford Hall Ambulatory Surgical Center
WPAFB	Wright-Patterson Air Force Base
	All Air Force hospitals and medical centers
Department of the Navy	
BMS	Bureau of Medicine and Surgery

CNA	Center for Naval Analyses
ECE	Entomology Center of Excellence
MMP	Marine Mammal Program
NA	Naval Academy
NAMI	Naval Aerospace Medical Institute
NAMRU	Naval Medical Research Unit
NAVAIR	Naval Air Systems Command
NAWC	Naval Air Warfare Center
NCCOSC	Naval Center for Combat and Operational Stress Control
NHRC	Naval Health Research Center
NMCPHC	Navy and Marine Corps Public Health Center
NMRC	Naval Medical Research Center
NO	Naval Observatory
NPS	Naval Postgraduate School
NRL	Naval Research Laboratory
NSMRL	Naval Submarine Medical Research Laboratory
NSSC	Naval Sea Systems Command
NSWC	Naval Surface Warfare Center
NUWC	Naval Undersea Warfare Center
NWC	Naval War College
ONR	Office of Naval Research
SPAWAR	Space and Naval Warfare Systems Center Pacific
USMC	US Marine Corps
USN	US Navy
	All Navy hospitals and medical centers

NGA	**National Geospatial-Intelligence Agency**
NRO	**National Reconnaissance Office**
NSA	**National Security Agency**
Department of Homeland Security	
DHS	**Department of Homeland Security**
CBP	Customs and Border Protection
CIS	Citizenship and Immigration Services
CSAC	Chemical Security Analysis Center
FEMA	Federal Emergency Management Agency
NBACC	National Biodefense Analysis and Countermeasures Center
PIADC	Plum Island Animal Disease Center
USCG	**US Coast Guard**
CGA	Coast Guard Academy
Department of the Interior	
DOI	**Department of the Interior**
BLM	Bureau of Land Management
FWS	Fish and Wildlife Service
NPS	National Park Service
USBR	US Bureau of Reclamation
USGS	**US Geological Survey**
ASC	Alaska Science Center
EROS	Center for Earth Resources Observation and Science

LSC	Leetown Science Center
NWHC	National Wildlife Health Center
PWRC	Patuxent Wildlife Research Center
Department of Commerce	
DOC	**Department of Commerce**
CB	Census Bureau
USPTO	Patent and Trademark Office
NIST	**National Institute of Standards and Technology**
CNR	Center for Neutron Research
NOAA	**National Oceanic and Atmospheric Administration**
AFSC	Alaska Fisheries Science Center
AOML	Atlantic Oceanographic and Meteorological Laboratory
ARL	Air Resources Laboratory
EMC	Environmental Modeling Center
ESRL	Earth System Research Laboratory
GFDL	Geophysical Fluid Dynamics Laboratory
GLERL	Great Lakes Environmental Research Laboratory
JCSDA	Joint Center for Satellite Data Assimilation
NCDC	National Climatic Data Center
NCEP	National Centers for Environmental Prediction
NCRI	National Coral Reef Institute
NEFSC	Northeast Fisheries Science Center
NESDIS	National Environmental Satellite, Data, and Information Service
NMFS	National Marine Fisheries Service

NOS	National Ocean Service
NSSL	National Severe Storms Laboratory
NWFSC	Northwest Fisheries Science Center
NWS	National Weather Service
OAR	Office of Oceanic and Atmospheric Research
PIFC	Pacific Islands Fisheries Science Center
PMEL	Pacific Marine Environmental Laboratory
SEFSC	Southeast Fisheries Science Center
STAR	Center for Satellite Applications and Research
SWFSC	Southwest Fisheries Science Center
Department of Energy	
DOE	**Department of Energy**
ALCF	Argonne Leadership Computing Facility
FERC	Federal Energy Regulatory Commission
GLBRC	Great Lakes Bioenergy Research Center
JBEI	Joint BioEnergy Institute
JGI	Joint Genome Institute
NCPV	National Center for Photovoltaics
NWTC	National Wind Technology Center
ORISE	Oak Ridge Institute for Science and Education
SIMES	Stanford Institute for Materials and Energy Sciences
SSRL	Stanford Synchrotron Radiation Lightsource
Y-12	Y-12 National Security Complex
National Laboratories	
AL	Ames Laboratory
ANL	Argonne National Laboratory

APS	*Advanced Photon Source*
BNL	Brookhaven National Laboratory
Fermilab	Fermi National Accelerator Laboratory
INL	Idaho National Laboratory
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
ALS	*Advanced Light Source*
LLNL	Lawrence Livermore National Laboratory
NETL	National Energy Technology Laboratory
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory
PPPL	Princeton Plasma Physics Laboratory
SLAC	SLAC National Accelerator Laboratory
SNL	Sandia National Laboratories
SRNL	Savannah River National Laboratory
TJNAF	Thomas Jefferson National Accelerator Facility
Department of Health and Human Services	
DHHS	**Department of Health and Human Services**
AHRQ	Agency for Healthcare Research and Quality
BARDA	Biomedical Advanced Research and Development Authority
FDA	Food and Drug Administration
FNLCR	Frederick National Laboratory for Cancer Research
HRSA	Health Resources and Services Administration
MCHB	Maternal and Child Health Bureau
NCTR	National Center for Toxicological Research

USPHS	United States Public Health Service
NIH	**National Institutes of Health**
FIC	John E. Fogarty International Center
GNL	Galveston National Laboratory
NCATS	National Center for Advancing Translational Sciences
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NEI	National Eye Institute
NHGRI	National Human Genome Research Institute
NHLBI	National Heart, Lung, and Blood Institute
NIA	National Institute on Aging
NIAAA	National Institute on Alcohol Abuse and Alcoholism
NIAID	National Institute of Allergy and Infectious Diseases
NIAMS	National Institute of Arthritis and Musculoskeletal and Skin Diseases
NIBIB	National Institute of Biomedical Imaging and Bioengineering
NICHD	Eunice Kennedy Shriver National Institute of Child Health and Human Development
NIDA	National Institute on Drug Abuse
NIDCD	National Institute on Deafness and Other Communication Disorders
NIDCR	National Institute of Dental and Craniofacial Research
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
NIEHS	National Institute of Environmental Health Sciences
NIGMS	National Institute of General Medical Sciences
NIMH	National Institute of Mental Health
NIMHD	National Institute on Minority Health and Health Disparities
NINDS	National Institute of Neurological Disorders and Stroke
NINR	National Institute of Nursing Research

NLM	National Library of Medicine
CNPRC	California National Primate Research Center
ONPRC	Oregon National Primate Research Center
SNPRC	Southwest National Primate Research Center
TNPRC	Tulane National Primate Research Center
WaNPRC	Washington National Primate Research Center
WNPRC	Wisconsin National Primate Research Center
YNPRC	Yerkes National Primate Research Center
CDC	**Centers for Disease Control and Prevention**
NCBDDD	National Center on Birth Defects and Developmental Disabilities
NCCDPHP	National Center for Chronic Disease Prevention and Health Promotion
NCEH	National Center for Environmental Health
NCEZID	National Center for Emerging and Zoonotic Infectious Diseases
NCHI	National Center for Health Statistics
NCIPC	National Center for Injury Prevention and Control
NCIRD	National Center for Immunization and Respiratory Diseases
NCHHSTP	National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention
NIOSH	National Institute for Occupational Safety and Health
Department of Veterans Affairs	
VA	**Department of Veteran Affairs**
NCPTSD	National Center for PTSD
VBA	Veterans Benefits Administration
VHA	Veterans Health Administration
	All VA hospitals, medical centers and healthcare systems
Environmental Protection Agency	
EPA	**Environmental Protection Agency**

GLNPO	Great Lakes National Program Office
NHEERL	National Health and Environmental Effects Research Laboratory
National Aeronautics and Space Administration	
NASA	**National Aeronautics & Space Administration**
APL	Astroparticle Physics Laboratory
ARC	Ames Research Center
GISS	Goddard Institute for Space Studies
GRC	Glenn Research Center
GSFC	Goddard Space Flight Center
JPL	Jet Propulsion Laboratory
JSC	Johnson Space Center
LaRC	Langley Research Center
MSFC	Marshall Space Flight Center
NAI	Astrobiology Institute
NExScI	Exoplanet Science Institute
SSERVI	Solar System Exploration Research Virtual Institute
Department of Transportation	
DOT	**Department of Transportation**
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FRA	Federal Railroad Administration
HTC	William J. Hughes Technical Center
NHTSA	National Highway Traffic Safety Administration
TFHRC	Turner-Fairbank Highway Research Center
VNTSC	John A. Volpe National Transportation Systems Center