4.2.1 Key indicators for evaluating science, technology, and innovation

HARA Yasushi[1]

First Published August 28, 2018 Final Updated April 26, 2019

Abstract

Bringing about science, technology, and innovation is a long term endeavor. Conventional research has focused on identifying events or fields of science and qualitatively analyzing the factors affecting them. However, with the development and opening up of databases, growth in computing power available to researchers, and theoretical developments in evaluation methods, it is becoming possible to quantitatively analyze aspects of science and technology by combining databases related to patents, papers, and various types of funding including competitive grants.

Keywords

Databases, patents, papers, Big Data

1 Key indicators for evaluating science, technology, and innovation

The evaluation of science and technology has tended to focus on specific phenomena and chronological or situational studies. However, with the increasing openness of data and growth in computing power, it is becoming possible to analyze macro data at the national level and micro data at the researcher level in greater detail, as well as combine data in conducting analysis over longer periods of time and with a view to structural change. Figure 1 presents science and technology indicators and the relationships between

1 Michelin Fellow, The School for Advanced Studies in the Social Sciences, Paris (Formerly with the National Graduate Institute for Policy Studies SciREX Center)

them. Data available for the analysis of scientific and technological innovation can be broadly classified into three levels: (1) micro level, (2) meso level, and (3) macro level. Bibliographic data of patents and academic papers are widely used as indicators for (1) and (2). These data make it possible to evaluate the output of research activities—that is, to examine published academic papers and patents and identify who they were written by and from which organizations; when, where, and with what funds the papers or patents were written; what existing research they refer to; how many other papers or patents reference them; and for what purposes. Recently, science linkage data, which cover patents' non-patent literature data, has also come into use. To produce these patents and academic papers, scientists—primarily those belonging to Japanese universities and research institutes—utilize competitive grants such as the grants-in-aid for scientific research (KAKENHI). This input data are one of the most important indicators for measuring and evaluating scientists' behavior.

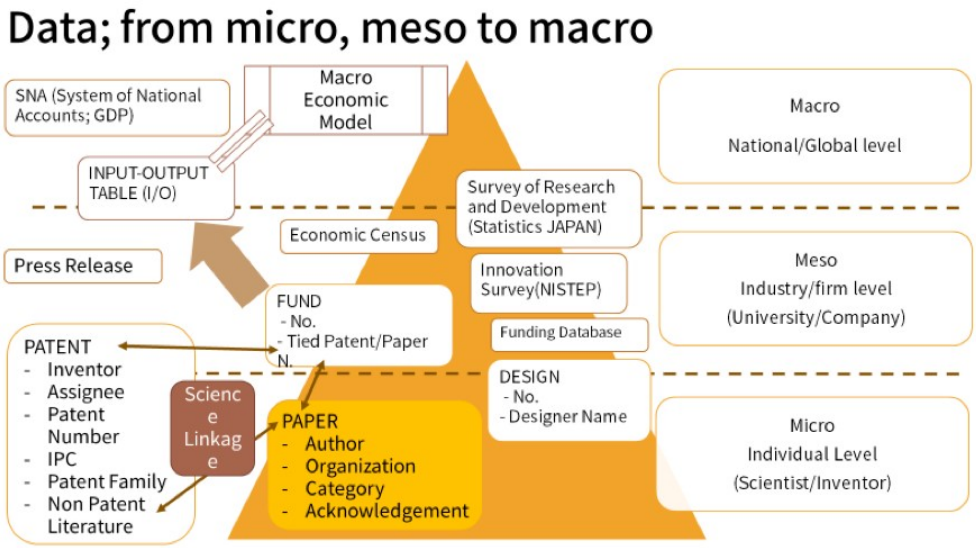

Figure 1. Relationships between science, technology, and innovation data.
Source: Created by the author

Moreover, recent studies have focused not only on these input–output relationships, but on indicators that can point to outcomes—such as detailed information on products in product and press release databases and the names of major developers, prices, and target consumer groups—with the aim of measuring the extent to which new knowledge generated by scientific and technological research activities has been translated into actual products, that is, into innovations that impact society. However, with the exception of the life science industry, and low-molecular drugs in particular, few industries have a 1:1 correspondence between patents and products. As such, there are still tremendous challenges to overcome in measuring the outcomes of innovation. Governments, universities, and research institutes have responded to this challenge by conducting surveys of inventors and companies to measure innovation output. Examples of such efforts include the National Innovation Survey of the National Institute of Science and Technology Policy (NISTEP) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT); the Survey on

Research and Development Activities of Firms in the Private Sector; the Science and Technology Research Survey by the Statistics Bureau of the Ministry of Internal Affairs and Communications; and the Economic Census. Only by combining such micro- and meso-level data at the industrial level can detailed macro-level estimates be made.

Additionally, using input-output tables as a methodology for measuring macro-level impact makes possible to measure the dynamics of each industry in detail and examine the interrelationships between them and the intellectual assets generated by science and technology research activities. In this way, when the analysis aggregates various micro-level data, it becomes possible to expand macro-level analysis; in turn, we can estimate and analyze the structure by which science and technology affect the macro economy by considering interdependencies among stakeholders within the national innovation system. However, such analysis requires (1) the aggregation of data across different databases, (2) combining of data, (3) cleaning and cleansing of data, and (4) the creation of interpolated data, on each occasion.

In order to begin science, technology, and innovation analysis on the micro level, it is necessary to comprehensively evaluate the research activities of the scientists, who are the main actors in these activities. However, as noted, the majority of existing studies measuring scientists' capabilities have focused on examining the bibliographic data of patents or papers, and have been unable to adequately analyze scientists' multifaceted effects and characteristics. Moreover, as suggested in the discussion about star scientists in another paper, scientists have more than just academic or product development capabilities. Indeed, they also make significant social impacts on the real world through educational activities, participation in business, participation in government committees, and indirect involvement legal system development. Therefore, in order to quantitatively and comprehensively evaluate scientists' impact, it is necessary to analyze their multifaceted academic and social impact by combining databases such as patents, papers, non-patent literature, and product information (e.g., press releases and award data) in a sophisticated manner.

1.1 The state of STI data development in Japan and abroad

Scientists' activities are complex and diverse, with a wide range of inputs (i.e., research funding), outputs (i.e., publication of papers and patent applications), and outcomes (e.g., social contribution activities and involvement in companies). For example, measuring the relationship between input and output enables discussions of the nature of the competitive funding system for researchers typified by the Grant-in-Aid for Scientific and Technological Research (KAKENHI). However, at present, only a few standardized tables, such as the NISTEP Dictionary of Company Names[2] and the NISTEP Dictionary of University and Research Institution Names[3], are available as tools that can link information between such databases according to the unit of analysis.

Foreign examples of such platforms providing multiple data sets include (a) the European Research Infrastructure for Research and Innovation Policy Studies (RISIS)[4] and (b) the NIH Research Portfolio

2 1NISTEP Company Name Dictionary, http://www.nistep.go.jp/research/scisip/rd-and-innovation-on-industry/
3 NISTEP Dictionary of Names of Universities and Research Institutes, http://www.nistep.go.jp/research/scisip/randd-on-university
4 RISIS, http://risis.eu/

Online Reporting Tools (RePORT)[5]. More specifically, (a) provides researchers with databases related to science, technology, and innovation through the RISIS Datasets Portal (datasets.risis.eu). For instance, PROFLE: The German Doctoral Candidates and Doctoral Holders Study provides individual data from a survey of doctoral candidates and doctorate holders at German universities and funding agencies to researchers who apply for access. RISIS has also developed the Semantically Mapping Service (SMS) Platform[6], and provides CorTEXT[7] data analysis platforms like Platform. The former allows researchers to combine and analyze science and technology indicator data provided by Web of Science, Scopus, PATSTAT, and the OECD; while the latter can automatically parse data in RDF, XML, and CSV formats, and perform graphing and network analysis. Meanwhile, (b) is a web interface that enables the analysis of principal investigators (PIs), amounts (direct and indirect costs), duration, related projects, and research outputs (e.g., papers and patents) in connection to NIH funds.

Conversely, Japan's scheme for developing and providing data related to science and technology is still in its infancy. In order to conduct research activities, scientists need to secure the necessary inputs for their research by acquiring funds through (1) research funds allocated by university administration grants, (2) competitive grants such as the Grant-in-Aid for Scientific Research and JST, and (3) joint research with companies. Of course, human resources such as postdocs and RAs are also important. However, with the exception of the database of Grants-in-Aid for Scientific Research provided by the National Institute of Informatics, information pertaining to the funds provided by national funding agencies (e.g., JST and NEDO) and ministries and agencies like MEXT and the Ministry of Economy, Trade and Industry has not been made open. Even where such data have been made available online, the databases themselves remain in their developmental stages. As a result, it is extremely difficult to determine how much money is being invested in a particular scientific area by the government or national research institutions. To rectify this situation, the GRIPS SciREX Center has established a web-based linkage system called the SciREX Policymaking Intelligent Assistance System (SPIAS), the details of which are described in the following section and discussed in greater detail in a later section.

In relation to scientists' output, JST is constructing a research article database, J-global, in collaboration with G-Search, Inc. In respect to patent databases, the JPO operates an SaaS-type system called J-Patplat, while the Institute of Intellectual Property provides a patent analysis database called the IIP Patent Database. However, only a few analyses linking these patents and papers together have been conducted, with Ikeuchi et al. (2017) pioneering such research in Japan. Additionally, in order to measure the relationship between both patents and papers and inputs and outputs in detail, data linkages between patents, papers, and competitive funding are essential. However, in terms of maintaining this kind of data, data linkages between papers, patents, and funds has only been carried out for some competitive grants, such as grants-in-aid for scientific research.

5 RePORT, https://report.nih.gov/
6 SMS Platform, http://sms.risis.eu/
7 CorTexT Platform, http://www.cortext.net/

As noted, in addition to patents and papers, the results of researchers' activities include their social contribution activities, media appearances, participation in councils of government-related organizations, degree of involvement in industry-academia collaboration, and participation in companies as outside directors or shareholders. However, due to the problem of data availability, the social effects of researchers are still qualitative and remain within the scope of case studies targeting specific researchers (Hara, Yasushi et al., 2017).

Meanwhile, in terms of leveraging text mining and natural language processing methods, approaches are being pursued to extract and analyze required information such as the names of researchers, organizations, and fields of contribution from atypical data (social media, newspaper articles, press releases) that record the social aspects of these researchers. In the SPIAS described above, the organization names are extracted from press release information collected by Nikkei Inc. and checked against companies, universities, and research institutes that have applied for patents and papers or have obtained funds. Accordingly, in addition to conventional methods of analysis assigning unique ID information to the name of the researcher or organization, it is becoming feasible to analyze not only the input and output of research activities but also outcomes using a combination of methods such as the approach of the RISIS SMS Platform, which measures likelihood based on natural language processing and links irregularly-shaped data.

1.2 Key science and technology indicators (microdata)

Widely-used indicators for scientific research at the researcher level include (a.) publications, (b.) patents, and (c.) funding data. Details are presented hereunder.

(a.) Research paper databases

- Scopus, provided by Elsevier, is a database of academic papers, mainly in English; researchers and organizations are given IDs, yielding slightly superior data consistency.
- Web of Science, provided by Clarivate Analytics, is a database of academic papers, predominantly in English. Data are provided from 1900, and the platform is essential for historical analysis; it is also widely used in existing research pertaining to star scientists.
- J-global is a database of academic papers and patents provided by the Japan Science and Technology Agency that covers an extremely wide range of Japanese journals.

(b.) Patent databases

- PATSTAT is a patent database provided by the European Patent Office (EPO) covering a wide range of patent databases in Europe, the United States, Japan, and three other major regions.
- PatentsView is a patent database provided by the United States Patent and Trademark Office (USPTO); it standardizes inventor and organization names.
- J-global is a patent database provided by the Japan Science and Technology Agency; it standardizes inventor and organization names.

- IIP Patent Database is patent database provided by the Institute of Intellectual Property; it covers Japanese patent data.

(c.) Funding information databases

- SPIAS is a data platform that comprehensively combines information on Japanese funds, patents, papers, and press releases; it was developed by the SciREX Center, NISTEP, and JST R&D Strategy Center.
- Grants-in-Aid for Scientific Research Database is a database provided by the NII and JST that combines detailed information on Grants-in-Aid for Scientific Research (KAKENHI), the resulting papers, and patent information.
- AMED Database is a database that publishes information on competitive grants and their results under the jurisdiction of AMED.
- Nanobank is a database of academic, publication, patent, and research funding information specific to nanotechnology; data are aggregated by organization name.
- COMETS is a database of patent and research funding information in all fields; data are aggregated by organization name.

As shown in the previous section, it is necessary to construct cohort data by joining these databases in order to obtain an encompassing picture of scientists' activities. However, several issues pertaining to data linkages emerge when creating cohort data, as follows: (1) there are fluctuations and misdescriptions of organization/researcher names in the original databases, and this information needs to be removed or cleaned; (2) it is necessary to develop a method to clean and identify information such as duplicate names or company names; (3) it is necessary to collate information such as organization names and researcher names written in different languages (e.g., names written in English and those written in Japanese) across different databases; and (4) when performing historical analysis, it is necessary to include information like the M&A information of companies. Meanwhile, in respect to patent and article data, as mentioned earlier, MEXT's National Institute of Science and Technology Policy is working on the development of relevant data tables, the use of which may facilitate the clarification of data. Similarly, issues with the management and operation of data include: (1) the vast size and complex structure of the bibliographic data of articles and patents, (2) the need for considerable computing resources to analyze data across multiple fields and years, and (3) the need for a high degree of security due to the data containing sensitive information, such as competitive funding information.

1.2.1 Tasks for the future

Going forward, it is necessary to examine methods for joining each database. As Figure 1 shows, each database should be joined based on the name of the organization or researcher. This requires the disambiguation and harmonization of researcher name data, as well as extracting and matching organization and researcher names using natural language processing techniques. Moreover, as extremely multivariate

and complex data structures will be processed, rather than relying on the RDBMS formats that have been traditionally used in the data preparation process of academic quantitative research, it will be necessary to consider and introduce analysis platforms that support irregularly shaped data, such as Neo4j and ElesticSearch (HARA Yasushi and KIUCHI Mitsutoshi, 2017). In order to conduct this analysis, it will also be necessary to develop a cloud platform infrastructure with abundant computing resources that researchers can freely utilize, as is being realized by the European research consortium through RISIS (SMS Platform, CorText Platform).

References

Zucker, G., Darby, M., and Armstrong, J. (2002). Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology. *Management Science*, 48(1). https://pubsonline.informs.org/doi/pdf/10.1287/mnsc.48.1.138.14274

Hara Yasushi, Kabeya Yukihiro, and Koizumi Amane. (2017). There is the light that never fades away: An empirical analysis for Nobel Prize Laureates. *Hitotsubashi Business Review*, 65(1):26–40. https://ci.nii.ac.jp/naid/40021245697/ [In Japanese]

Hara Yasushi and Kiuchi Mitsutoshi. (2017). Elasticsearch to kagaku gijutsu biggudēta ga kiri hiraku Nihon no chi no fukan to hakken [Elasticsearch and science and technology big data open up a bird's eye view of knowledge and discovery in Japan].
https://www.slideshare.net/yasushihara/elasticsearch-15-spias

Harada Hiroaki, Koshiba Hitoshi, Ikeuchi Kenta, Hara Yasushi, Kou Shunyou, Kuroda Masahiro, et al. (2017). Kagaku gijutsu inobēshon seisaku ritsuan no tame no dētapurattofōmu: Tekisutomainingu ni yoru kagaku gijutsu bun'ya no dōtei [A data platform for science, technology and innovation policy making: Identifying science and technology fields using text mining].
In Proceedings of the Annual Conference, volume 32, 344–347. Society for Research and Innovation.
https://dspace.jaist.ac.jp/dspace/bitstream/10119/15004/1/kouen32_344.pdf [In Japanese]

Ikeuchi Kenta, Motohashi Kazuyuki, Tamura Ryuichi., Tsukada Naotoshi, et al. (2017). *Kagaku gijutsu sangyō dēta no setsuzoku to sangyō no kagaku shūyaku-do no sokutei* [Connecting science, technology, and industry data and measuring the scientific intensity of industries].
Discussion Paper 142, Institute for Science and Technology Policy.
http://data.nistep.go.jp/dspace/bitstream/11035/3161/1/NISTEP_DP142_FullE.pdf [In Japanese]

Saito Hiromi and Maki Kaneta. (2017). Star scientists: The engine of innovation in Japan. *Hitotsubashi Business Review*, 65(1):42–56.
https://ci.nii.ac.jp/naid/40021245706/ [In Japan]