4.2.1 科学技術イノベーションに係る主な評価指標

原 泰史 *

2019 年 4 月 26 日

リード文

科学技術およびイノベーションの成立は長い時間を要する取り組みである。従来研究は事象あるいは科学領域を特定することで、定性的にその要因を解析することに主眼をおいていた。しかしながら、データベースの充実、オープン化、研究者が活用できるコンピューティングパワーの増大、および、評価手法の理論的な発展により、今日では特許、論文、あるいは競争的資金等各種の Funding に係るデータベースを接合し合うことで、定量的に科学技術の様相を解析することが可能になりつつある。

キーワード

データベース, 特許, 論文, ビッグデータ

本文

1　科学技術イノベーションに係る主な評価指標

科学技術の評価はこれまで、特定事象および年代、あるいは状況にフォーカスした調査が中心にして行われてきた。しかしながら、データのオープン化およびコンピューティングパワーの増大により、昨今では国レベルのマクロデータと研究者レベルのマイクロデータのより細密な分析、あるいはデータ間を接合し合うことで、より長期間かつ、構造的変化を見据えた上で解析することが可能になりつつある。図 1 に, 科学技術に関わる指標およびその関係性を示した。今日科学技術イノベーションを解析するにあたり利用可能なデータを, (1) マイクロ (micro) レベル, (2) メソ (meso) レベル, (3) マクロ (macro) レベルの三点に大別すると, (1) および (2) に係る指標としては、特許および学術論文の書誌情報データが広く活用されている。これらのデータでは、研究活動のアウトプットである学術論文および特許について、どのような組織の人間が、いつ、どこで、どのような

* パリ社会科学高等研究院ミシュランフェロー (前 政策研究大学院大学 SciREX センター専門職)

資金を活用して論文ないしは特許を執筆し、それがどのような既存研究を参照し、かつどれほど他の論文や特許で、どのような目的を持って活用されたかを、評価することが出来る。また昨今では、特許の非特許文献情報を網羅したサイエンスリンケージデータも活用されつつある。これら特許および学術論文を生産するにあたっては、主に日本の大学および研究機関に所属する科学者の場合、科学技術研究費 (科研費) に代表される競争的資金を活用している。こうしたインプットデータも、科学者の行動を測定・評価する上では重要な指標のひとつであるといえよう。

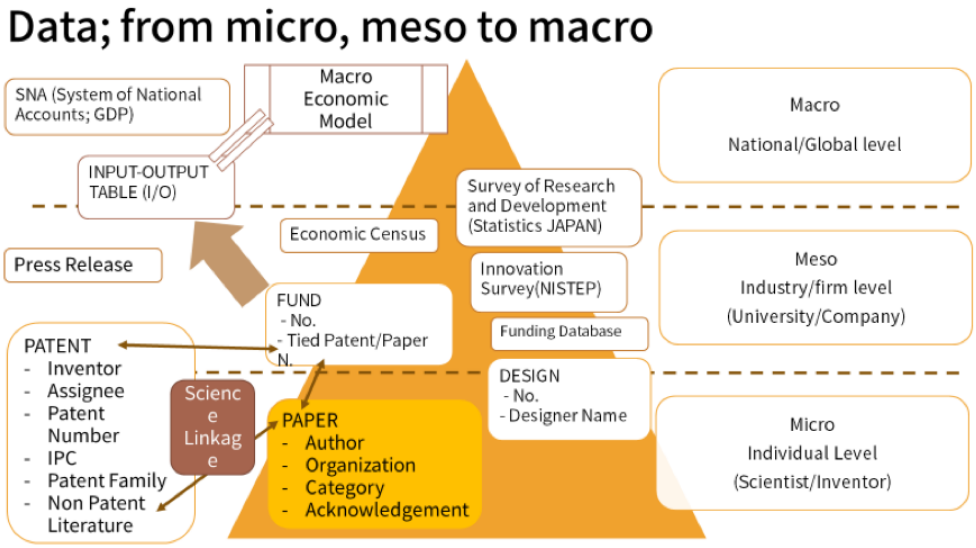

図 1　科学技術イノベーションデータの関係性
出所：筆者作成

また、近年の研究ではこれらインプット＝アウトプット間の関係性についてのみならず、科学技術の研究活動によって生み出された新たな知見がどの程度実際の生産物へと結びついたのか、すなわち、社会的な効果を持ちうるイノベーションへと結びついたかを測定することを目的として、製品データベース、プレスリリースデータベースなどに記載された製品の委細情報、主な開発者名、価格、対象消費者僧など、アウトカムを指し示しうる指標についても着目されつつある。しかしながら、ライフサイエンス産業、特に低分子医薬品を除き、多くの産業において特許と製品は 1:1 の対応関係にはない。このように、イノベーションのアウトカムを測定することには、多大なチャレンジが残されている。そのため、政府あるいは大学・研究機関では、イノベーションに係るアウトプットを測定するにあたり、発明者、あるいは企業に対してサーベイ調査を行うことでこうした課題に対応してきた。こうした取り組みの一例として、文部科学省 科学技術・学術政策研究所 (NISTEP) の全国イノベーション調査、民間企業の研究活動に関する調査 (民研調査)、総務省統計局の科学技術研究調査、経済センサスなどが挙げられる。こうしたマイクロあるいはメソレベルのデータを産業レベルで接合することで、はじめてマクロレベルでの推定を細密に行うことが可能となる。また、マクロレベルでの影響を測定するひとつの方法論として産業連関表を用いることで、科学技術研究活動から生まれた智的資産と産業ごとの動態および相互の関係性を細密に測定でき

る。このように、分析単位が様々なデータをマイクロレベルから集計することで、マクロレベルでの解析の精緻化、ひいては、科学技術がどのようにマクロ経済に影響を与えるのか、ナショナルイノベーションシステム内でのステークホルダー間の相互依存性を考慮しながら、その構造を推定し解析することが可能となる。しかしながらこうした分析を行うためには、(1) 異なるデータベース間でのデータ集計, (2) データの接合, (3) データのクリーニングおよびクレンジング, (4) 補間データの作成などの作業が都度要求される。

科学技術イノベーションの解析をマイクロレベルから開始するためには、その活動主体である科学者の研究活動を総合的に評価する必要がある。しかしながら、前述したように、科学者の能力測定に係る従来の既存研究の多くは特許あるいは論文の書誌情報のみを利用し解析をしており、科学者の多面的な効果や特性を十分に解析出来ていなかった。また、別節にて取り上げるスターサイエンティストに関わる議論でも示唆されているように、科学者は学術的な能力あるいは製品開発能力のみならず、教育活動や企業への参画、あるいは政府委員会への参画や法制度整備への間接的な関与など、実社会への社会的な影響 (social impact) も有している。このように、科学者の影響を定量的かつ総体的に評価可能とするためには、特許や論文、特許の非特許文献、製品 (プレスリリース, 表彰データ) 情報などのデータベースを複合的に組み合わせることで、科学者が学術的あるいは社会的に与える多面的な影響を解析する必要がある。

1.1 STI データ整備に関わる日本および海外の状況

科学者の活動は複合的かつ多彩であり、インプット (研究に繋がる資金調達) あるいはアウトプット (論文の公開および特許の出願)、アウトカム (社会的な貢献活動, 企業への参画) の手段も多岐に渡る。また、インプット＝アウトプット間の関係を測定することで、たとえば、科学技術研究費 (科研費) に代表される研究者に対する競争的資金制度の在り方を議論することも可能ではあるが、現時点では、こうしたデータベース間の情報を分析単位に応じて相互に接続できるツールとしてのマッチングテーブルは、NISTEP 企業名辞書[*1], NISTEP 大学・研究機関名辞書[*2]など一部に留まる。

こうした複数データセットの提供プラットフォームの海外における具体例として、(a.) 欧州 RISIS (Research Infrastructure for Research and Innovation Policy Studies)[*3], (b.) NIH Research Portfolio Online Reporting Tools (RePORT)[*4]が挙げられる。(a.) では、RISIS Datasets Portal (datasets.risis.eu) を通じて, 科学技術イノベーションに係るデータベースを研究者に対して提供している。一例として PROFLE - The German Doctoral Candidates and Doctoral Holders Study では、ドイツの大学およびファンディング機関で研究活動を行う博士号および博士候補生 (Doctoral Candidate) に対して行ったサーベイ調査の個票データを、利用申請を行った研究者が利用することが出来る。並行して RISIS では, SMS (Semantically Mapping Service) Platform[*5], CorTEXT

[*1] NISTEP 企業名辞書, http://www.nistep.go.jp/research/scisip/rd-and-innovation-on-industry/
[*2] NISTEP 大学・研究機関名辞書, http://www.nistep.go.jp/research/scisip/randd-on-university
[*3] RISIS, http://risis.eu/
[*4] RePORT, https://report.nih.gov/
[*5] SMS Platform, http://sms.risis.eu/

Platform[*6]などのデータ解析用プラットフォームも併せて提供している。前者では, Web of Science, Scopus や PATSTAT, OECD が提供する科学技術に係るインジケータ情報などを接合し分析することを可能にしている。後者では, RDF, XML, CSV 形式のデータを自動的に parse し、グラフ化やネットワーク分析などを実行できる。また (b.) では、NIH によるファンドの PI (Principal, Investigator; 研究代表者), 金額 (直接経費および間接経費), 期間, 関連プロジェクト、プロジェクトによる研究アウトプット (論文, 特許) などを Web インターフェースを通じ解析出来る。

翻って、日本のこうした科学技術に係るデータ整備および提供スキームは未だ発展途上の段階にある。科学者が研究活動を行うにあたっては、(1.) 大学運営交付金などから充当される研究資金、あるいは、(2.) 科研費, JST などの競争的資金、(3.) 企業との共同研究などを通じ資金を獲得し、研究に必要なインプットを確保する必要がある。無論、資金のみならずポスドク, RA などの人的資源も重要である。しかしながら、国立情報学研究所が提供する科研費のデータベースを例外として、JST, NEDO など国のファンディングエージェンシー、あるいは文部科学省および経済産業省など省庁による国家プロジェクトによるファンド情報はオープン化されておらず、あるいは、Web 上にデータが公開されていても、データベース化そのもは未だ中途段階にある。結果、特定の科学領域に関して政府あるいは国の研究機関がどの程度の資金を投入しているのか、把握することが極めて困難である。こうした状況を是正するため、政策研究大学院大学科学技術イノベーション研究センターでは SPIAS(SciREX 政策形成インテリジェント支援システム) と呼ばれる Web ベースのデータ接合システムを構築しているが、詳細については後節に記す。

科学者のアウトプットに関連して、論文データベースについては、JST がジー・サーチ社と共同で J-global を構築している。また特許データベースに関しては、特許庁が J-Patplat として SaaS 型のシステムを運用している他、知的財産研究所では IIP パテントデータベースとして特許解析用データベースを公開している。しかしながら、これら特許および論文間を接合した分析は一部に留まっている。日本における先駆的な取り組みとしては 池内健太 et al. (2017) を参照されたい。また、特許と論文間のみならず、インプット＝アウトプット間の関係を細密に測定するためには、特許＝論文＝競争的資金間のデータ接合が肝要となる。しかしながら、こうしたデータ整備については、論文および特許とファンド間のデータ接合が科研費などの一部競争的資金について行われているに留まる。

また前述したように、研究者の活動成果は特許および論文のみならず、社会的な貢献活動、メディアへの登場、政府関係機関の審議会への参画、産学連携への関与の度合い、企業への社外取締役あるいは株主としての参画など広義な要素を内包しているが、データの可用性の問題からこうした研究者の社会的効果については未だ定性的かつ特定の研究者を対象とした事例調査の範疇に留まる (原泰史 et al., 2017)。

しかしながら、テキストマイニングおよび自然言語処理の手法を活用することで、これら研究者の社会的側面が記録されている非定型データ (ソーシャルメディア, 新聞記事, プレスリリース) から研究者名、組織名、貢献分野など必要な情報を抜き出し解析するアプローチが採られつつある。

[*6]CorTexT Platform, http://www.cortext.net/

前述した SPIAS では, 日本経済新聞社が収集したプレスリリース情報から組織名を抽出することで, 特許、論文の出願あるいはファンドを獲得している企業および大学・研究機関との突合を行っている。このように、従来の研究者あるいは組織名に対して一意な ID 情報を付与することで解析する手法に加え、RISIS の SMS Platform が採るアプローチのような、自然言語処理にもとづき尤度を測定しながら不定形データを接合する手法を併用することにより、研究活動のインプットおよびアウトプットのみならず、アウトカムを包有して解析することは実現可能になりつつある。

1.2 主な科学技術指標 (マイクロデータ)

研究者レベルで科学的な調査を行うために広く用いられている指標として, (a.) 論文, (b.) 特許および (c.) ファンド情報が挙げられる. 詳細について以下に記す.

(a.) 論文データベース

- Scopus[エルゼビア社提供] (学術論文のデータベース、英語論文誌が中心; 研究者および組織名に ID が付与されていることにより, データの整合性に一日の長を持つ)
- Web of Science [Clarivate Analytics 社提供] (学術論文のデータベース、英語論文誌が中心; 1900 年のデータから提供されており、歴史的分析を行う上では必要不可欠である。スターサイエンティストに係る既存研究でも広く活用)
- J-global (科学技術振興機構が提供する学術論文・特許データベース; 日本の学術誌を極めて広くカバーしている)

(b.) 特許データベース

- PATSTAT (欧州特許庁 (EPO) が提供する特許データベース; ヨーロッパおよびアメリカ、日本と主要三地域の特許データベースを広くカバーしている)
- PatentsView (米国特許庁 (USPTO) が提供する特許データベース; 発明者および組織名について正規化が行われている)
- J-global (科学技術振興機構が提供する特許データベース; 発明者および組織名について正規化が行われている)
- IIP パテントデータベース (知的財産研究所が提供する特許データベース; 日本の特許データについてカバー)

(c.) ファンド情報データベース

- SPIAS (日本のファンド、特許、論文およびプレスリリース情報を総合的に接合したデータプラットフォーム; SciREX センター, NISTEP および JST 研究開発戦略センターが開発)
- 科研費 DB (NII および科学技術振興機構が提供する, 科学技術研究費 (科研費) の細目情報およびその成果論文および特許情報を接合したデータベース)

- AMED データベース (AMED が所管する競争的資金情報およびその成果を公表したデータベース)
- Nanobank (ナノテクノロジーの特化した学術、論文、特許、研究費情報のデータベース、組織名の名寄せ済み。)
- COMETS (全分野の特許、研究費情報のデータベース。組織名の名寄せ済み。)

前節に示したように, 科学者の活動を総体的に把握するためには、これらのデータベース間を接合しコホートデータを構築する必要がある。

コホートデータ構築にあたるデータ接合の課題として, (1) 元データベースの組織/研究者名情報に揺らぎや誤記載があり, そうした情報を除去あるいはクリーニングする必要があること, (2) 同姓同名や同一社名などの情報をクリーンアップし、識別する手法を開発する必要があること, (3) 英語名=日本語名など異なる言語で書かれた組織名, 研究者名などの情報を異なるデータベース間で突合する必要があること, (4) 歴史的な分析を行うにあたって, 企業の M&A 情報などを盛り込む必要があること、等が示唆できる。特許および論文データに関しては, 前述したように文部科学省科学技術・学術政策研究所が関連するデータテーブル整備を進めており、その成果を活用することで、データの精緻化を円滑に行える可能性がある。また同様に、データ管理・運用の課題として、(1) 論文あるいは特許の書誌情報データはデータサイズおよびその構造が極めて複雑であること, (2) 複数領域, 複数年度にまたがる解析を行うためには, 潤沢なコンピューティングリソースを必要とすること, (3) 競争的資金情報など機微性をともなうデータが含まれており, 高いセキュリティを担保する必要があること等が挙げられる。

1.2.1 今後の課題

今後の課題として、各データベース間の接合手法について検討する必要がある。図 1 に示したように, それぞれのデータベースは、組織名あるいは研究者名に基づき接合する必要がある。そのためには、研究者の氏名データの曖昧性除去 (disambiguation) および突合 (harmonization)、自然言語処理の技法を用いた組織名および研究者名の抽出およびマッチングを行う必要がある。また極めて多変量かつ複雑なデータ構造を処理することになるため、従来学術的な定量研究のデータ整備過程で利用されてきた RDBMS 形式ではなく、Neo4j、ElesticSearch などの不定形データに対応した解析プラットフォームについて検討・導入する必要がある (原泰史・木内満歳, 2017)。またこうした解析を行うためには, 欧州の研究コンソーシアムが RISIS (SMS Platform, CorText Platform) で実現しているような, 研究者が自由に活用できる潤沢なコンピューティングリソースを有するクラウドプラットフォームをコンピューティングインフラストラクチャとして整備する必要がある。

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.

原泰史, 壁谷如洋, and 小泉周 (2017). ノーベル賞受賞者の特性分析から見える革新的研究の特徴 (特集ノーベル賞と基礎研究: イノベーションの科学的源泉に迫る). 一橋ビジネスレビュー, 65(1):26–40. https://ci.nii.ac.jp/naid/40021245697/.

原泰史・木内満歳 (2017). Elasticsearch と科学技術ビッグデータが切り拓く日本の知の俯瞰と発見. https://www.slideshare.net/yasushihara/elasticsearch-15-spias.

原田裕明, 小柴等, 池内健太, 原泰史, 黄俊揚, 黒田昌裕, et al. (2017). 科学技術イノベーション政策立案のためのデータプラットフォーム: テキストマイニングによる科学技術分野の同定. In 年次学術大会講演要旨集, volume 32, pages 344–347. 研究・イノベーション学会. https://dspace.jaist.ac.jp/dspace/bitstream/10119/15004/1/kouen32_344.pdf.

池内健太, 元橋一之, 田村龍一, 塚田尚稔, et al. (2017). 科学・技術・産業データの接続と産業の科学集約度の測定. DISCUSSION PAPER 142, 科学技術・学術政策研究所. http://data.nistep.go.jp/dspace/bitstream/11035/3161/1/NISTEP_DP142_FullE.pdf.

齋藤裕美 and 牧兼充 (2017). スター・サイエンティストが拓く日本のイノベーション (特集ノーベル賞と基礎研究: イノベーションの科学的源泉に迫る). 一橋ビジネスレビュー, 65(1):42–56. https://ci.nii.ac.jp/naid/40021245706/.

関連データ・ソース

-

関連する拠点授業科目、関連する研究プロジェクトの情報

-