JAIST Repository
https://dspace.jaist.ac.jp/

Title	日本のスターサイエンティスト分析に係るデータプラットフォーム整備
Author(s)	原，泰史
Citation	年次学術大会講演要旨集，32：574-577
Issue Date	2017-10-28
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/14859
Rights	本著作物は研究・イノベーション学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Research Policy and Innovation Management.
Description	一般講演要旨

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

２Ｄ２０

日本のスターサイエンティスト分析に係るデータプラットフォーム整備

○原泰史（政策研究大学院大学 科学技術イノベーション政策研究センター）[1]

1. はじめに

スターサイエンティストの解析のためには、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、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。しかしながら、スターサイエンティストは学術的な能力あるいは製品開発能力のみならず、教育や企業への参画などを通じ、実社会への影響も同程度に有している。このように、スターサイエンティスト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、スターサイエンティスト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ためのプラットフォームを構築する必要がある。

本稿では、スターサイエンティストの解析に係るデータ構築の取り組みおよび、関連するデータセットの概要、予備的な分析結果および今後の課題について明らかにする。

2. スターサイエンティストデータプラットフォーム

2-1. 海外の動向および日本の現状

スターサイエンティストの解析を行うにあたっては、まず、「どこの（所属機関）」「誰が(研究者名)」スターサイエンティストであり、「どのようにして(判断尺度)」研究者群からスターサイエンティストを抽出できるか示す必要がある。また母集団としての研究者群を、どのよう特定出来るかも検討の余地がある。研究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[2]，NISTEP 大学・研究機関名辞書[3]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)[4]，(b.) NIH Research Portfolio Online Reporting Tools (RePORT)[5]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

[1] yasushi.hara1982@gmail.com, ya-hara@grips.ac.jp, hara@iir.hit-u.ac.jp
[2] NISTEP 企業名辞書, http://www.nistep.go.jp/research/scisip/rd-and-innovation-on-industry/
[3] NISTEP 大学・研究機関名辞書, http://www.nistep.go.jp/research/scisip/randd-on-university
[4] RISIS, http://risis.eu/
[5] RePORT, https://report.nih.gov/

出来る。並行して RISIS では，SMS（Semantically Mapping Service）Platform[6]，CorTEXT Platform[7] 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，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確認出来る。

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。研究者が科学的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、あるいはデータベース化そのものも中途段階にある。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（原田 et al. 2017）に譲りたい。

研究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 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 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。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た。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、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。

2-2. スターサイエンティスト・プラットフォームの構築

前節で示したように、スターサイエンティストは、論文数、被引用数、ひいては特許出願など、様々な点で優れた研究者のことを指し、単純に論文や特許数の多寡のみで定めることは出来ない。また、スターサイエンティストは必ずしも一意に定義されているわけではない。スターサイエンティストを検出するための評価軸は大別して（1.）経済的な影響度，（2.）研究的なインパクト，（3.）社会的なインパクトに分けられ、どの基準で評価するかによってスターサイエンティストの意味は異なる点にも注意しなくてはならない。既存研究では、論文データベースや特許データベース、および地域の企業のデータセットなどを結合し、企業名ないしは研究者名で突合することで、スターサイエンティストの特性や産業

[6] SMS Platform, http://sms.risis.eu/
[7] CorTexT Platform, http://www.cortext.net/

界へのインパクトを様々な観点から明らかにしてきた(Zucker & Darby; 2002)。スターサイエンティスト、特に日本の研究者に着目し解析するにあたり活用可能と考えうるデータセットとして、以下が挙げられる。

(1.) 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)

(2.) 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)

(3.) 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)
- 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。)

(4.) ベンチャー企業情報データベース
- Entrepredia (ベンチャー企業のデータベース、日本企業の情報が中心)
- Crunchbase (ベンチャー企業のデータベース、米国企業の情報が中心)

前節に示したように，スターサイエンティストの同定および、スターサイエンティスト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研究者のコホートデータを構築する必要がある。コホートデータを形成するために必要なデータベース間の接合可能性，接合手法などについて，図 1 にまとめた。

スターサイエンティストに係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。

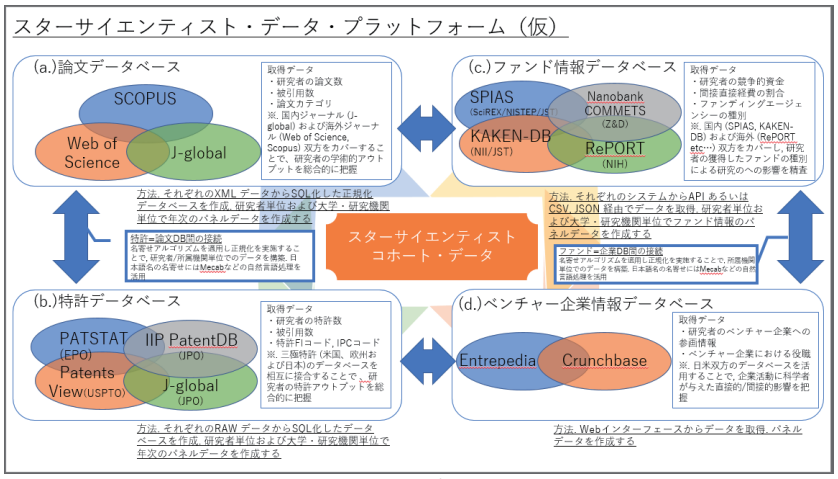

図 1. スターサイエンティスト・データ・プラットフォーム

3. データ整備の現状および今後の課題

最後に、スターサイエンティストデータの整備状況と今後の課題について記しておきたい。2017 年 9 月現在、各データベースからのデータ集約および予備的な解析を中心に実施中である。具体的には、Clarivate Analytics 社が提供する Web of Science Core Collection データについて，XML から MySQL 形式への変換作業を実施した。具体的には，1981 年から 2016 年までの書誌データについて，NISTEP 大学・研究機関名辞書との接合を行い，スターサイエンティストの特定および研究業績データの抽出を実施した。また SPIAS や科研費データベースを用いることで，研究のインプットとアウトプットを接合する作業を実施中である。また、Clarivate Analytics 社が提供する Highly Cited Researchers (HCR) データと、J-global および PatentsView データを照合することで、高頻度引用を有する研究者が、特許でも同様に高いパフォーマンスを示すか否かの調査を実施し、一部成果は（齋藤，牧 2017)にて公表した。また，Crunchbase を用いベンチャー企業の情報について収集中である。

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、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り，後者については Mecab (Yet Another Part-of-Speech and Morphological Analyzer) などの活用が考えうる。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・木内 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 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。

参考文献

池内，元橋，田村，塚田（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」，科学技術・学術政策研究所，DISCUSSION PAPER；142，http://doi.org/10.15108/dp142
齋藤，牧（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」，一橋ビジネスレビュー，2017 夏号，pp.42-57.
原，壁谷，小泉（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」，一橋ビジネスレビュー，2017 夏号，pp.26-41.
原，木内（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，July Tech Festa，産業技術大学院大学，2017.08.27
原田，小柴，池内，原，黄，黒田（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―」，研究イノベーション学会
Zucker，G.，Darby，M.R.，Armstrong，J.S. 2002. Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology. Management Science. 48(1) 138-153.