Science-Metrix

Patent Indicators and other intellectual property rights indicators for the Science and Engineering Indicators 2020

Technical Documentation

November 2019

Science-Metrix

Patent Indicators and other intellectual property rights indicators for the Science and Engineering Indicators 2020

Technical Documentation

November 20, 2020

Submitted to:
SRI International

Authors
Grégoire Côté
Guillaume Roberge
Alexandre Bédard-Vallée

Project Leader
Grégoire Côté

By:

Science-Metrix
1.514.495.6505 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com

Contents

Tables

1 Introduction

Science-Metrix has been commissioned by SRI International, on behalf of the National Science Foundation, to develop measures and indicators of research and patent activity using bibliometrics and patent data for inclusion in the Science and Engineering Indicators (SEI) 2020. 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) as produced in the SEI 2018, but also patent families with patents from dozens of patent authorities, trademarks, plant and design patents, and plant variety protections. The work done for the bibliometrics aspect is presented in a separate document. This documentation is accompanied by a collection of external files that are necessary complements to perform these tasks. The list of accompanying external files is as follows:

External File 1: IPC technology concordance table
External File 2: Patent number and uuid to Scopus ID
External File 3: Patent number and SEQ to countries and regions
External File 4: Patent number and SEQ to American states
External File 5: US applicant to sector
External File 6: Non-US applicant to academic sector

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

2 Patent indicators

USPTO

The patent indicators for the American market in this report were produced using an in-house implementation of the PatentsView patent database, a platform derived from the USPTO bulk data files. To accomplish such tasks, an in-house version of the database was built on an SQL server and was carefully conditioned for the production of large-scale comparative patent analyses. In addition to utility patents, which were the sole focus of the patent analyses in the SEI 2018, design patents and plant patents were also prepared this time.

Worldwide priority patents

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

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

The approach taken in Indicators 2020 broadens the economic and geographic scope of the patent analysis compared with earlier editions of SEI. Using PATSTAT data, which covers utility patents from close to 100 patent authorities, INPADOC patent families based on almost all data covered in PATSTAT were selected as the main unit of measure, following a methodology proposed by a team of researchers from academia and the OECD.[2] This method uses information within patent families to fill in gaps

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

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

regarding inventorship for patent offices where data are not complete, looking at related patents in other offices when information is not available for a patent. When no information on inventorship can't 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 fruit of local inventors. While this method is not perfect and can lead to errors when projecting inventorship at the level of individual patents, its level of precision is overall quite good when dealing with large-scale analyses such as the one prepared for this project. Details about the method can be found in the paper presented in footnote 3.

2.1 Data limitations

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

India

In the case of India, coverage in PATSTAT is lacking, resulting in an underestimation of its output. However, according to WIPO statistics,[3] India is a special case because only about 25% of patent applications to its national office were made by residents, out of an annual number hovering at around 45,000 in 2017. This means that about 12,000 patent applications are made by the residents each year (6,500 in 2008, 15,000 in 2017), with foreign companies and inventors dominating in terms of patent applications at the office (e.g., Qualcomm, Samsung, Huawei, Microsoft, Philips, General Electrics, Ericsson, Mitsubishi, BASF).[4]

The choice of INPADOC patent families as the indicator reduces the extent of the undercount 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 *Indicators* are measuring

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

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

granted patent families based on the year of the first grant in the patent family; however, the Indian patent authority is lagging behind in terms of processing of its patents, only granting about 11,000 patents to Indian residents over the last decade (less than 1,000 for most years). It takes an average 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.[5] While changes have been made to address this issue in recent years[6] and numbers of granted patents have started increasing, they are still relatively low in a global context (about 1,700 granted patents in 2017). As a result, this does not substantially alter India's ranking among leading countries.[7]

Italy

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

2.2 Kind codes

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

For the SEI 2020, USPTO kind codes were used to identify utility, design and plant 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[8] that select granted utility patents and applications, although the indicators were only computed on granted utility patents. Kind codes associated with utility patents at the USPTO were limited to three document types: A, B1 and B2. Kind code A applies to granted patents before 2001, while B1 and B2 replaced this kind code starting 2 January 2001. New to this edition, statistics on design patents (S) and plant patents (P, P1, P2, P3, P4, P9) were also prepared.

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.

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

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

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

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

2.3 Databases

PatentsView

Most of the patent analyses in *Indicators* were prepared using data from the USPTO indexed in PatentsView. The database provides details on patents such as full titles and abstracts, the country and state (when available) of the inventors and applicants, as well as names of the inventors and applicants. In most cases, applicants are organizations, although they are sometimes individuals when the patent is not assigned to any organization. The database also provides information on three classification schemes: the U.S. national classes (the USPC classes, although these are not available after 2015 as the system is no longer in use), the World Intellectual Property Organization's (WIPO) International Patent Classification (IPC), and the Cooperative Patent Classification (CPC). The CPC was produced in partnership between the USPTO and the EPO; it replaced the USPC classes after 2015, and the European Classification System (ECLA) after 2012. PatentsView is suitable for the production of technometric data dating from 1976, whereas patent data in the previous round of the SEI were largely prepared for the period 1996 to the present.

PatentsView tables were downloaded and uploaded into the Science-Metrix SQL server. The process is straightforward and does not require any treatment because the data are already parsed. Documentation[9] presenting the content of the tables is available on the PatentsView website.

PATSTAT

The European Patent Office Worldwide Patent Statistical database, better known as EPO PATSTAT or PATSTAT, is the database of reference in the field of international technometrics. Mainly developed for use by governmental organizations and academic institutions, it contains bibliographical and legal status patent data from most industrial and developing countries and covers major patent offices such as the EPO (Europe) and USPTO (United States). A conditioned in-house version of PATSTAT2019 Spring Edition, which consist of pre-defined tables with keys linking these together, was built on a SQL server and used to prepare statistics on worldwide priority patents.

2.4 Data standardization

2.4.1 Mapping of patents by technical fields

In SEI 2016, patents were matched on a classification scheme of 35 technical fields[10] developed by the World Intellectual Property Organization (WIPO). The main objective behind the development of such a classification was to provide a tool for country comparisons.[11] The technical fields defined by this classification are listed at Table I.

[9] http://www.patentsview.org/data/Patents_DB_dictionary_bulk_downloads.xlsx

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

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

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

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

Source: IPC Technology Concordance Table

This classification scheme is based on the IPC classification. Since the most recent U.S. patents are natively classified using the CPC, which replaced the USPC classification scheme at the national level, using this scheme as a starting point is more practical. In order to classify the patents by technology fields, a concordance table between CPC and IPC codes prepared by the USPTO, in collaboration with the EPO, was used.[12]

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

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

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

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

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

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

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

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

External File 1: IPC technology concordance table

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

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

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

Extracting references

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

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

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

- European Patent Office, Supplementary European Search Report dated Feb. 12, 2010 in European Application No. 04819909.5.

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

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

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

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

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

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

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

N-grams are contiguous sequences of *n* items from a given sequence. In this case, 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 which searched for hundreds of patterns used in bibliographic referencing were used to retrieve pages, issues, volumes and publication years 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. Compared to SEI 2018, the number of references matched to Scopus was much higher this time because Science-Metrix had gained access to scientific publications prior to 1996, resulting in millions of additional matches going back as far as the year 1828. Additionally, citations from design and plant patents were accounted for this time around, which was not the case in the SEI 2018.

Numerous techniques, including direct string matching, bag-of-words models, candidate clustering, and supervised machine learning were used to match the reference strings to actual articles. The matching algorithm was tuned to favor precision at the expense of recall. A total of 14,068,752 references were matched with high confidence to scientific literature in the Scopus database, going back to 1800s.

External File 2: Patent number and uuid to Scopus ID

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

- Cohen et al. Microphone Array Post-Filtering for Non-Stationary Noise, source(s): IEEE, May 2002.

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

At the end of the matching process, manual validations to estimate recall and precision were performed. Overall, the precision of the patent references matched to scientific publications stood at around 99%. Using a sample of 100 patent references that were not matched, recall within this sample was estimated at 95%—that is, only five of these references could be linked to scientific publications when searched for manually. This number is especially important because it makes it possible to estimate the number of references to scientific publications missed by the matching algorithms. In total, of the 37,113,970 references available in the "otherreference" table, 14,068,752 could be matched to a scientific publication indexed in Scopus. Since about 8.4 million references were filtered out in the pre-processing step (e.g., reference to patents, search reports) this left about 14.6 million references unmatched. Using the 95% recall estimated above on a sample of unmatched references, this means that approximately 5% of the 14.6 million references, or about 730,000 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 14.7 million, meaning that recall for the current exercise stands at about 95%. While it is expected that further improvement to the matching algorithm could be performed in the future, it will become extremely difficult to increase recall without compromising on precision because the missed cases are all hard to catch and will not be easily retrieved.

2.4.3 Data standardization: country, country groups, regions

To provide comparisons across countries and regions, data at the regional and national levels are presented in the SEI. It is fairly 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.[13] Science-Metrix matched country groups and regions using the USPTO conversion table, which enables quick identification of all countries included under each country group or region. A few corrections to country codes were performed to reassign outdated country codes to new codes reflecting geopolitical changes (e.g., Yugoslavia used for addresses in Serbia, Serbia and Montenegro, Slovenia).

Similar corrections were applied for data on Puerto Rico and the U.S. Virgin Islands. These were included under "Central and South America" in the SEI 2016 edition, but in the following rounds they were included under "North America", with the U.S. Virgin Islands being included under the United States and Puerto Rico being presented separately from the United States 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.

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

2.4.4 Data standardization: U.S. states

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

External File 4: Patent number and SEQ to American states

2.4.5 Data coding: U.S. sectors

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

- Government
- Private
- Academic
- Individuals
- Others

Automated coding was used to assign non-ambiguous forms of applicant names (e.g., "Univ" in the academic sector, "inc." in private) to the corresponding sector. After this first matching step, manual coding was performed to assign the remaining applicants' names that could not be automatically assigned. Coding forms extracted from the SEI 2018 exercise were also used to help during the coding exercise. In

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

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, 99.8% 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 144,000 patents in 2018), still represent a sizable number of patents at the level of the government and academic sectors (i.e., about 1,300 and 6,900 patents in 2018, 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 0.1% of all addresses that could not be classified. Overall, most categories were represented in accordance with their expected frequency based on occurrences in coded addresses, the only notable difference being the small over-representation of the "Others" sector in unclassified addresses. The "Others" sector represents 0.41% of all addresses in the database, but around 4% of all unclassified addresses. Yet, because unclassified addresses account for such a small number of cases, correcting for this does not change the proportion of addresses coded under the "Others" sector in the United States, because correcting for this would only add about 120 publications to this sector (or 0.006% of all publications).

External File 5: US applicant to sector

2.4.6 Non-U.S. academic institutions

As with the coding of U.S. sectors (see Section 2.4.5), automatic and manual coding of applicants from the academic sector outside the United States was performed. Generic forms of academic institutions in different languages were looked for in the applicants' names to retrieve all academic applicants across countries (e.g., Hochschule, ETH, Ecole). Coding from SEI 2018 was used to help and ensure comparability.

External File 6: Non-US applicant to academic sector

2.5 Indicators related to utility, design and plant patents

This section presents the patent indicators computed as part of this study. Compared to the SEI 2018 when only patent counts based on utility patents were prepared, patent counts on utility patents, design patents and plant patents were prepared this time.

2.5.1 Inventors versus applicants

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

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

2.5.2 Applications versus granted patents

All the statistics related to utility, design and plant 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.[14] 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.

[14] 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.

- The quality of data for applications is poor. Several applications do not have any information on the country and/or the state and/or the applicant name and/or the U.S. class. This information is sparse, and the quality varies from one provider to another. For instance, PatentsView appears to only have information regarding applications of granted patents.

2.5.3 Number of utility, design and plant 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.

2.6 Indicators related to worldwide priority patents

Patent counts based on worldwide priority patents using patent families were also prepared, accompanied by specialization indexes based on these counts. The specialization index is computed using fractional

counts indicating the level of involvement across categories, in this case technical fields. By definition, an entity cannot be specialized across all technical fields. Readers can find more details regarding the specialization index in the methodological report dedicated to bibliometrics.[15]

[15] http://www.science-metrix.com/?q=en/publications/reports#/?q=en/publications/reports/bibliometric-indicators-for-the-sei-2020-technical-documentation

3 Trademark indicators

In a spirit of broadening the scope of the SEI beyond traditional metrics based on patents, a decision to include statistics on trademarks in the SEI 2020 was reached by the NSF after consulting material prepared by Science-Metrix demonstrating the coverage of the data available. This decision was made possible by the recent addition of data sources covering trademark data, which were not available in the past. Given that data on trademarks from both the USPTO and the European Union Intellectual Property Office (EUIPO) were available, covering two of the largest markets in the world, it was decided that Science-Metrix would prepare statistics using both data sources independently because there is no means to satisfactorily create cross-office trademark families. Indeed, as opposed to patents—for which patent priorities can be used to link requests for inventions across multiple offices, thus avoiding counting the same invention multiple times—the concept of trademark priority, although it exists, is not widespread.

3.1 Building databases

Two databases were built to prepare statistics on trademarks, one covering USPTO trademarks and the other covering EUIPO trademarks. XML files containing data for both are freely available online[16] 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 inventors and assignees (to assign trademarks to countries, regions and U.S. states)
- Names of assignees (for sector analysis)
- Nice categories of goods and services (for comparison across categories)
- Registration year

3.2 International classification of goods and services

The international classification of goods and services, also known at the Nice classification, is a system used to register trademarks across categories of goods and services. It was adopted in 1957 following the Nice Agreement and comprises 45 classes, classes 1 to 34 covering goods and 35 to 45 covering services.[17] The system operates in close to 90 countries as of 2018.

3.3 Indicators related to trademarks

Around the end of 2018, the NSF requested a memo from Science-Metrix detailing which indicators could be prepared using trademark data. Below are the indicators that were selected for inclusion in the SEI 2020:

- Number of registered trademarks (USPTO and EUIPO), by region, country, or economy
- Number of registered trademarks (USPTO and EUIPO), by U.S. state

[16] USPTO : http://trademarks.reedtech.com/trademark-products.php; EUIPO : https://euipo.europa.eu/ohimportal/fr/open-data

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

- Number of registered trademarks (USPTO and EUIPO), by region, country, or economy, per Nice categories of goods and services
- Number of registered trademarks (USPTO and EUIPO), by U.S. state, per Nice categories of goods and services
- Specialization index of registered (USPTO and EUIPO) trademarks, by region, country, or economy, per Nice categories of goods and services
- Specialization index of registered (USPTO and EUIPO) trademarks, by U.S. state, per Nice categories of goods and services
- Number of registered trademarks (USPTO and EUIPO), by region, country, or economy, per industry sector (as defined by a mapping of Nice classes provided by Edital, a company specializing in trademark information)
- Number and share of registered trademarks (USPTO and EUIPO), by U.S. sector
- Share of registered trademarks (USPTO and EUIPO), by U.S. sector and category of goods and services
- Number of registered trademarks of the top trademarking companies at the USPTO
- Number of registered trademarks of the top trademarking companies at the USPTO, by category of goods and services
- Specialization index of registered trademarks of the top trademarking companies at the USPTO, by category of goods and services

3.4 Plant variety protections

Plant variety protections are a new addition to the SEI. They were added to provide additional depth to the IP rights analyses prepared because IPs related to plants were mostly not covered by traditional metrics based on patents.

3.4.1 Data source

Data for plant varieties were sourced from the Plant Variety Protection Statistics reports written by the International Union for the Protection of New Varieties of Plants (UPOV).[18] These reports are published every year and contain information about applications filed in each of the UPOV member states, both by residents and non-residents.[19] The numbers of titles issued by the member states, again presented separately for residents and non-residents, are also presented. Reports are composed of three main tables. The first is a list of all member states with their counts of applications and titles issued for the reporting year, as well as the four previous years (e.g., the 2016 report has data from 2012 to 2016). The second is a cross table showing the detail from foreign-filed and foreign-issued titles only for the reporting year. Each line corresponds to the member state in which the title was filed, and each column is the country

[18] Available online from UPOV's website : https://www.upov.int/meetings/en/topic.jsp?group_id=251

[19] UPOV receives data from around 75 offices. Data from the current extraction are from a UPOV document reporting on plant variety protections published for the 52nd ordinary session of the Council (November 2018). These data include official statistics reported by 75 national offices to the UPOV (data are complete for 43 of these offices, 11 offices are missing data for a single year, 5 offices are missing data for two years, the remaining 16 offices are missing more years).

that has filed the applications. There is also an "other" column. At the intersections, the number of titles either filed (in the top of the cells) or issued (in the bottom of the cell) are presented. The third and last table has the same format as the previous table and shows the detail from the "other" column of table 2 where applicable. The countries included as columns in this last table are most often states that are not members of UPOV.

3.4.2 Data extraction

The data extraction method used varied for the different tables and for the different report years. Data from 2012 to 2017 were the easiest to extract because reports from those years came with Excel files containing the data. A small amount of cleaning, such as removing titles and empty formatting lines from tables was nevertheless necessary. Additional columns containing the data year and filed/issued categories were also added to tables 2 and 3. Table 1 from the Excel file accompanying the 2017 report was also found to contain information for past years in hidden rows, some countries even having data from the 1980s available in this table. This finding removed the need to import table 1 from most other reports, except in the rare occurrences where data were missing from this table but present in some of the older reports. In those cases, the missing data were filled in by manually looking at the reports for the pertinent years.

For reports written from 2009 to 2011 extracting data from Tables 2 and 3 was more involved as reports were only available as PDF files. Still, those tables were in a format that could reliably be copied to Excel by Adobe Acrobat Pro X, where they were then cleaned in the same way as post-2012 tables.

The most involved cases arose from reports written pre-2009. The main difficulty with those tables is that data on filed titles and issued titles were not separated by a line as in other reports, but rather were put in the same cell. The table could still be copied from Adobe Acrobat Pro X to Excel as before, but it was only possible to reliably tell if a number corresponded to the "filed" or "issued" category when both counts in the same cell were non-zero. Otherwise, when in a given year and country titles were applied for but not granted or vice-versa, only a number with no other accompanying character was imported to Excel. Analysis of the PDF files seemed to indicate that the differences between both cases were due purely to cell formatting, which was not reliably reproduced in the Excel-import process and which could not be easily read from the PDF. What was done in those cases was to identify in Excel which cells had two numbers in them. Those could then be directly parsed to the correct cells in the final table. Then, the remaining cells containing data were identified but, as it was impossible to automatically tell in which category the numbers should fall, those cells were only highlighted in the final table using Excel conditional formatting. An analyst was then tasked with filling the highlighted cells manually by looking at the original PDF reports. Verification of the total counts of each category for each country was done to ensure no error arose from this step. The alignment of numbers from tables 2 and 3 with those from table 1 was also verified to ensure data integrity.

3.4.3 Indicators related to plant variety protections

Only counts were prepared using plant variety protection rights, and data only go back to 2008 as data prior to 2008 are not subdivided per resident countries for the office in charge of the European Union, which renders comparisons with European countries mostly impossible.