Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2024

Technical Documentation

February 2024

Science-Metrix

Patent and Trademark Indicators for the Science and Engineering Indicators 2024

Technical Documentation

February 15, 2024

Submitted to:
SRI International

Authors
Guillaume Roberge
Alexandre Bédard-Vallée

Project Leader
Guillaume Roberge

By:

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

Contents

Tables i
1 Introduction 1
2 Patent indicators 2
2.1 Data limitations 2
Kind codes 2
2.2 Database 3
2.3 Data standardization 4
2.3.1 Mapping of patents by technical fields 4
2.3.2 Mapping of patents under the Environment Technology Framework 5
2.3.3 Mapping of patents under the Environmental Technology Framework related to Agriculture and Forestry 6
2.3.4 Mapping of patents under key technology areas of the CHIPS and Science Act 7
2.3.5 Linking citations to non-patent literature to the bibliometric database 9
2.3.6 Data standardization: country, country groups, regions 12
2.3.7 Data standardization: U.S. states 12
2.3.8 Data coding: U.S. sectors 13
2.3.9 Data coding: U.S. counties 14
2.4 Indicators related to utility patents 17
2.4.1 Inventors versus applicants 18
2.4.2 Applications versus granted patents 18
2.4.3 Number of utility patents 19
2.4.4 Patent counts, publication output and patent citations related to Federally Funded Research and Development Centers (FFRDCs) 19
3 Trademark indicators 21
3.1 Data limitations 21
3.2 Database 21
3.3 Data standardization 22
3.3.1 International classification of goods and services 22
3.3.2 Data coding: industry sectors 22
3.3.3 Data coding: U.S. counties 22
3.4 Indicators related to trademarks 24
4 Annex 25

Tables

Table I WIPO classification scheme for the production of SEI patent indicators 4
Table II Example of a patent fractioned by technical fields according to IPC codes, following conversion from CPC codes 5
Table III Most frequent 2-grams in patent reference strings 10
Table IV Definition of industry sectors in trademark data 22

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) 2024. 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. The work done for the bibliometrics aspect is presented in a separate document. This documentation is accompanied by a collection of external files that are necessary complements to perform these tasks. The list of accompanying external files is as follows:

External File 1: IPC technology concordance table
External File 2: CPC mapping of Key Technology Areas
External File 3: Patent number and uuid to Scopus ID
External File 4: Patent number and SEQ to countries and regions
External File 5: Patent number and SEQ to American states
External File 6: US applicant to sector

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 accessible in an Elsevier Data Repository.[1]

[1] https://data.mendeley.com/datasets/vrg53tc5r2/1

2 Patent indicators

The patent indicators for the U.S. market in this report were produced using an in-house implementation of the PatentsView patent database, a platform derived from the USPTO bulk data files. To accomplish the tasks, 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.

2.1 Data limitations

There is no notable limitation regarding the 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[2] and by comparing with data prepared for the previous SEI.

However, 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, without more precise information that would be quite helpful for this geocoding exercise, such as zip codes, street names and street numbers. Although it is still possible to obtain a robust geocoding of U.S. counties using only state and city information, this adds a layer of uncertainty to the matching in cases 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.

Kind codes

Kind codes are a classification system used across patent offices to classify document types. Each patent office has its own classification system; although codes are often similar across offices, their implementation may differ across offices. For the SEI 2022, USPTO kind codes were used to identify utility patents from the USPTO.

The patent indicators for this study were produced using a set of kind codes[3] that returns 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.

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

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

2.2 Database

PatentsView

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,[4] and at a relatively high frequency, with around 90% of all U.S. patents being assigned a county in the data. However, this high level is not spread equally across the U.S., as only a little more than 50% of the approximately 40,000 distinct U.S. addresses in the database are assigned a county FIPS code, a reflection of the high imbalances observed in the U.S. regarding patent output. Additionally, PatentsView does not allow for multiple county assignments per address, which is sometimes expected given that patent data only contain state and city information. This can become especially problematic in the case of large cities, which are assigned to a single county in the data but should theoretically be linked to multiple counties given the uncertainty regarding the assignment (e.g., New York City is always forced under county FIPS code 36061 of New York County).

The database also provides information on three classification schemes: the U.S. national classes (the 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, whereas patent data in the previous round of the SEI were largely prepared for the period 1996 to the present.

PatentsView tables were downloaded and uploaded into the Science-Metrix AWS S3 and Databricks environments. The process is straightforward and does not require any initial treatment because the data are already parsed. Documentation[5] presenting the content of the tables is available on the PatentsView website.

One notable issue specific to this edition is an unusually high number of patents with missing country information for inventors and assignees, resulting in a larger share of unclassified content for 2022 (1% as opposed to less than 0.01% for other years). To address this issue, country information for this set of

[4] 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.

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

patents was complemented using in-house information from the LexisNexis TotalPatent One, a database from sister company LexisNexis.[6]

2.3 Data standardization

2.3.1 Mapping of patents by technical fields

In all SEI editions since 2016, patents have been matched on a classification scheme of 35 technical fields[7] developed by the World Intellectual Property Organization (WIPO). The main objective behind the development of such a classification is to provide a tool for country comparisons.[8] The technical fields defined by this classification are listed in Table I.

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

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

Source: IPC Technology Concordance Table

This classification scheme is based on the IPC classification. Since the most recent U.S. patents are natively classified using the CPC, which replaced the USPC classification scheme at the national level, using this scheme as a starting point is more practical. In order to classify the patents by technology fields,

[6] Our team made the PatentsView team aware of this potential data issue, but given tight schedule, in agreement with NCSES representatives, it was decided to go ahead with the ad-hoc solution instead of waiting for any correction to the source data.

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

[8] 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

a concordance table between CPC and IPC codes prepared by the USPTO, in collaboration with the EPO, is used.[9]

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 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 II details this process for one patent.

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

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

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

Source: Prepared by Science-Metrix using the IPC Technology Concordance Table

External File 1: IPC technology concordance table

or online at: https://www.wipo.int/ipstats/en/docs/ipc_technology.xlsx

2.3.2 Mapping of patents under the Environment Technology Framework

For SEI 2024, NCSES tasked Science-Metrix with the production of patent indicators at country, U.S. county, and U.S. state levels for environment-related technologies. To enable the production of metrics on these technologies, already existing classification schemes based on CPC codes developed by the OECD for the Green Growth Indicators[10] were used to identify relevant patents. In cases where patents were assigned to multiple categories under the 10 environmental technology categories, patents were

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

[10] https://read.oecd-ilibrary.org/environment/measuring-environmental-innovation-using-patent-data_5js009kf48xw-en#page1

fractionalized to avoid duplicating counts, so that the sum across all 10 categories equaled the total reported across all environmental technologies. Here are the 10 categories that were mapped:

- Environmental management
- Climate change mitigation technologies related to energy generation, transmission or distribution
- Capture, storage, sequestration or disposal of greenhouse gases
- Climate change mitigation technologies related to transportation
- Climate change mitigation technologies related to buildings
- Climate change mitigation technologies related to wastewater treatment or waste management
- Climate change mitigation technologies in the production or processing of goods
- Climate change mitigation in information and communication technologies [ICT]
- Climate change adaptation technologies
- Environment-related and adaptation technologies relevant to the ocean economy

2.3.3 Mapping of patents under the Environmental Technology Framework related to Agriculture and Forestry

Through reclassification, the patent technologies within the Environmental Technology framework allow for a direct linkage with agriculture and forestry categories. Relevant technologies in agriculture and forestry are found in "Climate change mitigation technologies in the production or processing of goods" and "Climate change adaptation technologies" and "climate change adaptation strategies." In the data presented in the SEI 2024 report, additional categories of patents are added that are not part of the Environmental Technology framework. These are plant hybrids in the following categories:

- plants tolerant to drought (Y02A 40/132)
- plants tolerant to salinity (Y02A 40/135)
- plants tolerant to heat - Y02A 40/138
- genetically modified [GMO] plants, e.g., transgenic plants (Y02A 40/146).

The Invention, Knowledge Transfer, and Innovation report provides agriculture and forestry patents in two experimental categories based on the following CPC codes:

Technologies relating to agriculture, livestock or agroalimentary industries, consisting of:

- Using renewable energies, e.g. solar water pumping (Y02P60/12)
- Measures for saving energy, e.g. in green houses (Y02P60/14)
- Reduction of greenhouse gas [GHG] emissions in agriculture (Y02P60/20-22)
- Land use policy measures (Y02P60/30)
- Afforestation or reforestation (Y02P60/40)

Technologies related to adaptation technologies in agriculture, forestry, livestock or agroalimentary production:

- In agriculture (Y02A40/10-58), abiotic stress (inclusive of plants tolerant to drought, salinity or heat); genetically modified [GMO] plants; fertilizer of biological origin; Improving land use; improving water use or availability;

controlling erosion; greenhouse technology, e.g. cooling systems thereof; specially adapted for farming or for storing agricultural or horticultural products; using renewable energies;

- Ecological corridors or buffer zones (Y02A40/60)

2.3.4 Mapping of patents under key technology areas 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 was tasked to develop a mapping of patents to create sets that could be used to generate patent activity indicators for these key technologies.

Because the act was only signed in the second half of 2022, providing data for the key technology areas was not part of the original scope of the project, and work on this was only requested a year later, in mid-August 2023, at the end of the originally planned timeline to produce the patent data for the SEI 2024. With limited time to act, Science-Metrix started working on defining the mapping, first using seed keywords defined according to priorities related to each of the 10 key technologies, to identify relevant patents using titles and abstracts. While this approach yielded interesting and relevant results, the team was concerned with recall, as although identified patents were relevant, the approach was missing a notable share of relevant patents. Furthermore, in recent SEI editions, the NSF moved away from keyword-based mappings as they can be quite challenging to maintain and update, more so as queries become more complex.

For these reasons, Science-Metrix moved to a second phase, focusing mostly on mappings of CPC codes. This presented with the advantage of aligning with other mappings used in the SEI (e.g., for Environmental Technologies), and relying upon a well-established mapping system, used at both the USPTO and EPO, would ensure perennity in the future. While using the CPC system proved to be advantageous, identifying the relevant CPC codes for each key technology area was still a challenge. At over half a million codes, making sure to capture 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, coming up with an effective definition for each technology was not trivial. To support in developing these definitions, SRI provided Science-Metrix with literature reviews for each category, filing in templated documents Science-Metrix had prepared to cover most of the relevant information needed to do a proper delineation of the key technologies. Science-Metric and SRI also received internal feedback from other groups at NSF with 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 provide additional validation of the mapping, our team investigated how it could take advantage of newly emerging generative AI models. In an ideal situation, 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 to re-adjust on this, the team had to compromise. 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 to be submitted to 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 separate 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 the current exercise.

However, the current approach also opens a new path for future improvements to the mapping. Indeed, with more budget to run the model fully on all CPC codes, estimated at about 10k-20k USD, it would be possible to get a full mapping from the generative AI model, potentially identifying additional codes that could have been missed by the current approach. Indeed, it would be naïve to think that the current mapping, with half a million potential codes, is not missing a single one at the moment, and while analysts focused on CPC codes linked to high levels of patents during validation to ensure that most content was properly captured, there is certainly room for improvement. With more time, combined with quicky diminishing costs to run generative AI models, it would seem reasonable to revisit the mapping in the coming months to further improve on it. Feedback from the research community would also be welcome, as it is expected that even experts in the same disciplines might not agree on some of the inclusions and exclusions to be made. Therefore, the current mapping only represents a first step into measuring patent activity under key technology areas of the CHIPs act, and future developments will ensure that this mapping can converge towards consensual definitions, and hopefully be used broadly in the field in the coming years. Given its size, the final mapping can be accessed in a separate external document.

External File 2: CPC mapping of Key Technology Areas

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

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

Extracting references

All references from patents indexed in the USPTO that were tagged as "non-patent literature" were first extracted from the PatentsView patent database (i.e., in table "Otherreference"). This represented 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 III.

Table III Most frequent 2-grams in patent reference strings

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

Source: SEI 2014 technical documentation

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

Matching references to scientific literature

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

The matching algorithm was tuned to favor precision at the expense of recall because increasing recall above the current rate attained (i.e., 94%) would greatly increase the number of false positive matches, with minimal impact on recall. A total of 20 million references were matched with high confidence to scientific literature in the Scopus database, going back to the 1800s.

External File 3: Patent number and uuid to Scopus ID

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

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

At the end of the matching process, manual validations to estimate recall and precision were performed. Overall, the precision of the patent references matched to scientific publications stood at around 99%. Using a sample of 100 patent references that were not matched, recall within this sample was estimated at 95%—that is, only five of these references could be linked to scientific publications when searched for manually. This number is especially important because it makes it possible to estimate the number of references to scientific publications missed by the matching algorithms. In total, of the 53.6 million references available in the "otherreference" table, 20 million could be matched to a scientific publication indexed in Scopus. Since about 8.7 million references were filtered out in the pre-processing step (e.g., reference to patents, search reports), this left about 24.9 million references unmatched. Using the 95% recall estimated above on a sample of unmatched references, this means that approximately 5% of the 24.9 million references, or about 1.25 million results, could potentially be references to scientific publications that the algorithm could not match. Therefore, the expected total number of matched references should stand at about 21.25 million, meaning that recall for the current exercise stands at about 94%. Although further improvement to the matching algorithm could be performed in the future, it will become extremely difficult to increase recall without compromising precision, because the missed cases, which mostly consist of exceptions and unstandardized ways of referencing literature, are hard to catch and will not be easily retrieved.

2.3.6 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.[11] Science-Metrix matched country groups and regions using the USPTO conversion table, which enables quick identification of all countries included under each country group or region. A few corrections to country codes were performed to reassign outdated country codes to new codes reflecting geopolitical changes (e.g., Yugoslavia used for addresses in Serbia, Serbia and Montenegro, Slovenia).

Similar corrections were applied for data on Puerto Rico and the U.S. Virgin Islands in the past. These were included under "Central and South America" in the SEI 2016 edition, but in the following rounds they were included under "North America", with the U.S. Virgin Islands being included under the United States and Puerto Rico being presented separately from the United States. For 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" being instead displayed in the state information. As a result, all country codes set to "US" for which the state code was displayed as "PR" 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 2024 edition sticks with changes from the 2022 edition described above.

External File 4: Patent number and SEQ to countries and regions

2.3.7 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 for those where state information was missing or invalid.

External File 5: Patent number and SEQ to American states

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

2.3.8 Data coding: U.S. sectors

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

- Government
- Private
- Academic
- Individuals
- Others

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 other category covers the remaining cases, including foundations, trusts and other non-academic non-profits entities that do not fit under the 4 other 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 first matching step, manual coding was performed to assign the remaining applicants' names that could not be automatically assigned. Coding forms extracted from the SEI 2022 exercise were also used to help during the coding exercise. In the end, tests were performed to ensure that distinct forms appearing in the database were always coded under the same sector, ensuring the absence of any ambiguous decisions. Of all U.S. addresses, 99.7% could be assigned a sector, the remaining cases being listed under a sixth sector, "Unclassified."

The academic and government sectors have far lower patenting output than the private sector. Because it was important for the SEI report to have accurate output estimates for these two sectors, Science-Metrix prioritized the crediting of patents to the academic and government sectors in the rare cases of multiple matches. If these sectors had not been prioritized, it is believed that slightly inaccurate and lower estimates of patenting activity for these two sectors would have been obtained, because these few cases, although almost unnoticeable at the level of output measured for the private sector (i.e., about 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 around 4% of all unclassified addresses. Yet, because unclassified addresses account for such a small number of cases, correcting for this does not change the proportion of addresses coded under the "Others" sector in the U.S., because correcting for this would only add about 120 publications to this sector (or 0.006% of all publications).

External File 6: US applicant to sector

2.3.9 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 level geographically. 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 2022, 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 which the country code is "US." U.S. addresses were geocoded to the 3,143 U.S. counties, 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.[12] Their results have been reproduced independently by Carlino et al.,[13] 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. Post Office reference file to match cities and states of residence of inventors to U.S. regional components,[14] Science-Metrix identified a more recent reference file from the U.S. Census

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

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

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

Bureau that linked place names with U.S. counties.[15] 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

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

This mapping file from the U.S. Post Office 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.[16] Most of the place names not matched were smaller units of cities (e.g., neighborhoods) or unincorporated places, which are not covered in the reference file.

Overall, the initial matching steps before manual coding enabled the geocoding of about 94% of all U.S. addresses to at least one U.S. county. About 38,000 combinations of U.S. states and cities, accounting for about 6% of all patents, remained unassigned prior to the manual step, but geocoding 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

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

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

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 given to Santa Clara County in that case. This also avoided drastically overestimating Contra Costa's output if the output was split equally across both counties.

Distribution of ambiguously assigned patents across counties and CBSAs

Although the proportion of ambiguously assigned U.S. patents is relatively low, at about 14%, this is nevertheless non-negligible. To account for this, a redistribution of the counts of the ambiguous cases was performed. At first, we envisioned redistributing the output following the proportions observed in the population that could be assigned unambiguously. However, it quickly became clear that doing so would lead to highly unreliable results. Indeed, for cities spread across more than one county, 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 declarative borough names, for which unambiguous assignment could be performed, would receive a larger share of the total output from the city.

Instead, it was decided that in the remaining cases of ambiguous assignment, each county would receive an equal share of the output from an entry, similarly to what is done by the PTMT team. This redistribution, although imperfect, should nevertheless be less biased than the other suggested approach. It is 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 considering the limitations associated with the data, manual and automatic validation was performed to check for the validity of the data obtained. A manual sampling approach, checking for a sample of 200 U.S. addresses on U.S. patents, was manually validated by analysts, looking for the addresses in Google to identify 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, which stood at 98%.

Alignment of the data with existing data sources

Two notable datasets with patent counts per U.S. counties were presented at the beginning of this report, the set from the USPTO PTMT and the data from Carlino's working paper. Since Carlino et al. mentioned in their paper that they compared their data with those from the PTMT and that they were highly similar, it was decided that data from the current exercise would only be compared with those from the PTMT, as 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. Post Office 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 performed demonstrated that the findings of the current project were aligned with those from the PTMT, reinforcing the assessment of robustness of the data prepared. Some discrepancies were observed for a few counties, which is to be expected given that some cities overlap with multiple counties, and Science-Metrix has no way to identify exactly to which county all cities were mapped by the PTMT.

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

2.4 Indicators related to 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.

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 were based on granted patents. One important distinction between patent applications and patent grants is the considerable time lag between the two. While an application is made closer to the time of invention, the granted patent is closer to the commercial return of the invention. Useful and complementary statistics can be derived from both approaches. However, several limitations in the quality of data on applications reduce their potential for the development of indicators. This is particularly true for U.S. applications, and Science-Metrix usually tries to avoid producing statistics for these. There are two main reasons for this:

- Applicants can ask that the application not be published.[17] Currently, only about 70% of patent applications are published. This proportion varies by type of industry, Patent Cooperation Treaty (PCT) versus non-PCT, size of company, country, and over time. Science-Metrix is not aware of any statistics on these variations. Importantly, once patents are granted, applications become public. So, this subsequently adds to the number of applications that were made public at the moment of application. Therefore, the exact number of applications for a given year is not known until at least 7–8 years later because of the time lapse between application and grant. These results have at least two implications: (1) statistics are always incomplete in more recent years, and (2) because of the variability in application-to-grant time, statistics for the most recent years are biased.
- The quality of data for applications is poor. Several applications do not have any information on the country and/or the state and/or the applicant name and/or the U.S. class. This information is sparse, and the quality varies from one provider to another.

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

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 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 2024 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 process will streamline their production and 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 being part of these. Overall, more than 200 entities are included under 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.

3 Trademark indicators

In a spirit of broadening the scope of the SEI beyond traditional metrics based on patents, a decision to include statistics on trademarks in the SEI 2020 was reached by the NSF after consulting material prepared by Science-Metrix demonstrating the coverage of the data available. This decision was made possible by the recent addition of data sources covering trademark data, which were not available in the past. Science-Metrix prepared statistics using trademark data from the USPTO for the SEI 2024.

3.1 Data limitations

Much like with patent data, there is no notable limitation regarding the USPTO data to be reported, because they provide complete coverage of U.S. trademarks. In addition, in comparison with patent data, USPTO trademark data are better suited to the geocoding exercise as addresses are much more complete, including not only U.S. states and U.S. cities, but zip codes, street names and street number as well. These more complete addresses are highly useful as they make it possible to differentiate cases that would be ambiguous if only cities were available, as is the case for patent data. Therefore, it was expected prior to the matching process that the percentage of U.S. patents assigned to more than one county would be drastically lower than the 14% measured for patents, and results presented later demonstrate that it is indeed the case.[18]

3.2 Database

One database covering USPTO trademarks was built to prepare statistics on trademarks. XML files containing data are freely available online[19] 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

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

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

files. Still, because of the more complete address data, the trademark data are much more suitable for a geocoding exercise.

To build the in-house version of the USPTO trademark database, Science-Metrix uploaded all the XML files from the USPTO website and reused a parser designed during the work performed for the SEI 2020 to extract the information needed and include it in Science-Metrix's Databricks environment. The process 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.[20] 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 IV 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.

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

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 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 discriminant enough to identify a single county using the same multi-step geocoding approach. Again, a manual step was performed to geocode the remaining addresses based on the highest frequency counts using Google Maps and ArcGIS online maps.

Overall, the initial matching steps before manual coding enabled the geocoding of about 94% of all U.S. addresses to at least one U.S. county. Geocoding a little more than 70 of these remaining state and city combinations increased the coverage of the geocoding to about 97% of all U.S. trademark counts. When dealing with patent data, this is where the matching process needed to stop, because all the available information had been used. However, zip codes are available for trademark data, so another round of matching was performed, this time using a crosswalk file between zip codes and U.S. counties, as defined in the 2010 Census, from the U.S. Office of Policy Development and Research of the Department of Housing and Urban Development.[21] 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

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

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

Alignment of the data with existing data sources

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

3.4 Indicators related to trademarks

Since SEI 2020, the list of indicators on trademarks has been continuously changing between editions. Below are the indicators that were selected for inclusion in the SEI 2024:

- 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), U.S. county, per business sector

4 Annex

This annex presents all the entities by Federal agency. Name variants and acronyms for each of these entities were searched for in addresses appearing in patents and publications to ensure a high recall of all their output. Cases in italic refer to sub-entities of specific organizations, and are part of the parent organization listed just above these 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
	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 naval 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