DISCUSSION PAPER No. 234

Database for Understanding Science-based Innovation by Linking Research Papers with Patent Information

学術論文と特許情報を用いたサイエンスイノベーション分析データベースの作成

2024 年 10 月

文部科学省　科学技術・学術政策研究所
第 1 研究グループ

池内 健太　塚田 尚稔　元橋 一之

本 DISCUSSION PAPER は、所内での討論に用いるとともに、関係の方々からの御意見を頂くことを目的に作成したものである。

また、本 DISCUSSION PAPER の内容は、執筆者の見解に基づいてまとめられたものであり、必ずしも機関の公式の見解を示すものではないことに留意されたい。

The DISCUSSION PAPER series is published for discussion within the National Institute of Science and Technology Policy (NISTEP) as well as receiving comments from the community.

It should be noticed that the opinions in this DISCUSSION PAPER are the sole responsibility of the author(s) and do not necessarily reflect the official views of NISTEP.

【執筆者】

池内 健太	経済産業研究所 上席研究員（政策エコノミスト）
	文部科学省科学技術・学術政策研究所 客員研究官
塚田 尚稔	新潟県立大学国際経済学部 准教授
	文部科学省科学技術・学術政策研究所 客員研究官
元橋 一之	東京大学先端科学技術研究センター 教授
	文部科学省科学技術・学術政策研究所 客員研究官

【Authors】

IKEUCHI Kenta	Senior Fellow (Policy Economist), Research Institute of Economy, Trade and Industry (RIETI)
	Affiliated Fellow, National Institute of Science and Technology Policy (NISTEP), MEXT
TSUKADA Naotoshi	Associate Professor, Faculty of International Economics, University of Niigata Prefecture
MOTOHASHI Kazuyuki	Professor, Research Center for Advanced Science and Technology, The University of Tokyo

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this paper.

池内健太・塚田尚稔・元橋一之(2024)「学術論文と特許情報を用いたサイエンスイノベーション分析データベースの作成」, *NISTEP DISCUSSION PAPER*, No.234, 文部科学省科学技術・学術政策研究所.

DOI: https://doi.org/10.15108/dp234

IKEUCHI Kenta, TSUKADA Naotoshi, and MOTOHASHI Kazuyuki (2024) "Database for Understanding Science-based Innovation by Linking Research Papers with Patent Information" *NISTEP DISCUSSION PAPER*, No.234, National Institute of Science and Technology Policy, Tokyo.

DOI: https://doi.org/10.15108/dp234

学術論文と特許情報を用いたサイエンスイノベーション分析データベースの作成

文部科学省 科学技術・学術政策研究所 第 1 研究グループ
池内 健太 塚田 尚稔 元橋 一之

要旨

本稿においては、学術論文（Open Alex）と特許情報（日本国特許庁公開情報）を用いたサイエンスイノベーションに関する分析用データベースを作成し、研究者のキャリアと生産性に関する分析を行った。まず、学術論文に関する公開情報である Open Alex から Web of Science の SCIE(Science Citation Index Expanded)に該当するジャーナル論文を抽出し、かつ著者情報から所属機関が日本に所在する著者による論文を取り出した。特許情報については、日本国特許庁の公開特許情報のうち、発明者住所が日本に所在する特許を抽出し、出願人情報等を用いて発明者の所属機関の推定を行った。次に上記の学術論文及び特許情報を用いて著者・発明者の識別作業を行った。具体的には、同一人物による論文・特許であるか否かの識別モデルを機械学習によって構築し、その情報を用いて同一著者・発明者による論文・特許をクラスター分析によって推定した。その結果については、科研費データベース（同一研究者による論文、特許情報）によって評価を行い、論文著者ペアについては 96.87%、特許発明者ペアについては 98.71%、論文著者・特許発明者ペアについては 81.64%の F1 スコアを得た。さらに、当該データを用いて著者・発明者ごとの経験年数と生産性に関する記述統計を作成した。その結果、大学に所属する著者についてはキャリア年数と論文数との間には上昇ののち、下降する逆 U 字型の関係が観察されたが、企業に所属する著者については明確な関係が見られなかった。特許出願数については、大学発明者、企業発明者ともキャリア年数の長い発明者については逆 U 字型の関係が観察された。

Database for Understanding Science-based Innovation by Linking Research Papers with Patent Information

First Theory-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP), MEXT
IKEUCHI Kenta, TSUKADA Naotoshi, MOTOHASHI, Kazuyuki

ABSTRACT

In this paper, we developed an analytical database for scientific innovation using academic papers (Open Alex) and patent information (from the Japan Patent Office), and analyzed the career trajectories and productivity of researchers. First, we extracted journal articles from Open Alex corresponding to the Science Citation Index Expanded (SCIE) of Web of Science, focusing on the authors who are affiliated with institutions in Japan. Second, we extracted patent information in which the inventors are listed in Japan as their addresses from the publicly available information of the Japan Patent Office, and estimated the inventors' affiliated institutions based on applicant details and other relevant information. Next, we identified authors and inventors by linking the academic papers with patent information. Specifically, we employed machine learning to build a model that identifies whether a given paper and patent were written by the same individual. Cluster analysis was then used to

estimate papers and patents written by the same researcher. We evaluated the results by using the Grant-in-Aid for Scientific Research Database, which contains both papers and patent information attributed to the same researchers. The model achieved F1 scores of 96.87% for paper author pairs, 98.71% for patent inventor pairs, and 81.64% for paper author/patent inventor pairs. Furthermore, using this database, we presented descriptive statistics on the career length and productivity of each author/inventor. As a result, we observed an inverted U-shaped relationship between the career length and the number of published papers for university-affiliated authors, with productivity initially rising and then declining. However, no clear relationship was observed for company-affiliated authors. Regarding the number of patent applications, an inverted U-shaped relationship was also observed for long-career inventors, both in universities and companies.

1．はじめに

新製品やサービスなどのイノベーションにおいて、科学技術の重要性が認識されて久しい。ライフサイエンスや材料科学などの多くの分野で産学連携が活発になっており、大学などの科学的知見をベースとしたイノベーションの重要性はますます高まっている（元橋、2014）。また、大学等における科学的知見を活用することによって、企業においてより画期的なイノベーションを実現することが可能であることが分かっている（Veugelers and Wang, 2019; Motohashi et. al., 2024）。多くの公的研究資金が投じられている大学等における研究成果が企業のイノベーションに与える影響を分析することはエビデンスベースの政策立案（EBPM）の観点からも重要である。

このようなサイエンスベースのイノベーションのプロセス（大学等における科学的発見、その製品技術への適用、市場普及による経済的効果）について分析するためには、論文、特許、企業活動データ等を相互に接続することが必要である。筆者らは、この作業を長年行ってきており、その成果をイノベーションプロセスデータベース（Innovation Process DataBase: IPDB）と称して公表してきている（Ikeuchi et. al., 2017; Ikeuchi and Motohashi, 2022; Motohashi et. al., 2023）。

ここでは、そのデータベース作成事業の一環として、学術論文と特許情報を活用した研究者レベルデータベースの作成とその内容について述べる。学術論文についてはウェブ情報公開情報を元に作成された Open Alex を、特許については日本国特許庁による公開情報を用いて日本における著者・発明者に関する網羅的なデータベースを作成した。そのプロセスにおいては、論文著者と特許発明者を同時に識別（Disambiguation）することにより、同一人物による論文と特許の特定を可能にした。IPDB の旧バージョンにおいては特許発明者の発明者識別を行い、その後に発明者と論文による著者情報を接続する二段階のプロセスを経て作成していた（Ikeuchi et. al., 2017; Ikeuchi and Motohashi, 2022）。今回の作業においては、これらを同時に行うことによって、より精度の高い研究者レベルの識別と著者発明者接続が行われたと考えられる。

本稿においては、まず次章では、分析に用いたデータ、すなわち Open Alex と特許情報の内容について説明する。第 3 章では、これらのデータの著者、発明者の識別作業の内容とその結果について述べる。第 4 章において、本データベースを用いた記述統計の結果として、研究者特性による科学的活動と特許活動の違いについて述べる。最後にまとめとして本データベースを用いることで可能となる研究の方向性について述べる。

２．データベース作成に用いた論文、特許情報

２－１．論文データ

２－１－１．Open Alex について

学術論文情報については商用データベースとして SCOPUS（Elsevier 社）と WoS: Web of Science（Clarivate 社）が有名であるが、これらのバルクデータの利用には多額の費用が発生する。また、基本的には一定期間内（例えば 1 年間）のデータ使用契約となっており、当然のこととして契約上データベースを利用できる人は制限されている。一方で、インターネット上で公開されている学術論文情報をクローリングで取得し、データベース化しているものに Google Scholar や Microsoft Academic Graph (MAG)などが存在する。特に MAG はバルクデータを一般公開しており、論文情報による Bibliometric 関係の研究者においても活用されている。

なお、MAG と WoS や SCOPUS などの商用データベースとの比較については、様々な研究が行われている（Paszcza, 2016; Harzing, 2016; Hug and Brändle, 2017）。総じていうと、MAG や Google Scholar などのウェブベースからの情報で作成されたデータベースは、WoS や SCOPUS などの商用データベースと比べてカバレッジは大きい。一方で、商用データベースはそれぞれの運営会社によって論文出版情報の整理が長年行われてきており、書誌情報の正確性の観点では、WoS や SCOPUS の方が優れている。なお、MAG と SCOPUS の論文単位の比較を行った著者らの分析によると、論文単位の著者数、出版年数等の両者間でほぼ整合的な内容であることが分かっている。その一方で、MAG においては論文著者の所属機関の情報等において大量の欠損が見られるという問題が分かっている（塚田・元橋, 2018）。

Open Alex は、Microsoft 社が MAG の運用を 2021 年末で停止すると決定したのを機に、NGO がその内容を引き継ぎ、データのアップデートと公開作業が行われているものである（Priem et. al, 2022）。Open Alex は works, authors, venues, institutions 及び concepts という 5 つの entry から構成されており、定期的にアップデートされている情報を API によって取得することが可能である（https://docs.openalex.org/）。また、過去情報も含めたバルクデータは Snapshot データということで自由にダウンロード可能となっている。

２－２－２．日本著者による論文データベースの作成

IPDB オープンにおける論文情報は、Open Alex における 2022 年 11 月時点の Snapshot データ（RELEASE 2022-10-10）をダウンロードし、①論文の RDB (Relational DataBase) 化、②Web of Science の SCIE (Science Citation Index Expanded)論文の抽出、③日本著者による論文のデータベースの作成、という手順で作成した。

まず、Open Alex データは works などの 5 つの entry それぞれに XML 形式でデータが格納されるため、データ全体の正規化（RDB 化）を行った。RDB のテーブルとしては、論文に関係するもの、著者に関係するもの、著者所属機関に関係するもの及び掲載誌に関するも

の合計 10 テーブルを用意した。これらの接続キーは論文 ID、著者 ID、機関 ID 及びジャーナル ID の 4 つである。なお、これらの ID については基本的に MAG のものを引き継いでおり、MAG との相互接続は可能である。論文数が約 2.4 億件、著者数が約 3.1 億件の大規模なデータベースである。ただし、すべての論文の著者情報、所属機関情報が存在するとは限らないことに留意が必要である。論文数約 2.4 億件に対して、著者情報が存在するものは約 0.9 億件にとどまる。また、著者情報があるものについても、その中で機関 ID が振られていない論文数は約 900 万件存在する。

次に、これらの論文から WoS の SCIE ジャーナルに該当する論文を取り出すことにした。Open Alex には、学術的ジャーナルにおける論文の他、学位論文等、ウェブ上で収集できる文献すべてが含まれている。科学的な研究活動の成果を正確に計測するためには、一定の質以上の研究論文にフォーカスすることが適当である。そこで、自然科学分野のジャーナルの基準である WoS の SCIE 論文に該当するものをデータベースに用いた。

具体的には Clarivate 社が公開している SCIE ジャーナルの ISSN 情報を Open Alex の論文データにマッチングを行い、該当する論文（及び関連する著者情報等）を抽出した（論文数：約 5500 万件、著者数：約 4200 万件）。その際には学術誌等の ISSN 情報を含む論文掲載元（Venue）の情報が必要となるが、Open Alex の総文献数である約 2.43 億件のうち Venue_id が付与されていない文献（ISSN 情報もなし）は約 1.15 億件存在する。ただし、文献のタイプが Journal Article となっている約 1.22 億件のうち、Venue_id がないものは 200 万件程度（約 2%）である。さらに、Open Alex において文献タイプ情報がないもの（doc_type なし）が約 7,000 万件存在し、これらの一部は学術研究論文である可能性はある。ただし、SCIE の対象となっているジャーナルの ISSN 数約 17,000 件に対して、Open Alex において対象文献が見つからなかったものが 270 件のみだったので、Open Alex 収録論文のうち SCIE 論文は概ね抽出できていると判断した。

前述したとおり、Open Alex 本体の約 2.4 億件のうち著者情報が存在するものは約 0.9 億件にとどまったが、SCIE 論文のみを抽出した結果、著者情報が存在する論文数は全体の約 5,500 万件のうち約 3,800 万件となった。

最後に、これらの SCIE 対象論文から、日本著者の論文の抽出を行った。具体的には、著者所属情報からその住所が日本である著者を特定し、該当する著者が一人以上いる論文すべてを取り出した。ここではあくまで論文ごとの著者所属機関の住所による国内判定を行ったものであり、国籍等の著者毎の属性で識別したものではないことに留意が必要である。例えば日本研究者であっても、一定期間海外の研究機関の所属している場合は、その期間中に公開された論文は対象外となっている。逆に、国内における外国人研究者の論文は含まれる。

著者所属機関は、Open Alex における institution_id を用いる。ただし、この institution_id もすべての文献著者に付与されているものではなく、SCIE 論文として取り出された文献のうち著者情報があるもの約 3,800 万件のうち、約 900 万件については institution_id が付与

されていない著者が存在する。その場合、情報は機関名情報のみで機関所在国情報はない。これらのレコードは日本著者の抽出作業から外して行った。

上記のプロセスを経て、IPDB オープンにおける論文データベースとして、論文数が2,379,659 件、論文・著者レコード数約 1,000 万件の RDB を作成した。テーブル構造やレコード数等については図 1 のとおりである。

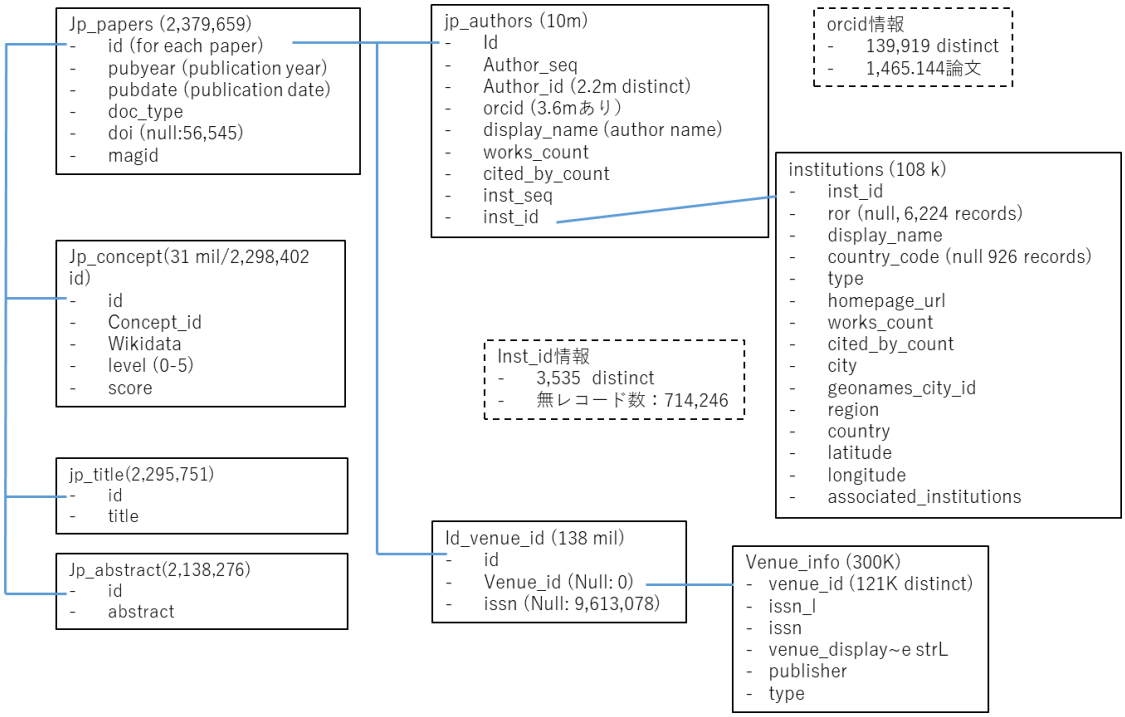

図 1．日本著者 SCIE 論文データベースの構成

2－2－3．WoS データベースとの比較

上記のプロセスによってオープンデータベースである Open Alex（OA）を用いて日本著者を含む WoS の SCIE 論文データベースを作成して、商用データベースである WoS と比較した。図 2 は出版年別の論文数であるが、全体的に OA の論文数が下回っている。

両者の乖離については、①OA 全体から WoS の SCIE ジャーナルを抜き出す作業、②OA における SCIE ジャーナル論文のカバレッジ及び③SCIE 論文として抜き出されたものから日本著者論文を抽出作業のそれぞれのプロセスで発生すると考えられる。まず、①については、前述したとおり ISSN ベースでの SCIE ジャーナルの抽出を行ったところ、総数約 17,000 ジャーナルのうち、OA 論文が見つからなかったものが 270 件のみであり、大きな格差にはつながらないと推測される。②については、OA がインターネット公開情報をベースにしていることから、より出版年が新しい論文ほど情報として入手出来ている可能性が高いと推測される。図 2 にみるように両者の乖離は出版年が新しくなるほど縮小しており、当該要因が想定程度影響していると推測される。なお、個々のジャーナル毎の論文収録件数を比較し

た際には、OA が WoS を上回っているものも見られた。その内容について詳細に見たところ、OA においてはジャーナル内の書評など学術論文ではないものも論文として含まれていることがあり、両者の違いについてはより詳細に検討することが必要である。最後の③であるが、前述のとおり、SCIE 論文全体でみると、約 5,500 万件の論文のうち約 1,700 万件（著者情報あり約 3,800 万件以外）について日本人著者論文の抽出ができていないことになる。この割合が日本著者論文にも当てはまるとすると約 3 割欠損していることになり、この差が図 2 の乖離の主な原因となっている可能性が高い。

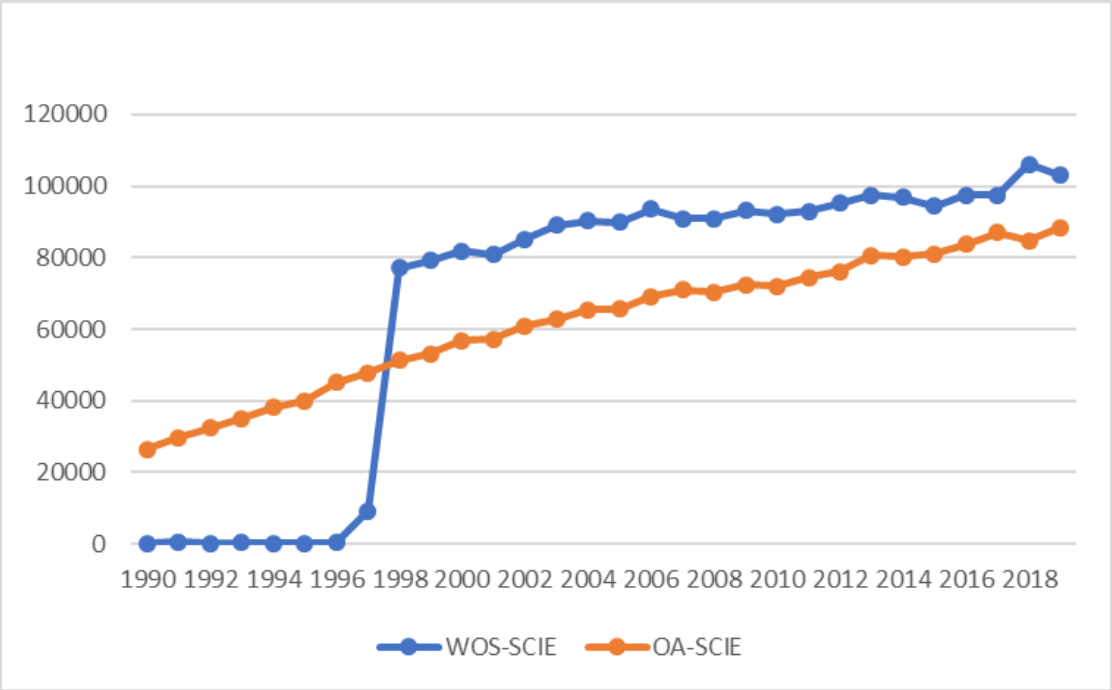

図 2．日本著者 SCIE 論文の WoS との比較

２－２．特許データ

２－２－１．分析に用いた特許情報

出願特許の公開制度によって、特許情報は出願があった日から 18 か月以内に公開される。したがって、特許情報は基本的には公開データであり、企業等の研究開発の状況を詳細に表すデータとして多くのイノベーションの実証研究に利用されている。日本特許に関する情報としては、大きく以下の 3 種類が存在する（その他に民間データベース会社によるものがある）。

- 公報情報（特許公開公報、登録情報等）：特許法に基づく特許公開情報
- 整理標準化データ：特許庁内の審査データベースが元となる公開データベース
- PATSTAT：EPO が公開する世界の出願特許データベース（各国特許庁の公開情報を集約したもの）

公報情報についてはインターネットで公開されており、整理標準化データも特許庁に登録することで適時アップデート情報の提供を受けることが可能である。PATSTAT は EPO からのサブスクリプションによって入手可能で、年 2 回データのアップデートが行われている。

公報情報は特許法に基づく公開情報でオフィシャルなものといえる。しかしながら、1990 年代以前の情報は電子化されていないという問題がある。一方で、整理標準化データは日本国特許庁に対する出願特許がすべて収録されており、同データを基に作成された研究者用特許データベースである IIP パテントデータベースもこちらの情報に準拠している（Goto and Motohashi, 2007）。ただし、2014 年以降の出願特許について、個人情報保護の観点から出願人や発明者の住所情報について、市町村レベルまでの公開に留まっているという問題がある。出願人等の場所は、後述する出願人名寄せの際にも用いており、最近の出願特許の詳細な住所情報を特定するためには、特許公報の情報を用いる必要がある。最後に PATSTAT における日本特許情報は、RDB 化された状態で年 2 回定期的にアップデートされており、データベースのバージョン管理が容易であること等の利点がある。ただし、発明の名称、要約、出願人名等の文字情報が英語に変換されているという問題がある。

このような状況に鑑み、今回のデータベース作成における特許情報としては以下の方針で上記 3 つのデータを組み合わせて作成した。

- 1990 年以降出願特許を対象として、PATSTAT2022 年秋バージョンにおいて収録されている特許を中心とする。日本語文字情報以外（出願日、公開日、登録日、IPC 分類、請求項数等）の情報については PATSTAT の情報を用いる。
- 出願人名・住所、発明者名・住所、発明の名称・要約等の日本語テキスト情報は、公報情報を用いることとし、1990 年代の出願特許については整理標準化データの情報で補完する。

２－２－２．出願人名寄せと出願機関タイプ情報の付与

上記のオリジナルの特許情報においては出願人の情報に表記ゆれが存在しており、研究者の所属機関に着目した分析（例えば、研究者の所属機関間の移動）を行うためには出願人情報の名寄せ作業が必要となる。また、機関の属性（いわゆる個法官コード）を特定することによってより分析の幅を広げることが可能となる（例えば、研究者の大学から企業への移動）。

そこで、以下の方法によって名寄せ作業と個法官コード付与を行った。まず、特許庁における出願人番号を使った名寄せを行った。ここで、出願人番号が付与されていない特許-出願人レコードについてはそれぞれ別の出願人として識別する。出願人番号は異なる出願人に対して同一番号が付与される Lumping Error は存在しないが、その逆の Splitting Error は存在する。したがって、ここでは名寄せの最小単位を作成して、その後の作業に

おいてこれらをさらに統合させていく方針とする。

その後、大学や公的機関、企業と出願人タイプ別にそれぞれ名寄せ情報が存在するので、出願人の機関タイプ識別を行った。ここでは個人（発明者）、営利企業、非営利機関及び大学の 4 種類に類型化をしたが、個人については、同一特許内において出願人名称＝発明者名称となっているものとした。それ以外については、ルールベース（例えば、「●●大学」については大学に分類）で割り振った。この内容をベースに以下の内容を加えて、出願人名称名寄せと機関タイプ識別の精緻化を行った。

- NIETEP 大学・公的機関名辞書（科学技術・学術政策研究所, 2022）：WoS SSCI ジャーナル掲載論文における日本における著者所属機関に関する辞書（大学、公的機関の他、企業の含まれる）：主に大学・公的研究機関に関する出願人については本データベースにおける大学・公的機関名と名称のファジーマッチを行うことで名寄せと機関タイプ分類の修正を同時に行った。
- NISTEP 企業名辞書(科学技術・学術政策研究所, 2023)：1970 年以降の累積出願特許数が 100 件以上の企業等、約 2.5 万社について企業情報を取りまとめたものであり、NISTEP における特許情報との接続作業も公開されている。当該辞書の企業 ID を用いて名称名寄せと機関タイプ分類の修正を同時に行った。

最後にこれらの情報によって特定できなかった特許-出願人情報については、出願人が同一名称でかつ住所が近接しているもの（住所情報の Geocoding 結果において、緯度・経度の差がそれぞれ 0.01 度以内のもの）を同一出願人によるものとした。

２－２－３．発明者所属機関情報の推計

著者・発明者の識別作業を行うためには、特許情報における発明者の所属機関に関する情報が必要である。ここでは上記のプロセスで得られた出願人情報を元に以下のルールによって発明者所属機関の識別を行った。

A） 同一特許内において発明者住所＝出願人住所となっている発明者については、当該出願人をその所属機関とする。

B） 単独出願人特許については、この出願人を当該特許の各発明者の所属機関とする。

ただし、B)の場合において、複数発明者のうち、一部の発明者の住所＝出願人住所となっているものがある場合は、それ以外の発明者（発明者住所≠出願人住所のもの）の所属は「不明（出願人を当該発明者の所属機関としない）」こととした。これは、国立大学法人化前において、国立大学が産学連携特許の出願人にはならないが（相手先企業の単独出願特許となる）、当該発明にかかわった大学研究者は発明者の一部を構成するケースに対応したものである。

したがって、上記のプロセスにおいて、複数出願人特許で発明者住所≠出願人住所であるケースと単独出願人特許で B)の例外として除かれたものについては、発明者の所属機関は不明として取り扱われることに留意が必要である。

３．特許発明者と論文著者の同一人物性の特定（Disambiguation）

論文情報は Open Alex において author_id が付与されているが、特許情報には発明者の氏名、住所情報のみで、同一氏名が同一人物とは限らない問題がある。この同姓同名の問題を処理し、発明者毎の id を作成する作業を発明者同定（inventor disambiguation）とよび、例えば米国特許情報については、米国特許商標庁（USPTO）が自ら行い、公表している。日本国特許庁は同様の作業を行っていないことから、従来の IPDB においては著者らが当該作業を行ってきた（Ikeuchi et. al., 2017; Ikeuchi and Motohashi, 2022）。

具体的には、共同発明者、特許の技術分類（IPC）、出願人の名称、発明者の所属先及び住所情報を用いて、異なる特許間における同姓同名発明者が同一人物であるか否かの判定を行う。この作業は機械学習モデルを使って行うが、教師データとしてはレアネーム情報（複数年の電話帳に 1 回もしくは 1 度も掲載されていない発明者の氏名）を用いている。次に、同一発明者氏名による特許グループを作成して、上記のモデルによって算出された対象特許ペアごとの同一発明者割合の情報を用いて、クラスタリングによって異なる発明者による特許に分割する作業を行った（詳細については、Ikeuchi and Motohashi, 2022 参照）。従来の IPDB においては、この結果得られた特許発明者の ID 毎に同姓同名で同一所属機関の論文著者情報をあてはめ、著者・発明者接続を行ってきた。

今回の作業においては、これまでのように発明者識別→論文著者情報とのマッチングという 2 段階を踏むのではなく，論文著者と特許発明者の識別を同時に行うことで両者の接続に関する情報も得る方法で作業した。機械学習モデルの用いるパラメータのうち、共同発明者（共著者）、出願人名称（著者所属機関名称）及び発明者の所属先及び住所（著者所属機関名称、住所）については同等の情報が利用可能である。特許の技術分類(IPC 分類)は個々の論文に存在しないため、両者の要約の内容類似度を用いることとしている。具体的には、論文・特許の要約の Embedding ベクトルのコサイン類似度で内容の近接性を図り、閾値を設けて同等の内容であるかどうかのパラメータを作成することとする。ここで得られた結果については、科研費データベースに収録されている特許及び論文の研究成果情報との整合性などを通じて、その評価が可能である。

ここで、分析対象とするデータは前節で述べた特許情報と論文情報のうち、発明者及び著者の住所が日本にあるデータである。また、特許については出願年、論文については出版年が 1990 年以降のものに限定した上で、英語表記の発明者及び著者の氏名にアルファベット以外の文字を含むデータは除外した。特許発明者は 18,181,590 件の特許×発明者単位のデータ、論文著者は 7,688,517 件の論文×著者単位のデータを分析対象となった。

特許発明者と論文著者の同一人物性の特定は、データの正確なマッチングを必要とする複雑な作業である。このプロセスでは、まずレアネームを用いて教師データを作成する。具体的には、特許発明者同士、論文著者同士、そして特許発明者と論文著者のペアを組んだレコードを用意する。これらのペアに対し、名前の一致度や所属機関の一致度などの類似度ベ

クトルを定義し、同一人物性を判定するための分類器を学習させる。

次に、クラスタリング手法を用いて、これらのレコードを同一人物グループに分ける。クラスタリングの精度を高めるため、氏名によるブロッキングを行い、学習済み分類器を用いてレコードペアの同一人物性の確率を予測する。最終的に、予測確率に基づいてクラスタリングを行い、同一人物性を高精度に特定する。

この手法の精度は、独立した検証用データセットを用いて検証し、精度、再現率、F1 スコアなどの指標で評価する。これにより、特許発明者と論文著者のデータを統合し、イノベーションの全体像を把握するためのデータベース構築が可能となる。以下では、特許発明者と論文著者の同一人物性を特定するための分析手順を詳細に説明する。

3.1. レアネームを用いた教師データの作成

3.1.1. 分類器用の教師データの作成（レコードペア単位）

教師データの作成には、特許発明者と論文著者のレコードをペアにして使用する。同じレアネームを持つ発明者及び著者は同一人物、名前の異なるレアネームの発明者及び著者は別人とする。以下の 3 つのペアの教師データを作成する。

- 特許発明者・特許発明者のペア
- 論文著者・論文著者のペア
- 特許発明者と論文著者のペア

なお、レアネームは 2000 年～2012 年の電話帳のデータに基づいて定義する。電話帳のデータを英語表記に変換して、姓と名を分離した上で、出現頻度の低い姓と出現頻度の低い名の組み合わせをレアネームとした。レアネームの基準はなるべく厳しい（なるべく出現頻度の低い姓と名の組み合わせ）ほど、精度は高いと考えられる。一方、レアネームの基準を厳しくすると教師データのサンプルサイズが小さくなる。そこで、基準を変えながら分類器の精度を検証したところ、サンプルサイズが 3 万件を下回ると分類器の汎化性能が低下する傾向がみられたため、サンプルサイズが 3 万件以上で、なるべく厳しいレアネームの基準を採用することとした。

最終的には、特許発明者のペア及び論文著者のペアについては出現頻度がそれぞれ 10 件以下の姓と名の組み合わせ、特許発明者と論文著者のペアについては出現頻度がそれぞれ 18 件以下の姓と名の組み合わせをレアネームと定義して用いた教師データとした分析結果を採用することとした。教師データに含まれる特許発明者のペアは 2,792,494 件、論文著者のペアは 92,338 件、特許発明者と論文著者のペアは 30,290 件である。なお、同一人物のペア（レアネームが一致するペア）と別人物のペア（レアネームが一致しないペア）の割合は半々になるように調整している（全ての可能な組み合わせを別人物のペアの割合が非常に大きくなるため、同一人物のペアの数と等しくなるように、別人物のペアは無作為に抽出した）。

3.1.2. クラスタリング用の教師データの作成

クラスタリングの教師データには、出現頻度がそれぞれ 10 件以下の姓と名の組み合わせのレアネームをもつ特許発明者と論文著者のデータを用いる。一方、クラスタリングは最終的には特許発明者及び論文著者の全体のデータではなく、あらかじめ氏名でグルーピング（ブロッキング）したデータに対して、その中で同一人物のグループを抽出する際に適用される。そのため、規模の異なる複数のグループに対して、バランスよく適用可能になるようなクラスタリングのモデルの選択及びハイパーパラメーターのチューニングが必要である。

そこで、電話帳のデータを用いて推定された英語に変換された日本人の氏名の同姓同名の人数の分布に基づいて、レアネームをグルーピングすることで、クラスタリング用の教師データを作成した。

1,446 人のレアネームを持つ特許発明者及び論文著者のデータを教師データとして用いる。そのうち、特許のみは 1,210 名、論文のみは 98 名、特許と論文の両方を持つのは 138 名であった。なお、教師データに含まれる特許は 8,366 件、論文は 2,273 件であった。

3.2. Similarity Vector の定義

レコードペアの同一性を判定するために、各ペアに対して類似度ベクトル（Similarity Vector）を定義する。このベクトルには、名前の一致度、所属機関の一致度、共著者の一致度など、複数の類似度指標が含まれる。具体的には以下のような変数を用いた。

- 特許・特許ペア
  - 共同発明者の氏名：Jaccard 係数
  - IPC クラス（3 桁）：Jaccard 係数
  - IPC サブクラス（4 桁）：Jaccard 係数
  - 出願人（NID・CID）：Jaccard 係数
  - 発明者の所属先の出願人（NID・CID）：所属が異なる→0、いずれかの所属が不明→0.5、所属が同一→1
  - 発明者住所の近接性＝1－（距離÷3,000km）
  - 概要文：300 次元の分散表現のコサイン類似度

- 論文・論文ペア
  - 共著者の氏名：Jaccard 係数
  - 所属先機関（NID）：Jaccard 係数
  - 発明者住所の近接性＝1－（距離÷3,000km）
  - 概要文：300 次元の分散表現のコサイン類似度

- 特許・論文ペア
  - 特許共同発明者・論文共著者の氏名：Jaccard 係数
  - 特許出願人・論文著者の所属機関（NID）：Jaccard 係数
  - 発明者の所属先の出願人・所属機関（NID）：Jaccard 係数
  - 発明者住所・所属機関住所の近接性＝1－(距離÷3,000km)の最大値
  - 概要文：300 次元の分散表現のコサイン類似度

ここで、Jaccard 係数とは 2 つの集合の類似度を示す指標であり、以下のように定義される。

集合 A・B の Jaccard 係数＝（A と B の積集合の要素数）÷（A と B の和集合の要素数）

また、NID は「NISTEP 大学・公的機関名辞書」の機関 ID、CID は「NISTEP 企業名辞書」の企業 ID を示す。

3. 3. レコードペア単位の教師データを用いた同一人物性を判定する分類器（Classifier）の学習及び手法の選択

以下の 3 つのペアごとに、同一人物性を判定するための分類器を学習させる。

- 特許発明者・特許発明者のペア
- 論文著者・論文著者のペア
- 特許発明者と論文著者のペア

候補としたのは以下の分類器である。

- ハイパーパラメーターなしの手法
  - Gaussian Naive Bayes（GNB）
  - 線形判別分析（LDA）
  - 二次判別分析（QDA）
- ハイパーパラメーターありの手法
  - Random Forest
  - AdaBoost（Adaptive Boosting）
  - Gradient Boosting
  - XGBoost（eXtreme Gradient Boosting）

分類器のハイパーパラメータのチューニングには、K 分割交差検証によるランダムサーチを適用した。全データを用いると膨大な時間がかかるため、教師データから無作為抽出した 3 万件の一部サンプルでチューニングを行う。5 分割交差検証×500 回のランダムサーチを行い、最終的にチューニングされたハイパーパラメータを用いて教師データ全体で学習

させた。

分類器の選択基準としては、確率閾値に依存しない ROC-AUC とする。精度評価のための検証用データは教師用データと同様に 10%ホールドアウトしておいたレアネーム由来のデータとする。

表 1 は検証用データに対する分類器の精度比較の結果を示す。結果として、3 種類のデータのいずれにおいても XGBoost の精度が最も高い結果となった。

表 1. レアネーム由来の検証用データによる分類器の精度比較：ROC-AUC

	発明者×著者	著者ペア	発明者ペア
GNB	89.82%	96.22%	98.23%
LDA	89.27%	97.27%	98.94%
QDA	51.99%	96.08%	98.26%
RandomForest	98.28%	99.33%	99.59%
AdaBoost	99.22%	99.36%	99.77%
GradientBoost	98.70%	99.35%	99.68%
XGBoost	**99.42%**	**99.46%**	**99.92%**

3. 4. 教師データを用いたクラスタリング手法の選択とパラメータのチューニング

クラスタリング手法として、グラフ分割法（Graph Partitioning）及び DBSCAN の手法を検討し、教師データを用いて最適な手法とパラメータを選択する。これにより、同一人物を含むグループを形成する。候補としたクラスタリングの手法は、以下とおりである。

- グラフ分割法（パラメータ：エッジをつなぐ 3 つのデータタイプ別の確率閾値）
  - Connected Components (CC)：隣接するノードの集合
  - Label Propagation - NetworkX (LP-Net)：ラベル伝搬法（隣接するノードにラベルを伝搬）
  - Greedy Modularity Maximization (GMM)：モジュラリティ指数（クラスタ内のエッジの割合）を最大化
- DBSCAN
  - パラメータ：
    - 確率（p）を距離（d）に変換する際の 3 つのデータタイプ別の指数（a）：d=ln(1/p^a)
    - eps：クラスターを探索する半径距離（eps）
    - min_samples：クラスターの最小データ数は 1 に固定
    - metric='precomputed'：距離行列を入力データとする

レアネーム由来の教師データでパラメータをチューニングする。データタイプごとの確率閾値もパラメータとして最適化する。評価基準は F1 スコアを用いる（クラスタリングのためには確率閾値を決める必要があるため)。表 2 に示すように、結果として LP-Net の精度が最も高くなった。

表 2. クラスタリング手法の精度評価：最大 F1 スコア

手法	F1 score
LP-Net	**99.21%**
CC	98.73%
DBSCAN	98.72%
GMM	97.03%

3.5. 氏名によるグループ分け（Blocking）

計算効率を向上させるため、まず氏名によってレコードをグループ分けする（ブロッキング)。これにより、同じ名前を持つレコードのみを対象に類似度計算を行うことができる。

3.6. 学習済み分類器を用いたレコードペアの同一人物性である確率の予測

学習済みの分類器を用いて、各レコードペアに対して同一人物性である確率を予測する。これにより、同一人物である可能性が高いペアを特定する。

3.7. 予測確率を用いたクラスタリング

予測確率を基に、レコードをクラスタリングして、同一人物と判断されるグループを形成する。このクラスタリングにより、特許発明者と論文著者の同一人物性を高精度に特定することができる。

なお、最後に、ファーストネームがイニシャルしか得られない論文著者については姓の一致するブロックの中での予測確率を計算し、予測確率が 95%以上でかつ最も予測確率の高いクラスターに紐づけた。

表 3 は最終的な特許発明者と論文著者の同一人物性の識別の結果を示している。25,477,918 件の特許及び論文データから合計で 3,229,025 人の発明者及び著者が識別された。

表 3. 特許発明者と論文著者の同一人物性の識別の結果

	合計	フルネーム			ファーストネームがイニシャル
		合計	複数レコード	単一レコード	
レコード数（特許＋論文）	25,477,918	25,444,389	24,989,363	455,026	33,529
ユニークな氏名の数（ブロック数）	1,425,143	1,425,143	970,117	455,026	-
同定されたクラスター数	3,229,025	3,229,025	2,773,999	455,026	-
1 ブロックあたりの平均クラスター数	2.266	2.266	2.859	1.000	-
1 クラスターあたりの平均レコード数	7.890	7.880	9.008	1.000	-

3. 8. 検証用データを用いた精度の検証

最後に、独立した検証用データセットを用いて、分類器の精度を検証する。精度の指標として、精度（Precision）、再現率（Recall）、F1 スコアなどを使用する。

レアネーム由来の検証用データに加えて、科研費由来の検証用データに対する精度を評価した。表 4 は最終的な分類器の精度評価の結果を示している。なお、検証用のデータについても同一人物であるペアと同一人物でないペアの比率が同一になるように最大 200 万件のデータを無作為に抽出した。レアネーム由来の検証用データについては ROC-AUC や PR-AUC が 99%以上、F1 スコアも 96%以上と非常に高い精度を示している。科研費由来のデータについてはやや精度が低くなり、特に発明者と著者のペアに対しては、F1 スコアは 88.75%と 90%を下回っている。特に、再現率（Recall）が 86.65%と低くなっている。特許発明者と論文著者という別の種類のデータにおける同一人物性の識別は同一種類のデータにおける識別と比べて難しい課題であると考えられる。

次に、表 5 は最終的なクラスタリングの後の結果について科研費由来の検証用データを用いて精度を評価した結果である。なお、科研費由来の検証用データに含まれる成果特許数は 18,257 件、成果論文数は 481,233 件である。全ペアにおける精度は F1 スコアが 98.41%と非常に高い精度を示している。ただし、発明者と著者のペアについては、F1 スコアが 81.64%と精度が低くなっている。特に、再現率（Recall）は 70%を下回ってしまっている。これは、本当は同一の科研費番号を持つ研究者である特許発明者と論文著者が別人物であると誤って識別してしまっているケースが 3 割の比率で生じていることを示している。つまり、いわゆる Splitting エラーが多く生じていることになる。

表 4. 分類器の検証用データに対する精度

データ	発明者同士のペア		著者同士のペア		発明者と著者のペア	
分類器	XGBoost		XGBoost		XGBoost	
検証用データ	レアネーム	科研費	レアネーム	科研費	レアネーム	科研費
N	310,278	58,368	10,260	4,000,000	3,366	443,002
ROC-AUC	99.93%	98.00%	99.46%	98.24%	99.42%	94.82%
PR-AUC	99.92%	98.41%	99.41%	98.63%	99.45%	95.83%
最大 F1 スコア	99.32%	94.21%	97.48%	95.15%	96.83%	88.75%
（最適な確率閾値）	32.53%	55.23%	46.65%	70.86%	38.04%	16.45%
Precision	98.92%	96.49%	96.25%	97.05%	96.43%	90.94%
Recall	99.72%	92.03%	98.74%	93.32%	97.23%	86.65%
Accuracy	99.32%	94.34%	97.44%	95.24%	96.79%	89.01%
Splitting error	0.28%	7.97%	1.26%	6.68%	2.77%	13.35%
Lumping error	1.09%	3.35%	3.85%	2.84%	3.60%	8.63%

表 5. クラスタリング後の科研費由来の検証用データの精度

	Recall	Precision	F1 score
全ペア	97.89%	98.93%	98.41%
発明者同士のペア	94.09%	99.82%	96.87%
著者同士のペア	98.49%	98.94%	98.71%
発明者と著者のペア	69.99%	97.94%	81.64%

このような Splitting エラーの原因には 2 つの可能性がある。英語氏名によるブロッキング時で生じるエラー（つまり英語氏名の表記ゆれに起因するエラー）と同一の英語氏名を持つブロック内での分類器及びクラスタリング時に生じるエラーである。そこで、両者の可能性を確認するため、科研費由来の検証用データのうち、英語氏名が一致しているペアに限定したデータにおいて精度を検証すると表 6 のような結果となった。英語氏名の表記ゆれがないケースでは、再現率（Recall）は 95.37%、F1 スコアも 96%以上となり十分に高い精度であることがわかる。

以上から、発明者と著者のペアの Splitting エラーは英語氏名の表記ゆれに主に起因していると考えられる。実際に確認してみると、科研費由来の検証用データに含まれる研究者 48,333 人のうち 2%にあたる 969 人において特許発明者及び論文著者の英語氏名の表記ゆれがあることがわかった。英語氏名の表記ゆれに対する対応については今後の課題である。

表 6. 同一ブロック内（英語氏名が一致）のペアにおける精度

	Recall	Precision	F1 score
全ペア	98.53%	98.93%	98.73%
発明者同士のペア	94.95%	99.82%	97.32%
著者同士のペア	98.60%	98.94%	98.77%
発明者と著者のペア	95.37%	97.94%	96.64%

4．研究者タイプ別論文・特許の状況

前節までの作業によって特許発明者と論文著者の同一人物性を識別したデータセットを用いて、識別研究者が発明して出願された特許件数や公表した論文件数を確認し、所属機関ごとの文献数、及び研究者のキャリアの長さやライフサイクルでの生産性について簡単に集計することで、データに不自然な点がないか確認する。

4.1. 特許数・論文数

最初に、特許発明者と論文著者の同一人物性を識別して研究者 ID（CID）を少なくとも一人に付与することができた文献数について確認する。図 3 に特許、論文それぞれについての時系列の推移を示した。特許の識別文献数は、2001 年の 33.6 万件をピークとしてその後は減少し、2017 年には 16.9 万件と 2001 年の約半分になっている。日本の全特許出願件数は 2001 年の 43.9 万件をピークとして減少している。2001 年以降に識別特許件数が減少していることは、マクロの出願動向も反映していると考えられる。一方、論文では 1990 年の 2 万件から 2021 年の 8.7 万件まで 4 倍以上に増加した。Open Alex への論文収録数が増加しているため識別論文数も増加したものと考えられる。

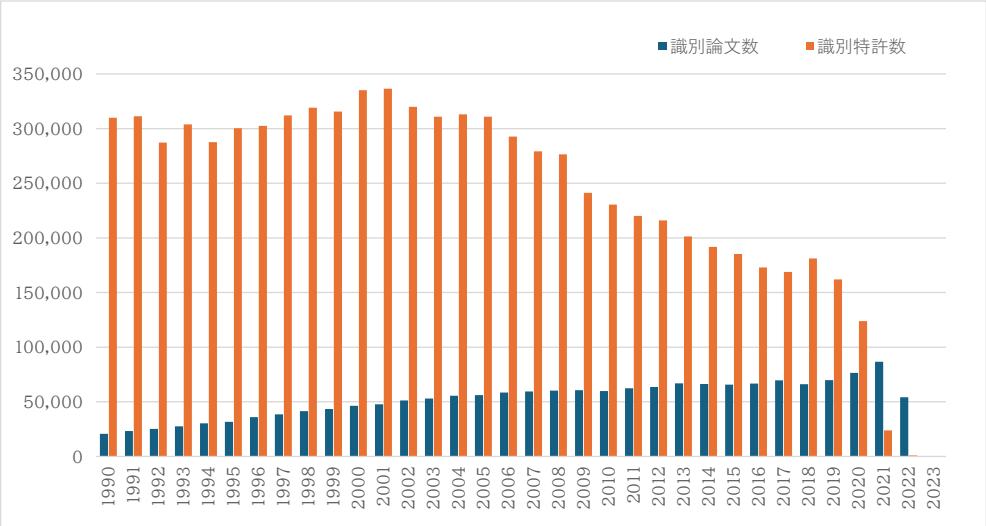

図 3．研究者の研究者 ID（CID）が付与された文献数

次に、研究者 ID が付与された研究者が生産した特許数・論文数を確認する。表 7 に特許・論文それぞれの件数を階級分けして、それに該当する研究者数を示した。研究者 ID が付与された研究者 323 万人のうちで、特許のみで論文件数はゼロである研究者は約 217 万人（67%）おりおよそ 2／3 を占める。特許はゼロで論文はある研究者は 93 万人（29%）、特許も論文もある研究者は 13 万人（約 4.1%）だった。特許・論文ともに 100 件を超える研究者は 109 人存在する。

なお、今回の研究者識別結果によると、研究者ごとの特許出願数の最大値は 19,936 件、論文数の最大値は 1,556 件だった（共同研究かどうかは考慮していない Whole count の件数）。特許出願数が 1,000 件以上である研究者は 35 人、論文数が 1,000 件以上である研究者は 6 人存在する。特許出願件数、論文数が最大である研究者の氏名をインターネット等で検索してみると、実際に研究成果が非常に多い研究者が存在することを確認している。

表 7. 特許数・論文数と研究者数

		論文数								Total	%
		0	1	2-5	6-10	11-20	21-50	51-100	101-		
特許数	0	-	469,528	291,403	73,571	46,709	33,037	10,333	3,896	928,477	28.8%
	1	924,892	7,406	9,669	4,597	4,234	4,482	1,830	899	958,009	29.7%
	2-5	671,823	8,307	9,159	4,387	4,329	5,272	2,587	1,554	707,418	21.9%
	6-10	227,744	4,819	4,023	1,605	1,538	1,975	1,223	758	243,685	7.5%
	11-20	167,614	5,348	4,381	1,413	1,203	1,350	981	703	182,993	5.7%
	21-50	125,861	6,597	5,901	1,712	1,225	1,183	740	668	143,887	4.5%
	51-100	38,054	3,034	3,333	1,008	721	540	223	233	47,146	1.5%
	101-	13,058	1,298	1,549	563	399	297	137	109	17,410	0.5%
Total		2,169,046	506,337	329,418	88,856	60,358	48,136	18,054	8,820	3,229,025	100%
%		67.2%	15.7%	10.2%	2.8%	1.9%	1.5%	0.6%	0.3%	100%	

4.2. 所属機関とセクター

前述したように、特許データについては、NISTEP 大学・公的機関名辞書と NISTEP 企業名辞書を出願人名に接続して用いている。NISTEP 企業名辞書で名寄せ対象になっていない出願人もあるため、NISTEP 機関 ID（NID）を付与できているのは発明者レコードのうちの 61.3%である。一方、論文データは、まず Open Alex から抽出した SCIE 対象論文の中で著者所属情報の住所が日本である著者を特定し、該当する著者が一人以上いる論文すべてを取り出して、NISTEP 大学・公的機関名辞書を接続したものを用いている。したがって、日本著者の所属機関については、NISTEP 機関 ID（NID）がほぼすべてのレコードに付与できている（99.7%）。また NISTEP 大学・公的機関名辞書には機関のセクター情報も収録されている。表 8 は今回整備したデータについてセクターごとの特許・論文のレコード数を示したものである。

表 8. 研究者×所属組織単位の特許・論文のレコード数

セクター	特許	論文
1:国立大学	65,306	4,043,983
2:国立短期大学	0	43
3:国立高等専門学校	63	11,921
4:公立大学	5,129	466,884
5:公立短期大学	0	371
6:公立高等専門学校	0	325
7:大学共同利用機関	661	34,125
11:学校法人	17,121	0
12:私立大学	105	1,504,435
13:私立短期大学	0	5,131
14:私立高等専門学校	0	56
8:国の機関	2,695	68,656
9:国立研究開発法人等	40,260	594,995
10:地方公共団体の機関	2,322	180,806
15:会社	12,646,976	272,069
16:非営利団体	28,148	216,749
17:その他の機関	0	706
小計	12,808,786	7,401,255
NIDなし、またはセクター不明	5,369,750	24,046
合計	25,603,837	

セクター大分類	特許	論文
大学等	88,385	6,067,274
国研等	45,277	844,457
会社	12,646,976	272,069
病院等	28,148	217,455

論文では、単一の論文においても一人の著者の所属機関が複数記載されていることがある。そのような著者は延べ 12 万人、著者×所属機関の単位で 24.8 万件あった。論文×著者×所属機関単位のレコード 740 万件のうちの 3.3%である。最大では 6 つの所属機関が記載されている著者もいる。表 8 はこのような複数組織への所属レコードも含めた件数を示している。また、以下のセクター単位の集計で用いるために、表 8 に示したようにセクターを大きく 4 つにまとめて「大学等」「国研等」「会社」「病院等」からなるセクターの大分類を定義した。

4.3. 研究者のキャリア年数とライフサイクルにおける生産性

次に、研究者のキャリアの長さを今回のデータセットで確認してみる。特許・論文の区別なしに、研究者の「最後の文献の出願年または公表年」と「最初の文献の出願年または公表年」の差に 1 を加えた値をキャリアの長さとする。文献が 1 件しかない研究者もキャリアの長さを 1 年とする。キャリア年数ごとの研究者数の分布を表 9 に示した。キャリア年数が 1 年のみの研究者が 44%と最も多いが、キャリア年数が 10 年を超えている研究者は 22%（約 57 万人）いる。

キャリアが長くなれば、所属組織を異動する頻度も増える。研究者のキャリア全体での所属機関数を、NID を接続できたレコードに基づいて求めて、件数の中央値や平均値なども示した。キャリア年数が 10 年以下の研究者では 1 機関しか経験していない研究者が多いが、31 年以上のキャリアを持つ研究者では平均で 3.5 機関に所属した経験を持つ。

表 9. キャリア年数の分布とキャリアにおける所属機関数

キャリア年数	研究者数		キャリア全体での所属機関数			
	人数	%	中央値	平均	p75	最大値
1年	1,161,860	44.1%	1	1.0	1	4
2-5年	547,624	20.8%	1	1.1	1	6
6-10年	354,217	13.4%	1	1.2	1	10
11-15年	229,106	8.7%	1	1.3	2	13
16-20年	157,714	6.0%	1	1.5	2	16
21-25年	101,442	3.8%	1	1.8	2	19
26-30年	63,411	2.4%	2	2.2	3	43
31-34年	21,193	0.8%	3	3.5	4	66
All	2,636,567	100%	1	1.2	1	66

研究者はしばしば所属機関を異動する。その異動は企業から大学などのように異なるセクターへの異動であることもある。研究者の異動や移動したときの生産性などは重要な研究テーマだがデータの取り扱いが難しい。研究者の所属セクターによる特徴を簡便に把握するために、今回は研究者の所属セクター（大分類：大学等・国研等・会社・病院等）ごとの文献数が最も多いセクターを、その研究者が主として所属したセクターとみなすこととした。

各セクターのキャリア年数と研究者数の分布は表 10 のようになる[1]。会社は比較的キャリア年数が長い研究者の割合が多く、約 4 割の研究者は 6 年以上のキャリアをもつが、30 年を超えるような長いキャリアをもつ研究者は大学等が最も多い。

表 10. 主たる所属セクターとキャリアの長さ

キャリア年数	大学等		国研等		会社		病院等		All	
	研究者数	%	研究者数	%	研究者数	%	研究者数	%	研究者数	%
1年	418,524	50.5%	55,373	49.0%	667,534	40.3%	20,429	54.2%	1,161,860	44.1%
2-5年	182,277	22.0%	21,245	18.8%	337,142	20.4%	6,960	18.5%	547,624	20.8%
6-10年	90,135	10.9%	13,015	11.5%	247,136	14.9%	3,931	10.4%	354,217	13.4%
11-15年	48,003	5.8%	8,218	7.3%	170,452	10.3%	2,433	6.5%	229,106	8.7%
16-20年	35,213	4.2%	6,387	5.6%	114,504	6.9%	1,610	4.3%	157,714	6.0%
21-25年	25,144	3.0%	4,485	4.0%	70,608	4.3%	1,205	3.2%	101,442	3.8%
26-30年	19,079	2.3%	3,036	2.7%	40,494	2.4%	802	2.1%	63,411	2.4%
31-34年	11,099	1.3%	1,290	1.1%	8,498	0.5%	306	0.8%	21,193	0.8%
Total	829,474	100%	113,049	100%	1,656,368	100%	37,676	100%	2,636,567	100%

図 4 は、各研究者が最初に文献を出願・公表した年を特許・論文別にそれぞれ調べて、各年の研究者数を主たる所属セクターごとにプロットしたものである。まず、特許（左図）でも論文（右図）でも、データセット期間の最初数年の人数が多い。これはキャリアの途中でデータが切断されている影響であると考えられる。つまり、実際にはある研究者が例えば 1980 年代からキャリアを開始していたとしても今回作成したデータセットではその期間の

[1] 文献数が最大という条件だけでは主として所属するセクターを決定できない場合は，各セクターで公表した文献の平均出版年が早い方のセクターを主として所属するセクターとする。平均出版年も同じ場合（8,534 人）は集計対象外とする。

文献が観測できず、キャリア終盤に公表した文献が最初の文献として特定されてしまっているケースが一定数あるものと考えられる。

最初の数年を除くと、論文では、Open Alex へのデータ収録数が増加していることを反映してか、「大学等」「国研等」「病院等」の各セクターではキャリア開始者数がおおむね右上がりとなっている。ただし「会社」セクターでは目立った増減がない。一方、特許のデータではセクターによって傾向が大きく異なる。特許出願の大半を占めるセクターである「会社」に注目すると、マクロの出願動向を反映して、最も特許出願数が多かった 2000 年頃に小さなピークがあり、その後はキャリアを開始した研究者数が減少傾向にある。「大学等」のセクターはそれ以外のセクターとは異なる傾向を示しており、初めて特許出願をした研究者数が 2000 年頃から増加して 2005 年にピークとなっている。これは 2004 年の国立大学法人化によって特許出願が増加したことを反映しているものと考えられる。その後は「大学等」のセクターでも特許についてのキャリア開始者数は減少傾向にあるが、法人化前よりも高い水準で推移している。

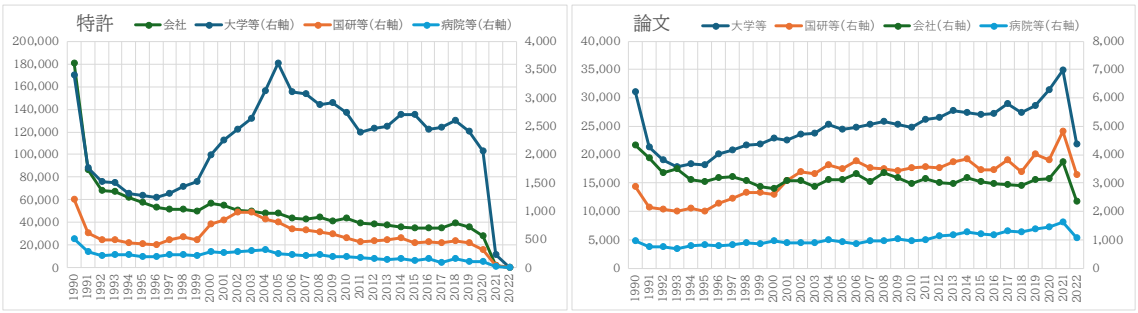

図 4．主たる所属セクターと研究者が最初に文献を公表した年
（左図：特許、右図：論文、縦軸単位：人）

最後に、今回作成したデータを用いて、研究者がそのライフサイクルにおいてどのくらいの特許及び論文を生産しているのか確認してみたい。以下では、キャリアの開始年が 1994 年から 1996 年である研究者に注目する。そのうえで、キャリアの長さが 25 年、20 年、15 年、10 年の研究者に分けて、そのライフサイクルにおける毎年の論文・特許の公表件数・出願件数の平均値の推移を確認する。主たる所属セクターが大学等である研究者と会社である研究者それぞれについて、論文数と特許数について集計した結果を図 5，図 6 に示した。

まず、主たる所属セクターが大学等である研究者に注目する。図 5 の左図は、キャリア開始後の毎年の平均論文公表数を示している。同時期にキャリアを開始した研究者同士を比較したとき、キャリアを長く続けている研究者の方がキャリア全体を通じた平均論文数がおおむね多い傾向にあることがわかる。また、キャリアの長さに関わらず、キャリア終盤には平均論文数は減少しており、全体として逆 U 字型の傾向がみられる。キャリア年数が 25 年・20 年の研究者についてはキャリア中盤過ぎまで平均論文数が増加し続けている。ただ

し、これは Open Alex に収録された論文数が増加していることを反映している可能性があるためデータの解釈に注意が必要である。図 5 の右図より、特許については、研究キャリアが 25 年ある研究者はキャリア初期もその後も平均特許件数が多いことが見てとれる。この世代の研究者の場合は、キャリア開始から約 10 年後の 2004 年に国立大学が法人化されている。キャリア 25 年の研究者については、その前後に平均特許数が増加している様子がグラフからは確認できる。

図 6 は、主たる所属機関が会社である研究者について同様の集計の結果を示したものである。長くキャリアを続けている研究者はキャリアが短い研究者と比較して、キャリア初期から平均特許件数が多く、その後も一貫してより多くの発明をしている傾向が顕著である。特許件数が最も多いタイミングは、キャリア 25 年の研究者の場合はキャリア開始後 6 ～12 年目頃となっている。同じ長さのキャリアをもつ大学研究者の場合に論文数が最も多いのはキャリア開始後 18～19 年目頃であることと比較すると、企業の研究者は生産性のピークが相対的に早い。ただし、日本の特許出願件数は 2001 年をピークに低下しているため、そのマクロの傾向を反映している可能性がある。企業研究者の論文数は全体として件数が少なく、論文を執筆する研究者が限定的であることをうかがわせる。キャリアの長さと正に比例して論文数が多い傾向は特許と同様である。

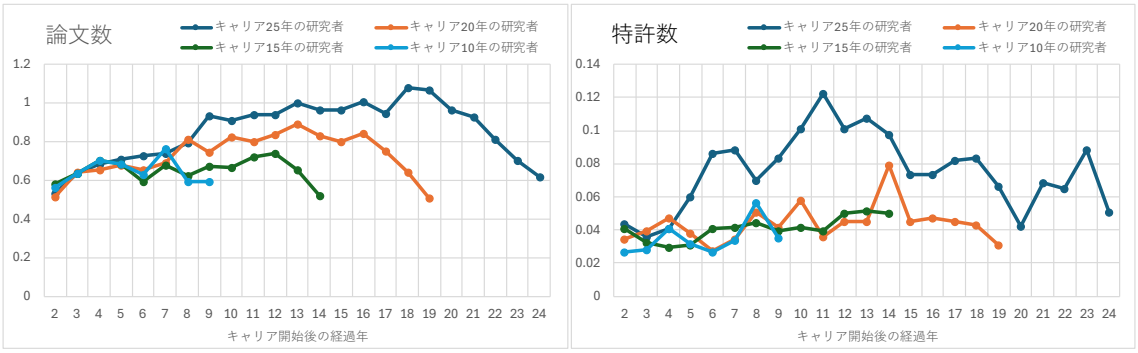

図 5．大学等： 1994-1996 年にキャリアを開始した研究者の平均論文数・平均特許数

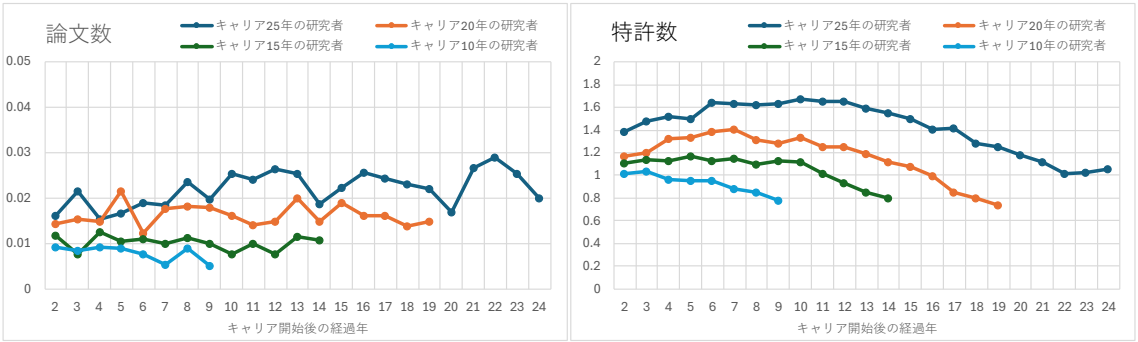

図 6．会社： 1994-1996 年にキャリアを開始した研究者の平均論文数・平均特許数

５．まとめと今後の研究方向性

本稿においては、学術論文（Open Alex）と特許情報（日本国特許庁公開情報）を用いたサイエンスイノベーションに関する分析用データベースを作成し、研究者のキャリアと生産性に関する分析を行った。まず、学術論文に関する公開情報である Open Alex から Web of Science の SCIE に該当するジャーナル論文を抽出し、かつ著者情報から所属機関が日本に所在する著者による論文を取り出した。特許情報については、日本国特許庁の公開特許情報のうち、発明者住所が日本に所在する特許を抽出し、出願人情報等を用いて発明者の所属機関の推定を行った。次に上記の学術論文及び特許情報を用いて著者・発明者の識別作業を行った。具体的には、同一人物による論文・特許であるか否かの識別モデルを機械学習によって構築し、その情報を用いて同一著者・発明者による論文・特許をクラスター分析によって推定した。その結果については、科研費データベース（同一研究者による論文、特許情報）によって評価を行い、論文著者ペアについては 96.87%、特許発明者ペアについては 98.71%、論文著者・特許発明者ペアについては 81.64%の F1 スコアを得た。

論文を公表したか特許を出願した研究者に限定されるが、日本における科学技術関係の研究者を網羅的にカバーしたものとして、その意義は大きい。ただし、科研費データベース（同一人物による論文、特許リスト）による検証結果について、論文と特許ペアについてはその他のペアと比べてやや F1 スコアが低くなった。これは特に Recall 比率が低いことによるもので、同一人物による特許と論文が違う研究者と識別されたものが多いことによる。その内容についてより詳細な調査を行ったところ、論文著者と特許発明者の間に標記の揺れが存在するレコードが多いことが分かった。これは原データの問題に起因するものであることから、対処することが困難な問題であるが、表記ゆれが起きやすい氏名について氏名ごとのブロックを広げる（表記ゆれの対象となる複数氏名を一つのブロックとする）ことが対策の一つとして考えられる。ただし、ブロックを広くとることによって、表記ゆれを起こしていないレコードに対して悪影響を及ぼすことも考えられるので、パフォーマンスを評価しながら対策方法を探っていく試行錯誤のプロセスが必要となる。

また、当該データを用いて著者・発明者ごとの経験年数と生産性に関する記述統計を作成した。その結果、大学著者についてはキャリア年数を論文数との間には上昇ののち、下降する逆 U 字型の関係が観察されたが、企業著者については明確な関係が見られなかった。特許出願数については、大学発明者、企業発明者ともキャリア年数の長い発明者については逆 U 字型の関係が観察された。

本データベースを活用することによって、研究者ごとの論文数や特許件数が明らかになるため、例えば、研究成果が極めて優れている研究者（スターサイエンティスト）に関して研究することが可能となる。また、同一研究者の所属情報から、研究者の所属情報の変化を識別できる。特に、企業と大学等の公的研究機関の双方を経験している研究者は、産学連携プロジェクトを有効に進めるための「橋渡し人材」としての役割が期待される。これらの研

究者が関与する論文や特許の質がコントロールグループと比べて高いのか、また「橋渡し人材」が企業に雇用されることによって、企業において、大学等における科学的成果に対する評価を適切に行えるようになり、当該研究者が関与しない産学連携発明（特許）についてもより商業価値の高いものが生まれる可能性がある。このように本データベースは、サイエンスと商業化技術の関係について研究者単位のミクロな情報を提供するものとして学術的価値が高いといえる。

（参照文献）

科学技術・学術政策研究所(2022). 「NISTEP 企業名辞書マニュアル（ver2022_1 対応)」, 文部科学省科学技術・学術政策研究所.

科学技術・学術政策研究所(2023). 「NISTEP 大学・公的機関名辞書マニュアル（ver2022.1 対応)」, 文部科学省科学技術・学術政策研究所.

塚田尚稔, 元橋一之（2018).「Microsoft Academic Graph の書誌情報データとしての評価」, NISTEP Discussion Paper, No.162.

元橋一之（2014）.『日はまた高く 産業競争力の再生』, 日本経済新聞出版社.

Goto, A., and Motohashi, K. (2007). “Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities,” *Research Policy*, 36(9): 1431-1442.

Harzing, A. (2016). “Microsoft Academic (Search): A phoenix arisen from the ashes?” *Scientometrics*, 108(3): 1637-1647. DOI: 10.1007/s11192-016-2026-y

Hug, S. E. and M. P. Brändle (2017). “The Coverage of Microsoft Academic: Analyzing the Publication Output of a University,” *Scientometrics*, 113(3): 1551-1571. DOI: 10.1007/s11192-017-2535-3

Ikeuchi, K., Motohashi, K., Tamura, R. and Tsukada, N. (2017). “Measuring Science Intensity of Industry Using Linked Dataset of Science, Technology and Industry,” RIETI Discussion Paper, 17-E-056.

Ikeuchi, K. and Motohashi, K.,(2022). “Linkage of Patent and Design Right Rata: Analysis of Industrial Design Activities in Companies at the Creator Level", *World Patent Information*, 70, 102114.

Motohashi, K., Koshiba, H. and Ikeuchi, K. (2023). “Measuring Science and Innovation Linkage Using Text Mining of Research Papers and Patent Information,” RIETI Discussion Paper, 23-E-015.

Motohashi, K., Ikeuchi, K. and Yamaguchi, A. (2024). “Absorptive Capacity for Science-Based Innovation Propensity: An Empirical Analysis Using Japanese National Innovation Survey,” *The Journal of Technology Transfer*, forthcoming.

Paszcza, B. (2016). "Comparison of Microsoft Academic Graph with Other Scholarly Citation

Databases," Thesis for the Degree of Master of Science, University of Southampton, September 2016, DOI: 10.13140/RG.2.2.21858.94405.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A Fully-open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts. https://arxiv.org/abs/2205.01833

Veugelers, R. and Wang, J. (2019). "Scientific Novelty and Technological Impact," *Research Policy*, 48(6) : 1362–1372.

DISCUSSION PAPER No.234

学術論文と特許情報を用いたサイエンスイノベーション分析データベースの作成

2024 年 10 月

文部科学省 科学技術・学術政策研究所 第 1 研究グループ
池内 健太 塚田 尚稔 元橋 一之

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階

Database for Understanding Science-based Innovation by Linking Research Papers with Patent Information

October 2024

IKEUCHI Kenta, TSUKADA Naotoshi, MOTOHASHI Kazuyuki

First Theory-Oriented Research Group
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

https://doi.org/10.15108/dp234

https://www.nistep.go.jp