STI Horizon 2024 Vol.10 No.2 https://doi.org/10.15108/stih.00367

ほらいずん

研究データのオープンアクセスに関する利用実態把握の試み

データ解析政策研究室 主任研究官 小柴 等

【概　要】

本稿では研究活動におけるオープンデータ・ソースの利用状況の把握を目的として、物理・情報系分野におけるメジャーなプレプリントサーバ（出版前の論文草稿共有サービス）である arXiv を対象に、オープンデータ・オープンソースに言及したプレプリント（原稿）数を調査した。

結果、オープンデータ言及原稿数（データ共有サービスである Zenodo、figshare への言及がある原稿の数）はほとんど伸びていないことが分かった。他方で、オープンソース言及原稿数（ソースコードの共有サービスである github への言及がある原稿の数）は 2023 年ベースでは全投稿の 3 割近くに達し、緩やかながら順調な伸びを見せていることが分かった。

キーワード：オープンデータ，オープンソース，プレプリントサーバ，第 6 期基本計画，指標

1. 背景と目的

研究活動も社会活動の一部であり、その様相も常に変化している。例えば、数十年前までは論文の媒体は紙が主体であったが、現在では電子版が主流になっている。また、読者が購読費用を負担することなく閲覧できるオープンアクセス（Open Access：OA）の論文や、査読を経て出版される前の原稿（草稿）であるプレプリントを共有するプレプリントサーバの活用も広がっている。研究成果についても、長らく一般的かつ、独占的地位にあった論文[注1]に加えて、実験データそのものや、プログラムのコードそのものも成果として認める事例・分野も出てきている。

こうした（常に変化し続ける）研究活動の様相を正しく捉えておくことは、適切かつ・若しくは積極的な科学技術・イノベーション政策立案を行うためにも重要であり、特にオープンサイエンス、オープンデータ、オープンアクセスなどの「オープン」を軸とした新たな研究動向の把握にも政府内で注目が集まっている。

例えば、2021 年度から 2025 年度を期間とする「第 6 期科学技術・イノベーション基本計画」[注2]では、オープンサイエンスやデータ駆動型研究等、昨今の新たな研究方法についての言及がある。更に、「第 6 期科学技術・イノベーション基本計画ロジックチャートと指標（2021 年 3 月時点）」[注3]（以下、第 6 期指標と言う。）では「2020 年度に実施した試行的取組をベースとして、DX による研究活動の変化等に関する新たな分析手法・指標の開発を行い、2021 年度以降、その高度化とモニタリングを実施する。【文】」との記載がある。

ここで、新たな研究動向は当然、これまでになかっ

注 1　ここでは、査読を経て出版される、いわゆる原著論文・ジャーナル論文。
注 2　第 6 期科学技術・イノベーション基本計画本文「2 章 2.（2）新たな研究システムの構築（オープンサイエンスとデータ駆動型研究等の推進）」https://www8.cao.go.jp/cstp/kihonkeikaku/6honbun.pdf (2022.11.27 last accessed.)
注 3　第 6 期科学技術・イノベーション基本計画ロジックチャートと指標（2021 年 3 月時点）
https://www8.cao.go.jp/cstp/kihonkeikaku/6chart.pdf (2022.11.27 last accessed.)

（2024.5.27 web 先行公開、2024.6.25 公開）

たものを含むため、どのようにすれば測定できるのか、そもそも何をどのように測定すべきか、は必ずしも明らかでない。したがって、第 6 期指標でも言及があるとおり、指標そのものの開発が重要である。

そこで本稿では研究活動におけるオープンデータ・ソースの利用状況の把握を目的として、物理・情報系分野におけるメジャーなプレプリントサーバである arXiv[2] を対象に、オープンデータ・オープンソース言及を行っているプレプリント（原稿）数を調査した。

なお、本報の大部分は既に報告書[3] において公表済みの内容であるが、今回の執筆に当たって、テキスト抽出方法を見直して再試行することによりデータの精度を向上させた。また、集計方法や単位の見直しなどにより、既報の内容を更改した。また、2022 年及び 2023 年分のデータを新たに追加した[注 4]。

2. 方法・データ

2.1 前提

第 6 期指標で言及されている「DX による研究活動の変化等に関する新たな分析手法・指標」すなわち「オープンサイエンスやデータ駆動研究による変化に関連しそうな指標」について、単純には例えば、ある分野の論文を専門分野における職業研究者以外の者が購読した量やその多様性、分析に際して用いられているデータの量や計算量、などが考えられる。

一方で、各指標については課題として、1. それらの指標が実際に計測できるか、2. 計測できたとして、安定的かつ低コストに収集・分析できるか、といった観点も存在する。

これらから、今回はオープンサイエンスやデータ駆動研究による変化のうち、オープンデータ・オープンソースの利活用に着目し、それらの度合いがどの程度変化しているか、その中で、日本がどういったステータスにあるか、を計測する方法について検討した[注 5]。

2.2 分析対象

まず、現状においてオープンデータ・オープンソースを利用した際、その引用の仕方は論文の引用と比較して全く確立していない。また、引用論文については OpenCitations[注 6] を始めとしてメタデータの形で整備されている一方、オープンデータ・オープンソースについては分析に使えるようなメタデータの蓄積はない。したがって、原稿の本文を直接分析する必要がある。

しかしながら、原稿の本文まで含めて収集分析するとなると、そのデータ容量や処理コストもさることながら、取得のための購入コストなども膨大となる。そこでまずは試行として、分野等の偏りはあるものの、本文の取得が容易であり、かつ、オープンデータ・オープンソースと親和性も高いと考えられるプレプリントサーバ arXiv[注 7] のデータを用いることにした。この際、後述する理由から PDF のデータを収集し、独自にテキスト変換したものを分析の対象とした。分析範囲は 2010 年 1 月から 2023 年 12 月末までの 14 年分で、分析対象の原稿数は 1,810,865 件である。

オープンデータ・オープンソースの利活用については、共有サービスの URL を用いることで検出を試みた。具体的にはオープンデータについて Zenodo, figshare、オープンソースとして github、を取り上げて、本文内での言及（URL 記載）を調査した[注 8]。

2.3 データ処理に関する補足

プレプリントの原稿は論文と異なり、バージョンを重ねることが想定されている。したがって、単に最新版を使うと計測時点によってデータ数が変化する問題がある。これを回避するため、分析対象の原稿は初版に固定する。

分析に関しては本文 PDF からテキスト抽出した結果を用いる。この際、PDF のテキスト変換が適切に実施できず、処理対象に含められないケースがごくわずか（全体約 20 万件に対して数件程度）存在する。また、分析に用いる URL が途中で改行されるなどして処理できないケースも想定される。更に arXiv の書誌情報（メタデータ）は論文のようには完備されていない。例えば、著者の所属機関やその国籍のデータなどは取得することが難しい。原稿の日付も、受付、公開など複数あり、かつバージョンによっては取得が困難なものもあるため、どれを使うかでカウントが異なることがある。

注 4　本報のフル版に相当する更改版のレポートについては 2024 年度中に公開することを予定している。
注 5　オープンソースもオープンデータに包含されるが、ここではオープンソースをオープンデータと別に整理する。
注 6　https://opencitations.net/ (2024.03.14 last accessed.)
注 7　https://arxiv.org/ (2024.03.14 last accessed.)
注 8　Zenodo, figshare にソースコードを配置することは可能であるし、github にソースコード以外のデータを配置することも可能であって、実際にそうした利用も見受けられる。ただしここでは便宜上、それぞれに振り分けた。関連して各サービスにはプレプリントを含む論文も多数配置されており、より正確な計測のためにはそれら種別も見ていく必要がある。

詳細については既報[3] に譲り、以下では特に注意が必要なもののみを説明する。

■国・地域の割当て

国籍・地域は各原稿において最初に検出されたメールアドレス 1 件を用い、主にそのトップレベルドメインの国別コードによって割り付けを行う。メールアドレスが検出できない場合や、検出できても国・地域が不明な場合は一部を除き「不明 (Unknown)」として扱う。以上より、各原稿は Unknown を含めて必ず一つの国・地域（あるいは edu, org などの組織）に所属することになり、国・地域別に集計した場合でも、その合計数は原稿総数に一致する。なお、米国については通常、メールアドレスに国別コードを用いないため、本手法では検出・判別できない。

一般的に論文を対象とした計量書誌分析では、著者所属機関の住所を用いて国・地域を割り当てる。したがって、ひとつの論文に 0 以上の国・地域が紐付き、これらを整数カウントで処理することが多い。翻って本報の処理方法はこれらと異なるものであり、比較の際には注意を要する。

■日時の割当て

後述の通り、年や月の単位でも原稿の集計を行っている。ここでは簡単のためファイル名をベースとして各原稿の年・月を定める。まず、解析用の arXiv 原稿は、任意のバージョンの PDF 原稿が取得できる Kaggle 用のデータ[注 9] を用いる。このとき arXiv のアーカイブは年・月単位でまとめられており、2024 年 2 月の原稿は 2402 というディレクトリに格納され、ファイル名も 2024 から始まる。厳密にはメタデータの公開日時データを用いるべきだが、ここでは作業の効率化のため、このファイル名を用いて日時を設定する。

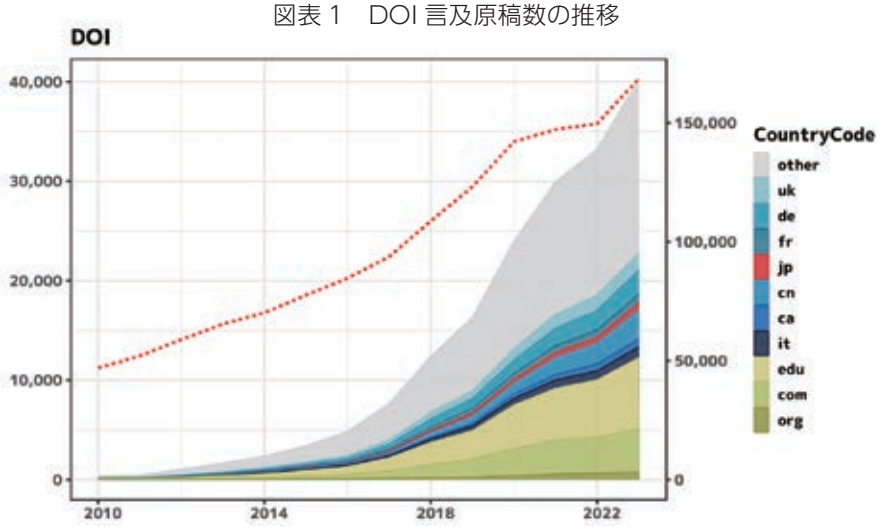
3. 結果

3.1 DOI 言及数

分析に先立ち、ベースラインを検討するために DOI（Digital Object Identifier, デジタルオブジェクト識別子）の記載のある原稿数を調べ、図表 1 にまとめた。カウント対象は URL の FQDN（Fully Qualified Domain Name：完全修飾ドメイン名。http(s):// の直後から、直近のスラッシュ (/) の間まで。）に「doi.org」を含むもの（が 1 件以上ある原稿の数。以下同じ。）である。図表中、第 2 軸として赤色の点線[注 10] で arXiv に投稿された原稿の総数を示した。

DOI は 1997 年に発案され、当初は論文やその図表に付与することが想定されていた[注 11]。日本では 2012 年に国内 DOI 登録機関が認可されており、こうした状況を鑑みると本稿の分析期間である 2010 年前後から普及し始めたと言える。DOI は論文引用という既存の「研究における基礎的な作法」を支援するもので、使用コストも少ないため、その低い敷居のもとで新規のサービスがどの程度受け入れられ広

図表 1　DOI 言及原稿数の推移



注 9　https://www.kaggle.com/datasets/Cornell-University/arxiv (2024.03.14 last accessed.)
注 10 紙媒体ではグレースケールですが、Web 版では図表をカラーで御覧いただけます。
注 11 現在では、例えば Zenodo, figshare に配置された場合も DOI が付与されるなど、論文以外でも積極的に利用される。したがって、DOI を全て検出し、そのメタデータを用いて分析する方法も考えられるが、PDF から URL 全体を正しく抽出することは困難であるため、今回は採用していない。

まってきたかを考察する指標として、ある程度妥当と考えられる。

図表 1 をみると DOI は 2016 年頃から特に利用数が伸びており、2023 年分に限って見ると全体の約四分の一、24.0% の原稿で、少なくとも 1 件の DOI が検出されている。国に着目すると、日本 (jp) は例えば 2023 年時点において 1.7%（694 件）で、仏国 (fr) の 2.3%（908 件）、加国 (ca) の 2.0%（800 件）、伊国 (it) の 2.7%（1,087 件）と同程度になっている。

なお、arXiv は 2022 年 1 月から、arXiv に投稿された各プレプリントに対して（過去搭載分についても随時、遡及（さかのぼって、過去のことにまで影響を及ぼすこと）的に）DOI の付与を開始した。これは 2022 年からの DOI 言及の増加傾向がわずかに改善していることの背景として考えられる。

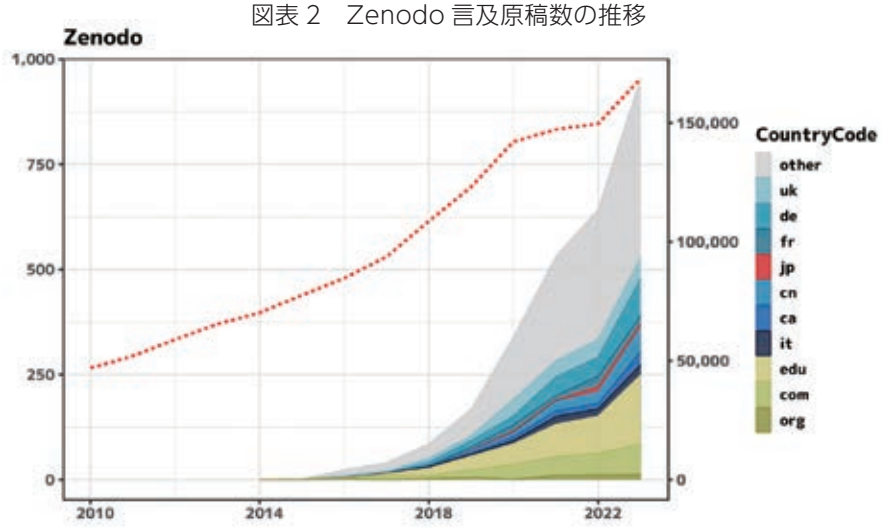
3.2　オープンデータ利用

既に述べたとおり、今回はオープンデータ利用の代理変数として、一般的なデータ共有サービスである Zenodo, figshare の URL 記載を用いた。それぞれ FQDN に「zenodo」「figshare」を含むものである。結果を図表 2、図表 3 に示す。

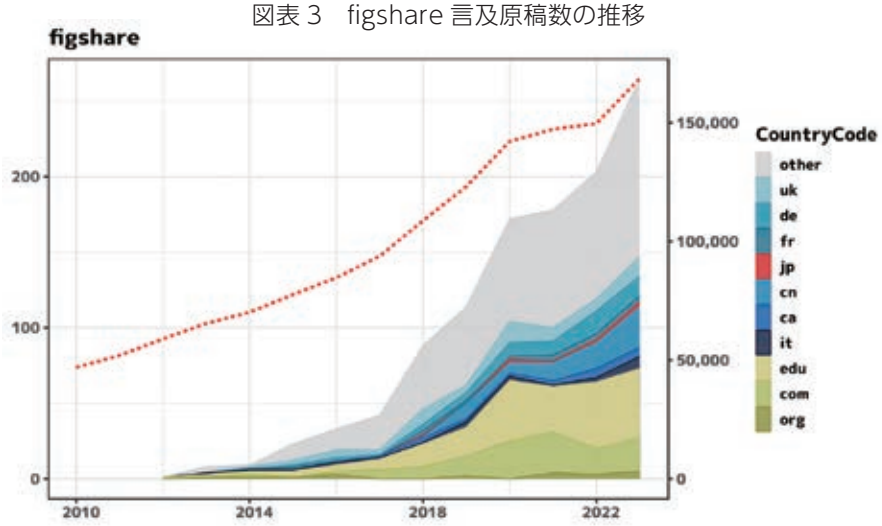
Zenodo は 2013 年、figshare は 2011 年にサービスを開始しているが、図表 2 及び図表 3 をみると、arXiv 上での登場は Zenodo が 2014～2015 年頃、figshare が 2012～2013 年頃であり、サービスのローンチから arXiv で観測されるまでの間に 2 年程度の遅れがあるように見える。また、サービス開始と arXiv 登場の両方において figshare の方が 2 年早いが、言及数の伸びは Zenodo の方が早く、2023 年では figshare が 250 件程度に対して Zenodo は 1,000 件程度と言及原稿数に数倍の差がある。

他方で、赤の点線で第 2 軸に示している arXiv 全体の月間投稿数に照らすと、figshare との比較で Zenodo が相対的に多いとはいえ微々たるものに過ぎず、（Zenodo，figshare を通じた）データ共有・活用はまだ普及したとは言い難い状況にある。国別で

図表 2　Zenodo 言及原稿数の推移



図表 3　figshare 言及原稿数の推移



は日本は仏国よりも少なく、取り上げた国の中では最少である。

3.3　オープンソース利用

既に述べたとおり、今回は研究データにオープンソースを含め、その代理変数として github の URL 記載を採用した。具体的には FQDN に「github」を含むものを調べた。結果を図表 4 に示す。

github は 2008 年にサービスを開始しているが、図表 4 を見る限り arXiv では 2012 年頃から登場し始め、2017, 8 年あたりから広まってきているように見える。2023 年では言及原稿数が 5 万件に迫り、DOI を越え、Zenodo や figshare を大きく引き離している。また、これらの全体傾向は同様の分析を行った先行研究[1] とも基本的に合致している。国別に見ると日本も仏国、加国、伊国と同程度の存在感を有している。

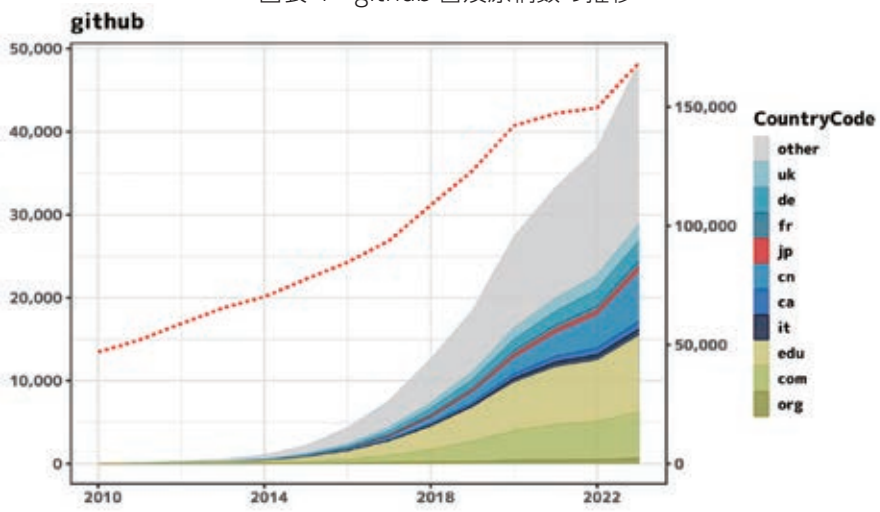
arXiv はプログラミングとの関係が深い情報系分野でも著名であり、特に AI 系の研究でもメジャーなプレプリントサーバである、という情報源の特性によるものとは考えられるが、DOI の言及数を上回る点は興味深い。

3.4　その他のデータ

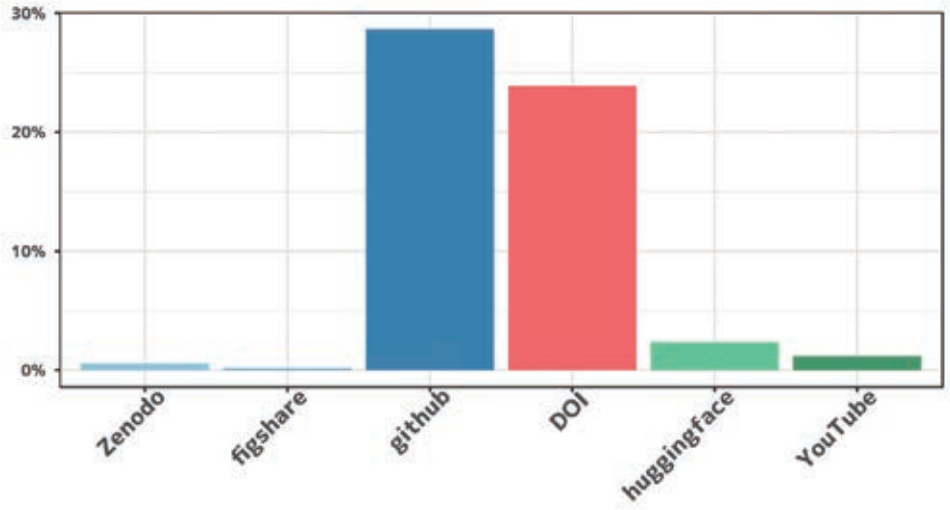
集計時点において、動画共有サービスである YouTube、大規模言語モデル（LLM）の共有サービスである huggingface の URL も多く見られたため、前述の、Zenodo, figshare などと併せて 2023 年分の言及原稿数を調べ、原稿数全体におけるシェアを図表 5 に示した。（検索条件は、YouTube が「youtube」あるいは「youtu.be」、huggingface は「huggingface」を含むものである。）

図表 5 をみると、YouTube，huggingface の言及原稿数は Zenodo などを上回っていることがわかる。huggingface は近年の LLM に関する盛り上がりと、arXiv が情報系分野と近いことを考えれば納得できる。YouTube の言及原稿が多いことも興味深いが、これらは例えば、実際のシステムのデモ動画など

図表 4　github 言及原稿数の推移



図表 5　各種サービスの言及原稿シェア（2023 年分）



を示している可能性が高い。

4. まとめ

arXiv の原稿を対象として、URL をベースにオープンデータ・オープンソースの利活用についての捕捉を試みた。特に「データ」の共有については、数自体は順調に増えているものの、少なくとも今回用いた方法の範囲では広まっていないことが示唆された。ただし、データ共有については、まだデファクト・スタンダード（業界の中で広く受け入れられ利用されるような実質的な業界標準・基準）となるサービスや利用方法、引用方法などが確立しているとは言い難い。したがって、この方法では十分に計測ができなかった可能性もある（なお、2024 年 3 月 14 日時点で Zenodo に登録のあるオープンなデータの数は 3,370,666 件あり、うち図表（images）は 863,846 件、データセット（dataset）は 301,467 件であった。figshare には 8,013,678 件のアイテムがあり、うち図表（figure）は 2,072,475 件、データセット（dataset）は 1,900,825 件であった。したがって、各サービスへの登録件数自体は少なくない）。

「オープンソース」については、arXiv の特性ともあいまって、DOI 以上に言及する論文が多く、arXiv の分野においては一般的になってきていること、更に、経年で数が伸びていることも確認できた。国別で見た場合に日本は仏国や加国、伊国などと比較的低い位置で同程度の状況を示している。他方で、英国（uk）や独逸（de）、中国（cn）は一定の存在感を示しており、この割合の違いについて別途の考察を行う余地がある。

その他、ある種のオープンデータである LLM の共有については一定数観測できること、動画についても一定数言及があること、も確認できた。オープンソースやモデル、動画の活用は特に、研究成果の生成・公開の変化を裏付けるものといえる。つまり、第 6 期指標にある「DX による研究活動の変化等に関する」指標として、今回採用した手法がある程度機能することが示唆された。他方で、より広く「データ」の観点で見た場合には十分に捕捉しづらい状況で、今後も計測方法を検討する必要がある。例えば LLM を用いることで、途中に改行が含まれていたり、空白が含まれていたりするなどして、現状では的確にテキスト変換ができていない URL を修復できる可能性がある。これが機能すれば URL 中に単に DOI や Zenodo を含むものがあるかどうかだけでなく、URL 全文を用いてメタデータを取得できるようになり、ここから論文や図、数値セットなどデータ種別を取得してより適切な分類ができる可能性がある。またジャーナル論文を対象とし、商用の論文データベースを用いる場合、データが JATS XML 形式で整備されているものであれば URL の抽出は更に容易であるため精度の課題はなくなる。一方で対象論文数が爆発するため、分散並列処理を行うなど別の課題に対処する必要がある。

また、そもそもの計測方法の課題に加え、今回対象とした arXiv の分野特性もあることから、より広範な分野を対象とした計測方法についても検討する必要がある。

参考文献・資料

1) Emily Escamilla, et.al. The Rise of GitHub in Scholarly Publications, arXiv, 2022. (preprint) DOI: https://doi.org/10.48550/arXiv.2208.04895
2) 林 和弘, 他. arXiv に着目したプレプリントの分析, Discussion Paper, No.187, 文部科学省科学技術・学術政策研究所, 2020. DOI: https://doi.org/10.15108/dp187
3) 林 和弘, 他. 研究活動におけるオープンソース・データの利用に関する簡易調査, 調査資料 (Research Material), No.324, 文部科学省科学技術・学術政策研究所, 2022. DOI: https://doi.org/10.15108/rm324