調査資料-324

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

2023 年 1 月

文部科学省　科学技術・学術政策研究所

データ解析政策研究室

林　和弘　　小柴　等

【調査研究体制】
林　　和弘　　文部科学省科学技術・学術政策研究所
データ解析政策研究室・室長

小柴　　等　　文部科学省科学技術・学術政策研究所
データ解析政策研究室・主任研究官

【Authors】
HAYASHI Kazuhiro　Director, Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

KOHISBA Hitoshi　Senior Research Fellow, Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this NISTEP RESEARCH MATERIAL.

林和弘，小柴等，“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”，NISTEP RESEARCH MATERIAL，No.324，文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/rm324

HAYASHI Kazuhiro, KOSHIBA Hitoshi, “Brief survey on the use of open source / data in research activities,” NISTEP RESEARCH MATERIAL, No.324, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/rm324

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

文部科学省 科学技術・学術政策研究所 データ解析政策研究室

要旨

本稿では、研究活動における新たな分析手法・指標の開発を念頭に，研究活動にけるオープンソース・データの利用状況の調査を目的として，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 を対象に，プレプリント（原稿）中のオープンソース・オープンデータ言及回数を調査した。

ここでは，オープンソースとして github，オープンデータに Zenodo, figshare を取り上げて調査した。また，比較のための基礎データとして DOI も取り上げて調査した。本文中に記載されたメールアドレスを手がかりとして，各原稿には（割り当て可能なものについては）国籍を割り付け，初版発行の年月ベースで整理した。

結果，Zenodo, figshare ベースでみると arXiv 上でのオープンデータ言及はほとんど進んでいないことが分かった。他方で，github ベースでみた arXiv 上でのオープンソース言及は緩やかながら順調な伸びを見せており，2022 年以降全体投稿の 2 割以上の原稿において言及されていることが分かった。

Brief survey on the use of open source / data in research activities

Research-Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

The purpose of this paper is to investigate the use of open source data in research activities, with a view to developing new analytical methods and indices for research activities. We surveyed the number of open source and open data mentions in preprints (manuscripts) on arXiv, a major preprint server in the field of physics and information systems.

In this study, github was selected as open source, and Zenodo and figshare were selected as open data. DOI was also included as basic data for comparison.

We assigned nationalities to each manuscript (if assignable) using the e-mail addresses listed in the text as clues, and organized them on the basis of the date of first publication. The results were organized on the basis of the year and month of the first publication.

As a result, it was found that there has been little progress in mentioning open data on arXiv in terms of Zenodo and figshare. On the other hand, the github-based open source mentions on arXiv have shown a slow but steady growth, and more than 20% of all manuscripts have been mentioned since 2022.

目次

図目次

表目次

1　はじめに

研究活動に関する指標には様々なものがあり，研究成果については，現在は原著論文と被引用数を中心として用いる影響度の分析と指標が主に用いられている。当研究所においても、サイエンスマップ[1] や論文ベンチマーク[2] などの活動を通じて、世界と日本、あるいは国内の大学の研究活動の動向を分析している。その一方で，情報通信技術 (ICT) や人工知能 (AI) の進展に伴うオープンサイエンスの潮流を踏まえ，研究データを軸とする新たな研究成果共有・公開様式にも注目が集まり、また、その様式をに基づく新たな指標の可能性が議論されている。

本稿では研究活動における新たな分析手法・指標の開発を念頭に，オープンソース・データの利用状況の調査を目的として，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 を対象に，プレプリント（原稿）中のオープンソース・オープンデータ言及回数を調査した結果について述べる。

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」[3] 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「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）」で言及されている。

関連して，「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）」[4] 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】」との記述がある。

「ＤＸによる研究活動の変化等」が意味するところは必ずしも明らかではないが，オープンサイエンスやデータ駆動研究によって，具体的にどこが，どのように，どの程度変化しているかを計測しようとするものと言える。また，そのために，そもそも何を測ればオープンサイエンスやデータ駆動研究による変化をとらえられるのかを検討しようとしているものとも言える。

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を単純に挙げるならば，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もの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，多様な指標を挙げることができる。一方で，独立した課題として，1. こ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。なども存在する。

これらの背景から，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，オープンソース・オープンデータの利活用に着目し，それらの度合いがどの程度変化しているか，その中で我が国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。

---

[1] サイエンスマップ https://www.nistep.go.jp/research/science-and-technology-indicators-and-scientometrics/sciencemap (2022.11.27 last accessed.)

[2] 論文ベンチマーク https://www.nistep.go.jp/research/science-and-technology-indicators-and-scientometrics/benchmark (2022.11.27 last accessed.)

[3] 第６期科学技術・イノベーション基本計画　本文 https://www8.cao.go.jp/cstp/kihonkeikaku/6honbun.pdf (2022.11.27 last accessed.)

[4] 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.)

結果として，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 を対象に，プレプリント（原稿）中のオープンソース・オープンデータ言及回数を調査し，実際に上記の問いに答えられそうであることを確認した。

本稿では，2 章で上述した背景や要件について再度確認する。3 章では，収集データの詳細と結果について述べる。4 章では，これらの結果についてまとめる。

なお，本文中で使用している図表の元となるデータについては別途，Zenodo を通じて公開している[5)]。

[5)] https://doi.org/10.5281/zenodo.7505229

2 要件

2.1 指標

すでに述べたとおり，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」[6] 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「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）」で言及されている。

ここでは，

- ビッグデータ等の多様なデータの収集や分析
- 計算機を活用したシミュレーションや *AI* を活用した研究
- 研究交流のリモート化や、研究設備・機器への遠隔からの接続、データ駆動型研究の拡大

などを「研究活動の DX（研究 DX）」の具体例として掲げた上で，さらに，

- 論文のオープンアクセス化
- 研究成果の迅速な公開の場の一つとしてのプレプリントの活用
- 研究データの公開・共有
- オープンサイエンス等の世界的な知の共有を目指した研究成果のオープン化

などの動きも加速していると指摘している。

たとえば，（自然科学系の）研究成果産出プロセスについて考えると，従来は，研究者が自ら実験等を行ってデータを収集し，論文としてとりまとめて出版する形態が主流であったと考えられるところ，近年では，論文のオープンアクセスが進展するとともに，論文化前にプレプリントをプレプリントサーバーに登録・公開するような動きも普及しつつあり，国立研究開発法人 科学技術振興機構 (JST) でもプレプリントサーバの運用を始めている[7] 。また今後，研究データの公開・共有が普及していくと，業績評価におけるデータ公開の価値が高まってデータ公開だけでも一定の評価が得られるようになり，さらにプロセスが変わっていく可能性がある。上記の指摘もこうした視点を示したものと考えられる。この（自然科学系の）研究成果産出プロセスについて図 1 に示した。

関連して，「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）」[8] 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】」との記述がある。

「ＤＸによる研究活動の変化等」が意味するところは必ずしも明らかではないが，前述の記載と合わせると，オープンサイエンスやデータ駆動研究によって，具体的にどこが，どのように，どの程度変

[6] 第６期科学技術・イノベーション基本計画　本文 https://www8.cao.go.jp/cstp/kihonkeikaku/6honbun.pdf (2022.11.27 last accessed.)

[7] https://jxiv.jst.go.jp/

[8] 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.)

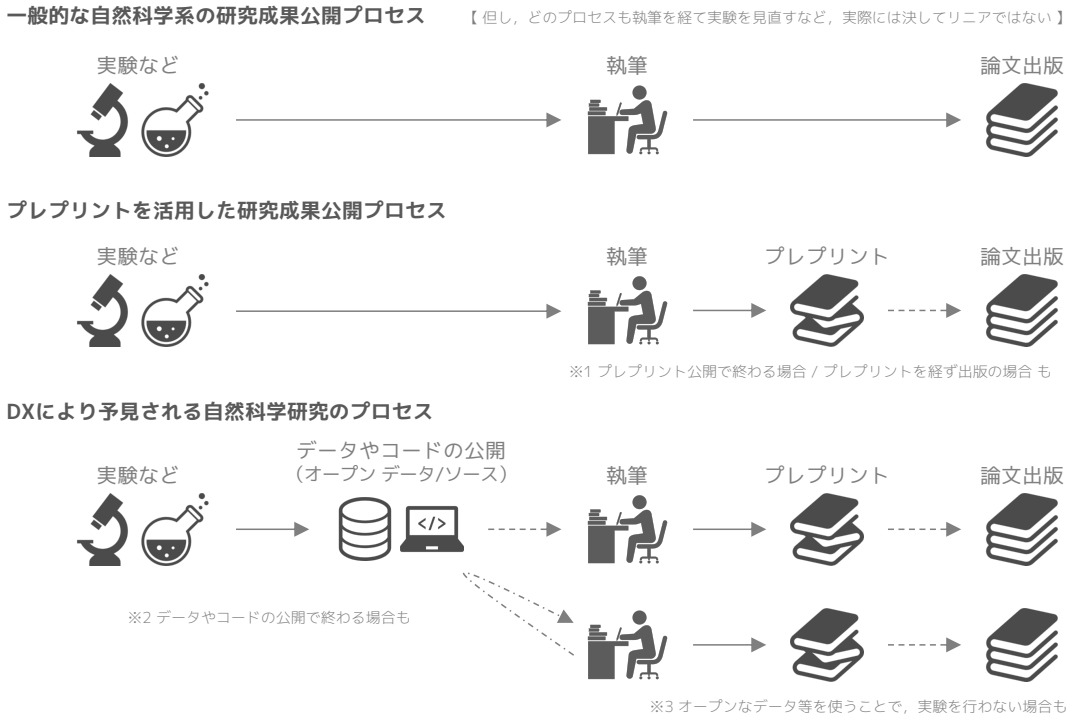

図 1: 研究活動（自然科学系の研究成果産出プロセス）の変化

化しているかを計測しようとするものと言える。また，そのために，そもそも何を測ればオープンサイエンスやデータ駆動研究による変化をとらえられるのかを検討しようとしているものとも言える。

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する要素は多岐にわたる。したがって指標にも様々なものが挙げられる。一方で，1. こ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。といった課題も存在する。

そこで，今回は「研究データの公開・共有」に着目し，かつ「オープンサイエンス」と「データ駆動型研究」の文脈から「研究データ」を広く捉え，「オープンソースソフトウェア (OSS: Open Source Software)」も含むものとして，“研究活動においてオープンな研究データがどの程度使われているのか” を計測し，我が国のステータスを明らかにすることを試みた。

この際，AI を含むデータ駆動型研究は相対的に時定数が短いと考えられること[9]，後述するように分析には論文の全文が必要なこと，などを考慮して，プレプリントサーバのデータを対象とすることにした。

ただし，おなじく後述の通り手法自体は特に特殊なことは行っておらず，単純に数を数えるに過ぎないので，Scopus や Web of Science など，いわゆる論文データベースにももちろん適用できる。

[9] 例えば，人工知能関連技術である深層学習関連では，囲碁で大きな成果を挙げた AlphaGo や，文書生成などの GPT-3，文章からの画像生成である Stable Diffusion など，（特に応用面や耳目を集めるというような観点で）分野内の勢力図を一夜にして一変するような新たな技術が次々と登場しており，1 年も経つと研究動向が変化している可能性が高い。

2.2 データソース：arXiv

すでに述べたとおり，本研究では “研究活動にオープンな研究データがどの程度使われているのか” を明らかにし，かつこれらを（特に科学技術イノベーション政策に活かすことを念頭に）継続的に計測していくことに興味がある。

この際，直近の状況を迅速に把握できるほど情報としての価値が高まる。そこで，今回はプレプリントサーバのデータを対象とすることにした。ここで，物理・情報系分野はその分野特徴からデータの公開・共有や活用と親和性が良いと考えられる。例えば，そもそもインターネット自体が黎明期において物理学の研究と関係が深く，物理・情報系の著名なプレプリントサーバである arXiv[林 20] は 1990 年頃から運営されてきている。人工知能関係の研究ではソースコードや機械学習のモデルを，github[10] や HuggingFace[11] などのプラットフォームを通じて（主に無償かつオープンに）公開，共有することも少なくない。こうした背景から，今回は物理・情報系の著名なプレプリントサーバである arXiv[12]を対象にデータを収集，分析することにした。

2.3 データ

arXiv のインストラクション[13] に従い，Kaggle [14]のプロジェクト[15] において，Google Cloud Storage (GCS) を通じて公開されている論文のバルクデータを取得して利用した。

このバルクデータは週次で更新されており，プレプリント各原稿の各バージョンに対応した PDF を取得できる。今回は 2022 年 11 月 8 日から 9 日にかけて，2008 年 1 月から収集時点における最新版までのデータを取得した。その上で，最終的に 2010 年 1 月から 2022 年 9 月いっぱいまでのデータを分析対象として設定した。

結果，対象期間中のデータ（原稿）件数は総数で 1,556,067 件，PDF のデータ量は約 3.7TB となった。

実際の解析はこの収集した PDF からテキストデータを抽出して行う。この際，PDF は基本的にレイアウトのための書式で，意味のデータを保持しないという点に注意を要する。例えば，「今日は良い天気です。」という文章があり，「今日は良い天」までの時点で紙面の右端に達して行折り返しが行われたとする。この場合，PDF の中ではレイアウト通り「今日は良い天（改行）気です。」のようにデータが保持され，改行後の文字が改行前からの続きか否かのデータは保持していない。したがって，PDF から抽出したテキストデータには改行に意味的な区切りと紙面上での物理的な区切りとの 2 種が混在し，かつ，各改行がどちらかを判別することは内容的にも作業量的にも困難なタスクである。さらに，ヘッダーやフッター（ページ番号など）の情報も挟まってくるし，図や表のデータはテキストを全く抽出できなかったり，正確に抽出できない場合が多い。この他，ごく少数，そもそもテキストを全く抽出できない場合も存在する。以上より，全数の完全かつ精密な調査にはならない。

[10] https://github.com/
[11] https://huggingface.co/
[12] https://arxiv.org/
[13] https://arxiv.org/help/bulk_data
[14] https://www.kaggle.com/
[15] https://www.kaggle.com/datasets/Cornell-University/arxiv

■対象とする原稿　原稿は常に第 1 版の内容を採用することにした。

プレプリントはその特性上，次々と更新し版を重ねることができる。第 1 版を登録した後，研究の進展や寄せられたコメントを元に改訂し，更新した版を登録することは珍しくない。

他方でこれはどの版を分析に用いるかという課題も生じさせる。最新版を用いると，内容的には推敲が重ねられ，より正確な記述が行われていると期待できる。しかしながら，今後新たに更新版が寄せられる可能性があることが問題を生じる。例えば，ある原稿 A について，2021 年 1 月に第 1 版が投稿され，2022 年 3 月に第 2 版，2022 年 10 月に第 3 版が登録されたとする。2022 年 5 月に分析する場合，2022 年 3 月版（第 2 版）が最新なのでこれを使う。その後，同様の分析を 2022 年 11 月に行う場合，2022 年 10 月版（第 3 版）を使う。このとき 2022 年 5 月の分析では 2022 年 3 月に計上されていた原稿が，2022 年 11 月の分析では 2022 年 10 月に計上されることになる。結果，2022 年 5 月と 2022 年 11 月で，2022 年 3 月分の論文の数が変わってくることになる。

総数としては変化がないものの，分析の時点によって過去分の原稿数が変化することは分析の一貫性，後方互換性の面から好ましくない。

更新版の有無にかかわらず，第 1 版を使うとこの問題は生じない他，定期的に分析を行う際に，以前の分析時点からの差分，新規投稿分のみを分析すれば良く手軽である。一方で，更新・改良された内容ではない可能性もある。

今回は定期的なモニタリングへの活用，後方互換性などを考慮して，原稿は常に第 1 版の内容を採用することにした。

■国籍の割り付け　国籍は先行研究を参考に，原稿 PDF からテキスト抽出した後，テキスト中に出てくる最初のメールアドレスをベースに，そのトップレベルドメインに基づいて割り付けることにした。したがって，研究者の国籍ではなく，あくまで所属機関の国籍であって，かつ，著者全員ではなくテキスト解析上最初に検出されたメールアドレス 1 件のみである点に注意が必要である。さらにこの際，「`XXX.com`」のようなものについて whois 情報から国籍を割り付けることはしない。

メールアドレスはアットマーク (@) をベースに検出するため，「`XXX[.at.]XXX.XX.XX`」のような形式の場合は検出できない。

米国は国を示すトップレベルドメインを用いないので，基本的には検出されない。

トップレベルドメインにおける国名 (ccTLD) は基本的には ISO 3166 に準じるが，英国は .uk を用いるなど例外がある。また，本報告書では簡単ため「国籍」としているが，ISO 3166 は「.hk（香港）」など地域も含む。

ccTLD 以外のものは原則として国籍不明 (NULL) とするが，「`.com`」「`.edu`」「`.org`」の 3 種類は参考情報として残す。

■URL の抽出　原稿 PDF からテキスト抽出した後，「`http`」で始まる一連の文字列を検出し，URL として採用する。

PDF からテキストを抽出しているので，URL の途中で改行が生じるケースも想定され，厳密には別途手当が必要になるが，今回はそれらは誤差として切り捨て，手当てしない。結果「`https://wwww`」などで終わるケースも一定数観察されている。

表 1 に，URL における FQDN (Fully Qualified Domain Name) [16] を原稿単位[17]で調べた分布を示す。5 位に「www」，12 位に「doi」などが入っており，これらは途中で途切れてしまったパターンと推定される。

表 1: FQDN ごとのカウント数（原稿単位）

#	FQDN	cnt	#	FQDN	cnt	#	FQDN	cnt
1	github.com	124,078	11	doi.acm.org	7,325	21	arxiv	4,882
2	doi.org	101,664	12	doi	6,780	22	www.kaggle.com	4,682
3	arXiv.org	100,083	13	github	6,598	23	archive.ics.uci.edu	4,469
4	dx.doi.org	42,557	14	sites.google.com	6,343	24	www.cosmos.esa.int	4,380
5	www	19,904	15	openreview.net	6,171	25	www.sdss.org	4,373
6	www.sciencedirect.com	17,099	16	www.jstor.org	5,472	26	link.springer.com	4,326
7	link.aps.org	13,113	17	creativecommons.org	5,158	27	ieeexplore.ieee.org	4,276
8	en.wikipedia.org	12,121	18	stacks.iop.org	5,124	28	www.youtube.com	4,203
9	pos.sissa.it	10,840	19	proceedings.mlr.press	4,973	29	youtu.be	4,153
10	dl.acm.org	7,441	20	onlinelibrary.wiley.com	4,909	30	papers.nips.cc	4,128

FQDN: Fully Qualified Domain Name　　* 2010.01〜2022.09まで。件数は原稿数（何件の原稿に出現したか）

■対象と，そのカウント方法　今回は「研究データの公開・共有」に着目し，かつ「オープンサイエンス」と「データ駆動型研究」の文脈から「研究データ」を広く捉え，「オープンソースソフトウェア (OSS: Open Source Software)」も含むものとして，“研究活動においてオープンな研究データがどの程度使われているのか” を計測し，我が国のステータスを明らかにすることを目的とした。

そこでカウント対象に, いわゆるオープンデータとして, Zenodo[18], figshare[19], OSS として github, を設定する。なお，上記は大まかな区分けであって，Zenodo, figshare がデータのみ，github がソースコードのみを共有するものというわけではない。Zenodo, figshare, github ともに，ソースコードを含む各種のデータを共有することもでき，例えば，Zenodo でソースコードを共有している例も，github で自治体一覧などのデータを共有している例もある[20]。

また, オープンデータ・OSS とは異なるが参考のため DOI(Digital Object Identifire) もカウントする。

github や Zenodo, figshare, DOI など，今回，カウント対象の抽出方法は基本的に単純な文字列マッチで行う。具体的には以下の通りである。

Zenodo	URL 中に「zenodo」の文字列を含むもの
figshare	URL 中に「10.6084/m9.figshare」の文字列を含むもの
github	FQDN 中に「github」の文字列を含むもの
DOI	FQDN 中に「doi.org」の文字列を含むもの

Zenodo は引用の際に DOI を用いるため FQDN ベースでは検出できないが，DOI が「10.5281/zenodo.XXXXXXX」の形式をとるため，上記の条件で抽出できる。

16) FQDN の抽出条件は “://” から直近の “/” もしくはスペースや改行など URL に使えない文字までとした。

17) 例えば，www.nistep.go.jp という FQDN が出現する原稿が何件あるか。「任意の FQDN が何回出現するか」**ではない**。

18) https://zenodo.org/

19) https://figshare.com/

20) さらに，非公開でデータをアップロードすることもできるため，搭載されている全てのデータがオープンというわけではないが，ここでは論文（プレプリント）の引用を対象とするため，ここで登場するものはオープンと捉えて差し支えないと考えている。

figshare は 搭載コンテンツに対して DOI を割り振るが，他にも機関レポジトリ側の URL，DOI などを利用することもできるため複雑である。ここでは figshare の付与する DOI が「10.6084/m9.figshare」で始まることを利用して，ひとまずこれに該当するもののみをカウントすることにした。

これらの説明から分かるとおり, DOI のカウントは Zenodo, figshare を含む。また, Zenodo, figshare は基本的にデータ等の公開に用いられるが，プレプリントやジャーナルも搭載でき，実際にサービス本体のデータを見ると，小数ながらそうした事例も観察できる。

ここでは，精緻な分析よりも迅速さを重視し，それらの点は特にケアせずに作業する。精緻な分析を行う場合は，DOI のメタデータなどを参照してコンテンツの種類を特定することが望ましい。

2.4 関連研究

関連研究としては，学術成果における OSS の普及度合いの調査を目的として，arXiv および PMC を対象に，github および関連サービス (GHP: Git Hosting Platforms) [21)]の URL 登場数を調べたもの [Escamilla22] が挙げられる。

手法等はほぼ類似だが，arXiv については第 1 版ではなく分析時の最新版を利用している点，2007 年 4 月から 2021 年 12 月までを対象範囲としている点，github のみならず GHP を対象としておりカバー範囲がやや広い点，国別の分析は行っていない点，などに本報告書との差異がある。

[21)] gitlab, bitbucket, SourceForge

3　結果

3.1　ベースライン（投稿数・DOI 記載数）

まず，分析に際して arXiv 自体のそもそもの傾向を確認するために，国別・全体の投稿件数推移や，DOI の記載がある原稿の国別・全体の投稿件数推移を確認した。結果を図 2 から図 5 に示す。なお，国については URL の記載のある原稿を国別に数え，上位のもの 7 件を採用した。

図 2 及び図 4 をみると，投稿件数及びその中における DOI 言及原稿数はおおむね着実に伸びている。ただし，投稿件数については 2020 年ごろからやや横ばい傾向とも言える。トリビアルには季節変動とみられる短期・周期的な増減傾向がある点は面白い。DOI については，2017 年頃から普及しはじめ，2022 年あたりでは月間約 3 千件の原稿に記載が見られる。2022 年辺りの月間投稿数が約 1.5 万件であるとすると，概ね 1/5 = 20% の原稿には DOI の記載があることになる。これは arXiv およびその周辺では引用対象が Web 上にあってかつ DOI が付与されている割合がおそらく高く，かつ/若しくは，引用先が DOI を持つ場合については原稿中で DOI を記載するという作法も広く普及していると考えられる。

図 3，図 5 は国別の比較を意識して，全原稿および DOI 言及原稿の各カテゴリにおける国別割合を示した。図 3 をみると朱色で示した日本をはじめ，英独伊など，多くの国で割合は安定しているように見える。一方で，中国や edu（一般にアメリカの大学等である割合が多い），com は徐々に割合を増やしている。また，フランスは長期的に見ると割合が徐々に減じてきている。図 5 については，2017 年頃までは数が少ないため安定しないが，2017 年以降を見ると図 3 と概ね似た傾向が見える。ただし，日本を示す朱色が全体傾向（図 3）ほどには目立っていない。図 3 の投稿件数に占める割合からの比較で考えると，日本の原稿における DOI 記載率は低い位置にとどまると言える。

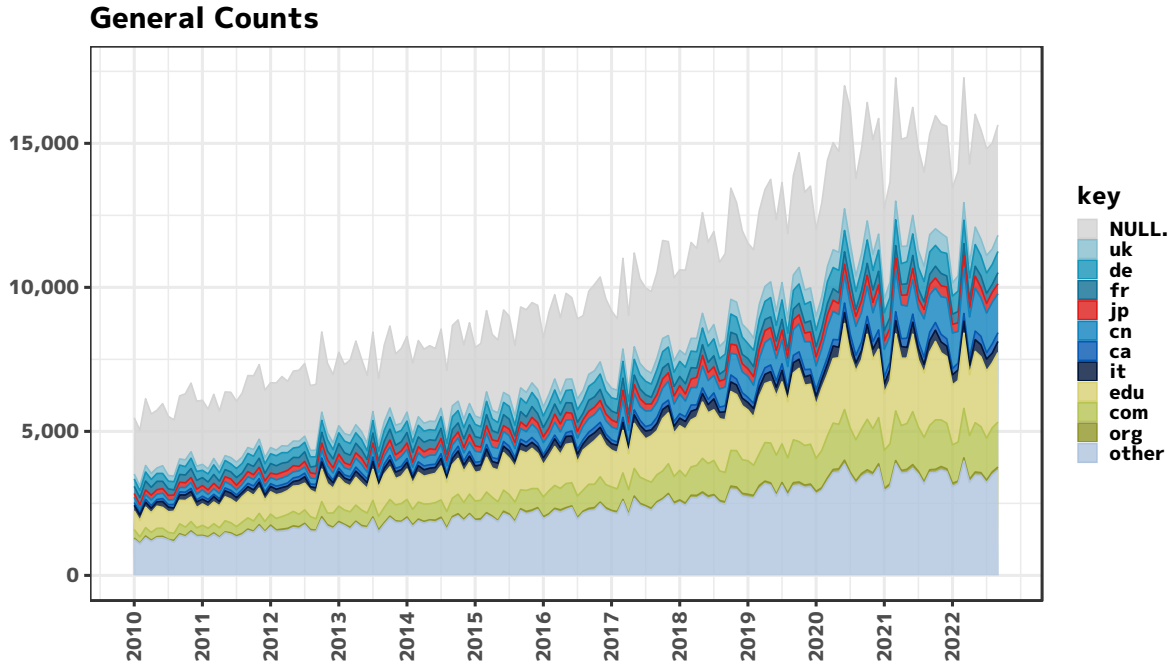

図 2: 原稿数の推移

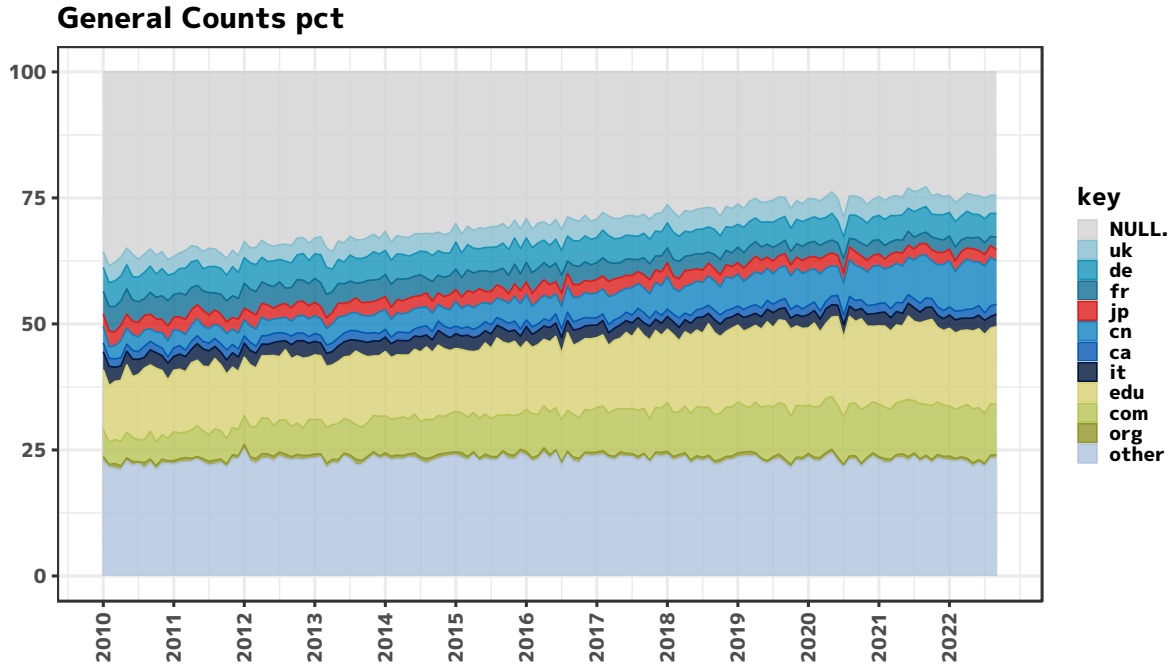

図 3: 原稿数の推移（割合）

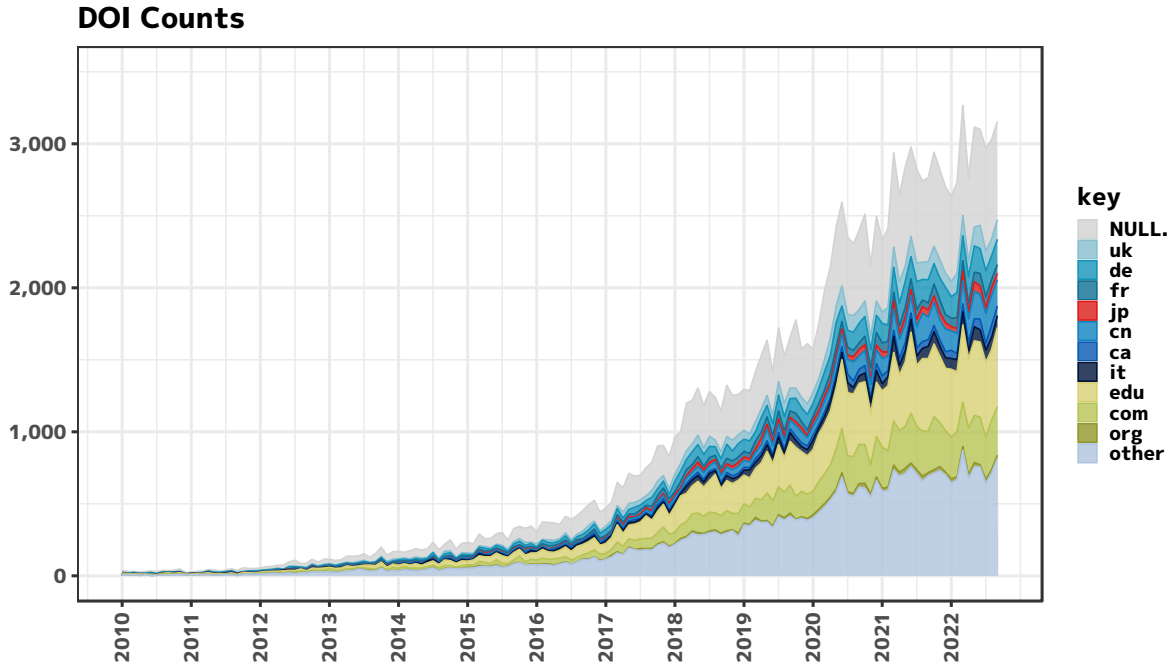

図 4: DOI 言及原稿数の推移

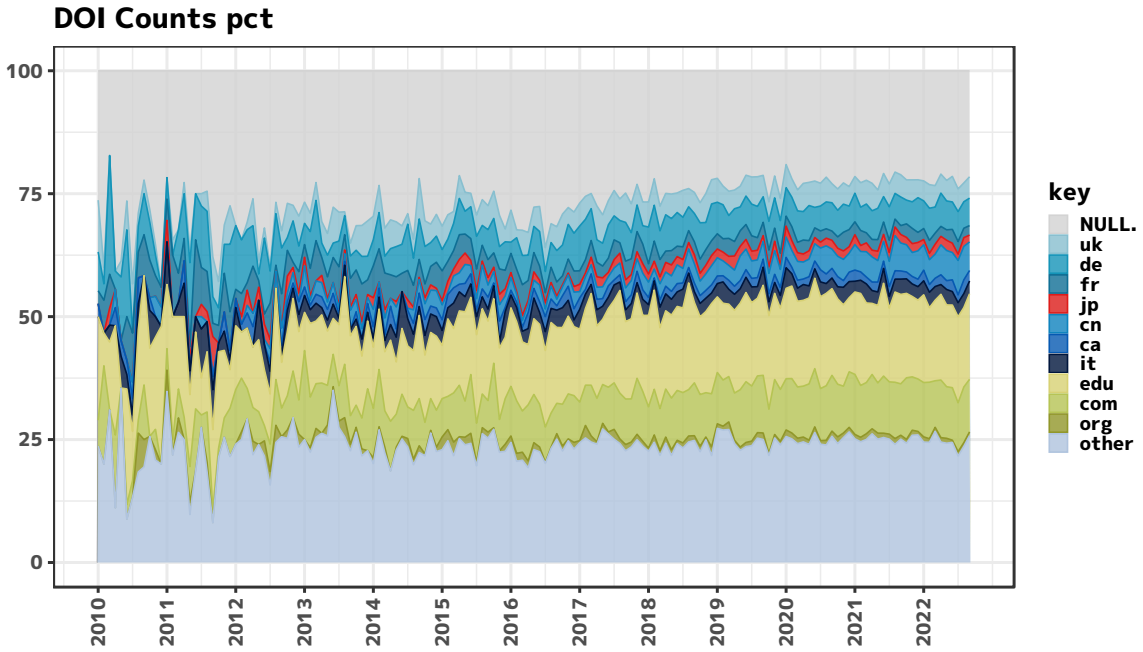

図 5: DOI 言及原稿数の推移（割合）

3.2 オープンデータ利用

すでに述べたとおり，今回はオープンデータ利用の代理変数として，Zenodo，figshare の URL 記載で代替した。結果を図 6 から図 9 に示す。

Zenodo は 2013 年，figshare は 2011 年のスタートだが，図 6 及び図 8 をみると，Zenodo は 2015 年頃，figshare は 2013 年頃から arXiv に登場し始めており，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。サービス開始，arXiv での登場の両方において figshare の方が 2 年速いが，言及数の伸びは Zenodo の方が早く，2022 年では figshare が月間 20 件未満に対して Zenodo は 200 件超と 1 桁程度の差がある。

他方で，図 2 に示すとおり arXiv 全体の月間投稿数が 1.5 万件程度であることを考えると，figshare との比較で Zenodo が相対的に多いとは言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がたい状況にある。

国別比較の観点では，Zenodo が DOI ベースであることともおそらく関係し，図 7 は図 5 とよく似た構造を示している。ここでも日本の存在感はかなり薄く，英国やドイツ，edu ドメインが一定の割合を占めている。

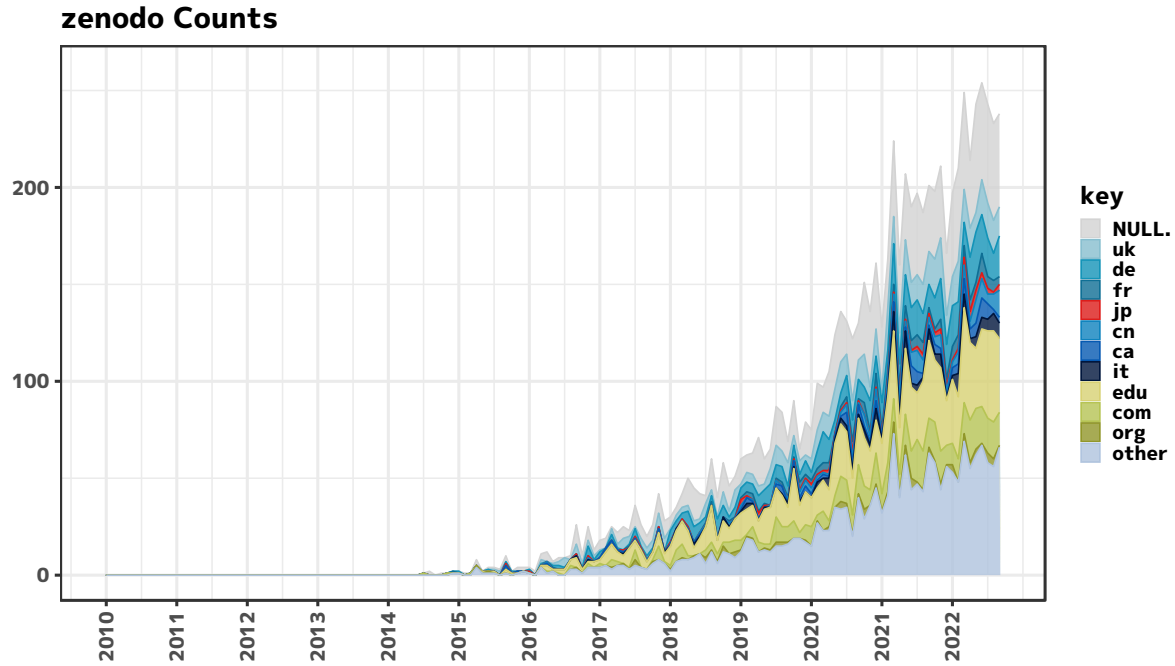

図 6: Zenodo 言及原稿数の推移

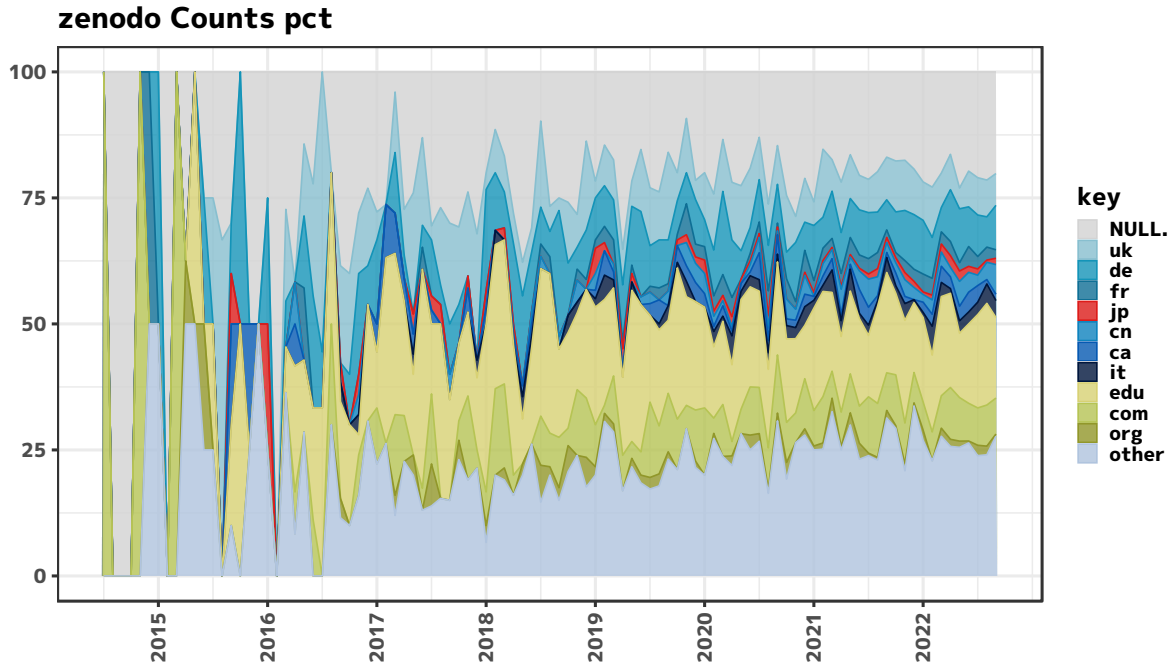

図 7: Zenodo 言及原稿数の推移（割合）

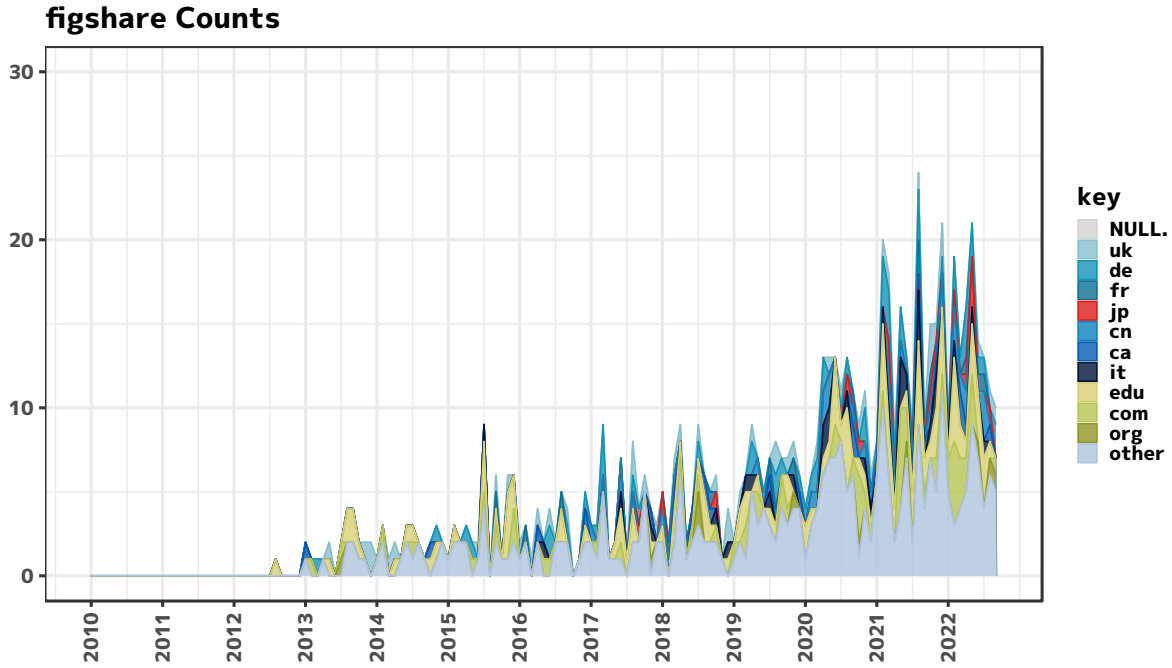

図 8: figshare 言及原稿数の推移

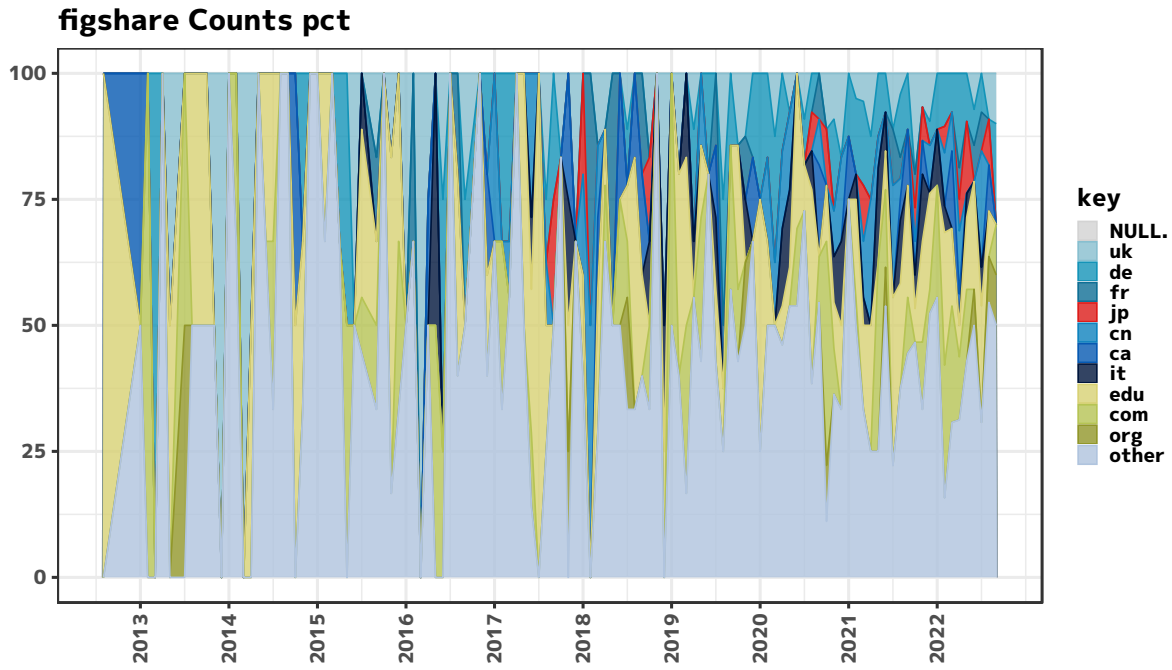

図 9: figshare 言及原稿数の推移（割合）

3.3 OSS 利用

■全体傾向 すでに述べたとおり，今回は研究データに OSS を含め，OSS 利用の代理変数として github の URL 記載で代替した。結果を図 10 及び図 11 に示す。

github は 2008 年スタートのサービスだが，図 10 を見る限り arXiv では 2012 年頃から登場しはじめ，2017,8 年あたりから広まってきているように見える。2022 年では月間言及数が 3 千件を越えており，Zenodo の 200 件からさらに一桁大きく，数百件とはいえ DOI 全体の言及数も越える。また，これらの全体傾向は同様の分析を行った先行研究 [Escamilla22] とも基本的に合致している。

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。

図 11 を見ると，日本，英国，ドイツなどは比較的安定，中国は割合を伸ばしつつあり，edu ドメインは微減傾向を示している。DOI 同様，投稿数から考えると日本の割合は相対的に小さい。

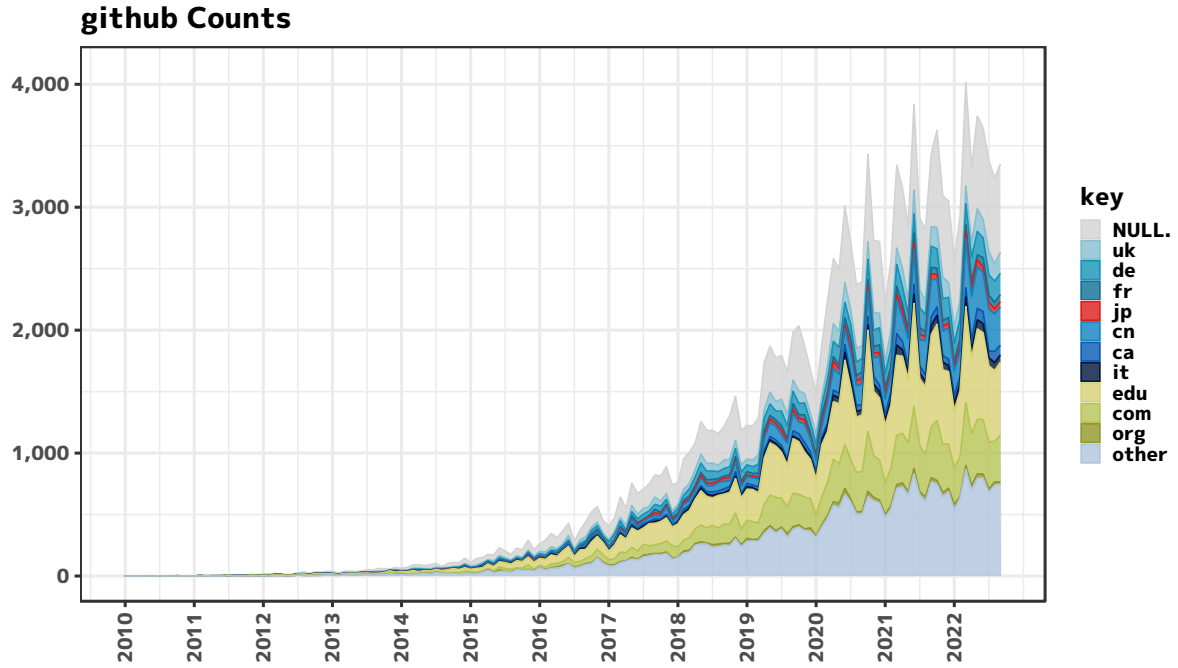

図 10: github 言及原稿数の推移

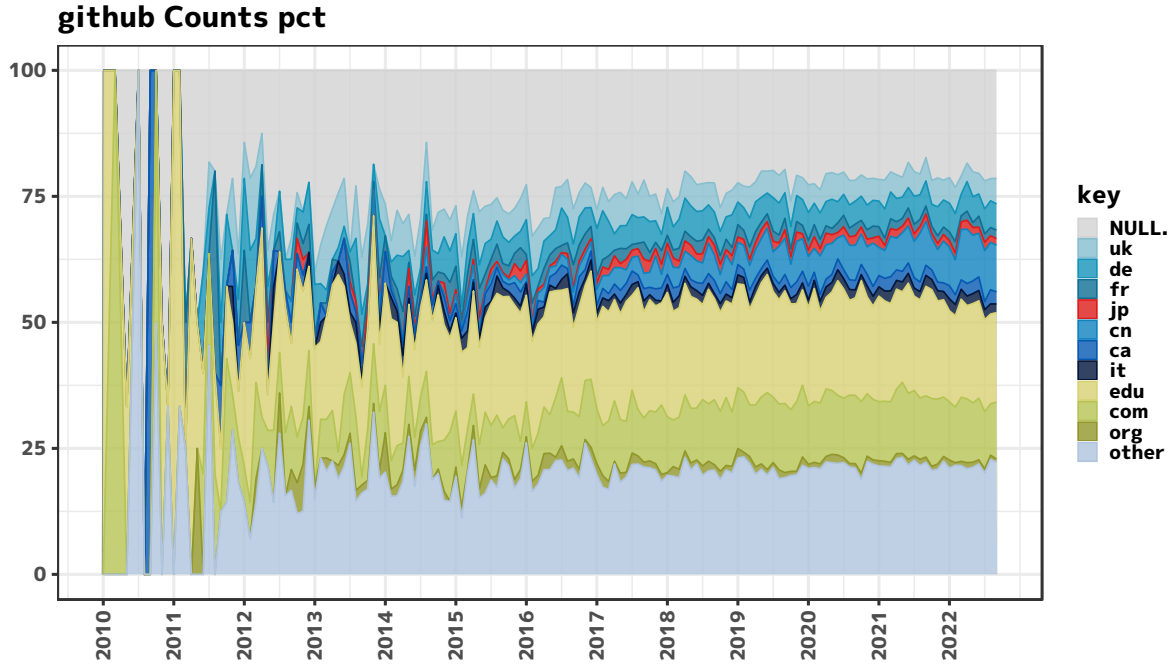

図 11: github 言及原稿数の推移（割合）

■日本の状況　ここで，日本の状況について詳細なデータを図 12 及び図 13 に示した。また比較のため英国について同様に図 14 及び図 15 に示した。

図 12 及び図 13 を見ると，日本からの投稿原稿のうち，github の言及を含むものの件数は少ないながらも順調に増加しており，割合で見ても日本からの投稿原稿全体の 1 割超には記載されている。

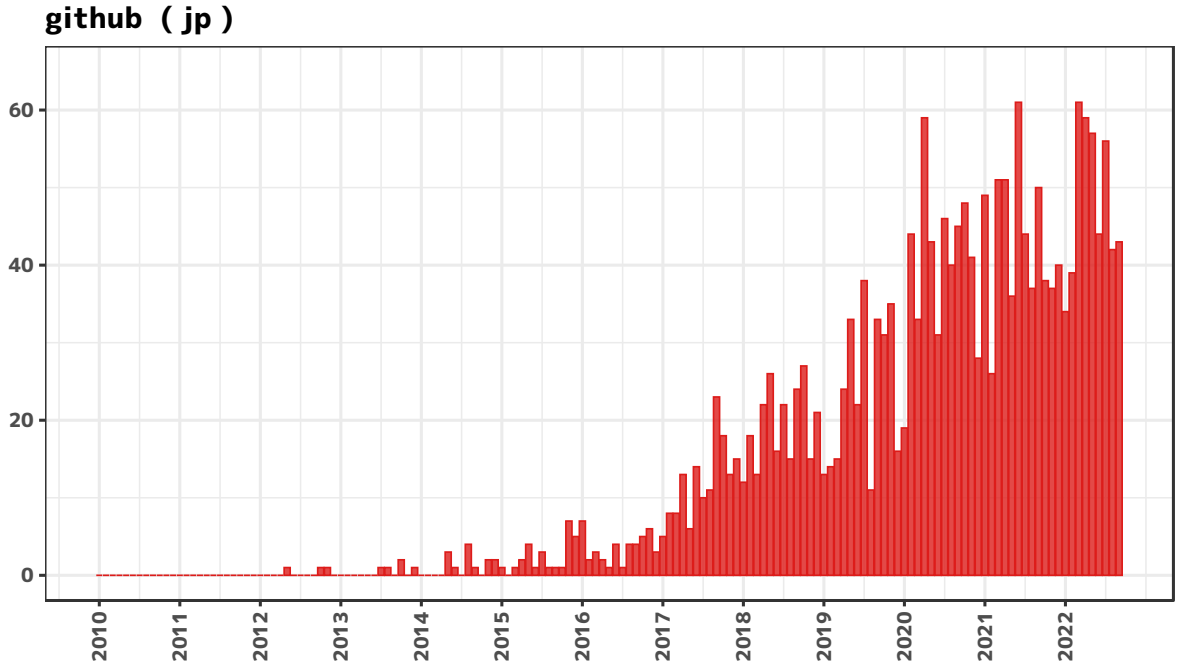

図 12: github 言及原稿数の推移（日本，実数）

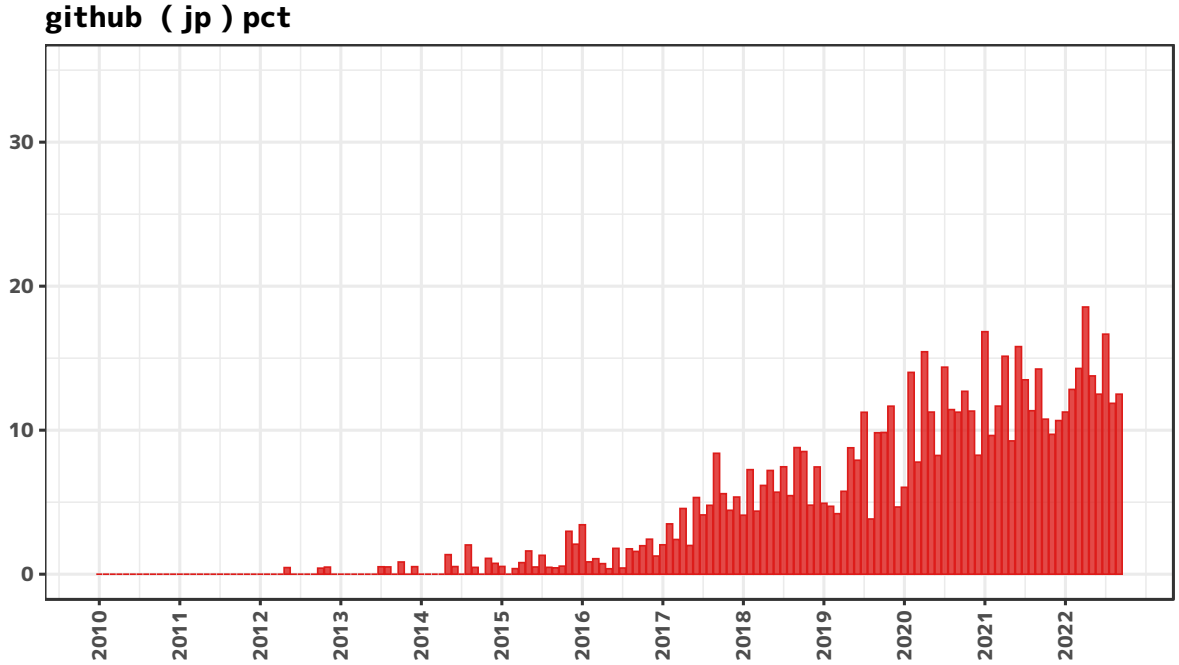

図 13: github 言及原稿数の推移（日本，割合）

比較対象に置いた英国のデータである図 14 及び図 15 を見ると，英国からの投稿原稿のうち，github の言及を含むものの件数は日本と同様に順調に増加している。他方，割合で見ると英国からの投稿原稿全体の 3 割弱には github の言及がなされ，日本との差が大きい。絶対数，割合ともに，日本の倍程度の状況と言える。

英国との比較をもう少し進め，2010 年 1 月を基準にそこからの累積投稿原稿数と，そのうちの

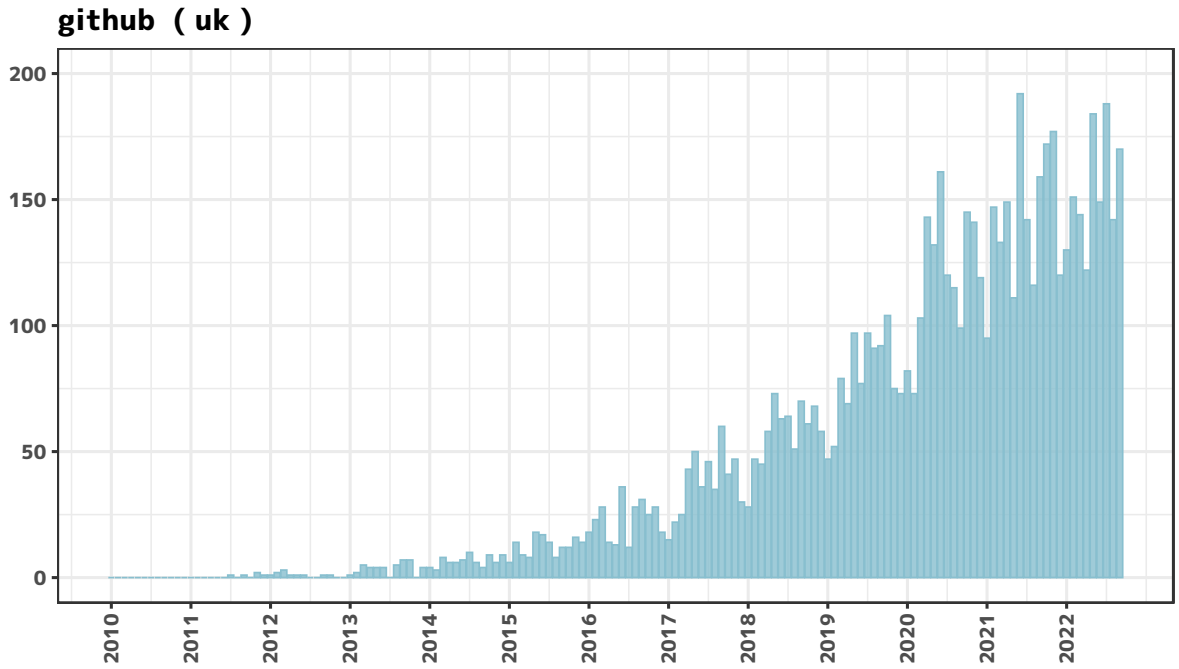

図 14: github 言及原稿数の推移（英国，実数）

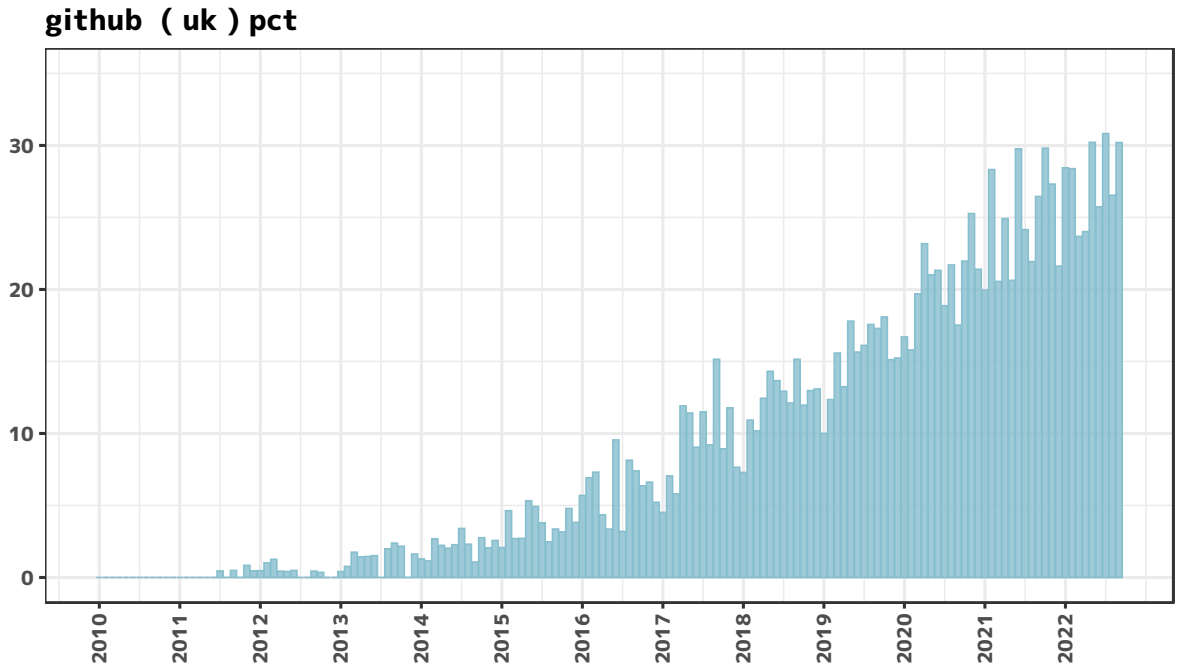

図 15: github 言及原稿数の推移（英国，割合）

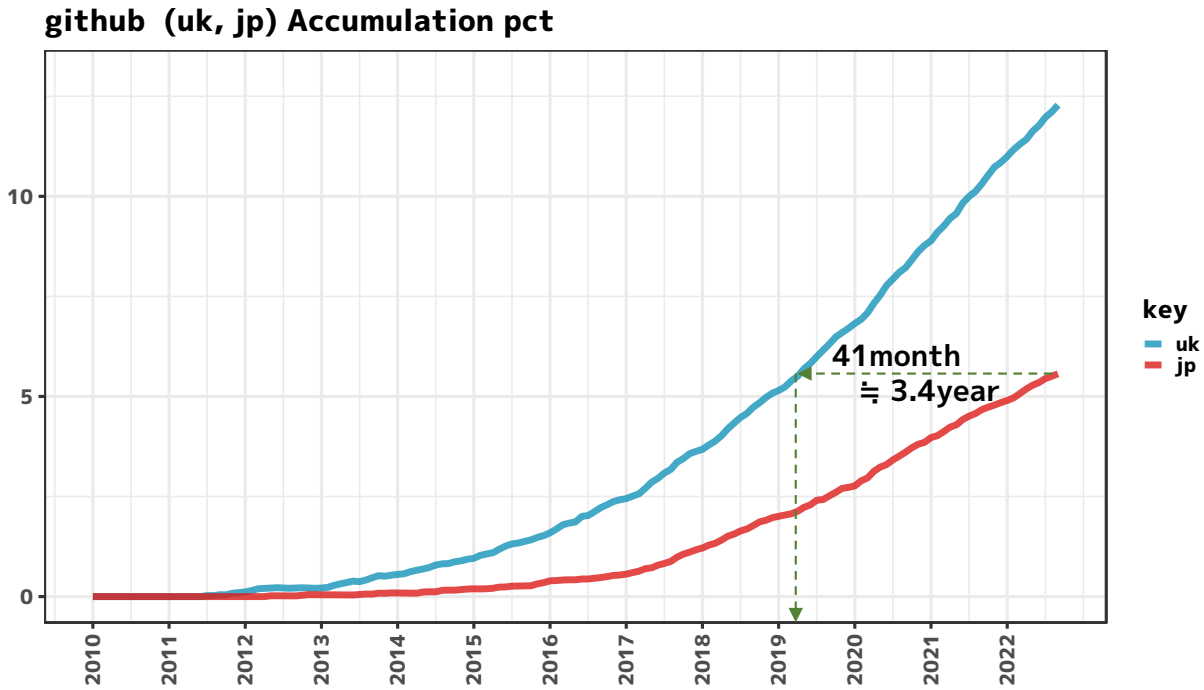

図 16: github 言及原稿数の推移（日英，累積割合）

github 言及原稿数を算出して，国別にその割合を求め，図 16 に示した。

図 16 をみると，日本・英国ともに似たような増加曲線を描いているように見受けられ，かつ英国が先行している状況と言える。そこで，現在の日本のステータス（累積 githib 言及割合）を英国に当てはめると，2022 年 9 月時点での日本のステータスは，41 ヶ月前の 2019 年 4 月の状況に相当し，英国におおよそ 3 年半後れをとっている状況と言える。

3.4 全体要約

2017 年から 2021 年までの 5 年分について，年，国，サービス単位での件数と，全投稿原稿に占める割合を表 2 から表 6 に示す。

なお，ここまで示してきた図の割合の多くは「DOI 言及原稿」「github 言及原稿」など各カテゴリにおける割合を示していたもの（上記の表では行単位で正規化したもの）に対して，これらの表にける割合は「期間中に投稿された全原稿数」に対する割合である点（上記の表では列方向に着目したもの）に注意が必要である。

表 2: 国/サービスの原稿数・割合 (2017 年)

Kind	Total	NULL	uk	de	fr	jp	cn	ca	it	edu	com	org	other
All Post	123,523	35,338	4,698	6,072	3,928	3,297	6,388	1,986	3,102	17,982	10,862	851	29,019
DOI	8,521	2,215	430	493	277	133	210	154	229	1,396	794	95	2,095
github	8,219	2,030	450	444	162	144	292	169	123	1,754	885	121	1,645
zenodo	310	77	43	17	1	4	1	10	2	69	23	6	57
figshare	51	0	4	6	0	1	5	1	2	7	2	1	22
DOI	6.9%	6.3%	9.2%	8.1%	7.1%	4.0%	3.3%	7.8%	7.4%	7.8%	7.3%	11.2%	7.2%
github	6.7%	5.7%	9.6%	7.3%	4.1%	4.4%	4.6%	8.5%	4.0%	9.8%	8.1%	14.2%	5.7%
zenodo	0.3%	0.2%	0.9%	0.3%	0.0%	0.1%	0.0%	0.5%	0.1%	0.4%	0.2%	0.7%	0.2%
figshare	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.1%	0.1%	0.1%	0.0%	0.0%	0.1%	0.1%

⋆2017年に投稿された原稿に基づく

表 3: 国/サービスの原稿数・割合 (2018 年)

Kind	Total	NULL	uk	de	fr	jp	cn	ca	it	edu	com	org	other
All Post	140,616	38,582	5,548	6,902	4,145	3,592	8,091	2,279	3,538	21,442	13,173	1,049	32,275
DOI	14,368	3,570	683	935	403	251	468	210	405	2,517	1,411	154	3,361
github	13,754	3,220	686	754	263	231	624	278	217	2,906	1,579	185	2,811
zenodo	540	126	45	63	5	3	5	4	7	124	43	14	101
figshare	60	0	6	1	3	2	3	4	2	10	4	2	23
DOI	10.2%	9.3%	12.3%	13.5%	9.7%	7.0%	5.8%	9.2%	11.4%	11.7%	10.7%	14.7%	10.4%
github	9.8%	8.3%	12.4%	10.9%	6.3%	6.4%	7.7%	12.2%	6.1%	13.6%	12.0%	17.6%	8.7%
zenodo	0.4%	0.3%	0.8%	0.9%	0.1%	0.1%	0.1%	0.2%	0.2%	0.6%	0.3%	1.3%	0.3%
figshare	0.0%	0.0%	0.1%	0.0%	0.1%	0.1%	0.0%	0.2%	0.1%	0.0%	0.0%	0.2%	0.1%

⋆2018年に投稿された原稿に基づく

表 4: 国/サービスの原稿数・割合 (2019 年)

Kind	Total	NULL	uk	de	fr	jp	cn	ca	it	edu	com	org	other
All Post	155,866	40,206	6,158	7,468	4,326	3,908	10,085	2,651	3,823	24,762	15,594	1,147	35,738
DOI	18,450	4,271	841	1,103	547	326	706	302	555	3,127	1,969	162	4,541
github	20,146	4,333	953	1,051	382	285	1,205	387	291	4,243	2,629	240	4,147
zenodo	855	173	73	79	19	11	14	20	12	185	71	11	187
figshare	77	0	6	7	3	0	1	2	4	9	8	1	36
DOI	11.8%	10.6%	13.7%	14.8%	12.6%	8.3%	7.0%	11.4%	14.5%	12.6%	12.6%	14.1%	12.7%
github	12.9%	10.8%	15.5%	14.1%	8.8%	7.3%	11.9%	14.6%	7.6%	17.1%	16.9%	20.9%	11.6%
zenodo	0.5%	0.4%	1.2%	1.1%	0.4%	0.3%	0.1%	0.8%	0.3%	0.7%	0.5%	1.0%	0.5%
figshare	0.0%	0.0%	0.1%	0.1%	0.1%	0.0%	0.0%	0.1%	0.1%	0.0%	0.1%	0.1%	0.1%

*2019年に投稿された原稿に基づく

表 5: 国/サービスの原稿数・割合 (2020 年)

Kind	Total	NULL	uk	de	fr	jp	cn	ca	it	edu	com	org	other
All Post	178,329	45,278	6,981	8,472	4,702	4,342	11,929	3,146	4,318	28,711	18,229	1,294	40,927
DOI	26,694	6,060	1,220	1,489	656	432	1,076	477	732	4,702	2,904	277	6,669
github	30,078	6,427	1,433	1,444	561	477	1,907	570	561	6,116	3,809	330	6,443
zenodo	1,467	307	141	120	40	14	32	32	37	245	118	19	362
figshare	118	0	6	10	1	2	9	2	7	18	8	1	54
DOI	15.0%	13.4%	17.5%	17.6%	14.0%	9.9%	9.0%	15.2%	17.0%	16.4%	15.9%	21.4%	16.3%
github	16.9%	14.2%	20.5%	17.0%	11.9%	11.0%	16.0%	18.1%	13.0%	21.3%	20.9%	25.5%	15.7%
zenodo	0.8%	0.7%	2.0%	1.4%	0.9%	0.3%	0.3%	1.0%	0.9%	0.9%	0.6%	1.5%	0.9%
figshare	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	0.1%	0.1%	0.2%	0.1%	0.0%	0.1%	0.1%

*2020年に投稿された原稿に基づく

表 6: 国/サービスの原稿数・割合 (2021 年)

Kind	Total	NULL	uk	de	fr	jp	cn	ca	it	edu	com	org	other
All Post	181,630	44,024	6,921	8,946	4,537	4,222	14,485	3,246	4,625	28,603	18,975	1,262	41,784
DOI	32,938	7,298	1,455	1,892	783	549	1,705	557	925	5,536	3,638	319	8,281
github	36,902	7,547	1,713	1,831	610	520	2,692	769	716	7,414	4,659	359	8,072
zenodo	2,232	422	190	191	53	20	63	41	53	384	181	31	603
figshare	176	0	11	16	3	4	11	4	12	28	13	1	73
DOI	18.1%	16.6%	21.0%	21.1%	17.3%	13.0%	11.8%	17.2%	20.0%	19.4%	19.2%	25.3%	19.8%
github	20.3%	17.1%	24.8%	20.5%	13.4%	12.3%	18.6%	23.7%	15.5%	25.9%	24.6%	28.4%	19.3%
zenodo	1.2%	1.0%	2.7%	2.1%	1.2%	0.5%	0.4%	1.3%	1.1%	1.3%	1.0%	2.5%	1.4%
figshare	0.1%	0.0%	0.2%	0.2%	0.1%	0.1%	0.1%	0.1%	0.3%	0.1%	0.1%	0.1%	0.2%

*2021年に投稿された原稿に基づく

4 まとめ

本稿では研究活動にけるオープンソース・データの利用状況の調査を目的として，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 を対象に，プレプリント（原稿）中のオープンソース・オープンデータ言及回数を調査した。

ここでは，オープンソースとして github，オープンデータに Zenodo, figshare を取り上げて調査した。また，比較のための基礎データとして DOI も取り上げて調査した。本文中に記載されたメールアドレスを手がかりとして，各原稿には（割り当て可能なものについては）国籍を割り付け，初版発行の年月ベースで整理した。

結果，Zenodo, figshare ベースでみると arXiv 上でのオープンデータ言及はほとんど進んでいないことが分かった。他方で，github ベースでみた arXiv 上でのオープンソース言及は緩やかながら順調な伸びを見せており，2022 年以降全体投稿の 2 割以上の原稿において言及されていることが分かった。オープンソース言及に関する我が国のステータスを英国と比べた場合，2022 年 9 月時点で英国の約半分であり，割合で見た場合，英国と同じ傾向をたどるとすれば，3 年半近く遅れた状態であることがわかった。

4.1 留意事項等

今回はデータソースに arXiv を選定しているため，主に物理・情報分野における状況についてのみの分析である点には留意が必要である。

また，速報性・簡易性を重視して，国籍や URL の抽出は単純な文字列マッチなどで行っている。結果として，概ねの傾向としては正しいと考えられるものの，精度には欠ける点がある。

国籍は最初に出現するメールアドレス 1 件（基本的には第 1 著者か連絡著者と期待される）のみに基づいて，そのトップレベルドメインで判定することになっているため，多くの計量書誌分析と条件が異なり，横並びでの比較は必ずしも適当でない。

URL ベースでのオープンデータ判別という手法に起因する本質的な課題として，「使った」のか，「作った」のかの判定が付かないという問題もある。原稿中に記載した手法の実装例として自身が作成したソースコードを github で公開することと，第三者が開発し，github で公開されていたツールを使って分析することの意味的な差は大きい。同じく，自身が原稿に関連して作成したデータを Zenodo に置くのと，Zenodo にあったデータを分析することも意味が異なる[22)]。

他にも，あくまで Zenodo，figshare，github のみを対象にした調査であり，測りやすいもののみを測っている点などにも留意が必要ではある。表 1 や，オリジナルのデータでこれらよりも目立つ情報源はなさそうなことは確認しているものの，例えば，github に類似するサービスとして bitbucket[23)]も存在する。分析上の手間は多くかかるが，DOI については 1 件づつ DataCite[24)]に問い合わせるなどし

[22)] ただし，この点は一般的な引用分析にも似た課題はあり，自己引用か否か，考察上の重要な引用か数ある関連手法の一つとしての軽い引用か，など，本来はそれぞれ重みが違うが，分析コストから同じものと割り切って分析することは決して珍しくない。

[23)] https://bitbucket.org/

[24)] https://datacite.org/

て，データかどうか判定することも考えられる。

最後に，例えば Zenodo や github に言及する原稿が多いほど良いのか，何割程度を占めれば適切かというような，評価の観点を本報は含んでいない。単純に何件あったのか，他と比較してどの程度多いのか，少ないのか，を示すにとどまっている。

参考文献

[Escamilla22] Emily Escamilla, et.al. The Rise of GitHub in Scholarly Publications, *arXiv*, 2022. (preprint) DOI: https://doi.org/10.48550/arXiv.2208.04895

[林 20] 林 和弘, 他. arXiv に着目したプレプリントの分析, *NISTEP 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

付録 A　参考データ

A.1　Dryad

本編ではデータ共有サービスに Zenodo，figshare を取り上げたが，Zenodo と類似する・関連が深いサービスとして Dryad[25]が存在する。そこで，参考までに Dryad の例も試行した。

Dryad も Zenodo と同様に形式としては DOI を有し，「`10.5061/dryad.XXXXXX`」の表記を行うため「`10.5061/dryad`」を検出することで分析できる。結果，2010 年 1 月から 2022 年 9 月の範囲で 82 件しか検出できなかった。そこで，単に「`dryad`」を含むものというレベルにまで粒度を粗くして検出を試みた。

結果を図 17 に示す。図中 DOI は DOI ベースのカウント，Other は単に「`dryad`」を含むものを検出したあと DOI 分を除いたもののカウントである。

結果を図 17 を見ると，Dryad は 2020 年頃から利用者がやや増えてきているものの，まだ利用されているとは言いがたい状況が伺える。

Zenodo と Dryad は 2021 年 2 月にシステム連携の声明を発表している[26] ことから，今後利用者が伸びる可能性があるが，現状ではモニタリング対象とするレベルにはないと言える。

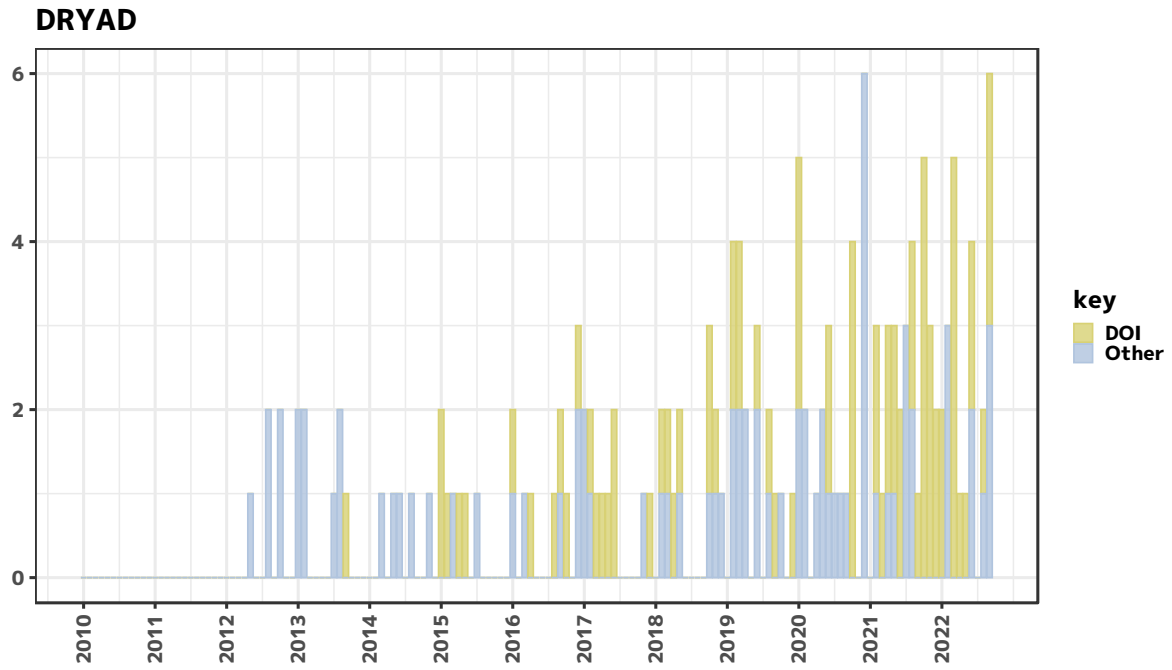

図 17: Dryad 言及原稿数の推移

[25] `https://datadryad.org/`

[26] Doing it Right: A Better Approach for Software & Data（Dryad,2021/2/8）`https://blog.datadryad.org/2021/02/08/doing-it-right-a-better-approach-for-software-amp-data/` Doing it Right: A Better Approach for Software & Data（Zenodo,2021/2/8）`https://blog.zenodo.org/2021/02/08/2021-02-08-doing-it-right/`

付録 B　bioRxiv

本編では物理・情報系の arXiv について調べたが，参考までに生物系の bioRxiv[27]についても調べた。

bioRxiv は 2013 年に開設された比較的歴史があるプレプリントサーバで，原稿のデータを xml で取得できるなど，arXiv にくらべてよりマシンリーダブルな点に特徴がある。結果，著者所属機関の国籍や，URL の取得も容易である。一方で，後述するとおり投稿数は arXiv に比べると少ない。

B.1　データの概要

収集対象や手法について，基本的には本編の arXiv 同様とし，以下では差分について述べる。

原稿のデータは bioRxiv の公式レポジトリ[28]から取得した。今回は比較のために，レポジトリ中の 2021 年分のデータのみを全件取得した。この取得データに含まれる原稿の xml データを具体的な解析対象とした。

xml データ中には，著者の名前や機関，機関のアドレス（国・地域）も記載されているため，国籍判定は xml 中に出現する最初の `country` タグ，すなわち第 1 著者の `country` タグの値を採用する[29]。関連して，メールアドレスを利用した本編と異なり，米国を明示的に判別できている。

URL については，原稿中の `body` 要素（論文本編）および，`ref-list` 要素（参考文献）に含まれる，`ext-link` 等そのうち，属性が `ext-link-type="uri"` のもののリンク情報を取得した。これにより，本編の arXiv のケースのように，途中で URL が途切れるなどの可能性はほぼなくなっている。

オープンソース・データの対象には，本編の各種サービスに加えて付録で述べた Dryad も加えた。

B.2　結果

結果について以下に示す。

なお，本分析における原稿と時点の紐付けは，原稿 xml にある「`date date-type="received"`」すなわち，bioRxiv への投稿日時を起点に行った。これに起因するものか定かではないが，bioRxiv の公式統計値[30] とは月あたり数十から最大 100 件程度のズレが生じている。bioRxiv の公式統計値はそもそもレポジトリの年月別ディレクトリにあるファイルの数ともズレており，このズレの原因は分析者側の原因による可能性は少ないと考えられる。本分析においては，時点は上記の通り投稿時点を採用し，かつ，レポジトリの原稿データに基づくカウントを正としてあつかう。

---

[27] `https://www.biorxiv.org/`

[28] `https://www.biorxiv.org/tdm`

[29] country タグは著者が自由記述するため, 表記には揺らぎがあり, 例えば米国について “US”,“USA”,“United States”,“U.S.A”, などが存在する。今回は図表中で取り上げる国についてのみ，目視で名寄せ作業を行っている。

[30] `http://api.biorxiv.org/reports/usage`

全体について簡易にまとめたものを表 7 に示す。これを見ると，DOI の付与率は 2 割程度で本編とほぼ同様，むしろ本編では 2022 年を取り上げて 2 割程度，2021 年では 1.8 割程度であることを考えると bioRxiv の方が高い[31]。一方で，オープンソース・データについては arXiv を下回っており，分野による差異が明確に現れている。

簡単な考察も行うならば，元々の投稿数が少ないことと分野的な特性も関連し，bioRxiv は現状では，我々が設定した手法でオープンソース・データ利活用の多寡を論じられる状況にない。ただし，マシンリーダブルなデータの性質には利点があり，またデータ利用の広がりの観察の点で観測対象としては適切と言える。

表 7: (bioRxiv) 国/サービスの原稿数・割合

	Total	NULL	USA	UK	Germany	France	Japan	China	Canada	Italy	Other
All Post	36,610	5,085	9,910	2,895	2,496	1,613	1,103	2,199	1,294	475	9,540
DOI	9,218	1,306	2,200	772	682	410	224	351	352	152	2,769
github	5,723	785	1,679	539	422	209	102	255	230	63	1,439
Zenodo	621	62	151	57	82	30	8	8	25	8	190
figshare	151	18	33	25	10	6	7	4	6	6	36
Dryad	92	14	25	5	7	3	1	4	6	0	27
DOI	25.2%	25.7%	22.2%	26.7%	27.3%	25.4%	20.3%	16.0%	27.2%	32.0%	29.0%
github	15.6%	15.4%	16.9%	18.6%	16.9%	13.0%	9.2%	11.6%	17.8%	13.3%	15.1%
Zenodo	1.7%	1.2%	1.5%	2.0%	3.3%	1.9%	0.7%	0.4%	1.9%	1.7%	2.0%
figshare	0.4%	0.4%	0.3%	0.9%	0.4%	0.4%	0.6%	0.2%	0.5%	1.3%	0.4%
Dryad	0.3%	0.3%	0.3%	0.2%	0.3%	0.2%	0.1%	0.2%	0.5%	0.0%	0.3%

⋆ 2021年に投稿された原稿に基づく

その他，本編と同様の処置を行った図表を含め，図 18 から図 30 に示す。

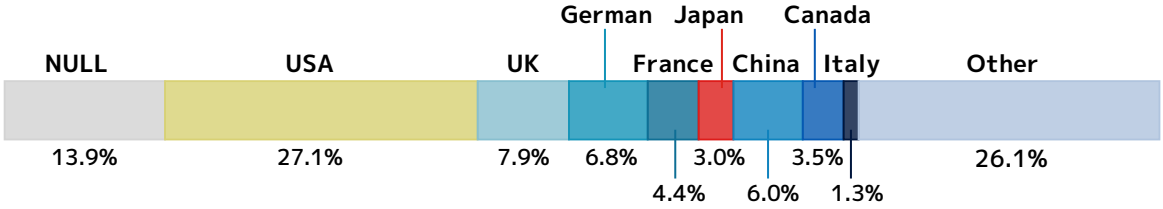

図 18: (bioRxiv)2021 年分原稿の国別内訳

[31] ただし，情報系では例えば大手学会である ACM (Association for Computing Machinery) は doi.org によらない DOI を用いているほか，arXiv も 2022 年に至るまで原稿に DOI は付かなかった。これにより ACM や arXiv を引用する際には DOI が付与されない。bioRxiv は以前より各原稿に DOI が付与されるなど，DOI フレンドリーな構造になっている。

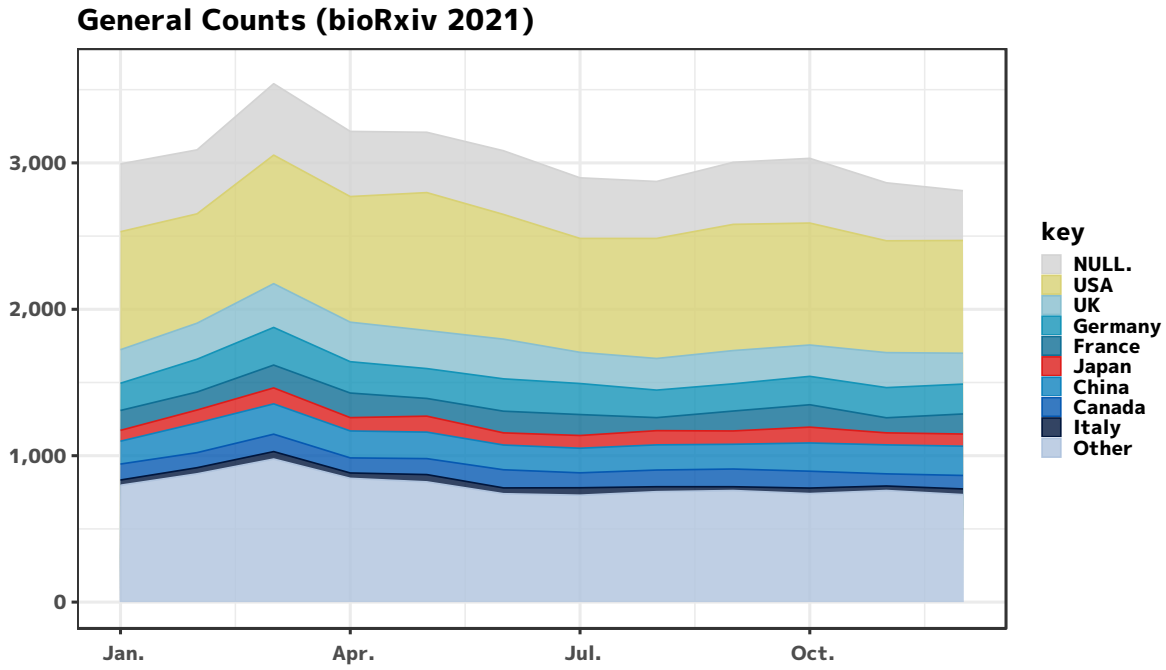

図 19: (bioRxiv) 原稿数の推移

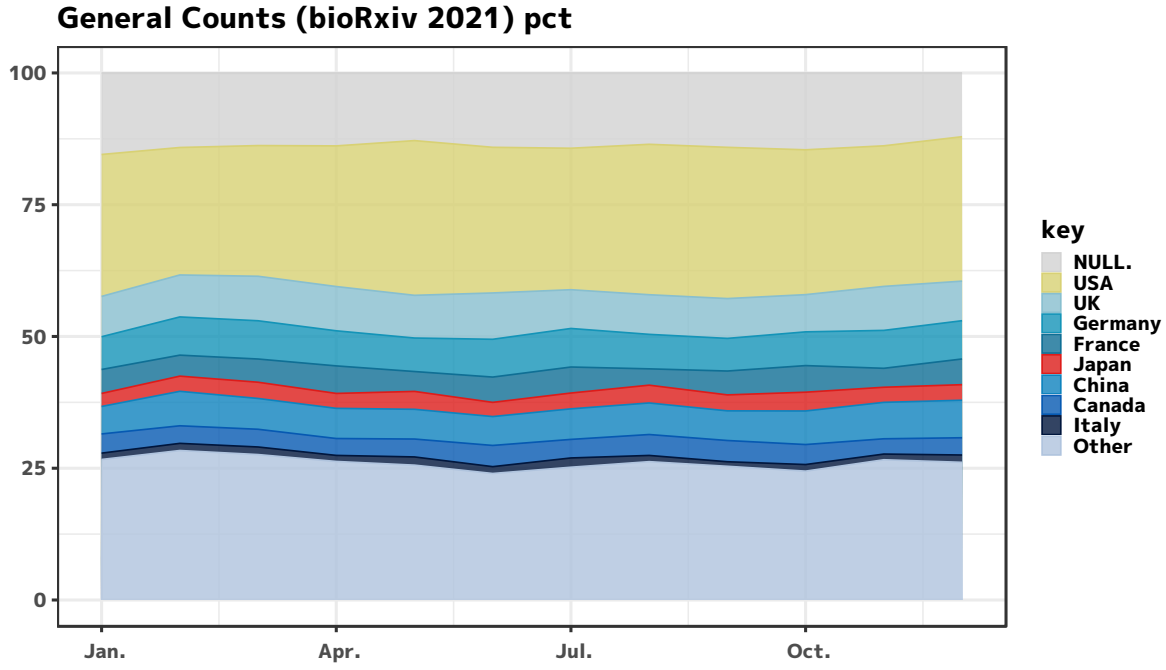

図 20: (bioRxiv) 原稿数の推移（割合）

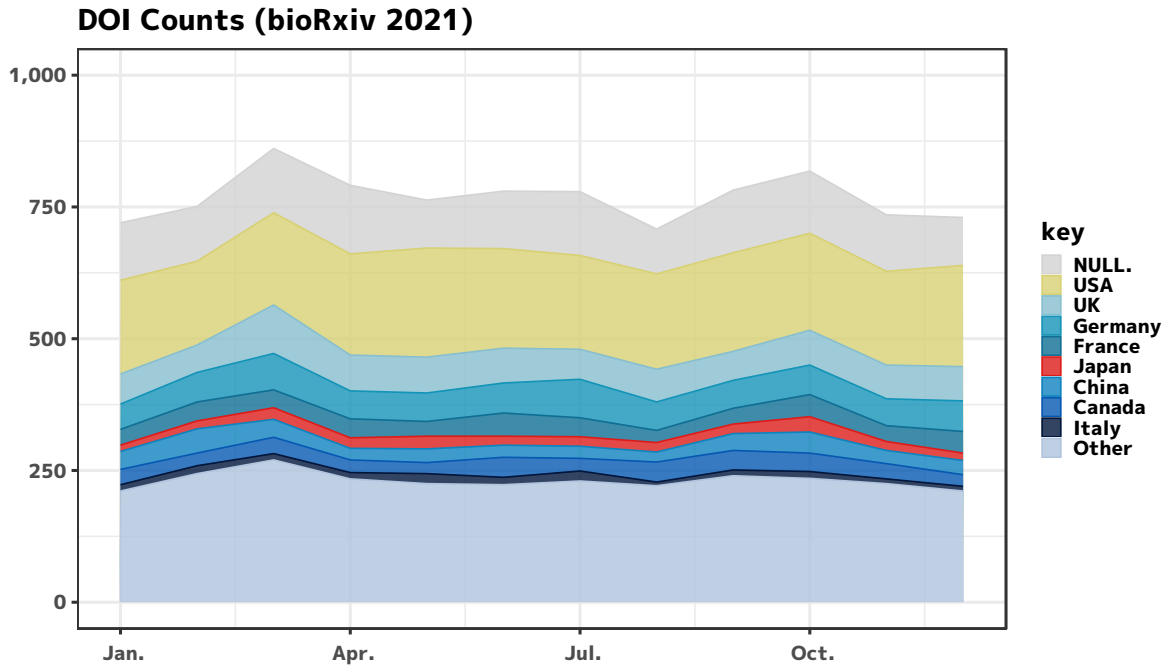

図 21: (bioRxiv)DOI 言及原稿数の推移

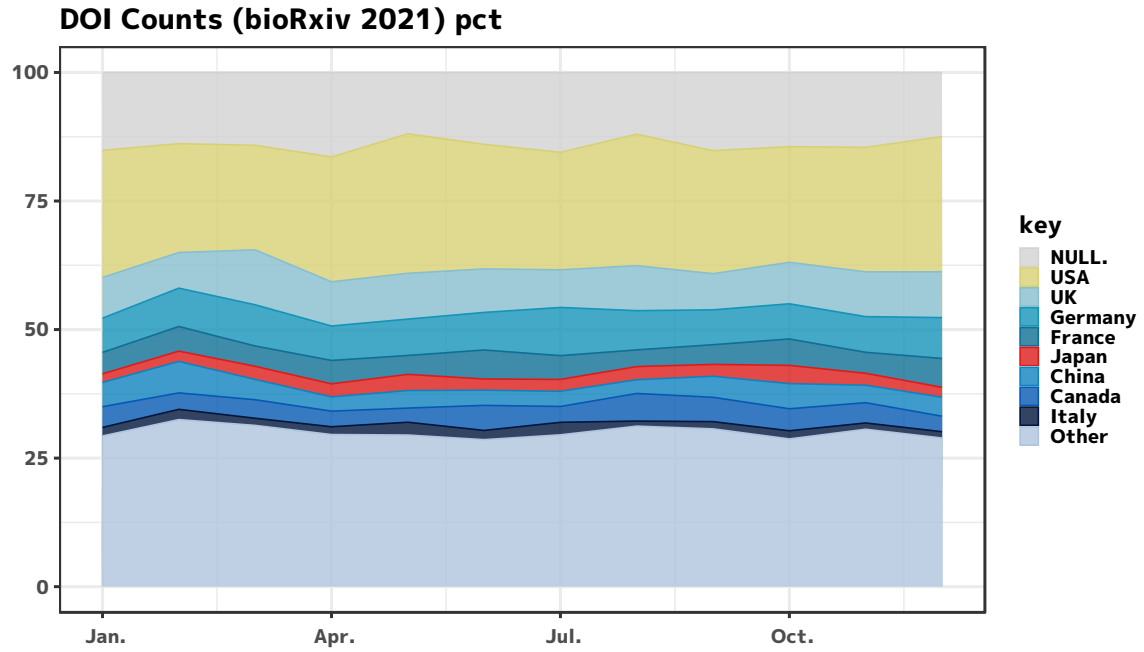

図 22: (bioRxiv)DOI 言及原稿数の推移（割合）

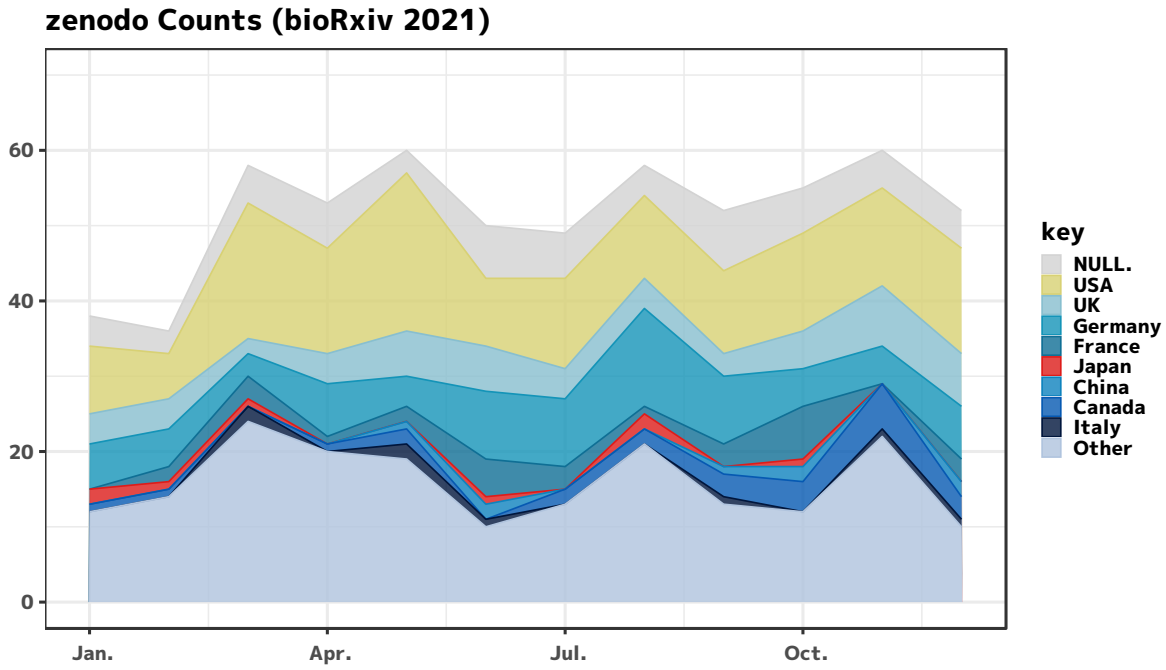

図 23: (bioRxiv)Zenodo 言及原稿数の推移

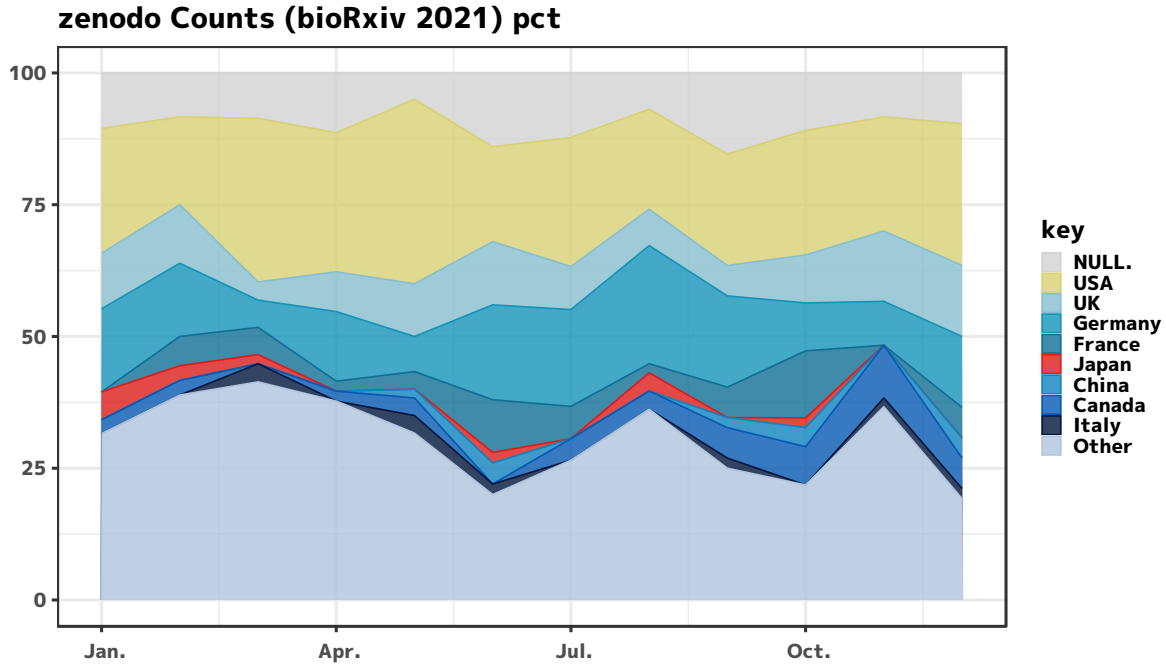

図 24: (bioRxiv)Zenodo 言及原稿数の推移（割合）

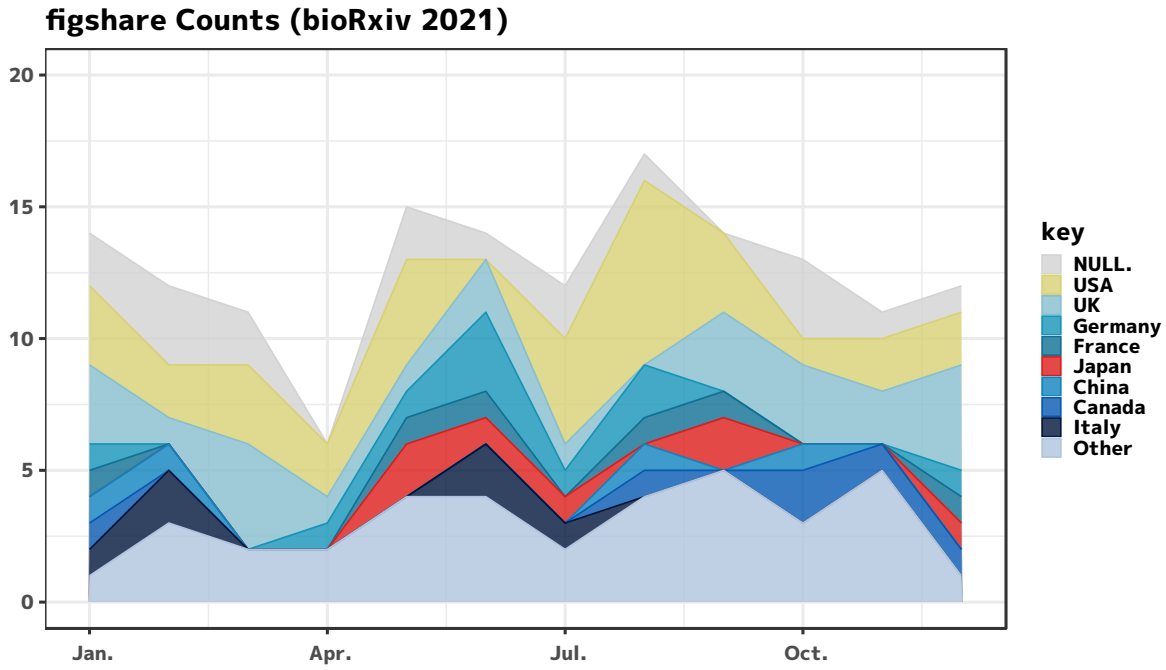

図 25: (bioRxiv)figshare 言及原稿数の推移

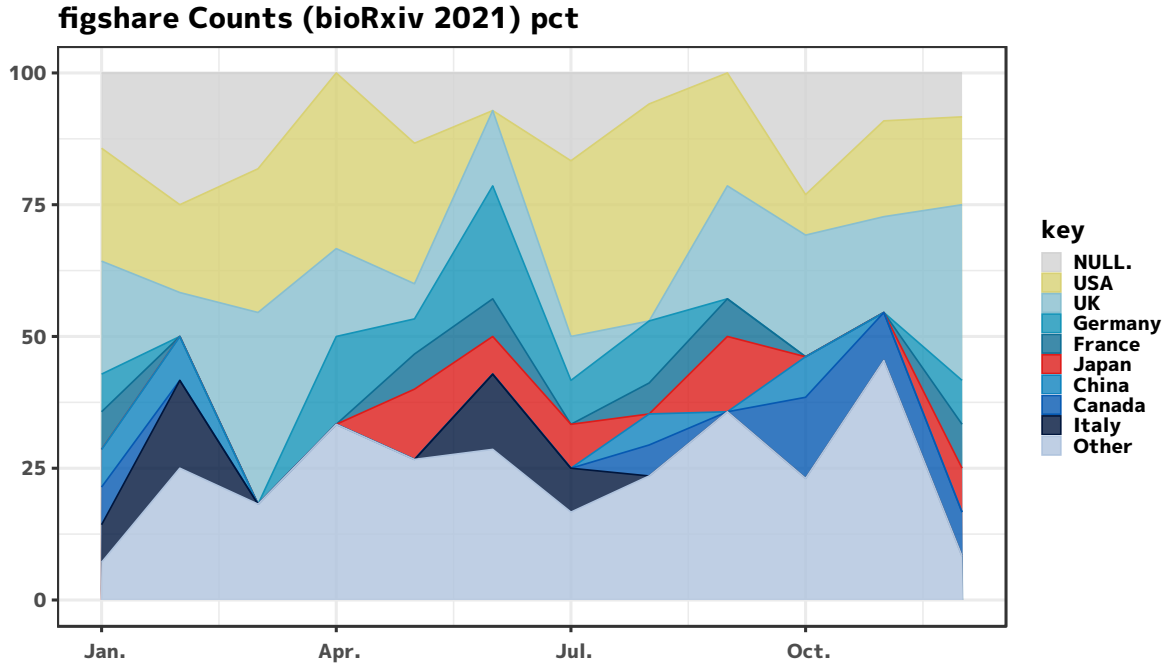

図 26: (bioRxiv)figshare 言及原稿数の推移（割合）

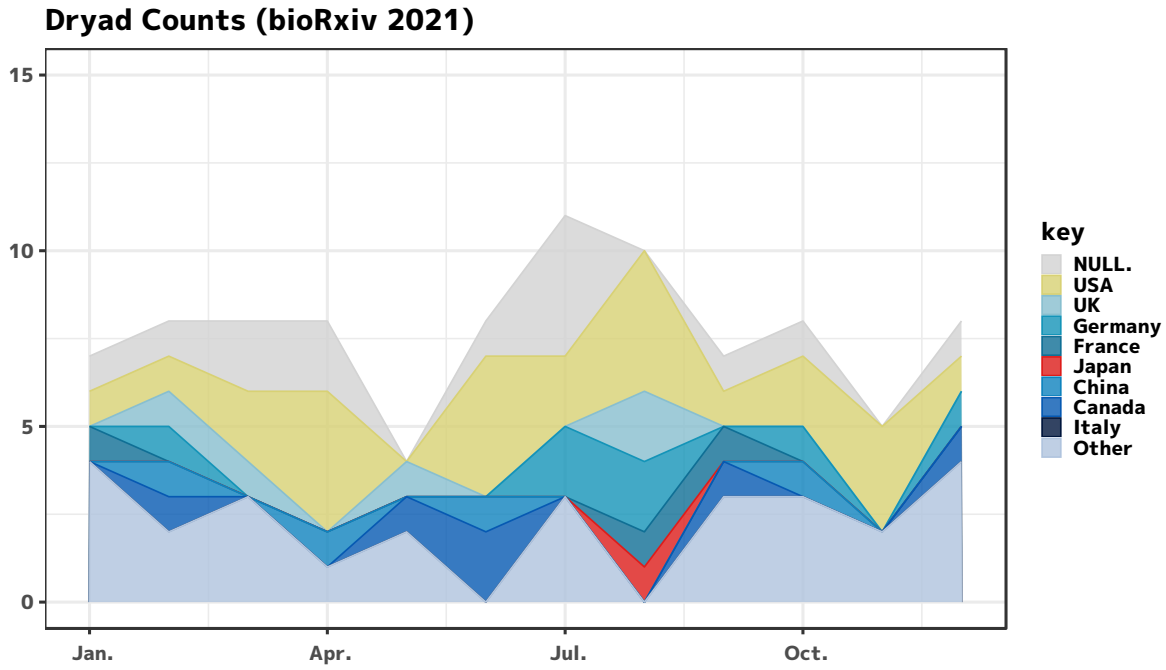

図 27: (bioRxiv)Dryad 言及原稿数の推移

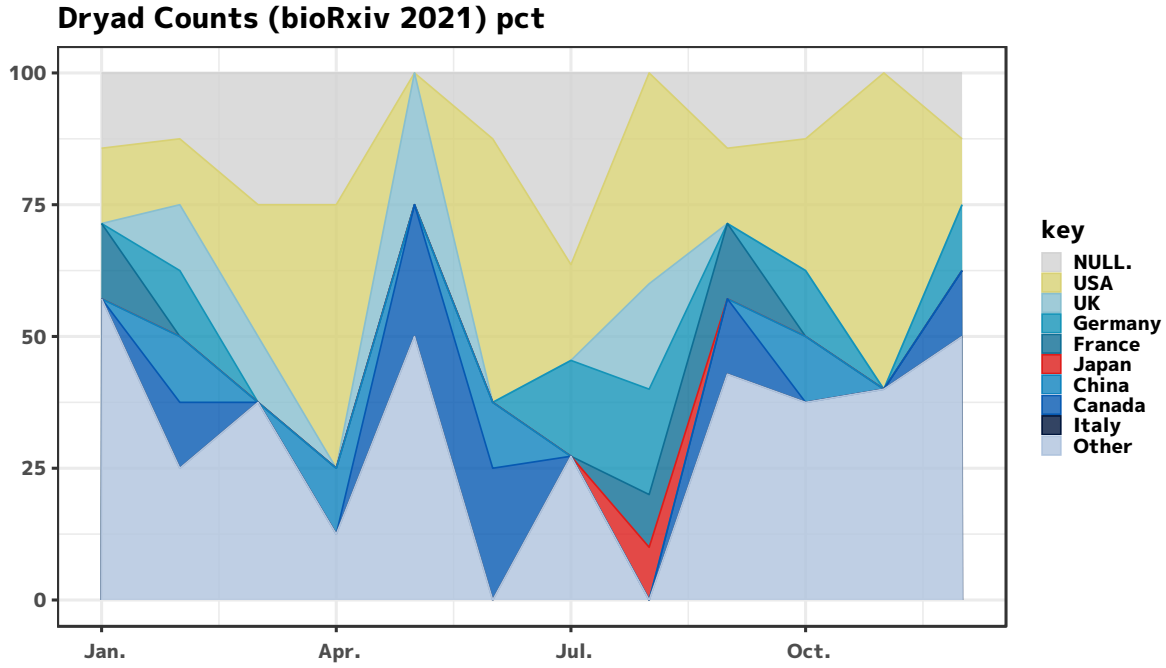

図 28: (bioRxiv)Dryad 言及原稿数の推移（割合）

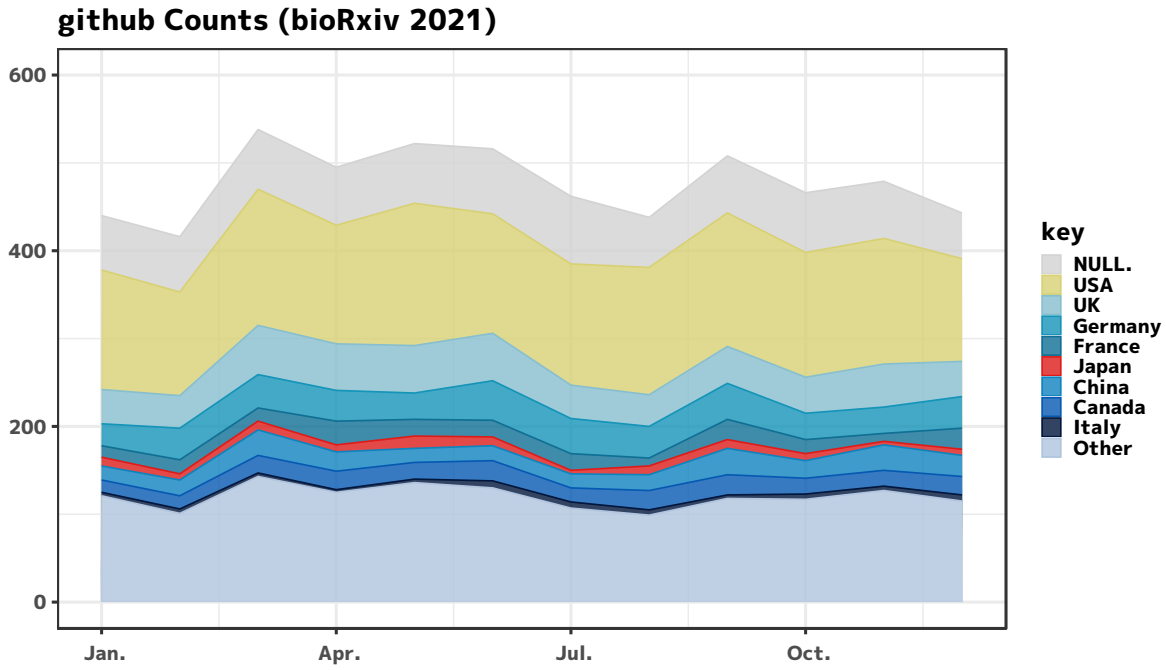

図 29: (bioRxiv)github 言及原稿数の推移

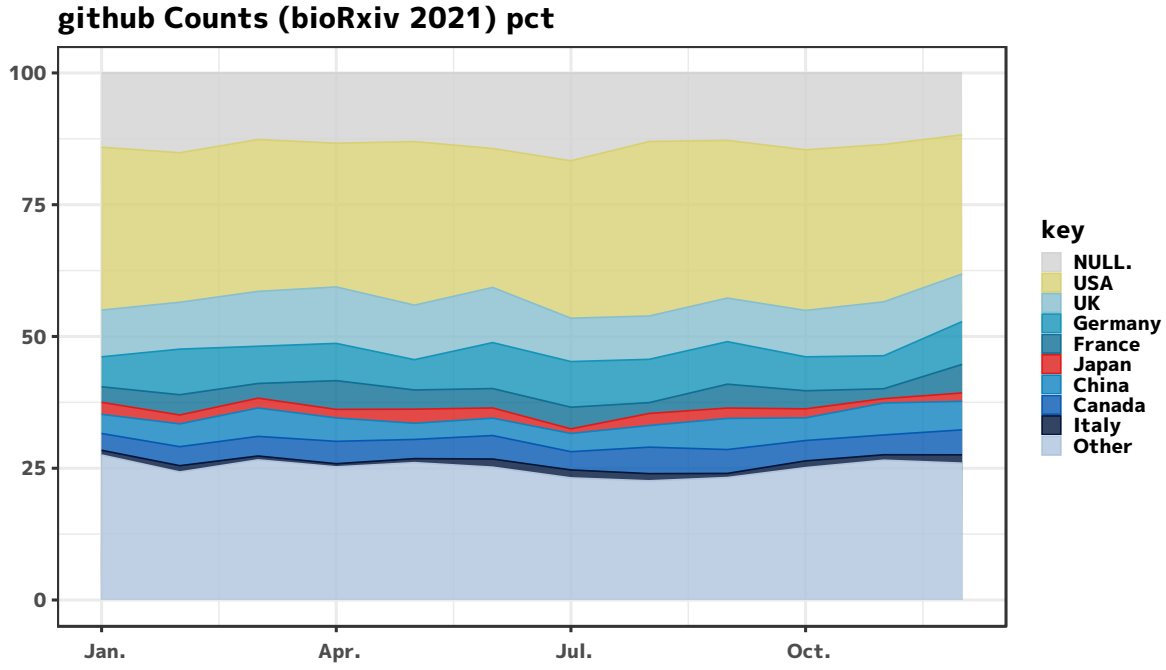

図 30: (bioRxiv)github 言及原稿数の推移（割合）

調査資料-324

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

2023 年 1 月

文部科学省 科学技術・学術政策研究所
データ解析政策研究室
林 和弘，小柴 等

〒100-0013　東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2393

Brief survey on the use of open source / data in research activities

Jan. 2023

HAYASHI Kazuhiro, KOSHIBA Hitoshi
Research-Unit for Data Application
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

https://doi.org/10.15108/rm324

https://www.nistep.go.jp