２Ｃ１７

オープンデータ時代の研究者：
研究内容と研究様式のパラダイムシフト

○沼尻 保奈美、竹之内 高志、林 隆之 (政策研究大学院大学)

1. はじめに：オープンデータがもたらす研究のパラダイムシフト

近年、オープンサイエンスの理念が世界的に広がりを見せており、科学的知識の生産と普及のあり方に根本的な変革をもたらしつつある(Baker, 2016; Borgman,2017; Couture et al., 2018)。オープンサイエンスとは、研究データ、手法、成果などを広く公開し、誰もが自由にアクセス・利用できるようにすることで、科学的知識の透明性と再現性を高め、研究プロセスの効率化を図る取り組みである(Roche et al., 2014; Molloy, 2011)。この取り組みの中核を成す要素の一つがオープンデータである。オープンデータとは、研究活動の過程で生成されたデータを、誰もが自由にアクセス、利用、再配布できる形で公開することを指す(Borgman,2017.; OECD, 2007)。オープンデータの重要性は国際的に広く認識されており、その推進に向けた具体的な取り組みが進められている。例えば、欧州委員会は 2019 年に、データの「Findable（見つけられる）」「Accessible（アクセスできる）」「Interoperable（相互運用できる）」「Reusable（再利用できる）」を目指す FAIR 原則を採用し(European Commission [EC], 2019)、ユネスコも 2021 年の総会で加盟国にオープンデータの推進を促している(UNESCO, 2021)。これらの動きは、オープンデータが科学研究の発展と社会的価値の創出に重要な役割を果たすという認識が広がっていることを示している。

このオープンデータの進展度と影響を理解するために、これまで多くの研究が行われてきた(池内ら,2020;Tenopir et al.,2011:Kim & Zhang、2015)。しかし、これらは主に研究者のデータをオープン化するモチベーションや障壁に焦点を当てており、オープンデータが研究活動そのものにもたらす変化については十分に検討されていない。特に注目すべき変化は、オープンデータの普及に伴って台頭してきた「データ駆動型研究」である。データ駆動型研究は、大規模なオープンデータを活用し、データ分析から仮説や知見を導き出す研究スタイルであり、従来の仮説検証型とは異なる、科学研究の新たなパラダイムとして注目を集めている(Hey et al. 2009;Kitchin, 2014; Gutierrez et al., 2021)。データ駆動型研究では、多様な分野から収集された大規模なデータセットを統合・分析することにより、これまで発見できなかった複雑な現象のパターンや関係性を明らかにする。この新しい研究スタイルは、研究者の研究手法や問題意識、さらには研究構造そのものに大きな影響を与える可能性がある(Kitchin, 2014)。

このように、オープンサイエンスとオープンデータの動向は、一見すると学術研究のパラダイムシフトを示唆しているように見える。しかし、別の見方をすれば、研究データの共有と公開という概念は、科学の歴史に深く根ざしたものであり、これらの動向は、むしろ長年の知識共有の理念を現代のデジタル技術によって再解釈し、拡張したものとも捉えられる。科学の発展は常に開放性と知識共有に支えられており、15 世紀の活版印刷の発明、17 世紀以降の学術雑誌の創刊、19 世紀の学術団体の設立などが、この原則を具現化し、科学の進歩を加速させてきた(Eisenstein, 1980; Stracke, 2020)。20 世紀には、マートンによる CUDOS 原理の提唱やギボンズらによる知識生産のモード論など、科学研究の規範や在り方に関する重要な概念が示された(Storer, 1973; Gibbons, 1994)。このような歴史的文脈において、オープンデータは、従来の科学研究の規範的基盤を踏まえつつ、デジタル技術の発展がもたらす新たな機会に対応する試みと位置づけられる。さらに現代におけるオープンデータの実践は、デジタル技術とインターネットの発展に支えられることで、その影響力と変革の速度において前例のないものとなる可能性がある(Berger et al., 2016)。情報技術の発展によってデータの収集、管理、共有が容易になったことで、オープンデータの実践は急速に広がりを見せている。こうした技術的な後押しを受けて、研究者はこれまでにない規模と速度でデータを共有し、協働することが可能になった(Hampton et al., 2013)。オープンデータの意義は、新型コロナウイルス感染症（COVID-19）のパンデミックへの対応において顕著に

示された。世界中の研究者が協力し、ウイルスの遺伝子配列データや疫学データを迅速に共有することで、ワクチンや治療法の開発が加速された(Nane et al., 2023)。この経験は、オープンデータが単なる理念ではなく、実際の危機対応において極めて有効な手段であることを実証したと言える。このようなオープンデータの実践とその影響の広がりは、既存の学問分野の枠組みを超えた新たな研究アプローチや協力体制を生み出している。例えば、生物学と情報科学の融合によるバイオインフォマティクスの発展(Kell & Oliver, 2004)や、複数の分野のデータを統合・分析することで形成されつつある持続可能性科学(Clark & Dickson, 2003)などが挙げられる。これらの事例は、オープンデータの進展が既存の学問分野の発展を促進し、新たな学問分野の創出にも寄与する可能性を示唆している。

このような議論を踏まえて、本研究では、科学技術社会論（STS）の分野における Hackett et al. (2017)による議論などを参照しつつ、新しい学問分野の創出要因とオープンデータの関係に焦点を置く。彼らの整理を以下の表 1 の左列「新しい学問分野、専門分野、研究領域を生み出す要因」に示し、この枠組みを基に、右列「オープンデータが新たな分野を生み出す要因」に整理した（表 1）。

表 1　新しい学問分野の創出要因とオープンデータの関係

要素	新たな分野を生み出す要因	オープンデータが新たな分野を生み出す要因
1	新しい科学的知識が、新しい使われ方をする(Schweber 2006; Shostak 2005)	・他の研究者が、新たな視点や分析手段を用いてデータを再分析し、新たな知見を得る。
2	新たな研究装置や研究技法が、これまで知り得なかったことを明らかにする(Bechtel 1986; Clarke 1998; Mulkay, Gilbert, and Woolgar1975)	・大量のデータや複数のデータを用いた新たな分析手法や分析技術を開発する。 ・オープンデータを再分析し、現在の技術では解決することのできない研究課題が将来的に解決可能になる(Tenopir et al., 2015)。
3	科学的な役割や職業が、既存の役割や職業と交差して形成される (Ben-David and Collins 1966; Frickel 2004)	・データ分析専門家とのチームサイエンスが生じ、データ分析が深化する。 ・市民が研究に参加するシチズンサイエンスが生じる。
4	新しい研究分野や知見を生み出す可能性のある問題や現象を中心に結束力のある研究者ネットワークが形成される(Griffith and Mullins 1972; Powell et al. 2005)	・異分野間のデータを活用した学際的な研究が促進される(Parsons et al., 2011)。 ・海外と共同研究をする際には複雑な手続きが必要であるが、データが公開されることで国際的な研究協力が容易になる(Tenopir et al., 2015)。
5	研究は、影響力のある利益集団の注目を集める (Clarke 1998; Frickel 2004;Lenoir 1997)	・複雑な社会現象や科学的課題に対する研究は政策立案者、資金配分機関や産業界からの支持をうけ、研究費が獲得しやすくなる。 ・社会のあらゆるデータが公開され、利用可能になることで社会問題解決に向けた研究が行いやすくなる。

表 1 に示すように、オープンデータが新しい学問分野の形成に与える影響は、主に 5 つの要因から考察することができる。これらの要因を踏まえ、以下のリサーチクエスチョンを設定する。

（1） オープンデータによって分野形成につながる効果が得られているか。オープンデータの導入によって研究者の研究手法や研究の問題意識、研究構造にどのような変化が生じているか（要因 1,2 に関連）

（2） オープンデータの普及に伴い、異分野間のコラボレーションはどのように活性化しているか？（要因 3,4 に異分野間のコラボレーションはどのように活性化しているか（要因 3,4 に関連）、

（3） 社会に直接役立つ知識の創出はどのように促進されているか（要因 5 に関連）。

（4）このようなオープンデータの効果が、データ駆動型研究への展開につながっているか？（要因

1,2 に関連）。

これらの問いに答えることで、オープンデータが学術研究のあり方をどのように変革しつつあるかを明らかにし、今後の学術研究の展望を示すことを目的とする。

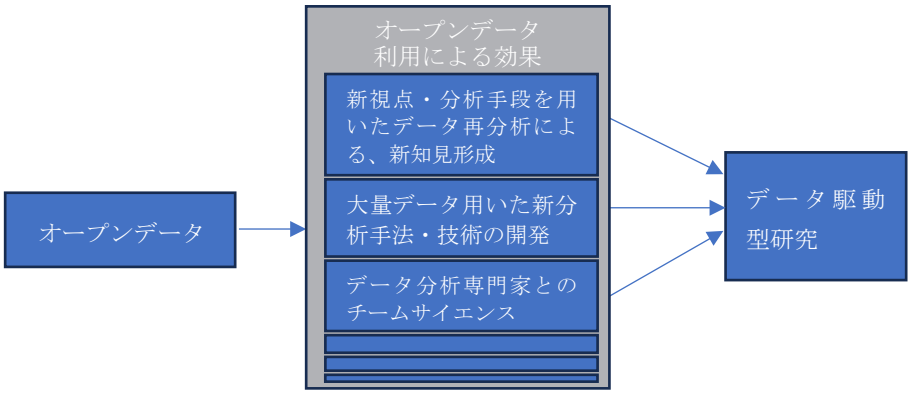

図 1　オープンデータの活用がデータ駆動型研究に与えるうる影響モデル

2. 分析：データ利用型研究者へのアンケート調査

2.1 分析対象：

本研究では、データ利用型の研究を行っている研究者を分析対象とする。ここでいうデータ利用型とは、研究データを活用し、データ分析を研究方法の中心に据えた研究活動を指す。このようなデータ利用型研究を行っている研究者として、Web of Science(WoS)に収録されている論文の中で、クラリベイト社の Data Citation Index (DCI)に収録されているデータを引用した日本の研究者を対象とする。

抽出においては、第一に、WoS に収録されている 2020 年から 2023 年の日本の研究者による論文のうち、DCI のアクセッション番号（DRCI）を含む参考文献がある 1,988 編を抽出した。これらの論文は、DCI に収録されているデータセット、ソフトウェア、Data Study、Repository の 4 つのタイプのいずれかを引用しており、いずれであってもデータを利用した研究を行っていることが期待されるため、分析対象とした。第二に、抽出された論文から著者を抽出した。著者数は延べ 10,303 名、固有の著者数は 6,612 名であった。第三に、固有の著者 6,612 名に対して、Web of Science からメールアドレスを取得した。取得できた著者は 1,497 名であり、アンケート調査の対象とした（ただし、実際にはメールが不達の場合があった）。

2.2 アンケートの設計と項目

本研究では、データ利用型研究を行う研究者の実態と意識を調査するため、以下の 5 つの主要セクションから構成されるアンケートを設計した：

1. 回答者の基本情報（研究分野、所属機関、職位等）
2. データ公開に関する経験と意見（公開経験、公開理由、公開場所、公開による効果）
3. オープンデータの使用経験と影響（使用経験、使用理由、研究活動への影響）
4. データ駆動型研究に関する質問（実施状況と認識）
5. オープンデータ政策に関する意見

ここでいう「データ駆動型研究」とは、オープンデータに限らず、大量のデータを分析することから、帰納的に新たな仮説や知見を導出する研究方法を指す(Kitchin, 2014)。従来の理論構築や仮設設定から始める演繹的アプローチとは異なり、データの探索的分析を通じてパターンや関係性を見出し、それらに基づいて仮説を構築する方法である。各セクションでは、選択式の質問と自由記述式の質問を組み合わせ、定量的および定性的データの収集を行った。なお、本報告では、オープンデータの利用に関する設問にのみ焦点をおいて報告する。

2.3 データ分析方法

オープンデータの活用がデータ駆動型研究に与えるうる影響を以下のモデルで説明する（図 1）。第一に、オープンデータの利用が、複数の効果（効果 1、効果 2、効果 3）を通じて研究に影響を与える。第二に、これらの効果が総合的に作用し、データ駆動型研究の実践につながる、各効果の強さや影響の経路は、研究分野や研究者の属性によって異なる可能性がある。具体的な分析としては、（1）各質問項目に対する記述統計分析（2）研究分野やデータ公開経験等の属性による回答の違いを明らかにする

ためのクロス集計分析（3）データ公開経験とデータ駆動型研究の実施度合いの関係等を調べるための相関分析を行う。

3 結果

3.1 回答状況

調査対象となった 1,497 名のうち、メールアドレスが既に使われていないなどの不達を除くと、アンケート送付の母集団は 1,368 名である。そのうち、197 名から回答を得た（回収率 14.4%）。この中で、有効回答数は 183 件であった

3.2 データ利用型研究者（回答者）の属性

本調査の回答者の基本属性について、研究分野と所属機関の分布を図 2 に示す。研究分野については科学研究費補助金の審査区分を用いた。研究分野については、B（数物系科学）が最も多く 38 名、次いで F（農学）が 33 名、G（生物系）が 32 名となっている。A（人文社会）は 24 名、K（情報学）は 18 名となっている。所属機関については、国立大学が最も多く 88 名、次いで私立大学が 28 名、国立研究開発法人が 33 名となっている。職位については、「教授、研究機関の部・室・グループ長」が 60 名、「准教授、主任研究員」が 52 名、専任講師が 10 名、助教が 21 名、ポストドクターが 20 名、その他が 17 名となっている。

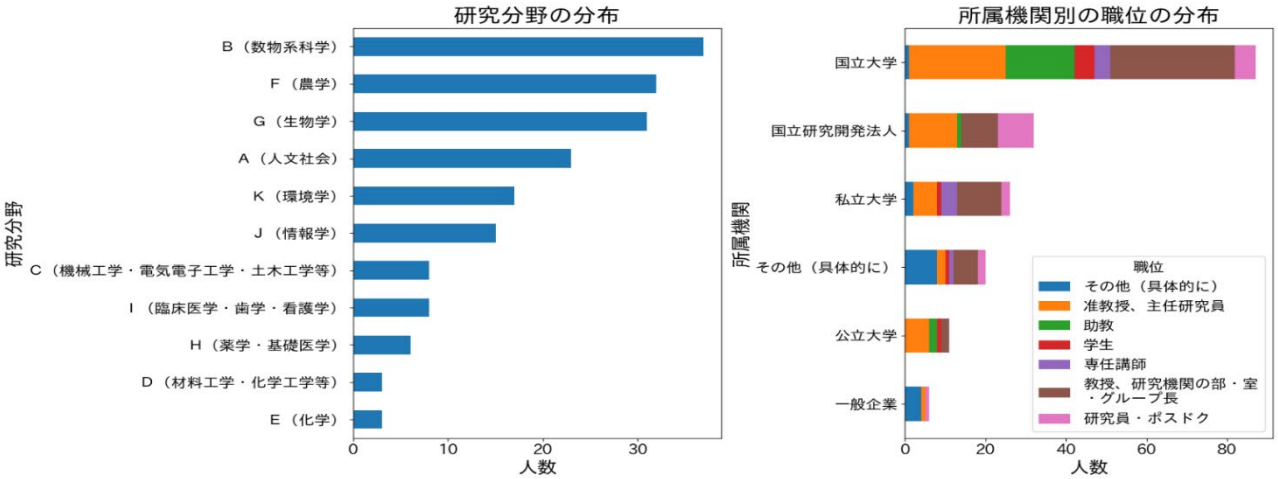

図 2　データ利用型研究者（回答者）の属性

3.3 オープンデータの使用状況

図 3 は、過去 3 年間の論文等の出版数とオープンデータを利用した論文等数の関係を示している。分析結果から、論文出版数は 1〜10 本の研究者が最も多く、そのうちオープンデータを利用した論文は 0〜3 本が主流であることが分かる。注目すべきは、論文出版数 4〜10 本の層で、オープンデータ利用に最も多様性が見られる点である。一方、論文出版数が多い（11 本以上）の研究者では、オープンデータを利用した論文数が比較的少ない傾向が観察された。これらの結果は、オープンデータの利用が特定の研究生産性層に偏っているわけではなく、様々な出版活動レベルの研究者によって行われていることを示唆している。

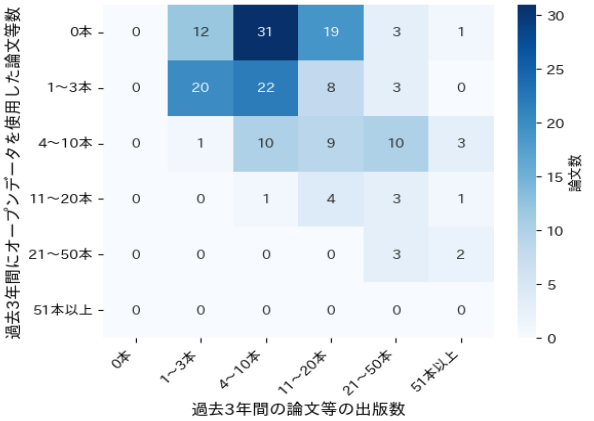

図 3　過去 3 年間の論文等の出版数とオープンデータを利用した論文等数の関係

3.4 オープンデータ使用による研究活動の変化

図 4 はオープンデータを使用することにより研究活動がどのように変化したかに関する回答の頻度分布を示している。最も顕著な変化は、異なるデータセットの比較分析（約 60 件）と多様なデータからの新知見獲得（約 55 件）であった。次いで、データ形成経費の削減による研究効率の向上（約 45 件）や、標準的情報の獲得による研究基盤の強化（約 40 件）が挙げられた。また、ビッグデータ分析の機会の増加（約 40 件）も同程度の頻度で挙げられ、オープンデータが大規模データ分析を促進していることが示唆された。学際的研究の実施（約 35 件）や研究の再現性向上（約 30 件）といった研究の質的向上に関する変化も一定数報告された。一方で、一般市民の研究参加体制の発展（約 5 件）など、市民参加に関する変化の報告は限定的だった。これらの結果は、オープンデータが研究活動に多面的かつ実質的な影響を与えていることを示している。

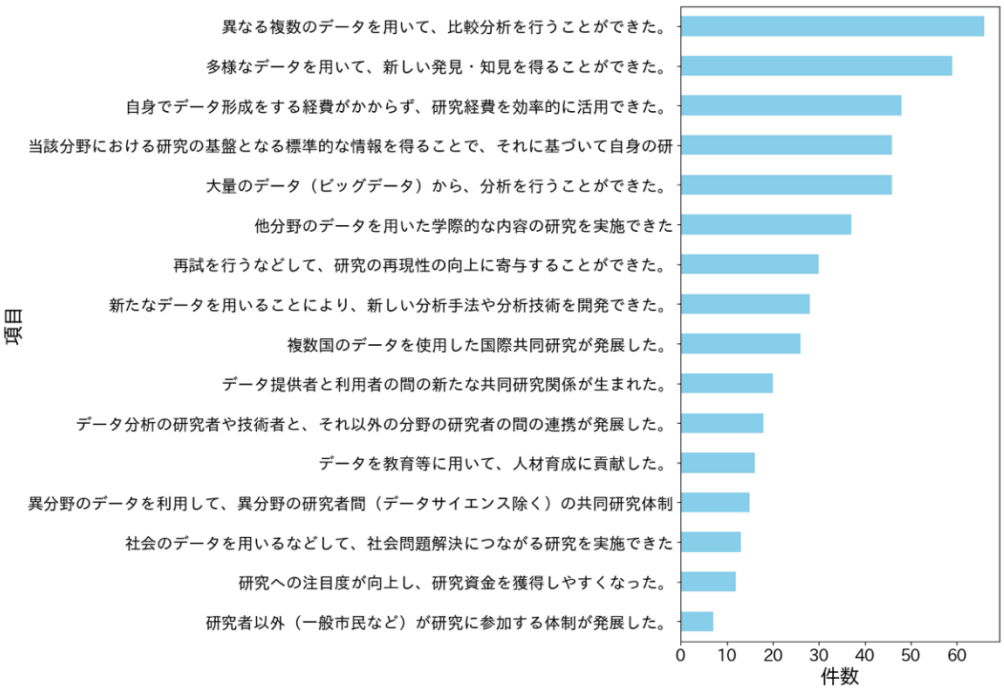

図 4　オープンデータ使用による研究活動の変化に関する回答の頻度

3.5 データ駆動型研究の実施状況

図 5 はデータ駆動型研究の実施度合いと研究分野の関係を示している。データ駆動型研究の実施度合いは、研究分野によって異なる傾向を示している。情報学が最も高い中央値（約 80%）を示し、農学、生物学、医学・基礎医学がそれに続く（約 70%）。一方、機械工学・電気電子工学・土木工学等は最も低い中央値（約 10%）を示している。しかし、ほとんどの分野において回答が 0 から 100 までの広範囲に分布しているため、回答者間での質問解釈の差異や、分野内での実践の多様性を示唆している可能性がある。

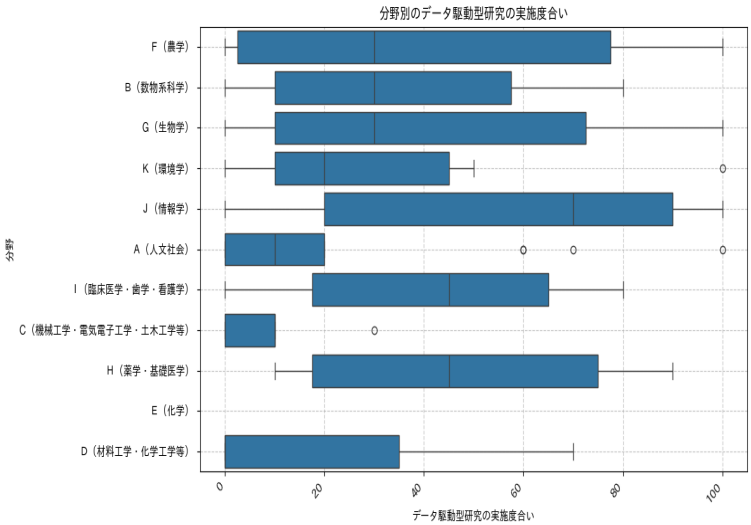

図 5　データ駆動型研究の実施度合いと研究分野の関係

3.6 データ駆動型研究の実施度合いに影響を与える要因分析

データ駆動型研究の実施度合いに影響を与える要因を明らかにするため、重回帰分析を実施した。従属変数には、個人のデータ駆動型研究の実施度合いを 0 から 1 までの連続値として用い、独立変数にはオープンデータ使用の有無と、オープンデータ使用による研究活動の変化を示す項目を用いて、ステップワイズ法を用いて変数の選定を行った。分析の結果を表に示す。いくつかの説明変数の影響は有意であるものの、モデルの R2 乗の値は 0.267 にとどまり、限定的なものとなった。

説明変数 (1)「「データ駆動型研究」へのオープンデータを使用か」が、実施度合いに最も強い正の影響を与えており、オープンデータ使用がデータ駆動型研究の重要な促進要因であることが確認された。(2)「大量のデータ（ビッグデータ）から、分析を行うことができた」という項目も、実施度合いに有意な正の影響を与えており、ビッグデータの活用も強く影響する。(3)「オープンデータによって基盤となるデータが得られやすくなったから」という項目は、実施度合いに対して負の影響を示す傾向が見られたが、わずかに統計的有意水準に達しなかった。(4)「再試を行うなどして、研究の再現性の向上に寄与することができた」という項目は、正の影響を示す傾向が見られたが、統計的に有意ではなかった。

表 2　データ駆動型研究の実施度合いの重回帰分析結果

説明変数	非標準化係数 (B)	標準誤差 (SE)	標準化係数 (ß)	t 値	有意確率 (p)
定数	14.491	7.934		1.826	0.073
オープンデータの使用度合い	27.453	8.057	0.387	3.407	0.001
研究基盤データへのアクセス向上	-15.426	7.743	-0.226	-1.992	0.051
ビッグデータ分析の可能性	17.177	7.277	0.258	2.360	0.021
再試などから研究の再現性の向上に寄与	14.259	7.983	0.195	1.786	0.079

(R^2: 0.267、調整済み R^2: 0.220、F 値: 5.642、p 値: .001)

4.議論

これらの結果は、オープンデータが新たな学問分野の形成に与える影響について、いくつかの重要な示唆を提供している。まず、ビッグデータ分析の実施が有意な影響を示していることは、オープンデータが大規模かつ複雑なデータセットの分析を可能にし、新たな科学的知見を生み出す潜在力を持っていることを示唆している。一方で、単にデータが得られやすくなっただけでは、必ずしもデータ駆動型研究の促進につながらない可能性も示唆された。また、研究の再現性向上への寄与が示唆されたことは、オープンデータが科学の信頼性と透明性を高め、より堅固な知識基盤の構築に貢献する可能性を示している。これらは、新たな研究分野の形成において重要な役割を果たす可能性がある。しかし、今回使用した統計モデルの説明力は中程度であり、データ駆動型研究の実施度合いに影響を与える他の要因も存在する可能性が高い。今後の研究では、ベータ回帰など、0 から 1 の範囲の連続値により適した分析手法の適用や、より多様な要因を考慮したモデルの構築が望まれる。

参考文献

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 2016 533:7604*. https://www.nature.com/articles/533452

Borgman, C. L. (2017). *Big data, little data, no data : scholarship in the networked world*. 383.

Berger, B., Daniels, N. M., & William Yu, Y. (2016). Computational Biology in the 21st Century: Scaling with Compressive Algorithms. *Communications of the ACM*, *59*(8), 72. https://doi.org/10.1145/2957324

Clark, W. C., & Dickson, N. M. (2003). Sustainability science: The emerging research program. *Proceedings of the National Academy of Sciences*, *100*(14), 8059–8061. https://doi.org/10.1073/PNAS.1231333100

Couture, J. L., Blake, R. E., McDonald, G., & Ward, C. L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLOS ONE*, *13*(7), e0199789. https://doi.org/10.1371/JOURNAL.PONE.0199789

Gibbons, M. (1994). *The new production of knowledge: the dynamics of science and research in contemporary societies*. https://philpapers.org/rec/GIBTNP

Gutierrez, R. R., Lefebvre, A., Núñez-González, F., & Avila, H. (2021). Towards adopting open and data-driven science practices in bed form dynamics research, and some steps to this end. *Earth Surface Processes and Landforms*, *46*(1), 47–54. https://doi.org/10.1002/ESP.4811

Hackett, E. J. (2017). The Social and Epistemic Organization of Scientific Work. In U. Felt, R. Fouché, C. A. Miller, & L. Smith-Doerr (Eds.), *The Handbook of Science and Technology Studies* (4th ed., pp. 500-520). Cambridge, MA: MIT Press.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future ofecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/10.1890/120103

Hey T, Tansley S, Tolle K (2009) Jim Grey on eScience: A transformed scientific method. In: Hey T, Tansley S, Tolle K (eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond: Microsoft Research, pp. xvii–xxxi.

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, *26*(1), 99–105. https://doi.org/10.1002/BIES.10385

Kim, Y., & Stanton, J. (2012). Institutional and Individual Influences on Scientists' Data Sharing Practices. *The Journal of Computational Science Education*, *3*(1), 47–56. https://doi.org/10.22369/ISSN.2153-4136/3/1/6

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1). https://doi.org/10.1177/2053951714528481

Nane, G. F., Robinson-Garcia, N., van Schalkwyk, F., & Torres-Salinas, D. (2023). COVID-19 and the scientific publishing system: growth, open access and scientific fields. *Scientometrics*, *128*(1), 345–362. https://doi.org/10.1007/S11192-022-04536-X/FIGURES/5

OECD (2015), Data-Driven Innovation: Big Data for Growth and Well-Being, *OECD Publishing, Paris,* https://doi.org/10.1787/9789264229358-en

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., Kokko, H., Jennions, M. D., & Kruuk, L. E. B. (2014). Troubleshooting Public Data Archiving: Suggestions to Increase Participation. *PLOS Biology*, *12*(1), e1001779. https://doi.org/10.1371/JOURNAL.PBIO.1001779

Storer, N. W. (1973). *RobertK. Merton Edited and with an Introduction by The Sociology of Science Theoretical and Empirical Investigations*.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/JOURNAL.PONE.0021101

池内有為, & 林和弘. (2020). *研究データ公開と論文のオープンアクセスに関する実態調査2020. 316*. https://doi.org/10.15108/RM316