２Ｄ１０

オープンデータによるデータ駆動型研究の推進

○沼尻　保奈美、林　隆之(政策研究大学院大学)

1. はじめに：データ駆動型研究について

研究活動で得られたデータや知見を、学術界で共有・活用するオープンサイエンスという新たな研究様式が各国で推進されている。オープンサイエンスは「研究成果およびデータの共有、研究手法の透明性の確保、研究者の社会的インパクトを増大させる費用対効果の高い取組み」と定義され（RIN/NESTA, 2010)、日本の第6期科学技術基本計画においてもオープンサイエンスを基盤とした研究活動の推進がされている。

オープンサイエンスの要素の一つである、「研究成果およびデータの共有」はオープンデータと称され、それらを集約したビックデータを駆使して行われる研究活動を「データ駆動型研究」という（OECD 2015）。例えば、遺伝学分野では、ICT の発展によるデータ取得の効率化やコスト低下により、ビックデータの取得および解析に基づく研究が可能になっている。「データ駆動型研究」は新しい概念であり、明確な定義は未だ存在しない。しかし、OECD のレポートでは、「データ駆動型イノベーション（DDI）」の一環として、その概念が研究活動にも活用される可能性が示唆されている。DDI は「製品、プロセス、組織的手法、市場を改善・育成するためにデータと分析を利用すること」と定義される。ICT の発展によりビックデータの収集と活用が可能になり、データそのものを扱うイノベーションが期待されている。近年、学術界では、科学の第四のパラダイムとされる「データ集約型科学発見」の到来に直面している。これまでデータ集約的ではなかった分野においても、データを利用したシミュレーションを行う研究が可能になった。オープンデータは、データ駆動型研究の基盤となり、研究活動の活発化や費用対効果を高めると期待され、その現状についての研究が行われている。例えば、データ共有・活用が盛んであるとされる遺伝学分野や材料科学についてのオープンデータの現状やその要因についての研究は行われてきた(Suhr,2020)。しかし、研究分野によってデータ引用の習慣に違いがあるため、その引用関係を明らかにするのは困難である(Nicolas ほか,2014)。そのため、データ公開やその活用が行われている分野についてデータの引用といった観点から横断的に調査した研究は存在しない。現在、2020 年に FORCE11（The Future of Research Communications and e-Scholarship）から発出された「データ引用原則の共同宣言（Joint Declaration of Data Citation Principles：JDDCP）」により、学術界のデータ引用の慣習化を進める動きがみられる。今後、データ駆動型研究のためのオープンデータ推進の効果を評価するためは、公開データの引用関係まで考慮した評価を行う必要がある。本研究では、研究データのデータベースを用いて、データ公開や引用の現状がどのように分析されるかを明らかにする。

2. 分析：オープンデータの現状とその活用

2.1 研究の問い

オープンデータの現状や活用状況が不明な現状を踏まえ、以下の二つの問いを立てる。(1)データ公開が盛んな国や分野はどこか。(2)データ公開によって「データ駆動型研究」が活性化している分野はどこか。問(1)に関しては、各国のオープンデータ政策やその速度によってデータ公開数の状況が異なっていることが想定され、同様に、研究分野によってもオープンデータの重要性の認識が異なるため、その現状に違いが発生している可能性がある。問(2)に関して、公開されたデータを利用して研究が行われる分野もあれば、公開データに依存しない研究分野も存在する可能性がある。データ駆動型研究を推進するための評価をおこなうためには、その分野ごとのデータ公開・利用の現状やその多様性を確認する必要がある。これまで先行研究では、個別の分野においてデータ公開がなされているかや、データ公開または理由の動機についての調査がなされてきたが、データ公開・利用の全体像は明らかでない(Bettina ほか, 2017: Nicolas ほか,2014,2020:Silvello)。

2.2 分析対象：DCI 登録データの状況

本研究では、クラリベイト社が発行する Data Citation Index（DCI）を分析に用いる。DCI とは 2012 年から同社で作成されている、401 以上のリポジトリからなる 900 万以上のデータセット等および、それらを引用する論文情報を収録したデータベースである[1]。さらに、DCI に登録されたデータ（以下、DCI 収録データ）を引用している論文の情報として同社の Web of Science（WoS）を使用する。これまで DCI データが公表された初期に、収録データの分野やその引用についての分析がなされたり(Nicolas ほか, 2014)、DCI の中のソフトウェアに関する分析は行われてきた(Park ほか, 2019)。しかし、近年のオープンデータの進展の中でデータ公開と活用がどのような現状になっているかは不明である。

DCI のデータを分析するにあたり、DCI のデータ収録内容の概要を示す。2020 年時点では DCI の総レコード数は 13,882,271 件であり、それらデータを引用している WOS 論文はのべ 1,371,848 件である。DCI 登録データは 4 種類のドキュメントタイプに分類される(図 1)。DCI に含まれるデータの種類として、データセットが全体の 88%を占めている。

表 1　DCI 収録データの基本情報

ドキュメントタイプ	データセット	データスタディ	ソフトウェア	レポジトリ
種類の概要	収集されたデータ、研究データまたはソフトウェアの単一または一貫したデータセット	研究データまたはソフトウェアとともにリポジトリに格納されている説明	研究を行う上で使用されたソフトウェア	研究データを含むデータベースおよび生データを格納してアクセスを可能にするメタデータ
データベース数(件)	12,227,647	1,400,409	253,772	443
引用(件)	1,012,547	294,251	53,453	11,592
引用割合(%)	73.8	21.4	3.8	0.8

実際に DCI の収録データを確認して明らかになることは、そもそも収録されるような「オープン化された研究データ」とはどのようなものであり、データは何を「1 件」と数えるべきであるのかが不明なことである。そのため、DCI において実際に収録されているデータセットがどのようなものかを理解するために、直近でデータ収録数が一番多い 2019 年において、多くのデータセットが収録されたレポジトリ上位 10 件を表 2 に示す（表 2)。

表 2　2019 年のデータセットが収録されているレポジトリの頻度上位 10 件の情報

	レポジトリ名	データ数	運営組織	開始年	分野	概要
1	Zenodo	550,688	欧州原子核研究機構(CERN)	2013	科学技術-その他/複合科学	研究者がだれでも自由に研究データを共有するためのオープンプラットフォーム
2	Figshare	224,254	ディジタル・サイエンス	2011	科学技術-その他/複合科学	研究者が図、データセット、画像、ビデオなどの研究成果を保存および共有できるオンラインのオープンアクセスリポジトリ
3	BacDive	162,953	Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures GmbH	2012	微生物学	細菌および古細菌の生物多様性に関する系統に関連した情報を提供する細菌メタデータベース
4	Barcode of Life Data System	102,910	生物多様性ゲノミクスセンター(カナダ)	2007	遺伝学、遺伝/生化学、分子生物学	DNA バーコード専用の Web プラットフォーム

[1] https://clarivate.com/ja/solutions/data-citation-index/https://clarivate.com/ja/solutions/data-citation-index/

5	European Nucleotide Archive.	68,550	欧州バイオインフォマティクス研究所	1980	遺伝学、遺伝	注釈付きの DNA および RNA 配列への無料かつ無制限のアクセスを提供するリポジトリ
6	Cambridge Structural Database	59,717	ケンブリッジ結晶学研究センター	1965	結晶学	幅広い有機、有機金属、有機金属分子を含む分子の 3 次元構造データのリポジトリ
7	U.S. Census Bureau TIGER/Line Shapefiles.	33,080	米国国勢調査局	2007	地理学	米国勢調査局の MAF/TIGER データベースから、道路、鉄道、河川などの地物や、法的・統計的な地理的エリアを抽出した空間ファイル
8	Mendeley Data	31,812	エルゼビア社	2015	科学技術-その他/複合科学	研究者のための研究データ公開リポジトリ
9	Data Archiving and Networked Services (DANS-KNAW)	20,103	オランダ王立芸術科学アカデミー（KNAW）/オランダ科学研究機構（NWO）の研究機関	2005	地球科学・総合/社会科学・学際的/地質学/人文学・総合/社会科学-その他/芸術・人文学-その他	研究者のための研究データ公開リポジトリ
10	MassBank	16,079	日本質量分析学会（MSSJ）	2006	分光学	化合物の同定や構造解明のための質量分析データを科学研究コミュニティで共有するために設計された、公開レポジトリ

レポジトリの頻度上位 10 件のうち、研究者向けの研究データ公開リポジトリが 4 件含まれている（1,2,8,9）。そこには、研究論文に付属する図表やデータが大量に収録されている。これは初期の 2014 年の先行分析(Nicolas ほか,2014)では見られなかった状況である。残りの 6 件は、ライフサイエンスやバイオ関係の個別の DNA 情報のデータや、米国の地理データ、化学構造データなどが万～数十万件単位で収録されているリポジトリである。この結果から、オープンデータといっても、少なくとも、研究者による研究論文に付随するデータ、特定の研究組織が運営するレポジトリにおける分野固有の大量なデータ、政府機関による調査データなど、異なる種類のデータが混在している現状にある。

さらに、DCI の引用の殆どが自己引用であるという指摘も上記先行研究で既にされているため、2015 年に引用されたデータセットのいくつかのサンプルについて、それを引用している WOS 論文を確認した。その結果、特に研究者個人によるデータ公開の場合には、当該論文に付随するデータや図表をリポジトリに格納したものも「引用」と扱われており、データの自己引用とみられるものが確認された（なお、自己引用か否かを系統的に確認できる情報は DCI には含まれていない）。収録されたデータセットのうちで一度でも引用された割合は 73.8%であるものの、自己引用が多くを占める可能性がある。ただし、たとえ自己引用が多くとも、そのことは論文に付随するデータが公開されている状況を反映していると考えれば、データ公開によりデータ駆動型研究の進展状況を反映する指標として見ることは可能である。

3 分析 1：国・分野ごとのデータ公開とその利用の現状

問 1 のデータ公開が盛んな国や分野を確認するために、「データセット」を対象として、主要 10 分野のデータベース数および時系列変化を集計した(図1,2)。図2に関しては 3年移動平均をとった。なお、分野はデータセットが収録されているレポジトリに対してつけられている。

DCI 収録のデータセットは、一番多い分野が遺伝学・遺伝 31.5%、その次に複合科学 23.3%、生化学・分子生物学 19.6%、結晶学 10.4%である。分野ごとの時系列的な変化を確認すると、どの分野も 2005 年以降から増加傾向であるが、2015 年から複合科学が急増し、それ以外の分野は減少傾向である。複合科学に分類されているものは表 2 の上位にも現れた、分野を問わずに研究者が論文付随のデータ等を公表できるレポジトリに公開されたデータである。つまり、2015 年ごろより研究者個人による、分野を特定しないレポジトリへのデータ公開が急速に増していることがうかがえ、それまで分野別のレポジトリに掲載されていたデータセットの一部もそれらのリポジトリに掲載されるようになってきている可能性がある。

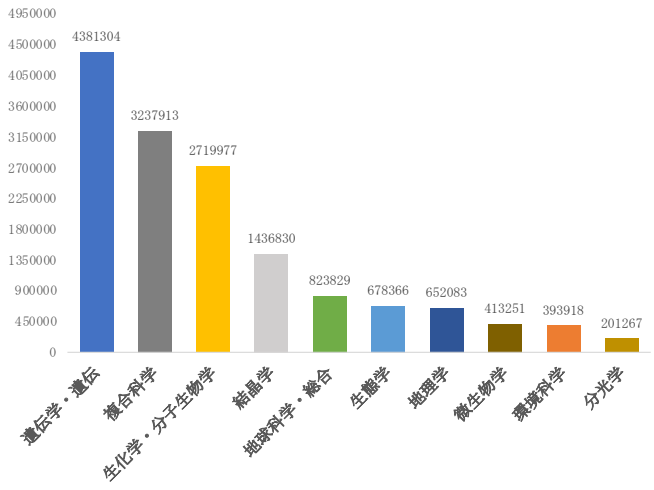


図 1　主要 10 分野のデータベース数

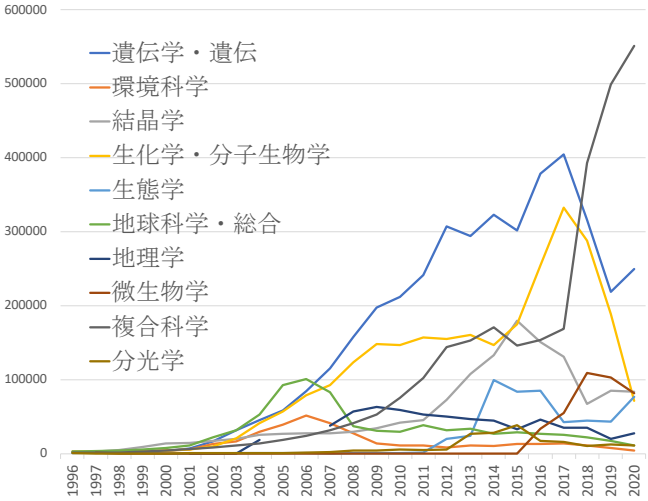


図 2 主要 10 分野のデータベース数の自系列変化

次に、主要国ごとに主要 10 分野のデータセット数を確認した(表 3)。なお、データセットの作成者の国が付与されているのはデータ全体で2,058,624件であり、全体の16.8%しか国名が付与されていない。そのため、各国の割合は全体的に大きくない。

表 3　主要国 5 ヵ国の主要 10 分野のデータセット数と占有率（%）

	アメリカ		中国		日本		ドイツ		ノルウェイ		他国/国なし	合計
遺伝学・遺伝	1,068,211	24	141,983	3	70,931	2	101,772	2	17,940	0	2,980,435	4,381,304
複合科学	38,943	1	17,330	0	5,675	0	12,714	0	87,065	3	3,076,184	3,237,913
生化学・分子生物学	927,307	34	48,520	2	61,967	2	81,463	3	8,653	0	1,595,026	2,722,977
結晶学	442	0	10	0	4	0	4	0	–	–	1,436,370	1,436,830
地球科学・総合	26,109	3	78	0	207	0	1,027	0	59	0	796,346	823,829
生態学	59,383	9	8	0	55	0	16	0	19	0	618,876	678,366
地理学	26,109	3	78	0	207	0	1,027	0	59	0	799,377	826,860
微生物学	–	–	–	–	–	–	–	–	–	–	413,251	413,251
環境科学	64,168	6	28	0	82	0	125	0	48	0	1,004,335	1,068,792
分光学	7,416	4	310	0	33,776	17	9,782	5	4	0	149,954	201,267
合計	1,446,658		181,639		121,204		140,018		110,637		12,870,153	

アメリカは生化学・分子生物学や遺伝学・遺伝分野において 2〜3 割程度のシェアを有する。中国は、上位 3 分野以外ではほとんどデータ公開が行われていない。分光学については、日本が上位 17%のデータセットを公開している。表 2 においても分光学分野リポジトリが日本質量分析学会（MSSJ）によって運営されていることが示されており、日本においてオープンデータが進んできた分野の一つであると言える。また、ノルウェーに関して、複合科学分野のデータが特出して多いが、これは「DataverseNO」というナショナルレポジトリを 2017 年より運用しており、そこに多数のデータが収録されているためである。主要分野を見てみると、政府によるデータ公開が進んでいるアメリカが多くのシェアを有するものの、その他の国もデータセットの公開の割合に違いがあり、国によって歴史的にデータ公開に力を入れてきた分野に違いがある可能性がある。ただし、上述のように DCI において国の付与が全体の 2 割ほどしかないため、実際の現状をつかみ切れない限界がある。

4 分析 2：分野ごとのデータ利用の現状

データ公開によるデータ駆動型研究が活発に行われている分野を、量的および質的な観点から分析する。分析では第一に、各分野における論文において、何らかのデータ引用がある論文の割合を、直近で最もデータセット搭載数の多い 2019 年を対象に算出する。対象としたのは 9 分野であり、DCI に登録されているデータセットが全年合計して WOS の論文に 2500 回以上引用されている分野である。分析の第二として、論文の分野ごとに、論文が引用している DCI データセット分野の多様度を計測する。これは、論文がどれほど多用な分野のデータを活用しているかを示すものであり、オープンサイエンスの一つの

特徴として、公開されたデータが他分野でも活用されるなどして学際的な研究を推進する触媒的機能を果たすことが期待されるためである(沼尻ほか,2021)。ここでは、多様度の指標として、経済計量学において市場の寡占度を測るのに用いられるハーフィンダール指標（HHI）を使用する。HHI は分母（対象論文数）が大きければ大きいほど多様度が高い結果になるという、規模の影響を受けやすい。そのため、下記に述べるようにランダムサンプリングを行い、HHI を 1,000 回計算し、その信頼度区間の誤差が一番小さいサンプル数を採用する。

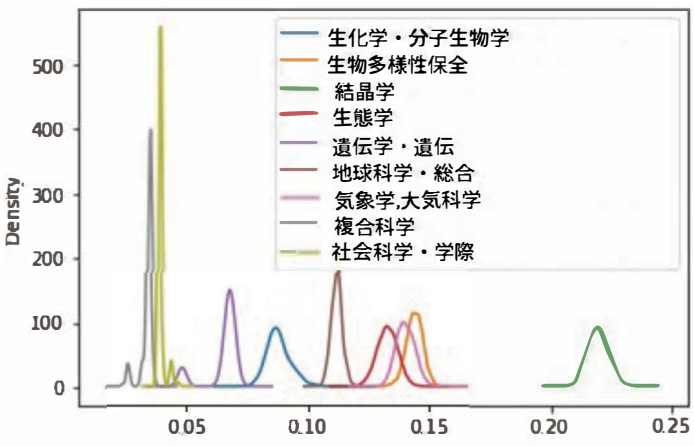
まず、各分野におけるデータ引用の割合（DCI 上で一つでもデータを引用している論文の割合）を、論文全体と、被引用数上位 10%論文の場合とで、表 4 に示す。ただし上述のように、ここには「自己引用」が含まれるため、純粋な引用よりは、データ公開の傾向を含めた、データ駆動型研究の進展度として解釈すべき指標である。

表 4 各分野におけるデータ引用がある論文の割合

	論文全体			top10%論文のみ			
	何らかのデータ引用がある論文	全論文	割合	何らかのデータ引用がある論文	全論文	割合	
生化学・分子生物学	1,801	56,419	3.19%	319	6,395	4.99%	**p<0.01
生物多様性保全	399	6,399	6.24%	57	679	8.39%	*p <0.05
遺伝学・遺伝	602	21,281	2.83%	129	1,999	6.45%	**p<0.01
結晶学	1,960	6,274	31.24%	196	634	30.91%	
地球科学・総合	592	26,965	2.20%	104	3,003	3.46%	**p<0.01
気象学、大気科学	441	14,536	3.03%	71	1,480	4.80%	**p<0.01
複合科学	1,612	61,635	2.62%	474	6,068	7.81%	**p<0.01
社会科学・学際	59	6,684	0.88%	6	789	0.76%	
生態学	1,635	20,369	8.03%	248	1,981	12.52%	**p<0.01

結果、結晶学は全論文と被引用数上位 10%論文双方で、データを引用している割合が 30%以上あり、その他の分野に比べるとデータ駆動型研究が進んでいると考えられる。次には生態学や生物多様性が 6～8%と高く、ライフサイエンスや地球科学関係が 2～3%という状況である。また、9 分野中 6 分野において論文全体でのデータ引用割合にくらべ被引用数上位 10%論文の割合が統計的に有意に上回る傾向が示された。引用数が高く注目されるような論文はデータ引用（自己引用を含む）を行っている傾向があることが示された。

次に、WOS 論文 9 分野が引用するデータベースの HHI の分布を図 3 に示す。1000 から 25000 までの範囲でデータセットランダムサンプリングを行った結果、サンプル数が 1000 程度だと全体に対するばらつきの誤差が大きくなってしまうため、その誤差がより小さい 2500 サンプル数で 1000 回計算を行った。



全体の結果をみると、各分野によって引用されるデータベースの分野の使われ方に違いがあることがわかる。結晶学は 0.20～0.25 と他分野と比べると集中度が高く、表 4 の結果と合わせると、2019 年の論文分野においてデータセットの引用率は高いものの、引用されるデータセットの分野は集中していると言える。その次に生態学や生物多様性が続き、これらも特定分野のデータを活用している傾向がみられる。一方で、多様度が高いのは複合科学と社会科学・学際分野であり、論文の分野自体が学際的であるために利用しているデータも多様である。ただし、本分析で用いた HHI では、分野間の潜在的な距離の考慮はできていないため、更なる多様性の測定が必要である。

5. 議論

本研究では、データ駆動型研究の推進の現状を分野横断的に把握するために、DCI のデータをさまざまな観点から分析した。 その結果、オープンデータが盛んな国や分野は異なる傾向にあることが明らかになった。国を挙げてオープンデータを推進しているアメリカは複数の分野においてデータセット公開数が多い傾向にある。日本は、特に分光学分野データ公開が進んでいる。その一方、近年、分野を問わずに研究者がデータを搭載できるレポジトリ（複合科学分野に区分）のデータセットが急激に増加し、研究者のデータ公開が進んでいることが分かった。

論文によるデータ引用率の観点から、とくに結晶学分野においてデータ引用が盛んであるが、引用されるデータセットの分野は集中化している。その一方で、データ引用率は高くないものの、使われるデータベースは多様であるといった分野も存在する。今後は、データベース分野の多様性の評価をする際に、分野間の距離を測る指標を採用し、国ごとや分野ごとの分析を行うなど、より実態にせまった解析を行うことが必要である。

また、今回の分析から DCI のデータベースを分析する際の制約も明らかになった。これは DCI の問題というよりは、データを引用した際の論文等への記載方法や、レポジトリにおける引用情報の整備が統一されていないなど、データ引用の慣習の未確立が、その一因となっていると考えられる(Sivello 2018)。また、論文とその付随する公開データの関係についても、これを「自己引用」と考えて引用分析から削除する発想もあるが、一方で、データのほうを主体にみれば、公開されたデータを用いている論文が、著者が誰であれ存在しているという見方をすることもでき、その場合には、データの利用価値という観点からは自己引用を削除する必要はないかもしれない。データ引用の考え方自体をさらに検討することで、指標も再度検討していくことが必要である。オープンデータが研究活動へ与える影響は、論文数のみではなく研究内容にまで影響を与えうるものである （沼尻ほか, 2021)。オープンデータが研究活動にどのような影響を与えるかを様々な角度から検証していくことが今後の課題である。

参考文献

・Bettina Suhr, Johanna Dungl, Alexander Stocker.(2020). Search, reuse and sharing of research data in materials science and engineering—A qualitative interview study. PLOS ONE
・Gianmaria Silvello.(2018). Theory and practice of data citation. Journal of the Association for Information Science. and Technology
・H Park, D Wolfram.(2019). Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse . Journal of Informetrics. Elsevier
・Nicolas Robinson-Garcia, Evaristo Jiménez-Contreras, Daniel Torres-Salinas.(2015). Analyzing data citation practices using the Data Citation Index. Journal of the Association for Information Science and Technology
・OECD (2015), Data-Driven Innovation: Big Data for Growth and Well-Being, Paris: OECD
・RIN/NESTA. (2010). *Open to all? Case studies of openness in research*. URL http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies
・カレントアウェアネス・ポータル.(2018).「FORCE11、研究データの引用のための新しい基準を公表」https://current.ndl.go.jp/node/35966 (2020.9.14 参照)
・沼尻保奈美, 林隆之(2021). 「オープンサイエンスが研究活動へ与える影響：ナショナルフォレストインベントリの事例」研究イノベーション学会発表予稿