

レポート

新型コロナウイルス感染症に対する研究開発資金配分 － OECD による Fundstat データ基盤を用いた 分析の紹介－

科学技術予測・政策基盤調査研究センター 主任研究官 村上 昭義

概要

本稿では、経済協力開発機構（OECD）が 2023 年 10 月に公表した、新型コロナウイルス感染症に対する各国政府の研究開発資金配分における動向を分析・報告したワーキング・ペーパーを紹介する。13 か国及び EU の計 27 の資金配分機関からデータを抽出した Fundstat データ基盤内において、2019 年から 2021 年における新型コロナウイルス感染症に関連する研究開発プロジェクトの総数は 11,886 件であった。研究開発資金配分の総額は約 125.9 億ドル（約 1.34 兆円）であり、プロジェクト 1 件当たりの資金配分額は約 120 万ドル（約 1.28 億円）であった。また、トピックモデリング分析等の自然言語処理を用いた分析から、新型コロナウイルス感染症に関連する研究開発プロジェクトを 8 つのトピック・クラスター等に分類し、新型コロナウイルス感染症への政府資金配分では、どのような研究開発が行われていたかが明らかになった。

キーワード：新型コロナウイルス感染症，政府資金配分，研究開発（R&D）プロジェクト，トピックモデリング，大規模言語モデル

1. はじめに

経済協力開発機構（OECD）では、各国の資金配分機関の研究開発資金配分データベースに収載されているデータを収集し、研究開発プロジェクトレベルでの政府資金配分の動向を分析する Fundstat プロジェクトを実施している^{1,2)}。OECD は、Fundstat データ基盤を新型コロナウイルス感染症の研究開発に対する政府支援に適用し、研究開発の動向を把握する実験的な研究を行い、2023 年 10 月にワーキング・ペーパーを公表している³⁾。

本稿では、そのワーキング・ペーパーの分析及び取組を紹介し、分析から得られた示唆を報告する。

2. OECD における Fundstat データ基盤について

Fundstat データ基盤は、分析時点（2022 年 12

月）において、13 か国及び EU（EC）の計 27 の資金配分機関の研究開発資金配分データベースを収集し、データ分析の基礎としている。13 か国は、3 文字の国コードのアルファベット順に、AUS（オーストラリア）、AUT（オーストリア）、BEL（ベルギー）、CAN（カナダ）、CHE（スイス）、DEU（ドイツ）、FRA（フランス）、GBR（英国）、JPN（日本）、LVA（ラトビア）、NOR（ノルウェー）、SWE（スウェーデン）、USA（米国）である。また、日本のデータベースとしては、科学研究費助成事業データベース（KAKEN）の公開データが収集されたのに加えて、国立研究開発法人日本医療研究開発機構（AMED）から、研究開発課題データベース（AMEDfind）^{注 1}のバルクデータが OECD に提供されている。

構築された Fundstat データ基盤において全ての国・地域の資金配分機関の研究開発プロジェクトを合計すると、2019 年から 2021 年の期間において 607,098 件あり、資金配分額の合計は、約 2,876

注 1 研究開発課題データベース（AMEDfind），<https://amedfind.amed.go.jp/amed/index.html>

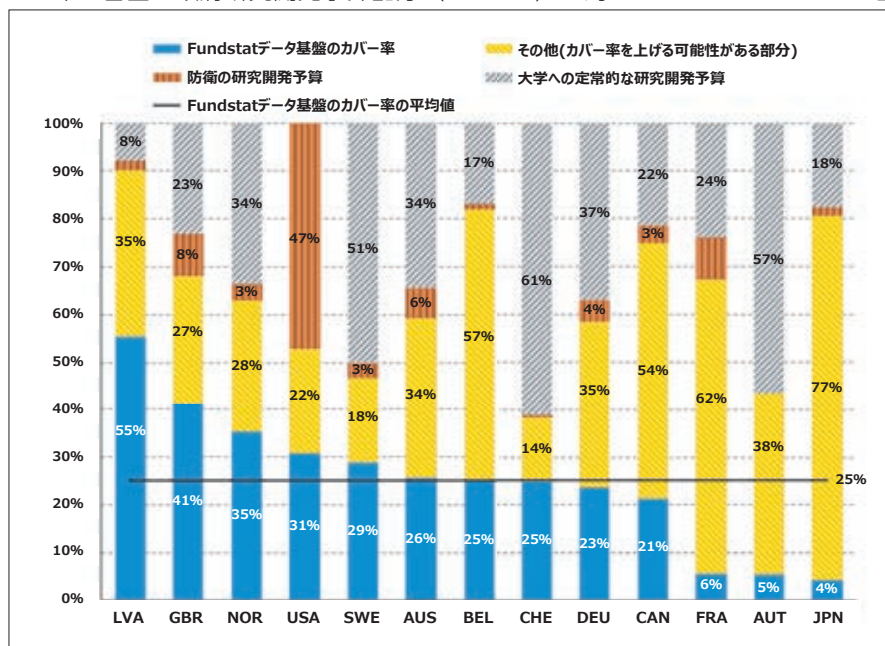
億ドル（約 31 兆円^{注2}）であった^{注3}。研究開発プロジェクト 1 件当たりの資金配分額は、各国・地域を横並びで平均すると約 50 万ドル（約 5,340 万円）であった。Fundstat データ基盤内の資金配分額の上位国・地域は、米国、EU、ドイツであり、それらで 8 割以上を占めている。日本の資金配分額は、Fundstat データ基盤内で約 3.3% を占める。また、日本の研究開発プロジェクト 1 件当たりの資金配分額は、これらの国・地域の中では最も小さく、約 9 万ドル（約 961 万円）であった。これは科研費の 1 件当たりの資金配分額が相対的に小さい（そして日本のデータに占める科研費の割合が大きい）ことの影響を受けているためと考えられる。ワーキング・ペーパーにおいても、各国で研究開発プロジェクトの資金規模の違いが生じるのは、プロジェクトの対象範囲や性質、データ報告構造などが各国で異なるためと言及されている。

図表 1 では、2020 年の各国の政府研究開発予算

配分（GBARD）に対する Fundstat データ基盤の資金配分額の割合（カバー率）を示す。各国の政府研究開発予算配分に対する Fundstat データ基盤のカバー率は、国によって異なっているが、平均すると約 25% をカバーしている。オーストリア（AUT）の場合、Fundstat データ基盤のカバー率はわずか 5% であるが、GBARD の 57% が大学への定常的な研究開発予算によって占められているため、Fundstat データ基盤の実質的なカバー率は高いと言える。

日本（JPN）の場合、Fundstat データ基盤では科研費と AMED のみが対象となっており、また大学への定常的な研究開発予算の割合も 18% と相対的に高くないため、Fundstat データ基盤の実質的なカバー率は高くないことが図表 1 から読み取れる。この点に関して、2020 年の政府研究開発予算配分には、比較的規模の大きい補正予算が含まれている点に注意を要するものの、それらを勘案しても^{注4}、政府の研究開発予算は、研究開発プロジェクト以外の多様な項

図表 1 2020 年の各国の政府研究開発予算配分（GBARD）に対する Fundstat データ基盤のカバー率



図注 1：本図表は、OECD ワーキング・ペーパーの Figure 1 を基に、筆者が日本語訳を加筆し、加工・作成したものである。

図注 2：各国の GBARD を、大学への定常的な研究開発予算、防衛の研究開発予算、それら以外に分けた後、それら以外（大学への定常的な予算及び防衛関連を除く）の部分が Fundstat で把握されるべきものと仮定している。その上で、本来 Fundstat で把握されるべき、大学への定常的な予算及び防衛関連以外の部分について、実際に Fundstat データ基盤でカバーされている部分とその他に分けている。その他は、Fundstat データ基盤のカバー率を上げる可能性がある部分と考えられる。

出典：参考文献・資料の 3), <https://dx.doi.org/10.1787/4889f5f2-en>

注 2 本稿における為替レートは、1 ドル = 106.775 円（2020 年の為替レート）を用いた。為替レートの出典：OECD (2024), Exchange rates (indicator), <https://doi.org/10.1787/037ed317-en> (Accessed on 18 March 2024)

注 3 収集時点において、2021 年のデータは一部の資金配分機関でデータがないため、注意を要する。日本の AMED のデータも 2021 年のデータは、収集のタイミングで不完全であったため分析対象から除外している。また、Fundstat データ基盤の一部の研究開発プロジェクト（全体の 5% 程度）には、資金配分の情報が欠落している点にも注意を要する。

注 4 2020 年の日本の GBARD から、比較的規模の大きい補正予算である「グリーンイノベーション基金事業（2 兆円）」及び「10 兆円規模の大学ファンド（0.5 兆円）」の 2.5 兆円を除いて、Fundstat データ基盤（KAKEN と AMED）のカバー率を計算しても、1~2% 程度の上昇にとどまり、平均の 25% に比べてもまだ低い値を示す。

目を含んでいることが分かる^{注5}。

3. 新型コロナウイルス感染症に関連する研究開発プロジェクトの特定と特徴

ワーキング・ペーパーでは、新型コロナウイルス感染症に関連する研究開発プロジェクトの特定には、キータームマッチングを用いている。その手順としては、まず、10のキータームをベースとして、それらに他のデータベースを用いて24のキータームを追加・拡張し、全部で34のキーターム^{注6}を選定した。その後、研究開発プロジェクトのタイトル及び概要のテキスト群にキータームマッチングを適用することで、新型コロナウイルス感染症に関連する研究開発プロジェクトを特定している。また、タイトル及び概要へのキータームマッチングで抽出された研究プロジェクトには、文脈上新型コロナウイルス感染症に触れていないだけで、研究対象としては取り扱っていないものも含まれているため、それらのプロジェクトをコンテキスト分析の実施及び目視によって除外している。

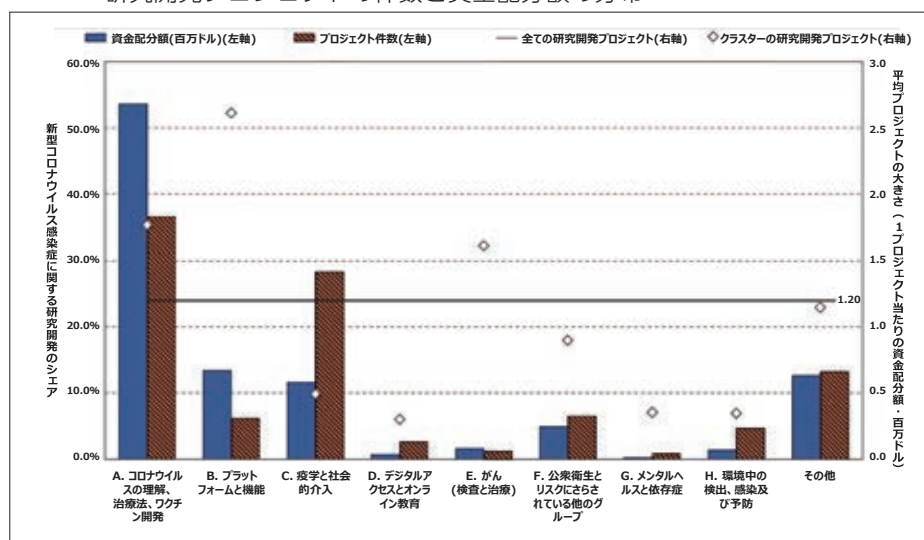
上記の結果、2019年から2021年における新型コ

ロウイルス感染症に関連する研究開発プロジェクトの総数は11,886件であった。それらの資金配分額の合計は約125.9億ドル（約1.34兆円）であり、プロジェクト1件当たりの資金配分額は約120万ドル（約1.28億円）であった。新型コロナウイルス感染症に関連する研究開発プロジェクトは、Fundstatデータ基盤内の全体に対して、資金配分額の約4%、プロジェクト件数の約2%に相当している。

新型コロナウイルス感染症に関連する研究開発プロジェクトのテキストにトピックモデリング分析^{注7}を適用すると、34の異なるトピックが抽出され、8つのより大きいレベルのトピック・クラスターにグループ化された。8つのトピック・クラスターは、「A. コロナウイルスの理解、治療法、ワクチン開発」、「B. プラットフォームと機能^{注8}」、「C. 疫学と社会的介入」、「D. デジタルアクセスとオンライン教育」、「E. がん（検査と治療）」、「F. 公衆衛生とリスクにさらされている他のグループ^{注9}」、「G. メンタルヘルスと依存症」、「H. 環境中の検出、感染及び予防」であった。

図表2には、上記8つのトピック・クラスターにおける新型コロナウイルス感染症に関連する研究開

図表2 8つのトピック・クラスターにおける新型コロナウイルス感染症に関連する研究開発プロジェクトの件数と資金配分額の分布



図注：本図表は、OECD ワーキング・ペーパーの Figure 8 を基に、筆者が日本語訳を加筆し、加工・作成したものである。
出典：参考文献・資料の 3), <https://dx.doi.org/10.1787/4889f5f2-en>

注5 日本の GBARD の値は、内閣府がまとめている科学技術関係予算に対応しており、科学技術関係予算には研究開発以外の経費も含まれている。内閣府 科学技術関係予算 <https://www8.cao.go.jp/cstp/budget/index2.html>

注6 キータームは、次に示すベースの10のキーターム (Coronavirus disease, covid, covid19, ncov, novel coronavirus, sars coronavirus 2, sars cov 2, severe acute respiratory syndrome coronavirus 2, wuhan coronavirus, wuhan seafood market pneumonia virus) に加えて、ベースと似た意味で用いられている24のキータームも追加・拡張されている。

注7 Fundstat データ基盤には、英語以外の言語も含まれるため、ワーキング・ペーパーでは、トピックモデリング分析において、多言語に対応する大規模言語モデル (BERT) を用いた分析に挑戦している。

注8 横断的な性質を持つ生物医学分野のプラットフォームとネットワークを指し、これらには、治験、大規模集団研究、バイオインフォマティクス、プロトコルの開発、トレーニングなどの要素が含まれる。

注9 これには、医療従事者、妊婦、糖尿病、高齢者、その他の社会的弱者に関連する研究開発が含まれているだけでなく、医療提供からの除外に関する側面も含まれている。

発プロジェクトの件数と資金配分額の分布を示す。生物医学的な研究開発である「A. コロナウイルスの理解、治療法、ワクチン開発」と社会科学的研究開発である「C. 疫学と社会的介入」は、研究開発プロジェクトの件数では上位 2 つで件数が多いが、資金配分額では、A のシェアが大きく、生物医学的な研究開発プロジェクトに政府資金配分が集中した様子が分かる。

また、プロジェクトのタイトルや概要の類似性から約 1 万件の研究開発プロジェクトをマッピングし、8 つのトピック・クラスターごとに色付けすると、「B. プラットフォームと機能」が中心にあることから、ワーキング・ペーパーでは新型コロナウイルス感染症に関連する研究開発の多様な分野にわたって重要な役割を果たしていることが示唆されると指摘している（図表 3）。このことは、将来のパンデミックに対して医療や研究開発のシステムを備える場合や、新型コロナウイルス感染症に基づく発見や技術を他の健康課題に適用する場合に、「B. プラットフォームと機能」に関連する研究開発への資金配分が重要な観点であるとしている。

これらに加えて、政府資金配分の動向を探る重要な観点として、市場で使用される製品（商品又はサービ

ス）及びプロセスの研究開発であるかという観点もある。ワーキング・ペーパーでは、ビジネス及び市場に関連する語彙を用いて、研究開発プロジェクトのテキスト群にキータームマッチングを行い、市場指向の研究開発プロジェクトを特定している注 10。それらの市場指向の研究開発プロジェクトは、新型コロナウイルス感染症に関連する研究開発プロジェクト全体において、資金配分額の 34% を占め、プロジェクト件数の 22% を占めていた。Fundstat データ基盤内の資金配分機関に注目すると、イノベーションに関する資金配分機関（例：英国 Innovate UK、スウェーデン Vinnova など）は、市場指向の研究開発プロジェクトの件数と資金配分額のシェアが相対的に大きいことが示されている。

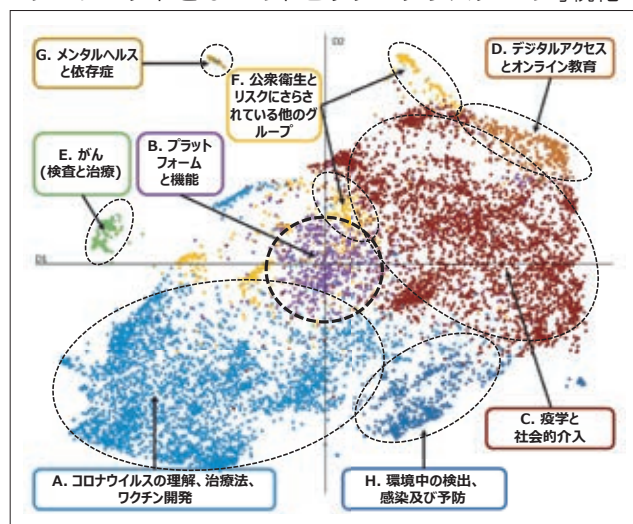
4. WHO の優先トピックへの専門家による分類結果や他のデータソースとの比較

ワーキング・ペーパーでは、Fundstat データ基盤で特定された新型コロナウイルス感染症に関連する研究開発プロジェクトについて、WHO（世界保健機関）の新型コロナウイルス感染症における 9 つの優先トピックへの専門家による分類結果や他のデータソースを用いた場合との比較検討を行っている。

まず、WHO の優先トピックへの分類として、UKCDR（UK Collaborative on Development Research）の「COVID-19 tracker」データベース内の情報を用いている。「COVID-19 tracker」は、収集された研究開発プロジェクトを、専門家が手作業によって、WHO の優先トピックへタグ付けしているため、これを教師データとした機械学習モデルを開発し、比較検討の手段に用いている。

図表 4 は、[A]UKCDR「COVID-19 Tracker」の報告値、[B] 機械学習モデルを UKCDR「COVID-19 Tracker」に適用した予測値、[C] 機械学習モデルを Fundstat で得られたデータに適用した予測値、の 3 つについてプロジェクト件数と資金配分額を比較した結果を示す。ここで、[A] のプロジェクト件数と資金配分額は、「COVID-19 tracker」内でタグ付けされた WHO 優先トピックに基づいているが、1 つのプロジェクトは 1 つ以上の WHO 優先トピックにタグ付けされているため、[A] を合計した値は、「COVID-

図表 3 新型コロナウイルス感染症に関連する研究開発プロジェクトと 8 つのトピック・クラスターの可視化



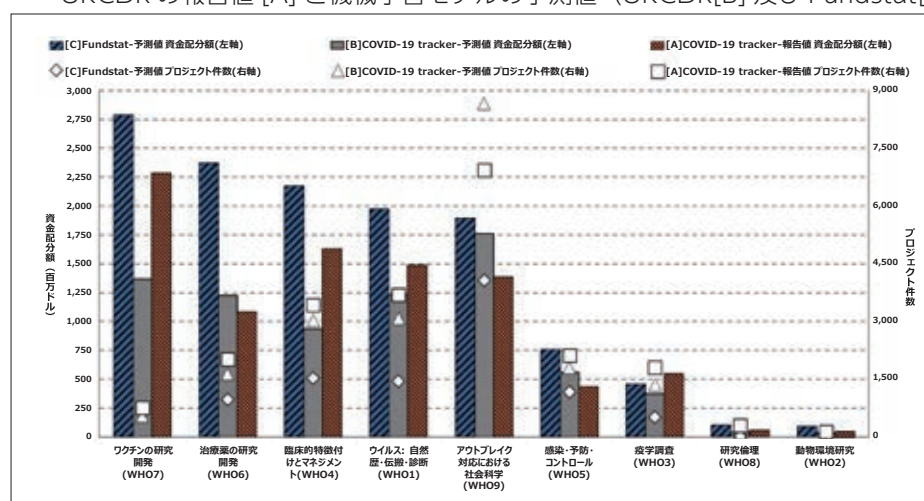
図注：本図表は、OECD ワーキング・ペーパーの Figure 9 を基に、筆者が日本語訳を加筆し、加工・作成したものである。

出典：参考文献・資料の 3)。

<https://dx.doi.org/10.1787/4889f5f2-en>

注10 ここでのキータームは、33 のベースのキーターム (adoption, business, commercialisation, competitive, consumer, corporate, corporation, diffusion, enterprise, entrepreneurship, expenditure, feasibility study, firm, industry, innovation, intellectual property, investment, IPR, joint venture, license, management, manufacture, market, model, process, product, production, service, spinoff, start-up, strategy, technology, technology transfer) に加えて、類似する 13 のキータームを拡張・追加しているが、ここでのキータームマッチングでは、これらのキータームを 2 つ以上持つ研究開発プロジェクトを、市場指向の研究開発プロジェクトであると特定している。

図表 4 WHO 優先トピックにおけるプロジェクト件数と資金配分額の分布
-UKCDR の報告値 [A] と機械学習モデルの予測値 (UKCDR[B] 及び Fundstat[C]) -



図注 1：本図表は、OECD ワーキング・ペーパーの Figure 16 を基に、筆者が日本語訳を加筆し、加工・作成したものである。

図注 2：UKCDR「COVID-19 tracker」データベースには、65 億ドル（約 0.7 兆円）の資金投資を伴う 17,955 件のプロジェクトが含まれている。Fundstat の予測値は、Fundstat データ基盤内の資金配分情報を持つ 10,472 件のプロジェクトに基づいている。

出典：参考文献・資料の 3), <https://dx.doi.org/10.1787/4889f5f2-en>

19 Tracker」内のプロジェクト件数と資金投資額の合計値を超える。そのため、[B]において、1つのプロジェクトに対するトピックへの所属確率を機械学習モデルによって算出し、プロジェクト件数と資金配分額を案分することで、「COVID-19 tracker」内のプロジェクト件数と資金配分額の合計値に一致させた予測値を示している。

[A]と[B]の資金配分額を比較すると、「ワクチンの研究開発 (WHO7)」では差が見られるが、「治療薬の研究開発 (WHO6)」では差がほとんど見られない。これは、ワクチンに関連するプロジェクトは、他のトピックに関連付けられるテキストの要素やパターンが含まれているため、機械学習モデルによって確率的に優先トピックに割り当てると、その部分が他のトピックの資金配分額になり、「ワクチンの研究開発 (WHO7)」で、[B]の資金配分額は小さくなっていると考えられる。この結果を踏まえると、[A]で示されるように、「治療薬の研究開発 (WHO6)」が「ワクチンの研究開発 (WHO7)」に比べて比較的資金不足であったと結論付けることはできないことをワーキング・ペーパーでは指摘している。ただし、機械学習モデルでの予測誤差による違いも影響している可能性があるため、結果の解釈には注意を要としている。

[B]と[C]の資金配分額を比較すると、「ワクチンの研究開発 (WHO7)」、「治療薬の研究開発 (WHO6)」、「臨床的特徴付けとマネジメント (WHO4)」で、[C]の資金配分額が大幅に大きくなっている。この違いは、Fundstat データ基盤の中に、米国の連邦調達研究開発データ (Federal

procurement R&D data) が含まれている一方、「COVID-19 tracker」にはそれが含まれていないため、そのようなデータ収録の違いに起因しているとワーキング・ペーパーでは指摘している。

また、他のデータソースとして、論文データベースとの比較を行っている。まず、Fundstat データ基盤に用いたキータムマッチングに基づく分類方法を用いて、新型コロナウイルス感染症に関連する論文を特定した。次に特定された論文を、上記の Fundstat データ基盤で構築された機械学習モデルや UKCDR で構築された機械学習モデルに適応し、8つのトピック・クラスターや9つの WHO の優先トピックと関連付けることで、両者との比較も実施している。その結果、論文数は、研究開発プロジェクトの件数には近い値を示すが、資金配分額では大きな違いが生じていることを明らかにしている。この論文数に対するプロジェクト件数と資金配分額の違いは、特にワクチンや治療薬に関する研究開発で見られている。ワーキング・ペーパーでは、これらを踏まえ、新型コロナウイルス感染症に関連する研究開発の状況を把握するには、論文数などのアウトプットの情報に加えて、資金配分データを考慮することの重要性を指摘している。

5. 今後に向けた示唆

<大規模言語モデルの政策研究への適用可能性>

ワーキング・ペーパーでは、テキストデータを処理するために大規模言語モデルを使用しており、統計分析と政策分析における大きな方向性を示唆している。ま

た、分析手法は、今後、新型コロナウイルス感染症以外の政策課題にも適用できる可能性がある。さらに、Fundstat データ基盤に、最近注目されている生成系 AI ツールを組み込むことができれば、迅速なモニタリングや、他分野での研究開発ニーズを予測するのに役立つはずであると指摘されている。

NISTEP においても、新型コロナウイルス感染症関連のプレプリントを対象に、自然言語処理を用いたエマージング領域の把握を試行している⁴⁾。今後、自然言語処理や大規模言語モデルを用いた分析手法が政策研究においても大きく展開される可能性が高いと筆者は考える。

＜複数のデータを用いて総合的に判断することの重要性＞

ワーキング・ペーパーでは、特定の政策課題に関連する研究開発を把握する際に直面する課題への対処や、複数のデータソースからのエビデンスを比較・統合して解釈するなど、幾つかの方法論的な洞察を提供している。また、自然言語処理の教師なし学習で行った機械的な分類手順を実装し、専門家による分類と比較検討を行うことで、それら全体の相乗効果を見いだしている^{注11)}。このような分析アプローチは、資金配分情報と定性的なプロジェクトの詳細説明を組み合わせることで総合的に判断することの重要性を示唆している。プロジェクト件数やその他のアウトプット指標である論文数などは、資金配分額の情報が無い場合、一部のトピックの相対的な重要性を誇張する傾向がある点に注意が必要であるとワーキング・ペーパーでも改めて指摘されている。

＜研究開発プロジェクトにおける世界的な資金配分

データの標準化の重要性＞

Fundstat データ基盤内の研究開発プロジェクトの概要には、信頼できるトピック予測をする上で、十分なテキストデータや情報が欠落している場合があった。また、そのような情報の欠落によって、分析では、新型コロナウイルス感染症のような特定の社会的課題に対処する研究開発プロジェクトの境界を定義することが難しい点も示されている。これらを踏まえ、各国の資金配分機関が、説明責任と分析のために、より調和のとれた形で研究開発プロジェクトの資金配分データを標準化し、公開することは、今後、このような分析・取組を行う上でも大きな意義があると、ワーキング・ペーパーでは指摘している。

また、ワーキング・ペーパーの Fundstat プロジェクトは、OECD の科学技術指標各国専門家作業部会 (NESTI) 内のプロジェクトであるが、NESTI では、各国の資金配分機関において管理データの運営又は分析を担当する者が多く参加する専門家グループ (MARIAD) を設置している。MARIAD には、基礎となるデータ基盤開発とこれらの方法論の適用を今後も促進するための活動が期待されているとワーキング・ペーパーで言及されている。

謝辞

本稿の執筆では、OECD 科学技術・イノベーション局科学技術政策課の松本久仁子政策アナリスト、NISTEP 科学技術予測・政策基盤調査研究センター伊神正貫センター長、山下泉主任研究官から支援を頂いた。お礼申し上げます。

注11 本稿では紹介していないが、ワーキング・ペーパーでは、3章で示した自然言語処理の教師なしで行った機械的な分類である8つのトピック・クラスターと、4章で言及した専門家の分類である9つの優先トピックへの専門家による分類結果を、研究開発プロジェクトで対応付け、比較検討した結果も示されている。

参考文献・資料

- 1) 菱山 豊 (2024), OECD/CSTP の活動に関する座談会－持続可能な未来に向けた科学技術・イノベーション政策の変革を目指す OECD/CSTP を内部と外部から論じる－. STI Horizon. Vol.10, No.1, <https://doi.org/10.15108/stih.00361>
- 2) 村上 昭義 (2023), 経済協力開発機構 (OECD) における研究開発資金配分データベースを用いた分析の取組－Fundstat プロジェクト及び MARIAD の紹介と日本への示唆－. STI Horizon. Vol.9, No.1, <https://doi.org/10.15108/stih.00323>
- 3) Leonidas Aristodemou, Fernando Galindo-Rueda, Kuniko Matsumoto, Akiyoshi Murakami. (2023), "Measuring governments' R&D funding response to COVID-19: An application of the OECD Fundstat infrastructure to the analysis of R&D directionality", OECD Science, Technology and Industry Working Papers, No. 2023/06, OECD Publishing, Paris, <https://dx.doi.org/10.1787/4889f5f2-en>
- 4) 小柴 等, 林 和弘, 伊藤 裕子 (2020) 「COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析」, NISTEP DISCUSSION PAPER, No.186, 文部科学省科学技術・学術政策研究所. <http://doi.org/10.15108/dp186>