調査資料-342

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 2024

2024 年 09 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

小柴 等　　林 和弘

【調査研究体制】

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
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

【Authors】

KOSHIBA Hitoshi	Senior Research Fellow, Research Unit for Data Application, National Institute of Science and Technology Policy (NISTEP), MEXT
HAYASHI Kazuhiro	Director, 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.

小柴 等，林 和弘「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 2024」，NISTEP RESEARCH MATERIAL，No.342，文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/rm342

KOSHIBA Hitoshi, HAYASHI Kazuhiro “Brief survey on the use of open source / data in research activities 2024,” NISTEP RESEARCH MATERIAL, No.342, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/rm342

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 2024

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，過去の調査を踏襲し，「ＤＸによる研究活動の変化等」の把握を念頭に，研究活動にお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）国籍を割り付けた。

本調査では年の単位で 2010 年から 2023 年までの 24 時点を調査した。2022 年 9 月までのデータを対象とした前回調査と比較し，対象となる原稿数そのものは増加しているものの，Zenodo, figshare, github に関する言及原稿の割合は 2022 年以前と比較して相対的に大きな変化は認められなかった。

本調査では新たに分野差についても観察するために，arXiv の各原稿に割り付けられた分野，さらに，生物学，医学の分野においてそれぞれメジャーな 2 つのプレプリントサーバ，bioRxiv, medRxiv の 2023 年分原稿についても調査した。結果，DOI の普及数は arXiv, bioRxiv, medRxiv の単位ではほぼ同等で全体の 2 割程度，github 言及では差があるものの， bioRxiv で 2 割近く，medRxiv でも 1 割近くの原稿で言及が見られるなど，分野間での特徴がいくつか観察できた。また，arXiv 内でも物理，数学，情報系では差があることも確認できた。

数値や図表といった種別のオープンデータ利用が進んでいない点については，別途，実施したアンケート調査の結果なども念頭に読み取っていく必要がある。他方，オープンソースの代理変数として用いている github の言及の増加は「ＤＸによる研究活動の変化等」を表しており，業績評価などで考慮することなどが考えられる。

Brief survey on the use of open source / data in research activities 2024

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

ABSTRACT

This paper presents the results of a survey conducted to understand "changes in research activities due to DX (digital transformation)". The specific purpose of this survey is to investigate the use of open source data in research activities.

For this purpose, we investigated the number of mentions of open source and open data in manuscripts on arXiv, a major preprint server in the fields of physics and information sciences. In the survey, github was set as a proxy variable for open source, and Zenodo and figshare as proxy variables for open data. The DOI was also investigated as basic data for comparison. Using the email address given in the text as a clue, each manuscript was assigned a nationality (where assignable) and organized based on the year and month of first publication. In terms of years, the survey covered 24 time points from 2010 to 2023.

Compared to the previous survey, which covered data up to September 2022, the proportion of manuscripts mentioning Zenodo, figshare, and github did not change relatively significantly compared to the pre-2022 period, although the number of manuscripts covered itself increased. The survey further compared each arXiv manuscript by discipline. We also examined manuscripts from 2023 on the two major preprint servers, bioRxiv and medRxiv, in the fields of biology and medicine respectively, to investigate larger disciplinary differences.

As a result, we observed that the number of disseminated DOIs was almost equal in units of arXiv, bioRxiv, and medRxiv, around 20% of the total, while there were differences in github mentions, with nearly 20% of manuscripts in bioRxiv and nearly 10% in medRxiv having mentions, showing some characteristics between the fields. It was also confirmed that even within the arXiv, there were differences among physics, mathematics, and information sciences. The lack of progress in the use of open data for data types such as numerical values and charts needs to be interpreted in conjunction with the results of a separate questionnaire survey. On the other hand, the increase in the number of references to github, which is used as a proxy variable for open source, indicates "changes in research activities due to DX", which could be taken into account in performance evaluation.

目次

図目次

表目次

1　はじめに

本稿では「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）」における，「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，過去に実施した調査 [林 22] を更新し，2023 年までの状況を分析した結果について述べる。

社会のさまざまな活動と同じく，研究活動も時代とともに変化しており，それらの動向を把握しておくことは重要である。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」[1] では，「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）」において，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）」[2] 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，先行調査 [林 22] において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。その結果として，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。

今回は上記の先行調査 [林 22] から 1 年以上が経過したことを受け，前回調査が 2022 年 9 月時点までであったところ，データを更新して 2023 年いっぱいまでの状況を年単位で把握する。また，arXiv の原稿に割り振られた「分野 (Category)」の単位での推移を確認した他，おなじくプレプリントサーバである bioRxiv, MedRxiv でも同様の手法で単年 (2023 年) 分についても分析を試み，分野特性の把

---

[1] 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.)

[2] 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.)

握を試みた。さらに，過去分についてもテキスト抽出が困難だった事例があったことを念頭に手法を見直し，精度の向上を試みた。

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。

2　要件

基本的に先行調査 [林 22] を踏襲して調査を実施した。また，これらに加えて分野別の分析を新たに追加した。ここでは，先行調査 [林 22] の内容についても簡単に言及しつつ，新たに加えた分野別の分析について述べる。

2.1　データ

データソースには主に物理・情報系の分野を対象に 1990 年代から運用されてきたプレプリントサーバ arXiv [3] [林 20] に搭載された原稿を対象とする。

プレプリントは基本的に適時更新されるものであり，arXiv に搭載される原稿も複数のバージョンが存在する。これにより原稿の「最新版」を利用すると分析のタイミングによって結果が変化する (図 1) ことから，初版を分析の対象とする。

実際のデータは arXiv のインストラクション[4] に従い，Kaggle [5]のプロジェクト[6] において，Google Cloud Storage (GCS) を通じて公開されている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 2024 年 3 月にデータを収集し，2008 年 1 月から収集時点における最新版までの

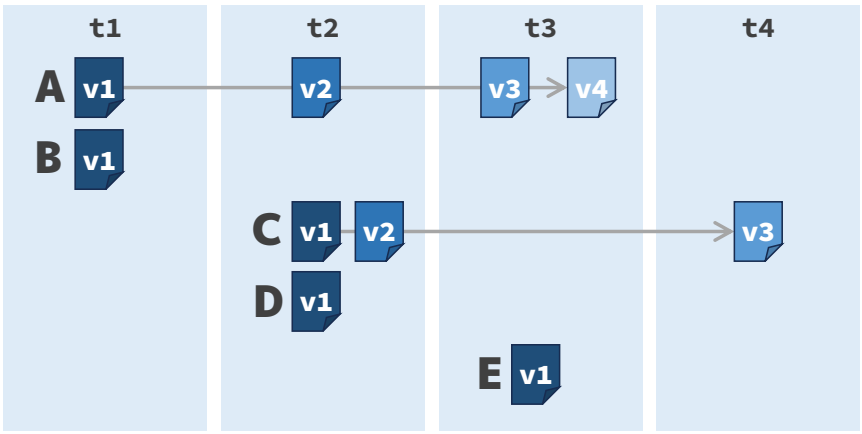

図 1: 観測時点とバージョンによる原稿数の変化

3) 読み：あーかいぶ，https://arxiv.org/ (Last accessed: 2024.05.21)
4) https://arxiv.org/help/bulk_data
5) https://www.kaggle.com/
6) https://www.kaggle.com/datasets/Cornell-University/arxiv

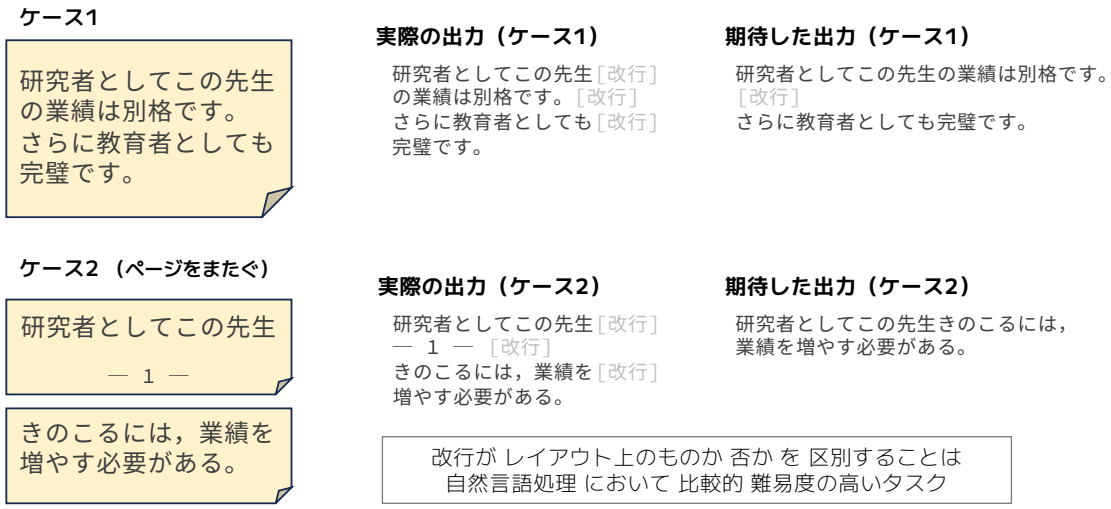

図 2: PDF からのテキスト抽出における課題

データを取得した。その上で，最終的に 2010 年から 2023 年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 1,814,100 件となった。

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[7]（図 2）。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，図や表のデータの問題もあり，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。

国籍の割り付け

先行調査 [林 22] と同様に，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

[7] 大規模言語モデル (LLM: Large Language Model) を用いることで判別の精度をかなり向上できる見込みがあるが，データ量を鑑みると現状においては主に金銭的コストの観点において現実的ではない。

図 3: URL における FQDN

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

先行調査 [林 22] と同様に，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，ここで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) [8] を原稿単位[9]で調べた分布を示す。5 位に「www」，10 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	Count	#	FQDN	Count	#	FQDN	Count
1	github.com	176,109	11	pos.sissa.it	10,853	21	proceedings.neurips.cc	6,512
2	doi.org	147,075	12	github	9,844	22	huggingface.co	6,447
3	arxiv.org	119,532	13	dl.acm.org	8,820	23	creativecommons.org	6,144
4	dx.doi.org	47,779	14	proceedings.mlr.press	8,465	24	link.springer.com	5,783
5	www	24,982	15	doi.acm.org	7,742	25	ieeexplore.ieee.org	5,545
6	www.sciencedirect.com	22,998	16	sites.google.com	7,564	26	archive.ics.uci.edu	5,398
7	link.aps.org	15,543	17	www.jstor.org	7,454	27	www.cosmos.esa.int	5,378
8	en.wikipedia.org	13,847	18	onlinelibrary.wiley.com	7,397	28	www.youtube.com	5,368
9	openreview.net	11,130	19	arxiv	6,702	29	stacks.iop.org	5,247
10	doi	10,870	20	www.kaggle.com	6,700	30	youtu.be	5,218

FQDN: Fully Qualified Domain Name

* 2010.01〜2023.12まで。件数は原稿数（何件の原稿に出現したか）

[8] 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 (図 3)。複数の点から厳密には FQDN とは言えないが，本報告書では便宜上これを FQDN と呼称している。

[9] 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**。

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。

そこで先行調査 [林 22] と同様に，カウント対象を以下の通り設定した。いわゆるオープンデータとして，Zenodo[10]，figshare[11]を設定する。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[12]。

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「`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

figshare について，先行調査 [林 22] では “ URL 中に「`10.6084/m9.figshare`」の文字列を含むもの ” という条件であったところ，“ URL 中に「`figshare`」の文字列を含むもの ” と，幅が広がっている点に注意を要する。

なお，先行調査 [林 22] にもあるとおり，Zenodo, figshare はそれぞれ DOI も発行しているため 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，少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。

2.2 分野別の分析

今回は国・地域に加えて分野ごとの分析も考慮する。分野については，arXiv にあらかじめ設定された分野を用いる。

arXiv では cs.DL (Digital Libraries) など，155 の Category が用意されており[13]，投稿時に各原稿に

10) https://zenodo.org/

11) https://figshare.com/

12) 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。

13) arXiv Category Taxonomy, https://arxiv.org/category_taxonomy (Last Accessed 2024.05.22)

ついて，ひとつ以上の Category を割り付けることになっている。

155 の Category はさらに，以下に示す 8 つの分野にまとめられている。ただし，分野が内包する Category の数には大きな偏りがある点に注意が必要である。

- Computer Science (cs) 40 分野
- Economics (econ) 3 分野
- Electrical Engineering and Systems Science (eess) 4 分野
- Mathematics (math) 32 分野
- Physics (physics) 51 分野
- Quantitative Biology (q-bio) 10 分野
- Quantitative Finance (q-fin) 9 分野
- Statistics (stat) 6 分野

Category は 155 と多いため，本調査ではこの 8 分野の単位で分野別の状況把握も試みる。この際，カウントは分野単位で整数カウントとする。たとえば，cs（計算機科学）分野の Category である cs.DL, cs.LG, cs.AI と stat（統計学）分野の Category である stat.ML の 4 Category が登録された原稿があったとする。この場合，分野としては cs（計算機科学）と stat（統計学）の 2 分野になるので，cs に 1 件，stat に 1 件を計上する。

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，全体および国別，分野別の原稿数推移などの基礎的性質や，その中で DOI の記載がある原稿数推移を確認した。結果を図 4 から図 12 に示す。なお，国については先行調査 [林 22] で採用した 7 カ国を踏襲した。

3.1.1　国別の状況

今回は各原稿について最大ひとつの国（あるいは org, com などの属性）を割り振っているため，積算は原稿総数に一致する。

そこでまず，単純な原稿数についての国別（および積み上げ結果である総数）について図 4 及び図 6 をみると，原稿数及びその中における DOI 言及原稿数は前回調査 [林 22] と変わらず着実に伸びている。DOI 言及原稿の割合については，2023 年の原稿数が約 20 万件，うち DOI 記載のある原稿数は 4 万件であることから約 2 割の原稿が DOI 言及を行っていることになる。これは 2022 年とほぼ同程度の割合で，実数としては伸びているものの割合は横ばいと言える。

図 5，図 7 は国別の比較を意識して，全原稿および DOI 言及原稿の国別割合を示した。図 5 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，中国は 2022 年 9 月までを分析した前回調査 [林 22] との差分で見ても着実に割合を増やしており躍進が目立つ。図 7 についても前回調査 [林 22] と同じく図 5 と概ね似た傾向が見える他，日本を示す朱色が全体傾向（図 5）ほどには目立っていない。

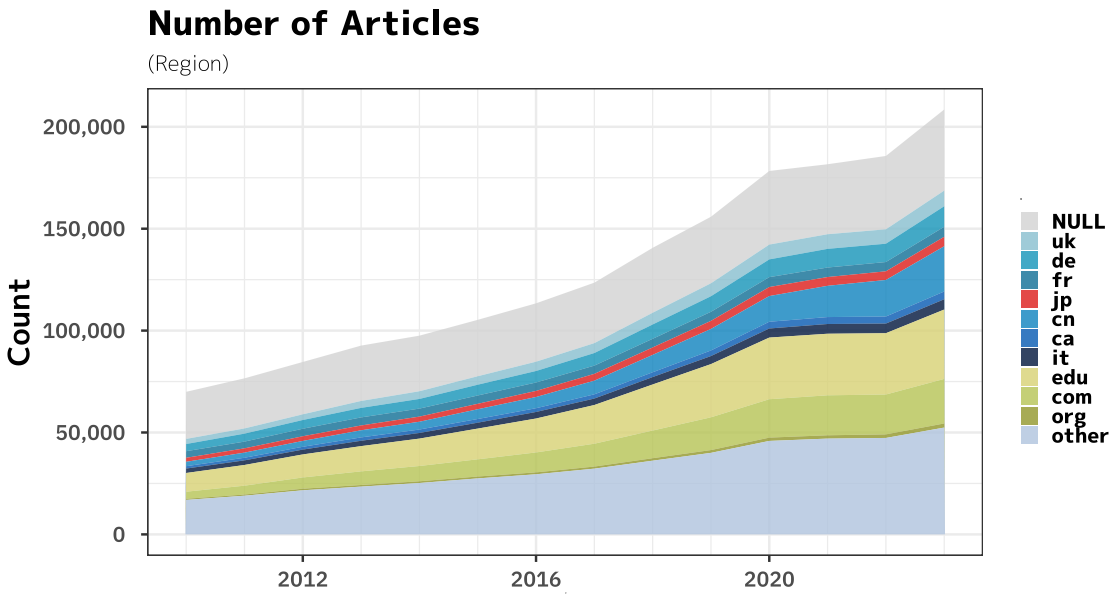

図 4: 原稿数の推移（国別）

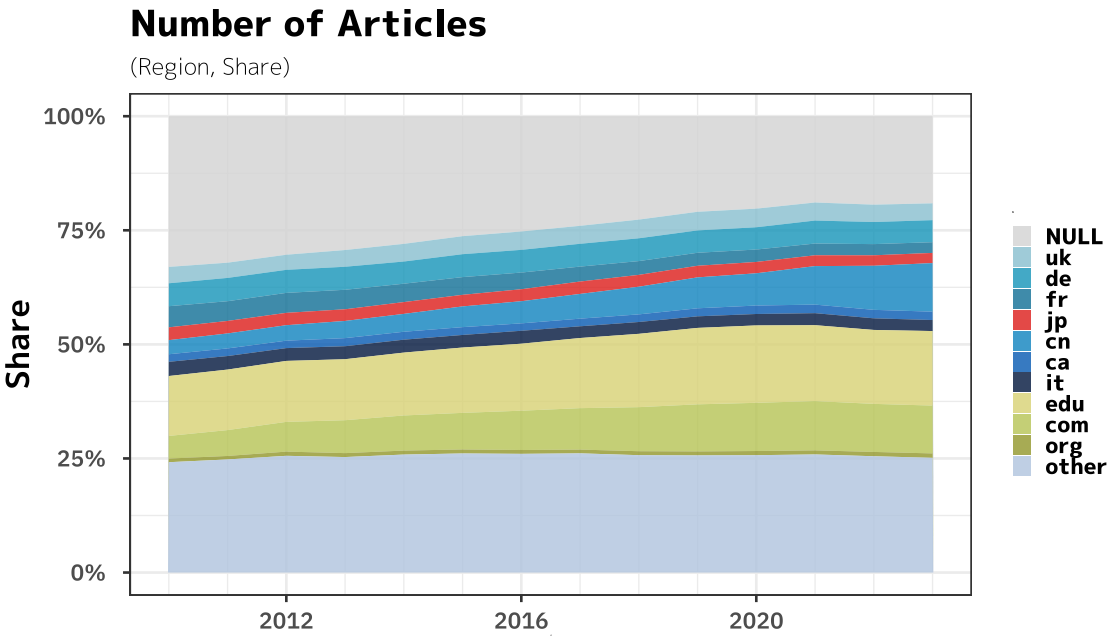

図 5: 原稿数の推移（国別，割合）

表 2: 原稿数の推移（国別）

Whole	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	1,161	1,280	1,384	1,645	1,694	1,800	1,841	2,068	2,349	2,742	3,308	3,408	3,454	3,824
China (cn)	2,160	2,512	2,861	3,487	3,816	4,781	5,500	6,694	8,523	10,564	12,596	15,326	17,933	22,191
Germany (de)	3,514	3,907	4,271	4,673	4,738	5,287	5,644	6,175	7,023	7,635	8,700	9,151	8,984	10,081
France (fr)	3,241	3,313	3,705	3,959	3,931	4,089	4,145	4,046	4,250	4,447	4,857	4,675	4,555	4,901
Ilaly (it)	2,177	2,270	2,372	2,635	2,761	2,905	3,212	3,175	3,646	3,950	4,461	4,781	4,725	5,037
Japan (jp)	1,965	2,099	2,292	2,369	2,523	2,674	2,978	3,349	3,649	3,962	4,433	4,309	4,259	4,646
UK (uk)	2,515	2,558	2,802	3,400	3,777	4,159	4,565	4,868	5,763	6,362	7,281	7,187	7,070	7,683
com	3,472	4,339	5,519	6,684	7,505	8,444	9,770	11,219	13,579	16,081	18,837	19,619	19,560	21,854
edu	9,181	10,153	11,299	12,384	13,434	15,071	16,644	18,982	22,594	26,106	30,251	30,203	30,084	34,045
org	559	589	779	798	853	925	956	1,021	1,223	1,365	1,628	1,631	1,755	2,024
other	16,915	18,978	21,640	23,444	25,210	27,472	29,488	32,261	36,163	40,010	45,862	47,002	47,324	52,409
NULL	23,097	24,576	25,679	27,163	27,275	27,673	28,637	29,665	31,854	32,642	36,115	34,338	35,989	39,797
Total	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

表 3: DOI 言及原稿数の推移（国別）

DOI	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	8	6	17	23	50	73	101	138	187	273	442	532	678	799
China (cn)	5	1	17	39	61	93	104	207	450	699	1,078	1,707	2,117	2,932
Germany (de)	32	45	87	83	148	198	278	420	807	979	1,324	1,735	1,951	2,378
France (fr)	22	33	41	87	118	147	167	251	347	487	587	710	775	908
Ilaly (it)	12	23	31	39	89	117	150	211	360	490	680	856	914	1,087
Japan (jp)	1	7	17	14	21	53	45	121	221	297	390	512	542	694
UK (uk)	16	19	34	81	114	149	222	409	606	744	1,123	1,350	1,483	1,677
com	21	29	76	151	197	301	436	726	1,285	1,797	2,684	3,392	3,631	4,306
edu	69	78	134	236	331	535	779	1,317	2,229	2,875	4,380	5,213	5,744	7,164
org	6	4	18	30	45	73	86	154	217	261	432	570	647	785
other	94	118	289	472	610	929	1,254	2,120	3,321	4,535	6,716	8,446	9,156	10,975
NULL	116	126	270	426	568	752	1,229	1,604	2,357	2,892	4,118	4,939	5,566	6,601
Total	402	489	1,031	1,681	2,352	3,420	4,851	7,678	12,387	16,329	23,954	29,962	33,204	40,306

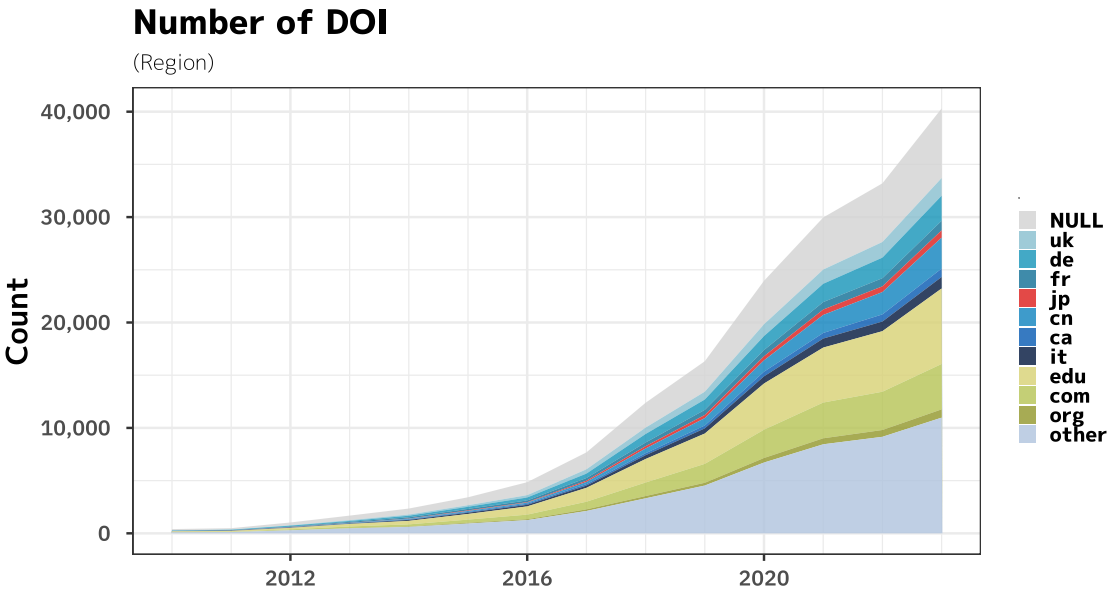

図 6: DOI 言及原稿数の推移（国別）

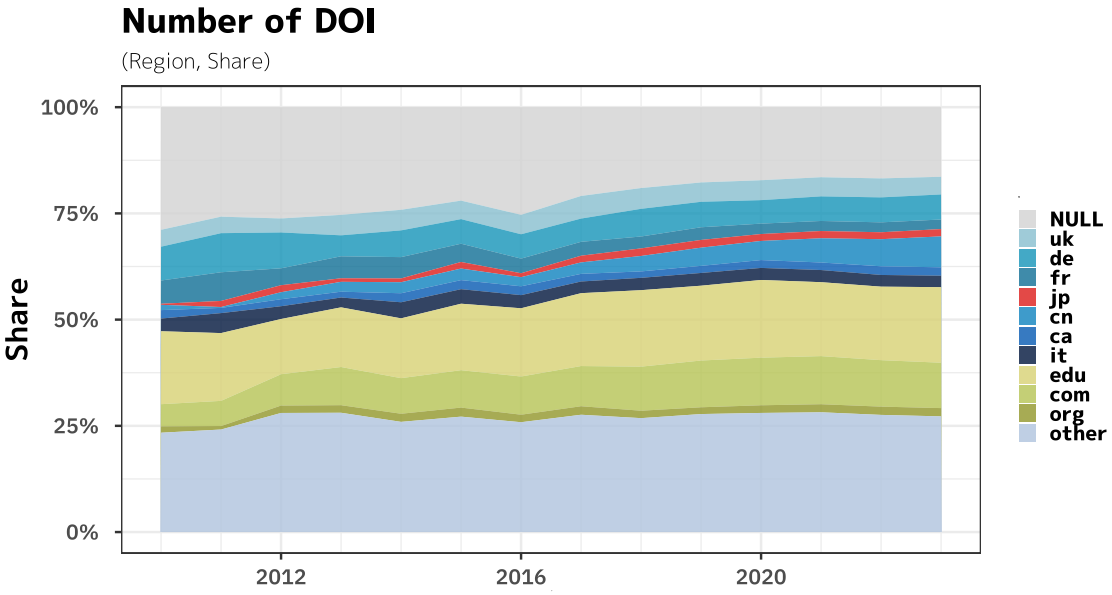

図 7: DOI 言及原稿数の推移（国別，割合）

3.1.2　分野別の状況

次に分野別の原稿数，うち DOI 言及原稿数について示す。

前述とおり，分野については整数カウントのため積算は総原稿数を超える。割合 (Share) についても積算は 1 を超える。

なお，2010 年から 2023 年までの全原稿を対象に，原稿辺りの平均分野数を算出すると約 1.2 件である。また，分野間の共起関係（ひとつの原稿に対して同時に割り振られる分野の組み合わせ）について図 8 に示した。図 8 中，1 分野のみに属する原稿については当該分野から当該分野へ接続することで表現している。これを見ると，econ（経済学）, q-fin（金融学）などの元々の規模が小さい分野においては cs とのつながりは相対的に大きい。また，規模の大きな physics（物理学）, cs（計算機科学）, math（数学） を見たときに，physics（物理学）は相対的には多分野との共起割合が小さく独立性が高い。

前述の通り，分野については整数カウントを採用しており，各分野が必ずしも独立ではなく，分野の規模も異なるため，結果を読み取る際にはこれら規模と共起に関する考慮が不可欠と言える。

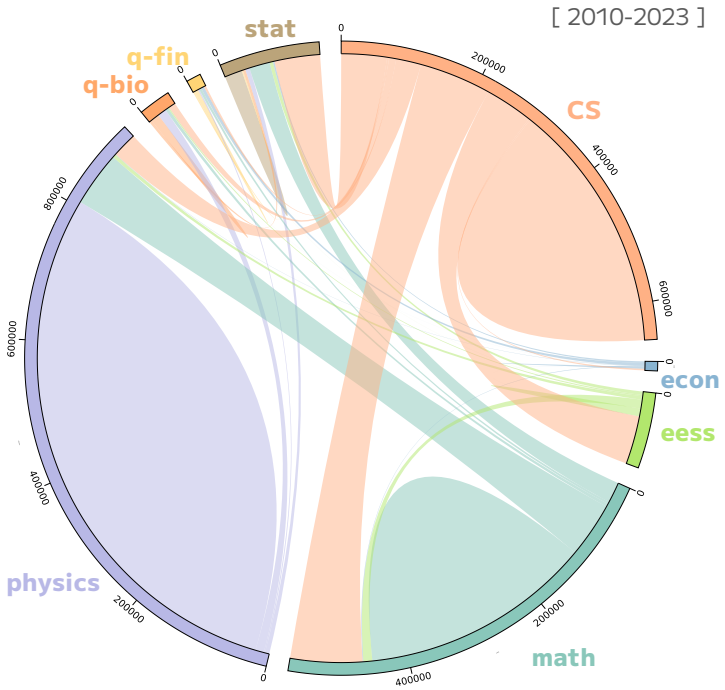

図 8: 分野の共起関係（2010 年〜2030 年まで積算）

単純な分野の傾向に目を移し，図 9，図 10 をみると，2015 年くらいから cs（計算機科学）が数を伸ばし，2021 年には arXiv の元々の発祥分野である phisics（物理学）を超え，その後も順調に増加している。

DOI について，図 11，図 12 をみると，各分野で順調に数・割合を伸ばしており，増加の割合も安定しているように見受けられる。q-fin（金融学）, q-bio（計量生物学）については相対的に付与割合が高く 3 割程度となっているが，もともと全体数が少なく，かつ，cs や math と共起しやすいことなどから，これが分野の特性を表すものかどうかの考察は難しい。

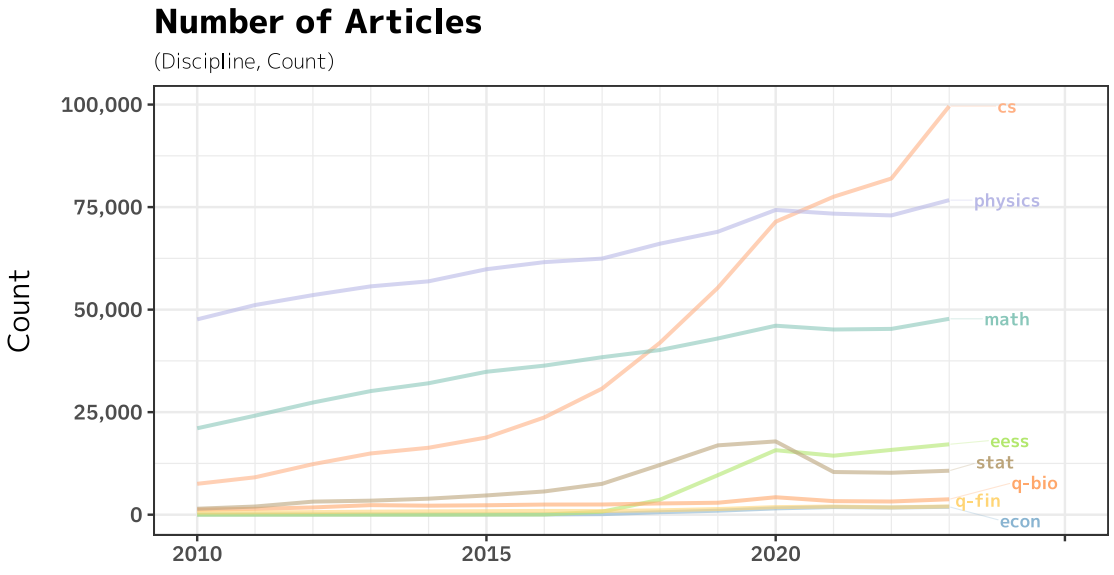

図 9: 原稿数の推移（分野別）

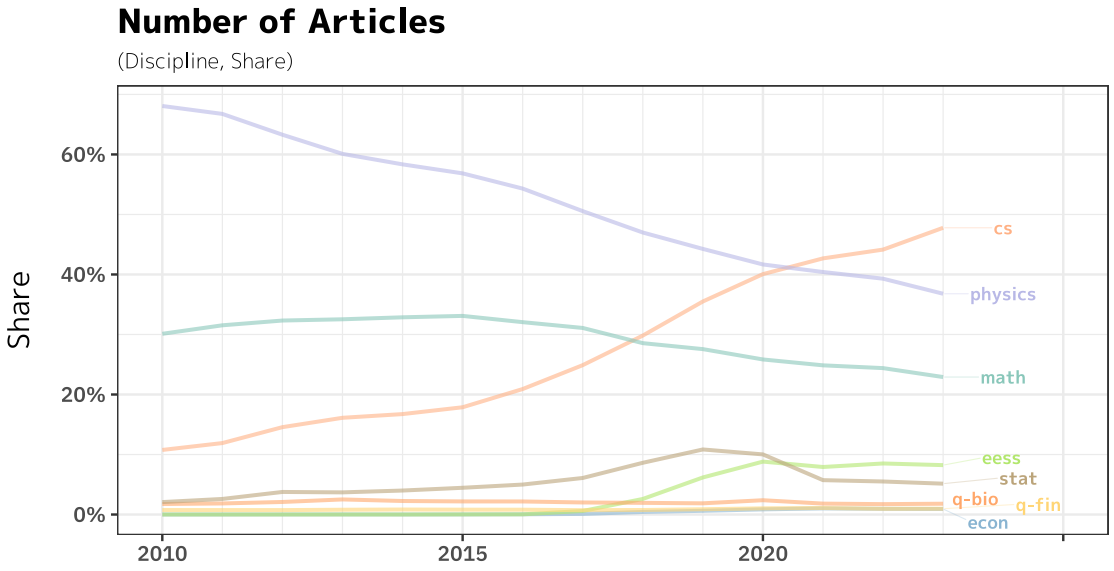

図 10: 原稿数の推移（分野別，割合）

表 4: 原稿数の推移（分野別）

Whole	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	7528	9118	12317	14933	16321	18819	23709	30749	41930	55332	71433	77520	81974	99614
econ	3	4	10	15	10	20	47	126	603	964	1537	1859	1746	1921
eess	3	4	5	7	3	23	36	730	3652	9632	15727	14385	15798	17161
math	21067	24145	27347	30139	32046	34845	36342	38400	40153	42960	46077	45151	45300	47760
physics	47622	51116	53542	55673	56898	59846	61577	62445	66077	68985	74305	73400	72970	76699
q-bio	1228	1414	1771	2330	2198	2300	2470	2482	2737	2911	4254	3302	3216	3755
q-fin	523	571	627	741	815	852	914	897	1063	1373	1845	1983	1777	2019
stat	1451	1988	3183	3420	3899	4694	5664	7526	12125	16902	17866	10405	10224	10718
all	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

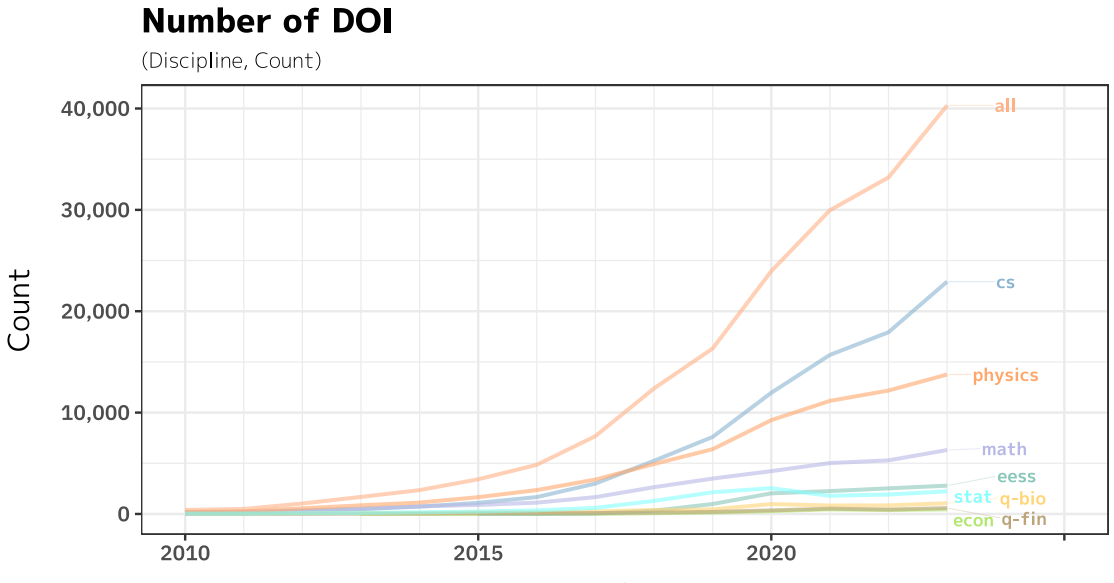

図 11: DOI 言及原稿数の推移（分野別）

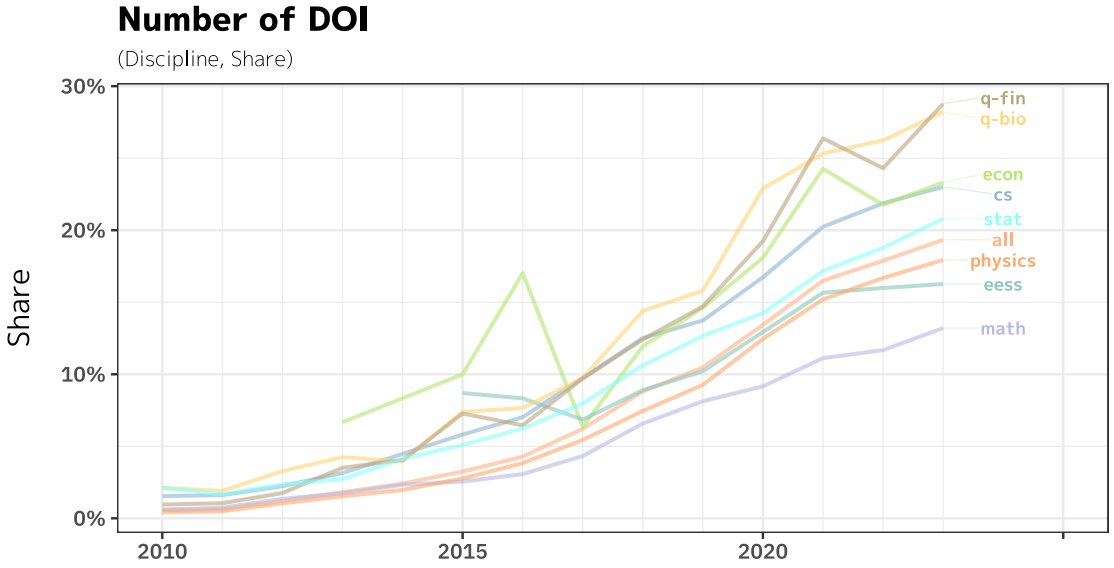

図 12: DOI 言及原稿数の推移（分野別，割合）

表 5: DOI 言及原稿数の推移（分野別）

DOI	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	116	147	275	470	730	1,095	1,664	2,997	5,249	7,599	11,954	15,690	17,929	22,923
econ	0	0	0	1	0	2	8	8	72	141	278	451	380	448
eess	0	0	0	0	0	2	3	50	325	983	2,034	2,253	2,526	2,791
math	128	177	371	524	746	889	1,115	1,661	2,642	3,492	4,222	5,021	5,289	6,306
physics	198	252	557	851	1,122	1,656	2,356	3,396	4,927	6,390	9,255	11,155	12,175	13,758
q-bio	26	27	58	99	87	170	189	242	394	460	974	836	844	1,059
q-fin	5	6	11	26	33	62	59	87	132	202	355	523	432	581
stat	31	33	75	93	161	239	353	600	1,287	2,140	2,546	1,786	1,920	2,230
all	402	489	1,031	1,681	2,352	3,420	4,851	7,678	12,387	16,329	23,954	29,962	33,204	40,306
Whole	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

表 6: Zenodo 言及原稿数の推移（国別）

Zenodo	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	0	0	0	0	0	3	1	10	4	20	33	41	62	97
China (cn)	0	0	0	0	0	0	2	1	5	15	34	68	111	190
Germany (de)	0	0	0	0	0	4	14	17	63	79	120	192	231	349
France (fr)	0	0	0	0	1	0	1	1	5	19	44	54	82	85
Ilaly (it)	0	0	0	0	0	0	2	2	7	13	38	55	81	104
Japan (jp)	0	0	0	0	0	1	3	4	3	11	14	21	39	41
UK (uk)	0	0	0	0	0	3	22	48	45	74	144	193	189	251
com	0	0	0	0	2	1	6	23	44	73	119	184	231	285
edu	0	0	0	0	0	6	25	70	124	187	251	395	477	640
org	0	0	0	0	0	2	3	7	15	15	23	40	43	62
other	0	0	0	0	1	15	25	74	124	223	425	701	855	1,106
NULL	0	0	0	0	3	10	34	56	101	129	228	303	459	567
Total	0	0	0	0	7	45	138	313	540	858	1,473	2,247	2,860	3,777

3.2 オープンデータ利用

3.2.1 国別の状況

前回調査 [林 22] を踏襲し，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 13 から図 16 に示す。これらの結果を見ると，基本的に前回調査と同じ傾向と言える。まず，figshare については言及原稿数自体は着実に増えているものの，絶対数が少ないため安定せず，傾向を読み取りにくい。次に，Zenodo に目を向けると，中国がわずかずつではあるが割合を伸ばしている傾向にあり，かつ，これらは DOI における傾向と類似している。表 8 に示したとおり，2023 年分を対象として各国単位で全原稿に占める DOI・Zenodo 言及原稿の割合を見た場合，中国の割合は日本とほぼ同レベルを示しており，これまでの増加率を維持すると 2024 年には国単位での言及割合で日本を抜く可能性が高い。

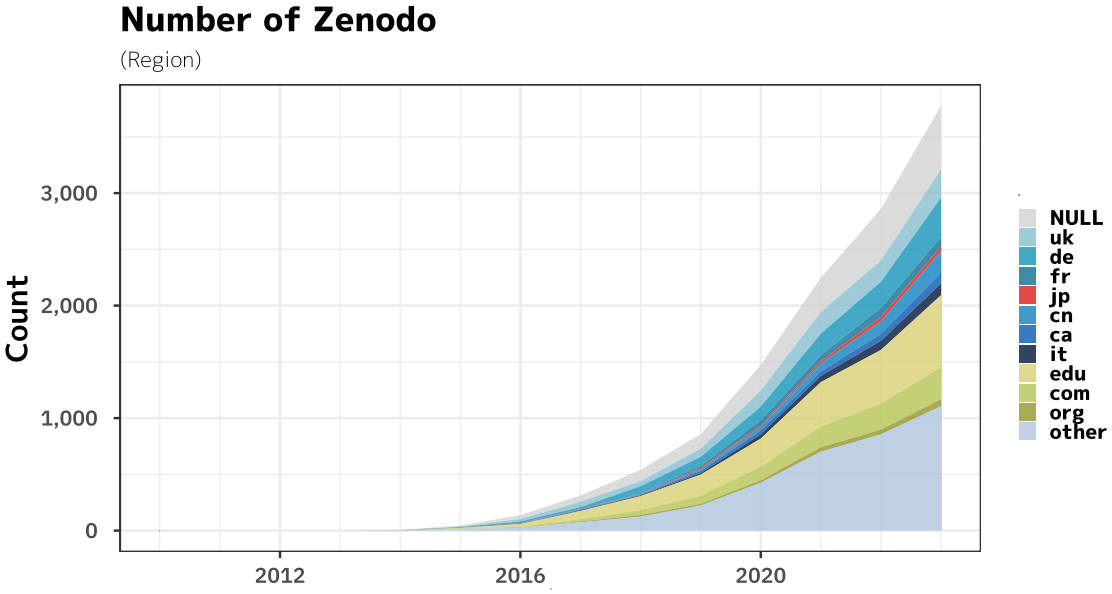

図 13: Zenodo 言及原稿数の推移（国別）

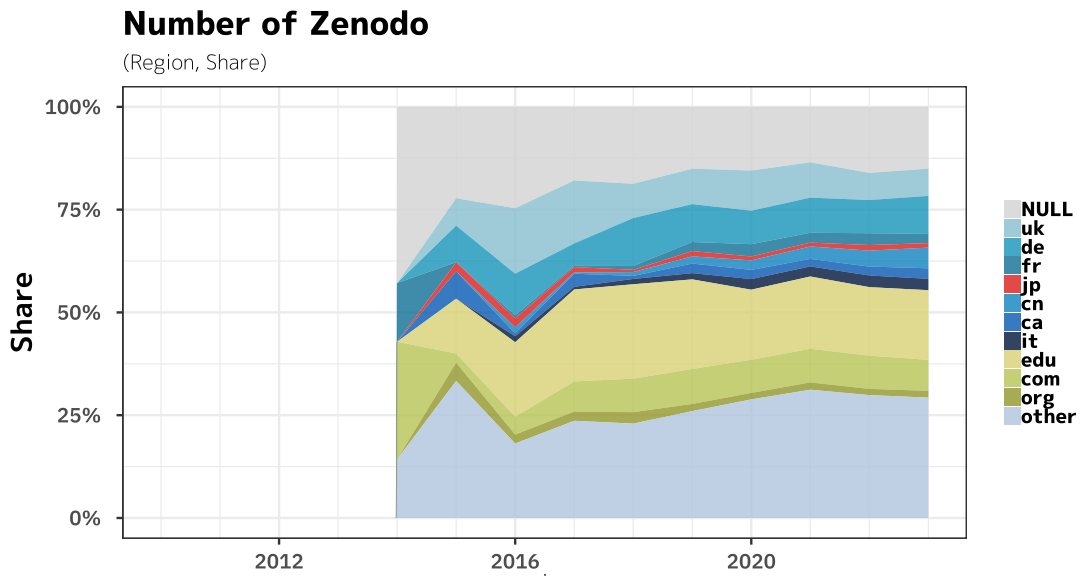

図 14: Zenodo 言及原稿数の推移（国別，割合）

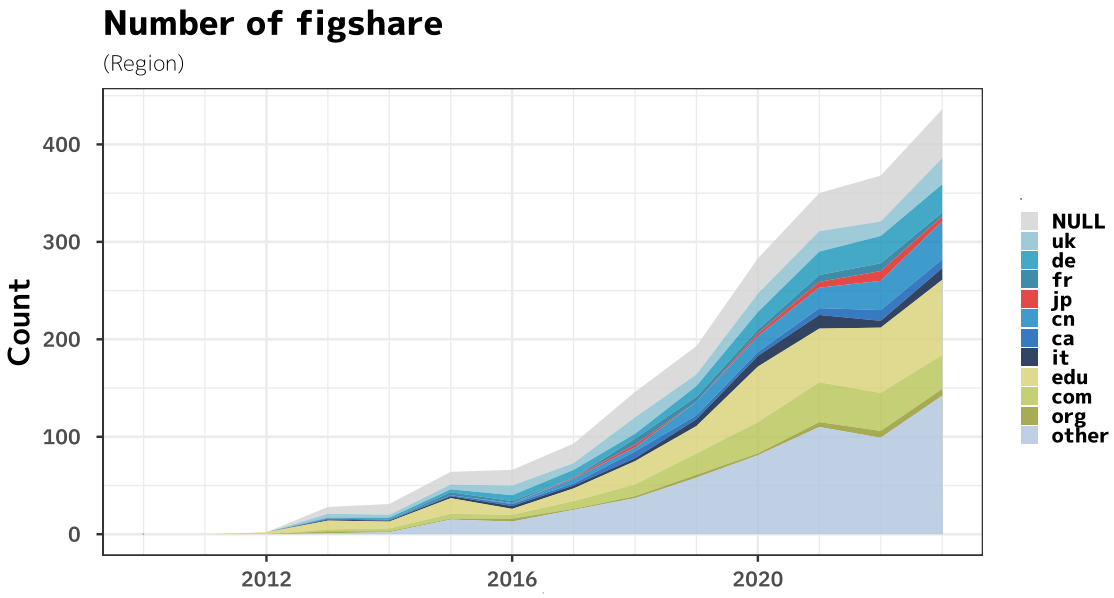

図 15: figshare 言及原稿数の推移（国別）

表 7: figshare 言及原稿数の推移（国別）

figshare	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	0	0	0	1	1	1	2	2	7	4	4	7	11	9
China (cn)	0	0	0	0	0	0	1	5	4	14	16	21	30	39
Germany (de)	0	0	0	1	2	3	6	8	5	11	18	24	28	29
France (fr)	0	0	0	0	0	3	2	0	6	5	4	7	8	4
Ilaly (it)	0	0	0	1	1	2	3	3	3	7	11	14	7	12
Japan (jp)	0	0	0	0	0	0	0	1	3	0	3	6	10	5
UK (uk)	0	0	0	4	3	5	10	7	17	12	19	21	15	27
com	0	0	0	2	3	5	4	8	12	22	32	41	39	35
edu	0	0	2	9	7	16	6	13	24	28	57	55	67	77
org	0	0	0	2	1	1	3	1	2	3	2	5	7	7
other	0	0	0	1	2	15	13	25	37	58	81	110	99	142
NULL	0	0	0	7	11	13	16	20	26	29	36	39	47	50
Total	0	0	2	28	31	64	66	93	146	193	283	350	368	436

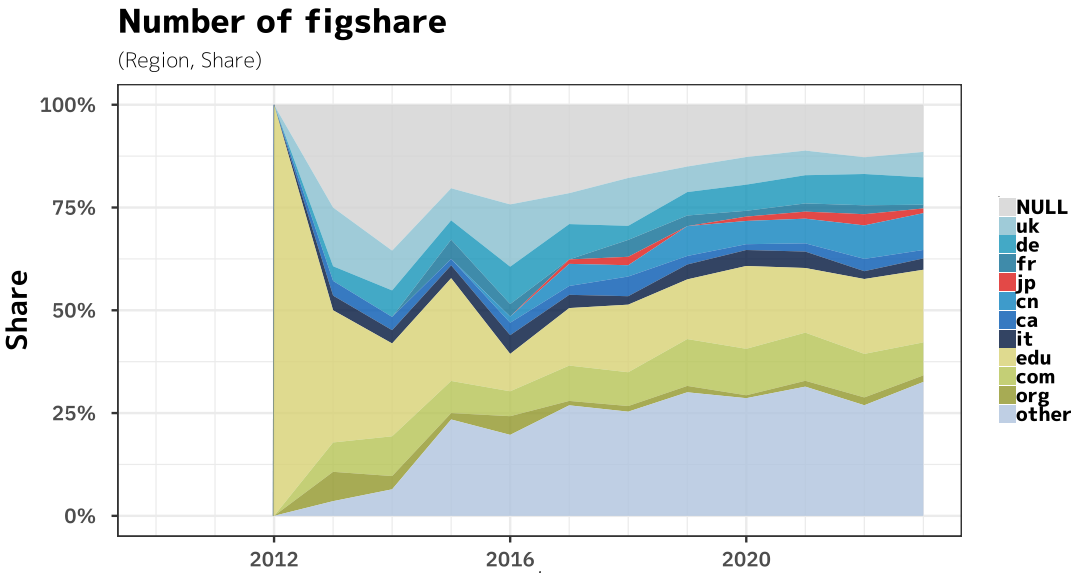

図 16: figshare 言及原稿数の推移（国別，割合）

表 8: 国別の原稿数と DOI，Zenodo 言及原稿数（2023 年分）

	All	DOI	Zenodo
ca (Canada)	3,824	799 (20.9%)	97 (2.5%)
cn (China)	22,191	2,932 (13.2%)	190 (0.9%)
de (Germany)	10,081	2,378 (23.6%)	349 (3.5%)
fr (France)	4,901	908 (18.5%)	85 (1.7%)
it (Italy)	5,037	1,087 (21.6%)	104 (2.1%)
jp (Japan)	4,646	694 (14.9%)	41 (0.9%)
uk (UK)	7,683	1,677 (21.8%)	251 (3.3%)
com	21,854	4,306 (19.7%)	285 (1.3%)
edu	34,045	7,164 (21.0%)	640 (1.9%)
org	2,024	785 (38.8%)	62 (3.1%)
other	52,409	10,975 (20.9%)	1,106 (2.1%)
NULL	39,797	6,601 (16.6%)	567 (1.4%)

* 2023年分，丸括弧内は割合

3.2.2 分野別の状況

分野別での結果を図 17 から図 20 に示す。分野の単位で見ると，Zenodo，figshare のどちらでも，q-bio（定量生物学）のシェアが大きい。次に，Zenodo については，physics（物理学）, cs（計算機科学） が同程度の割合となっている。physics（物理学） と cs（計算機科学） の共起はそれほど大きなものではないため，physics（物理学）, cs（計算機科学）それぞれの分野で同程度に利用が進んでいると考えられる。一方，math（数学）はそもそも学問分野の特性としてデータを直接生成・分析するようなケースが多くはないためか，言及原稿のシェアは小さい。

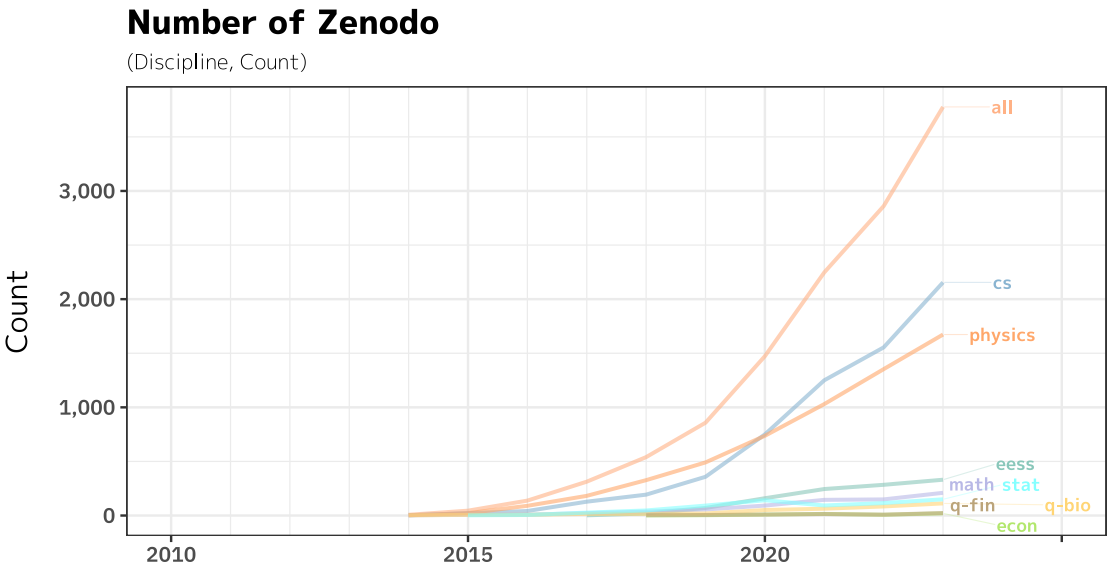

図 17: Zenodo 言及原稿数の推移（分野別）

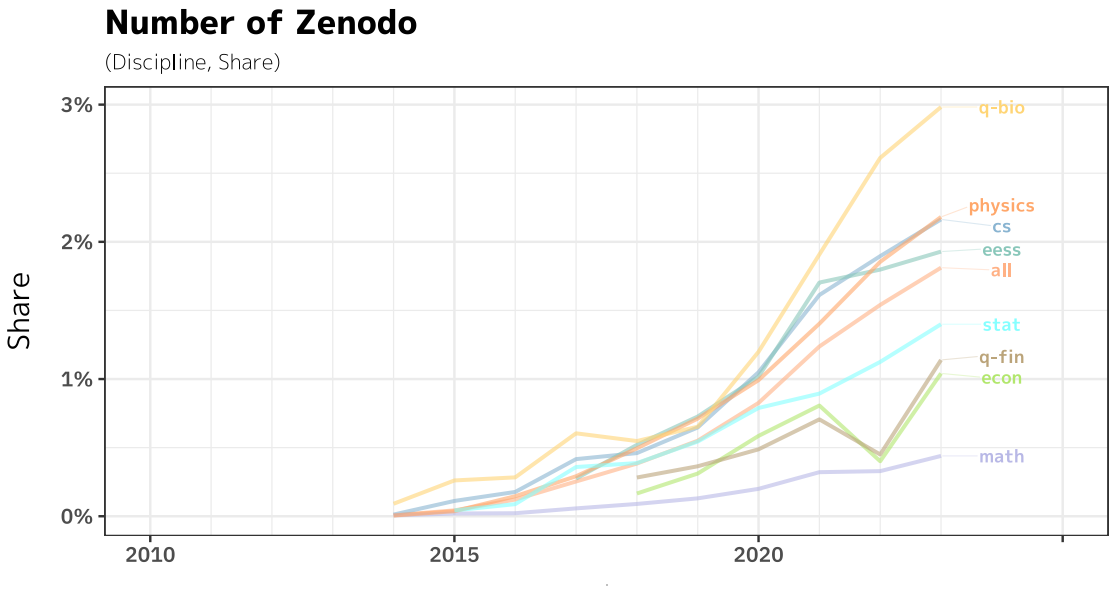

図 18: Zenodo 言及原稿数の推移（分野別，割合）

表 9: Zenodo 言及原稿数の推移（分野別）

Zenodo	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	0	0	0	0	2	21	42	128	193	358	750	1,250	1,554	2,155
econ	0	0	0	0	0	0	0	0	1	3	9	15	7	20
eess	0	0	0	0	0	0	0	2	19	70	160	245	284	331
math	0	0	0	0	2	7	8	22	36	56	92	145	149	210
physics	0	0	0	0	3	22	90	182	327	491	737	1,030	1,353	1,673
q-bio	0	0	0	0	2	6	7	15	15	19	51	63	84	112
q-fin	0	0	0	0	0	0	0	0	3	5	9	14	8	23
stat	0	0	0	0	0	2	5	27	47	92	141	93	115	150
all	0	0	0	0	7	45	138	313	540	858	1,473	2,247	2,860	3,777
amount	0	0	0	0	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

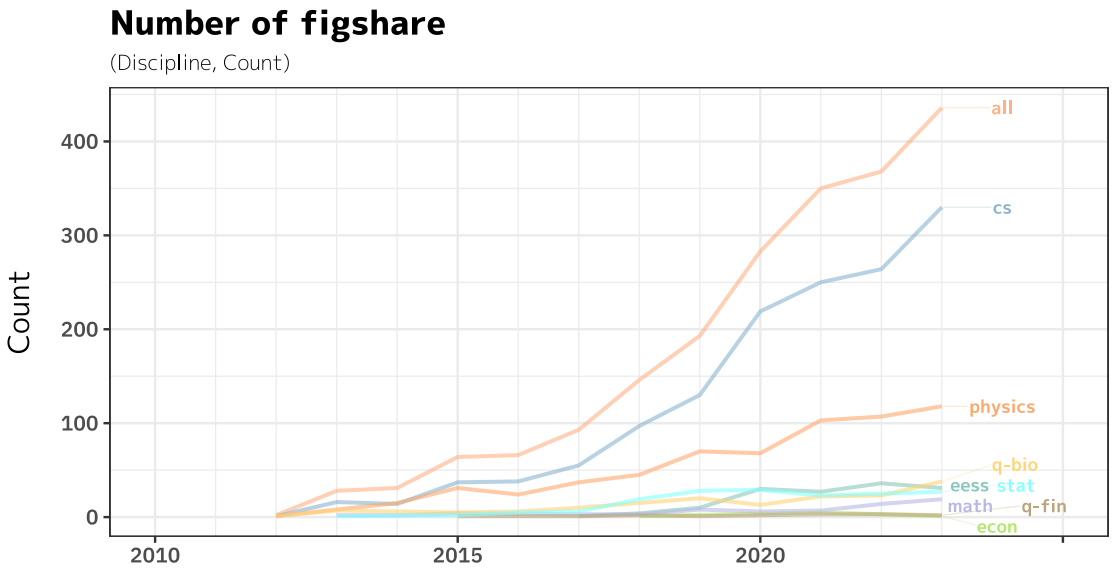

図 19: figshare 言及原稿数の推移（分野別）

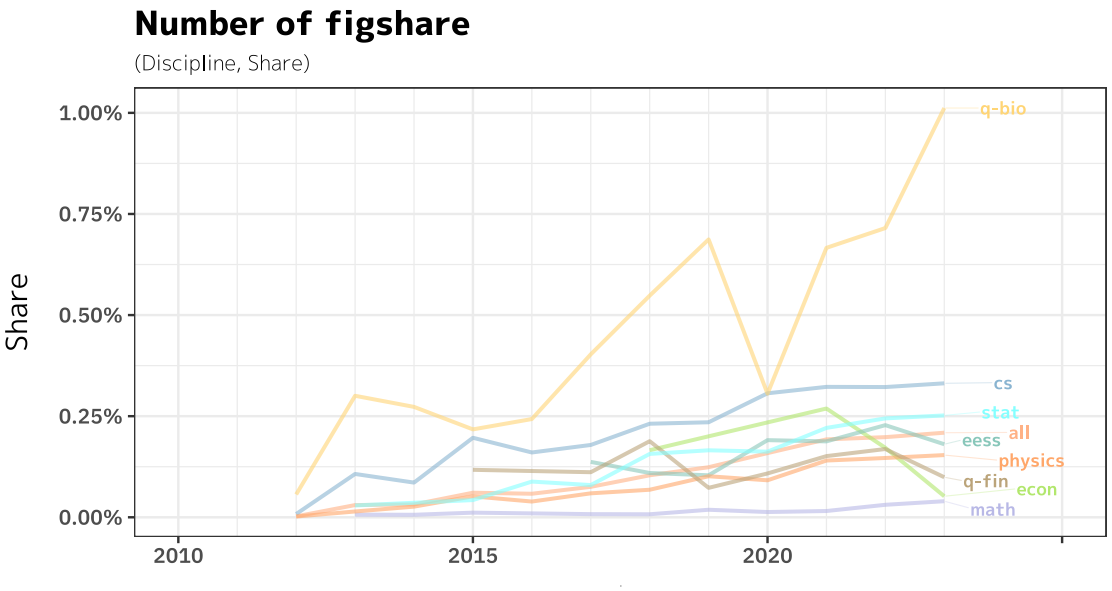

図 20: figshare 言及原稿数の推移（分野別，割合）

表 10: figshare 言及原稿数の推移（分野別）

figshare	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	0	0	1	16	14	37	38	55	97	130	219	250	264	330
econ	0	0	0	0	0	0	0	0	1	0	0	5	3	1
eess	0	0	0	0	0	0	0	1	4	10	30	27	36	31
math	0	0	0	2	2	4	0	3	3	8	6	7	14	19
physics	0	0	1	8	15	31	24	37	45	70	68	103	107	118
q-bio	0	0	1	7	6	5	6	10	15	20	13	22	23	38
q-fin	0	0	0	0	0	1	0	1	2	1	2	3	3	2
stat	0	0	0	1	0	2	5	6	19	28	29	23	25	27
all	0	0	2	28	31	64	66	93	146	193	283	350	368	436
amount	0	0	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

3.3　OSS 利用

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。

3.3.1　国別の状況

結果を図 21 及び図 22 に示す。

2022 年と 2023 年を比較した際に，単純な原稿数が 1.1 倍に対して github 言及原稿数は 1.3 倍であり，投稿原稿数の伸び以上に増えてきている。シェアで見たときには中国の存在感がますます大きくなっている一方，多くが米国の大学と考えられる edu の割合が低下している点が興味深い。その他については割合に大きな変化が見られず安定している。

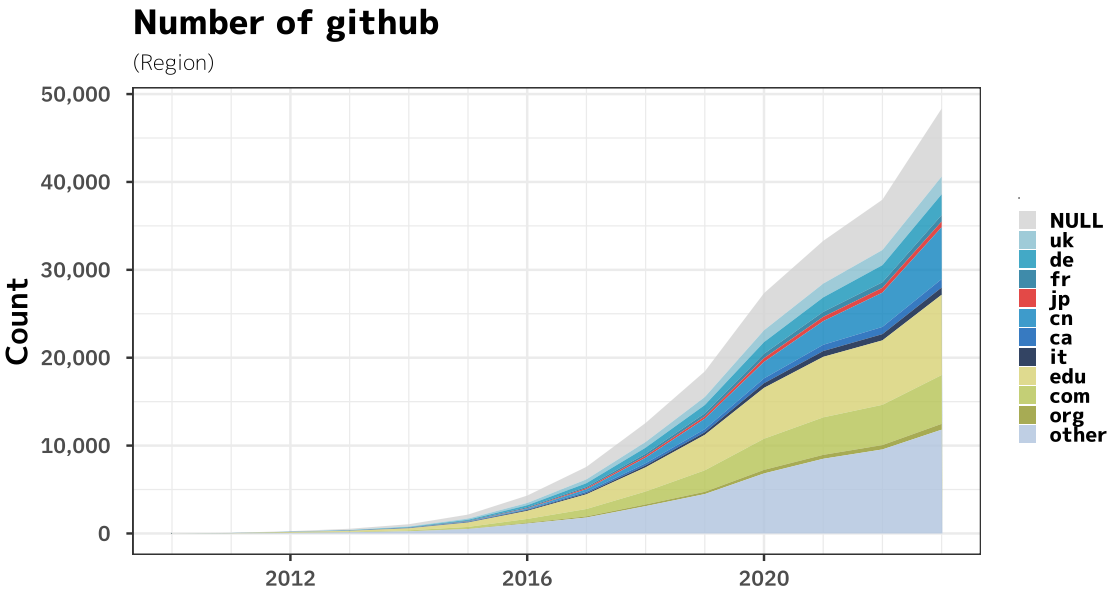

図 21: github 言及原稿数の推移（国別）

表 11: github 言及原稿数の推移（国別）

github	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	1	3	7	10	24	47	75	165	254	357	546	715	822	958
China (cn)	0	0	1	4	6	32	96	282	625	1,185	1,877	2,688	3,903	5,901
Germany (de)	0	5	16	30	54	103	217	415	691	946	1,318	1,646	1,945	2,374
France (fr)	0	3	9	12	33	70	114	145	245	349	507	570	632	728
Ilaly (it)	0	1	2	9	18	40	92	122	213	288	530	660	710	825
Japan (jp)	0	0	3	5	11	22	44	129	211	264	428	486	547	650
UK (uk)	0	6	11	42	71	143	263	438	648	909	1,349	1,570	1,725	2,011
com	2	8	31	44	86	194	434	837	1,472	2,447	3,543	4,266	4,583	5,552
edu	3	15	63	107	222	496	915	1,675	2,747	4,036	5,787	6,885	7,332	9,135
org	0	2	12	11	26	46	87	135	196	260	407	448	496	665
other	4	17	56	137	245	505	1,132	1,812	3,110	4,482	6,831	8,498	9,561	11,810
NULL	5	14	52	125	255	450	841	1,422	2,163	2,919	4,219	4,864	5,728	7,740
Total	15	74	263	536	1,051	2,148	4,310	7,577	12,575	18,442	27,342	33,296	37,984	48,349

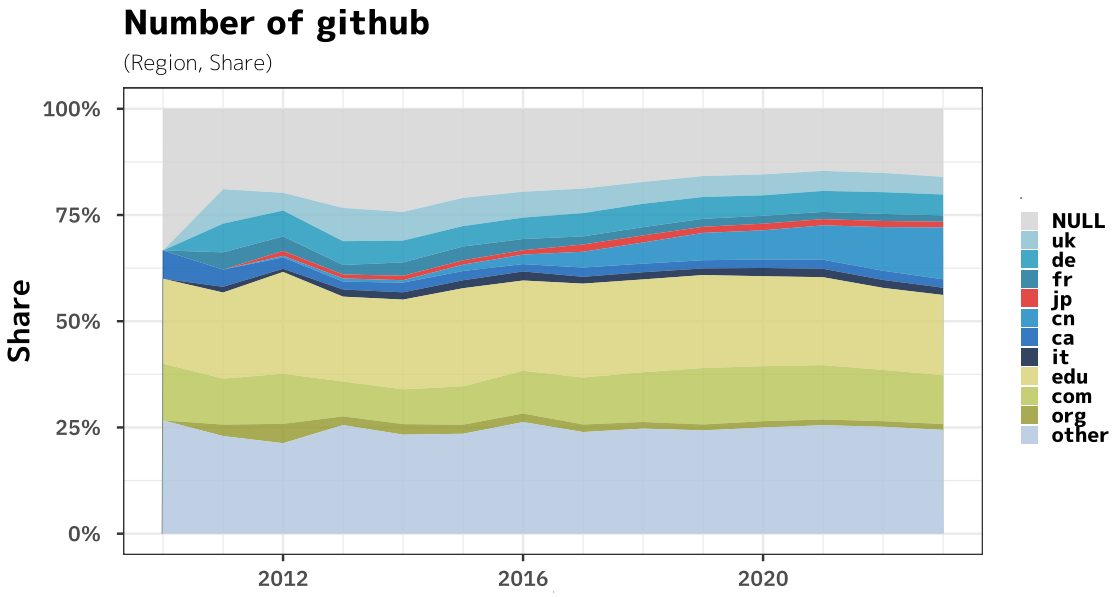

図 22: github 言及原稿数の推移（国別，割合）

3.3.2 分野別の状況

結果を図 23 及び図 24 に示す。

分野別では予想の通り，cs（計算機科学）における数・割合がもっとも多く，2023 年分で見ると cs（計算機科学）全体の 4 割において github の言及が見られる。本報告の調査手法では原稿の著者らのアルゴリズムなどを紹介しているのか，あるいは自己・他者のソースコードを引用・参照したのか，利用の方向性については検知できない。しかしながら，原稿の外にある（当該原稿内で紹介したアルゴリズムの実装などを含め広い意味で）ソースコードを利用する，という研究スタイルが少なくとも cs

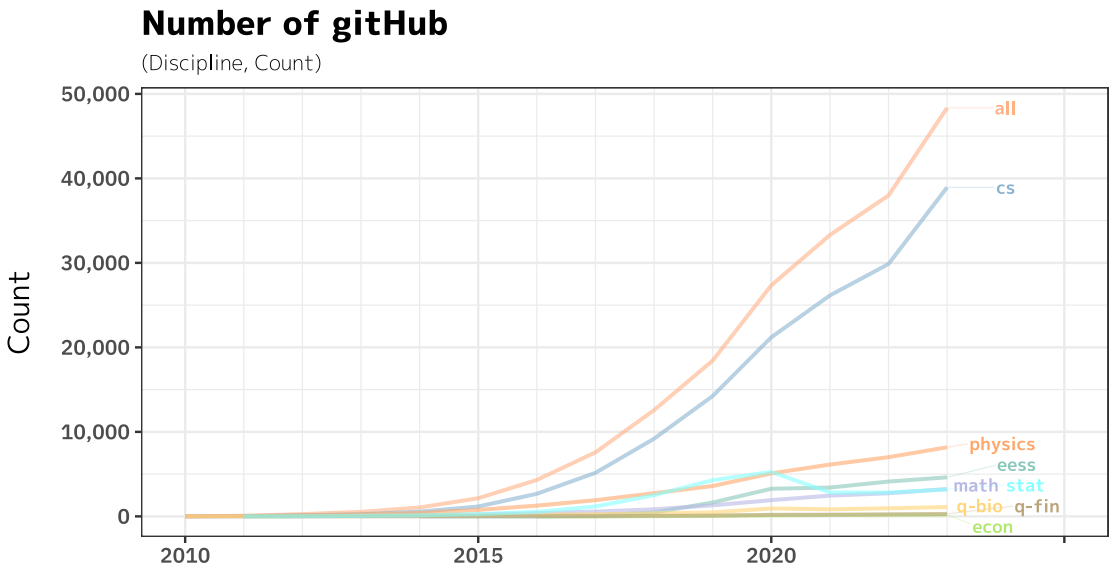

図 23: github 言及原稿数の推移（分野別）

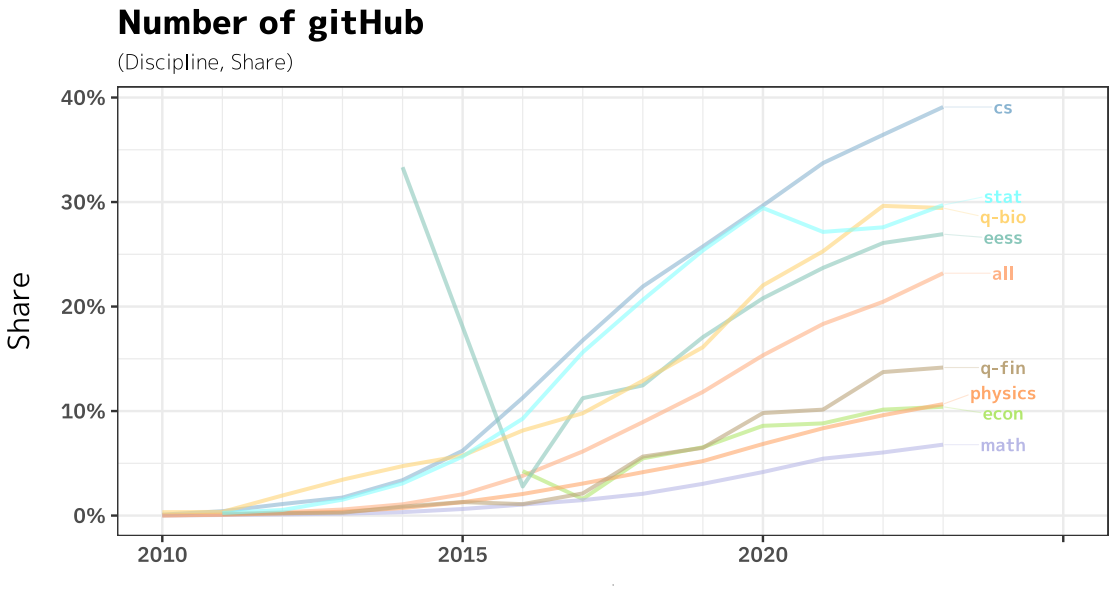

図 24: github 言及原稿数の推移（分野別，割合）

表 12: github 言及原稿数の推移（分野別）

github	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	7	38	136	257	553	1,168	2,672	5,156	9,189	14,245	21,192	26,147	29,867	38,943
econ	0	0	0	0	0	0	2	2	33	63	132	164	177	200
eess	0	0	0	0	1	0	1	82	455	1,644	3,270	3,409	4,120	4,621
math	1	13	30	55	105	219	383	567	837	1,305	1,920	2,458	2,735	3,241
physics	4	38	116	203	411	768	1,272	1,913	2,744	3,595	5,089	6,132	7,007	8,179
q-bio	4	5	34	80	104	132	201	243	353	469	937	835	953	1,105
q-fin	0	1	0	2	7	11	10	19	60	89	181	201	244	286
stat	0	5	17	52	120	265	525	1,176	2,502	4,288	5,258	2,824	2,820	3,184
all	15	74	263	536	1,051	2,148	4,310	7,577	12,575	18,442	27,342	33,296	37,984	48,349
amount	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

分野においては一般化していることが示唆される。また，physics（物理学）, math（数学）の分野でもそれぞれ 10%, 5% は github 言及原稿が見られる点も興味深い。

3.4　その他データの利用

ここまで，「データ一般」の代理指標として zenodo, figshare の言及を，中でも「ソースコード」特に OSS の代理指標として github の言及を紹介した。

ところで，表 1 を見ると，22 位に huggingface, 28 位，30 位には youtube が確認できる。Hugging Face は YouTube と比べて知名度が低いと考えられるが，主に生成 AI などのコアである大規模言語モデル (Large Language Model) をはじめ各種のモデルや評価用データセットなどを共有するサービスで，たとえば SNS サービスの Facebook を運営する Meta 社の LLM，“llama” などもここから取得できる。

本調査の本来の目的は「ＤＸによる研究活動の変化等」の検出であることに鑑み，これらの言及数についても簡単に調査した。

条件は github とほぼ同様で，以下の通りである。

Hugging Face	FQDN 中に「`huggingface`」の文字列を含むもの
YouTube	FQDN 中に「`youtube`」もしくは「`youtu.be`」の文字列を含むもの

3.4.1　Hugging Face

Hugging Face についての結果を図 25 から図 28 に示す。

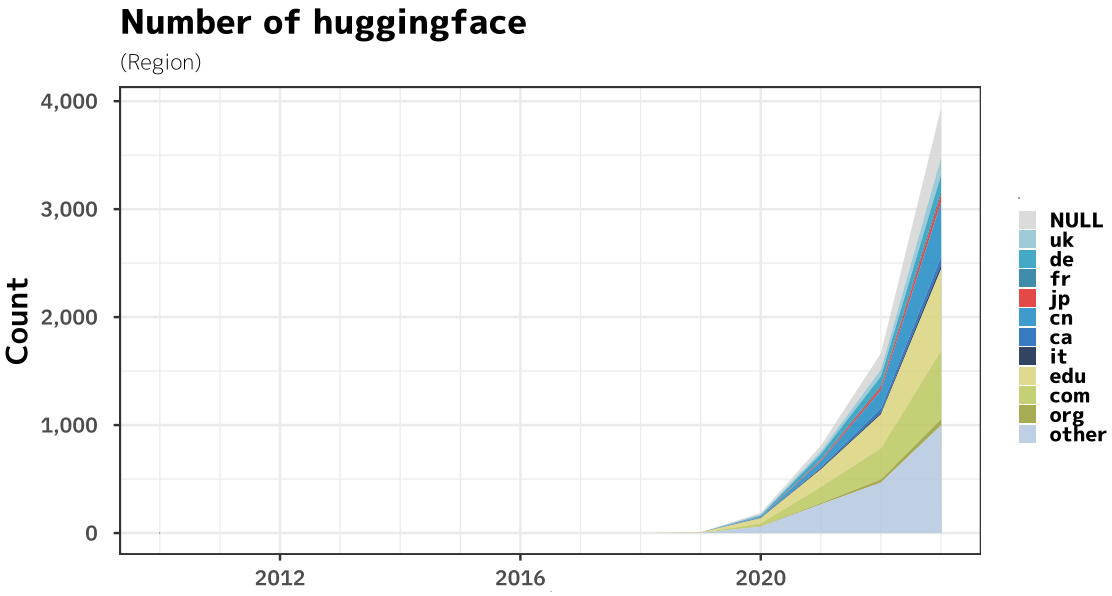

図 25: Hugging Face 言及原稿数の推移（国別）

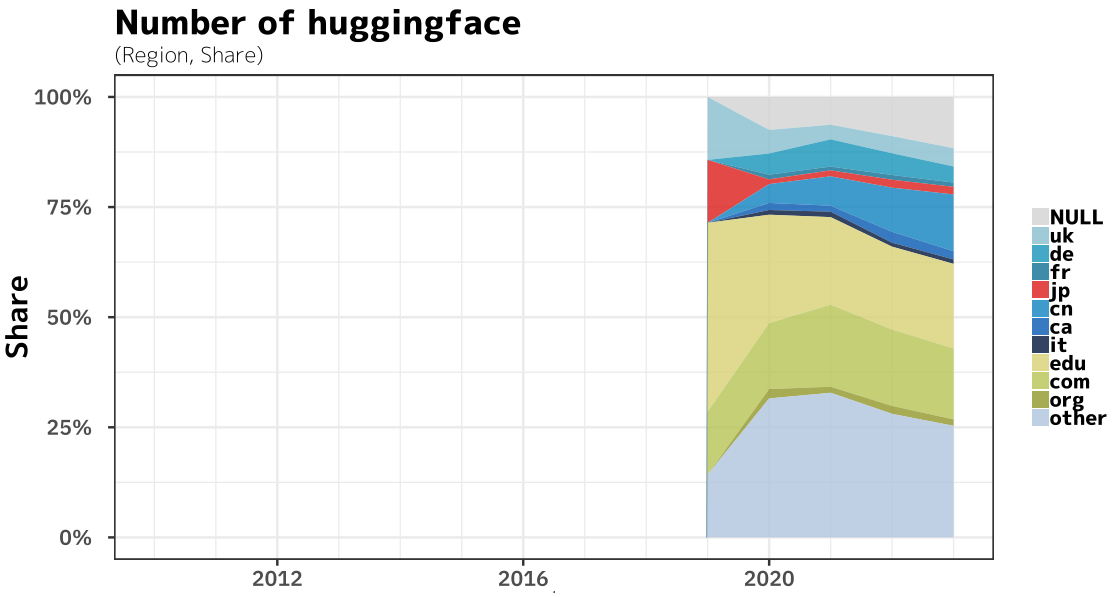

図 26: Hugging Face 言及原稿数の推移（国別，割合）

表 13: Hugging Face 言及原稿数の推移（国別）

huggingface	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	0	0	0	0	0	0	0	0	0	0	3	11	41	74
China (cn)	0	0	0	0	0	0	0	0	0	0	8	54	167	507
Germany (de)	0	0	0	0	0	0	0	0	0	0	9	50	82	144
France (fr)	0	0	0	0	0	0	0	0	0	0	2	7	18	38
Ilaly (it)	0	0	0	0	0	0	0	0	0	0	2	10	15	39
Japan (jp)	0	0	0	0	0	0	0	0	0	1	2	11	30	68
UK (uk)	0	0	0	0	0	0	0	0	0	1	10	27	65	163
com	0	0	0	0	0	0	0	0	0	1	28	151	289	633
edu	0	0	0	0	0	0	0	0	0	3	46	161	313	757
org	0	0	0	0	0	0	0	0	0	0	4	11	30	57
other	0	0	0	0	0	0	0	0	0	1	59	266	467	997
NULL	0	0	0	0	0	0	0	0	0	0	14	51	148	458
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**7**	**187**	**810**	**1,665**	**3,935**

企業としての Hugging Face が 2016 年にスタートで，当初はモデルの共有サービスなどは提供していなかったこともあり，arXiv 原稿中での出現は 2019 年からとなっている。ただし，LLM・生成 AI の急激な流行を反映したものか言及原稿数は短期間で増加しており，2023 年には 2022 年の約 2.4 倍，Zenodo を上回る 4000 件程度の言及原稿を観察できている (図 25)。分野的には図 27 のとおり，全原稿数と cs（計算機科学）分野の原稿数がほとんど一致していることから，基本的に cs（計算機科学）

表 14: Hugging Face 言及原稿数の推移（分野別）

huggingface	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	0	0	0	0	0	0	0	0	0	7	185	806	1,651	3,909
econ	0	0	0	0	0	0	0	0	0	0	1	1	1	8
eess	0	0	0	0	0	0	0	0	0	0	0	38	133	314
math	0	0	0	0	0	0	0	0	0	0	2	2	6	19
physics	0	0	0	0	0	0	0	0	0	0	0	3	15	35
q-bio	0	0	0	0	0	0	0	0	0	0	0	3	10	23
q-fin	0	0	0	0	0	0	0	0	0	0	1	3	7	18
stat	0	0	0	0	0	0	0	0	0	1	14	16	24	51
all	0	0	0	0	0	0	0	0	0	7	187	810	1,665	3,935
amount	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

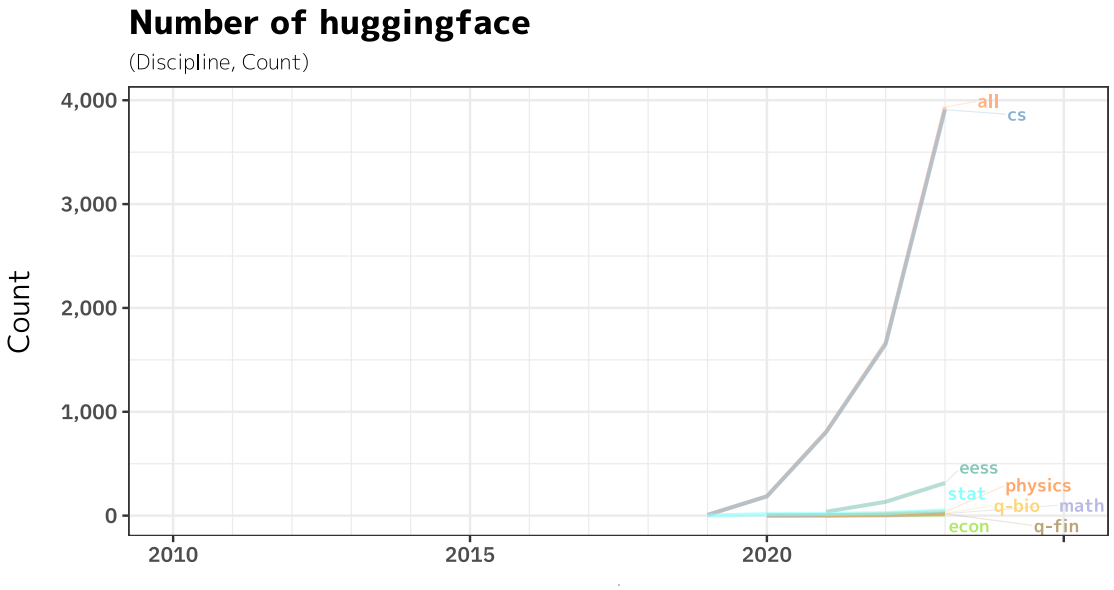

図 27: Hugging Face 言及原稿数の推移（分野別）

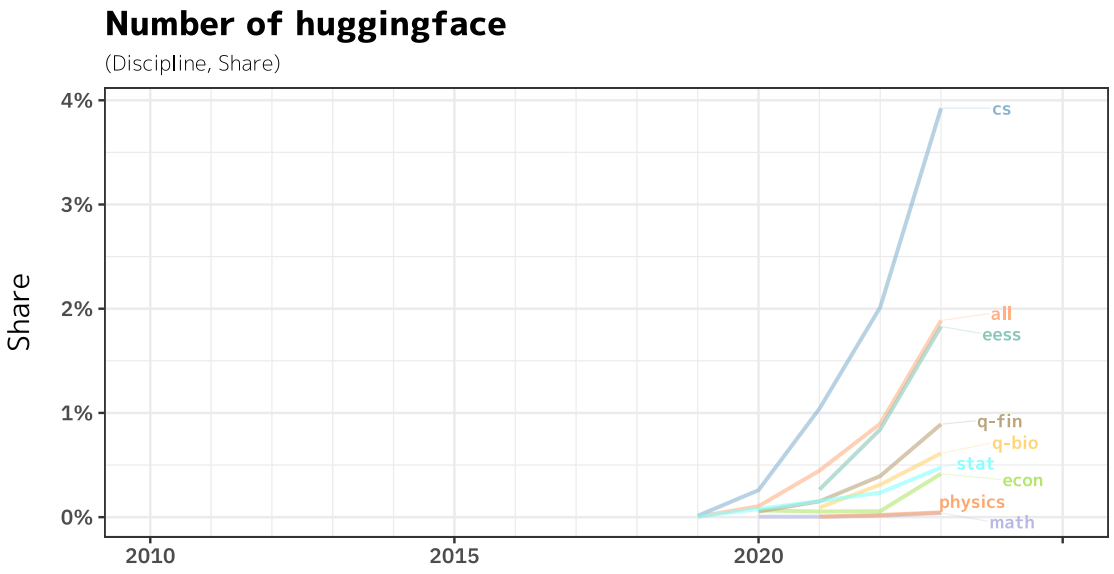

図 28: Hugging Face 言及原稿数の推移（分野別，割合）

分野で利用されていることがわかる。

3.4.2 YouTube

YouTube についての結果を図 29 から図 32 に示す。図 29 を見ると 2023 年時点で言及原稿数 2.000 件程度となっている。Zenodo の言及数には及ばないものの，一定の存在感を示している。

国別に見たときには目立った特徴はみられない。分野別では cs の数，シェアが目立っており，Hugging Face に近い印象を持つ。

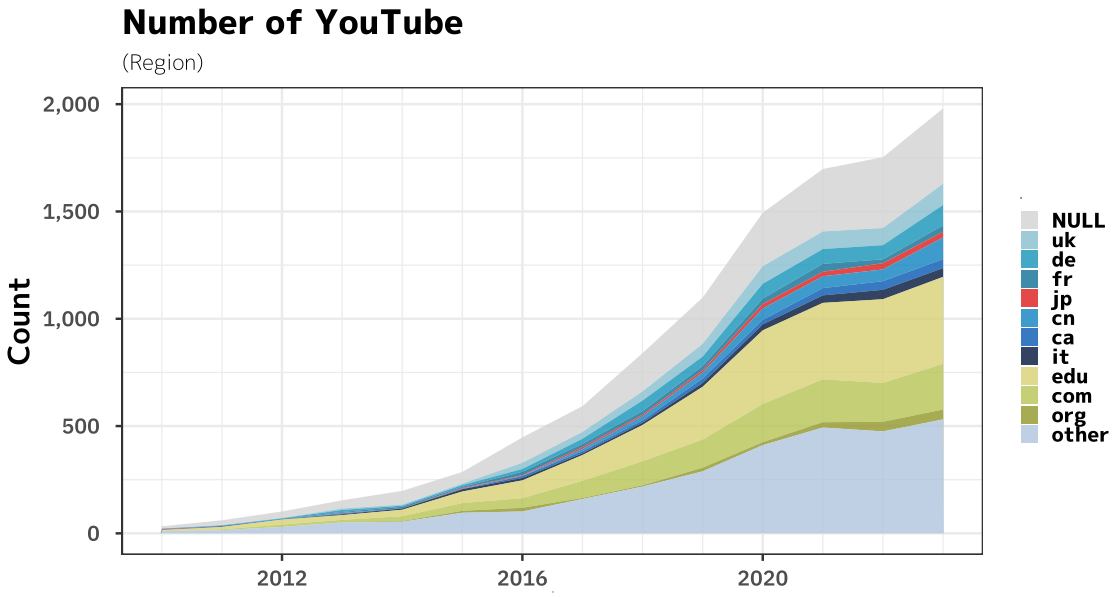

図 29: YouTube 言及原稿数の推移（国別）

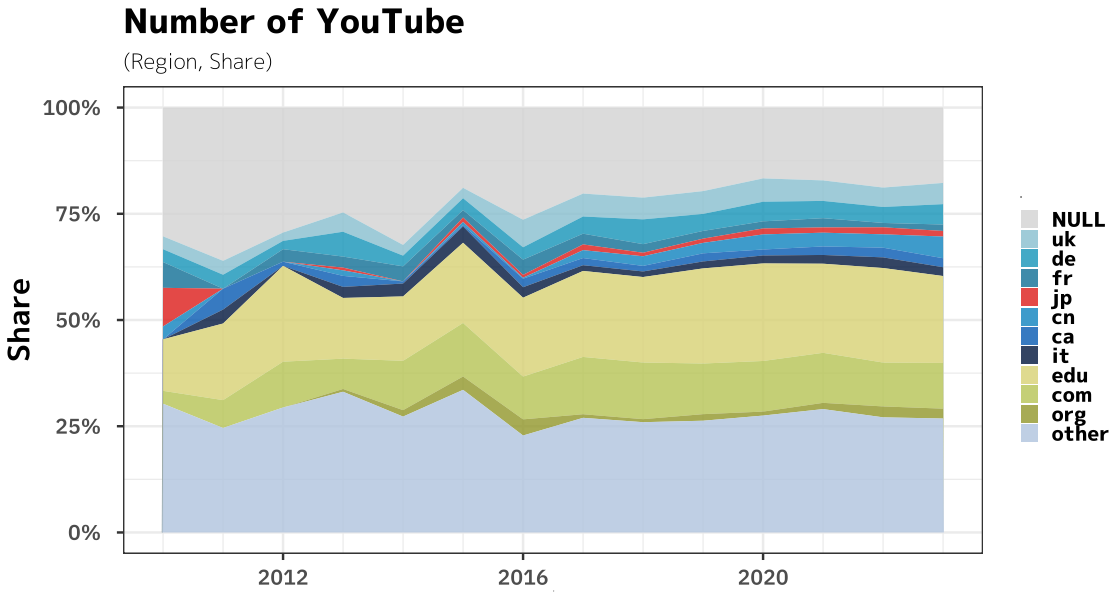

図 30: YouTube 言及原稿数の推移（国別，割合）

表 15: YouTube 言及原稿数の推移（国別）

YouTube	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Canada (ca)	0	3	1	4	1	1	9	10	11	21	21	34	40	42
China (cn)	1	0	0	2	0	2	1	11	19	27	53	55	55	101
Germany (de)	1	2	2	9	5	8	13	24	49	44	69	69	66	95
France (fr)	2	0	3	4	7	5	16	15	17	20	25	37	18	30
Ilaly (it)	0	2	0	4	6	11	11	8	11	18	28	35	44	41
Japan (jp)	3	0	0	1	0	3	3	8	7	11	21	21	29	26
UK (uk)	1	2	2	7	5	7	29	32	43	59	82	82	80	99
com	1	4	11	11	23	36	45	80	112	131	178	200	181	213
edu	4	11	23	22	30	54	83	120	169	246	343	356	390	405
org	0	0	0	1	3	9	17	5	6	17	14	25	45	46
other	10	15	30	51	54	96	102	160	218	289	411	493	475	531
NULL	10	22	30	38	64	54	118	120	178	216	249	291	330	351
Total	33	61	102	154	198	286	447	593	840	1,099	1,494	1,698	1,753	1,980

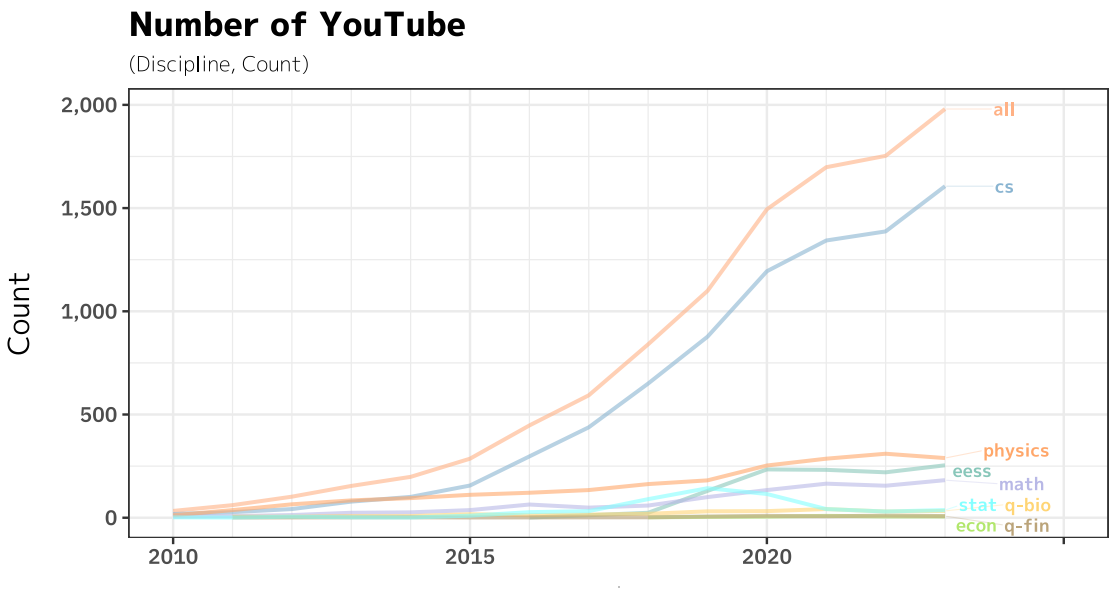

図 31: YouTube 言及原稿数の推移（分野別）

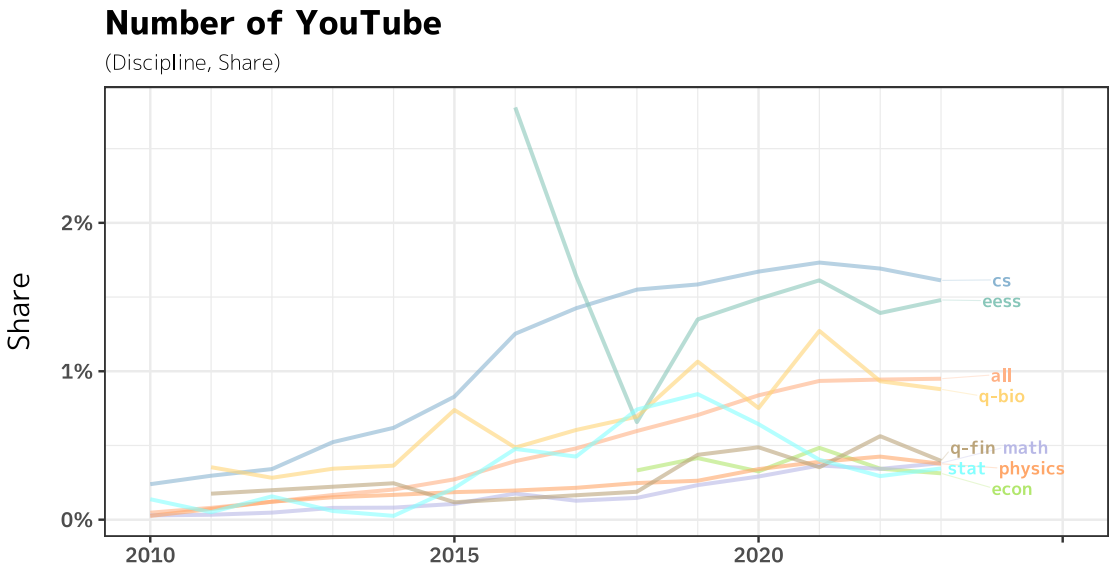

図 32: YouTube 言及原稿数の推移（分野別，割合）

表 16: YouTube 言及原稿数の推移（分野別）

YouTube	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
cs	18	27	42	78	101	156	297	438	650	877	1,194	1,343	1,387	1,606
econ	0	0	0	0	0	0	0	0	2	4	5	9	6	6
eess	0	0	0	0	0	0	1	12	24	130	234	232	220	254
math	6	8	13	24	26	37	64	49	59	100	134	165	155	182
physics	12	37	65	84	95	111	121	134	163	181	253	286	310	289
q-bio	0	5	5	8	8	17	12	15	19	31	32	42	30	33
q-fin	0	1	0	0	2	1	0	0	2	6	9	7	10	8
stat	2	1	5	2	1	10	27	32	90	143	115	42	30	37
all	33	61	102	154	198	286	447	593	840	1,099	1,494	1,698	1,753	1,980
amount	69,957	76,574	84,603	92,641	97,517	105,280	113,380	123,523	140,616	155,866	178,329	181,630	185,692	208,492

3.5 異データソースを用いた追加分析

ここまでで，国別の他，分野によっても利用実態に差があることが確認できた。これらの結果は池内らによるアンケート調査の結果とも合致する [池内 22]。

一方で，データソースである arXiv の特性により，複数分野あるとはいえ主な対象は物理・情報系であり，一般化可能性には疑問も残る。

そこで，生物学系でメジャーな bioRxiv[14] 医学系でメジャーな medRxiv[15][MEXT20] の 2 つのプレプリントサーバについて，2023 年分原稿を対象に類似の分析を行った。データは公式のリポジトリ[16] を通じて 2024 年 5 月に取得した。この際，初版のみに限って取得することが困難であったため，bioRxiv, medRxiv は arXiv と異なり，取得時点でリポジトリにある最終更新時点のものが対象となっていることに注意を要する。リポジトリ上の分析対象原稿総数はそれぞれ 39,095 件，11,005 件である。

なお，bioRxiv, medRxiv は原稿を xml 形式で取得できるため，これを用いて分析する。これにより，PDF を解析する arXiv と異なり，基本的には完全な形で URL を取得できる。したがって，URL はこの要素[17]を採用した。また，著者所属についても xml のメタデータとして整備されており，その中に ‘country’ タグも含まれているため，これを用いることでメールアドレスに頼らない国の把握が可能である。今回は arXiv の分析に合わせて，第 1 著者の country タグの値を原稿の国籍とした。メールアドレスに頼らないため，bioRxiv, medRxiv では com, edu, org のカテゴリは生じない。一方で「United States」「US」など，米国についての検出が可能となっている。

結果を図 33 に示す。図 33 のとおり，2023 年分の原稿についてみると，arXiv と bioRxiv の間で DOI 言及原稿の割合は同程度であることが確認できる。medRxiv は相対的にはやや少ないものの，こちらも 2 割に近い状態である。

OSS の github に目を向けると，arXiv には及ばないものの bioRxiv でも DOI と同程度，2 割程度の原稿で言及が見られ，生物系においても一定の存在感を示していることがわかる。また，medRxiv においても 1.5 割程度の原稿で言及が見られる。バイオインフォマティクスなどが影響している可能性が考えられるが，情報系分野に限らず OSS の利用が広がっている点は興味深い。

一般的なデータシェアの観点で Zenodo,figshare に目を向けると，わずかではあるが，bioRxiv のシェアが arXiv を上回る。研究者らへのアンケートを通じ，日本における分野ごとでのデータ公開などについての実態を調査した関連研究 [池内 22] を見ると，2022 年調査での分野別データ公開率は生物科学 (64.7%, n=150), 計算機科学 (55.1%, n=49), 数学 (50.0%, n=14), 医学 (47.7%, n=88), 物理・天文学 (45.7%, n=81), となっており，Zenodo,figshare の言及率について bioRxiv のシェアが arXiv を上回る傾向に合致する。

まったく異なる観点からの示唆としては，arXiv と bio/medRxiv を比較した際に，com, edu, org と

---

14) 読み：ばいおあーかいぶ，https://www.biorxiv.org/ (Last accessed: 2024.05.21)

15) 読み：めどあーかいぶ，https://www.medrxiv.org/ (Last accessed: 2024.05.21)

16) Amazon Web Service (AWS) の S3 にある以下のリポジトリ。bioRxib: s3://biorxiv-src-monthly, medRxiv: s3://medrxiv-src-monthly

17) 正確には ‘ext-link’ タグ中 ‘ext-link-type’ 属性が ‘uri’, ‘xlink:href’ の値が “http” で始まるときの，‘xlink:href’ の値。

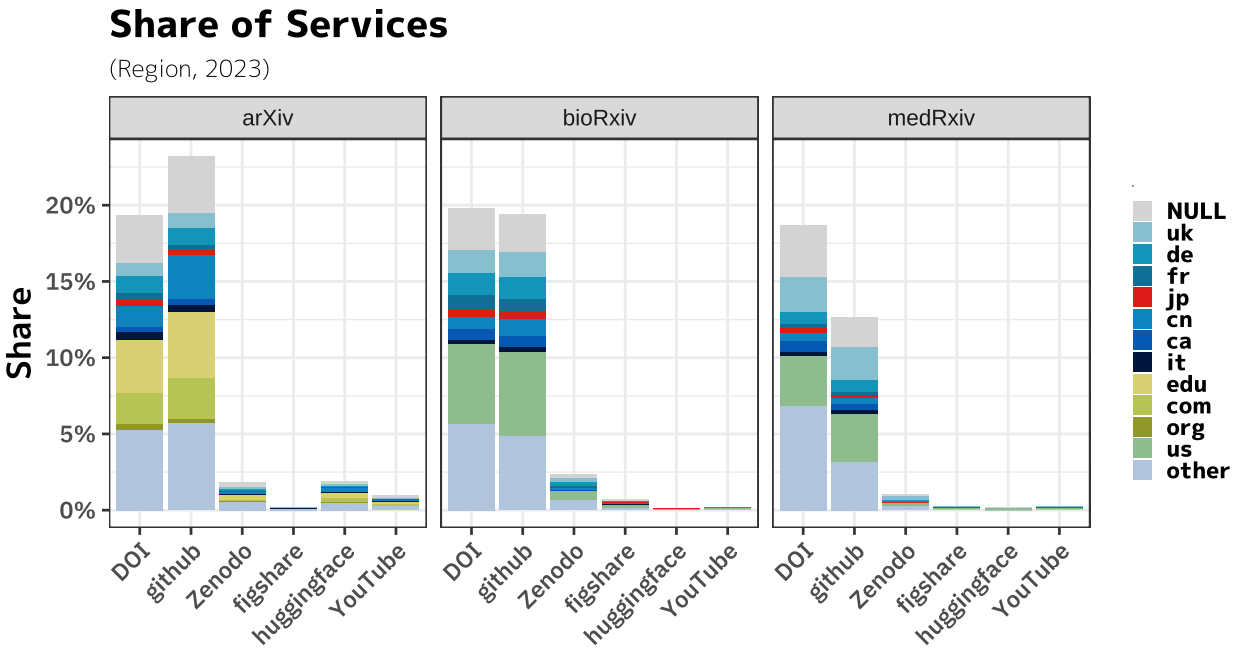

図 33: プレプリントサーバ別の言及原稿数（2023 年分，国別，割合）

us の占める比率はある程度近しいものになっており，米国分を推定するための参考として，com, edu, org を記載することについて，一定の妥当性が得られた。

3.5.1 bio/medRxiv の分野別 github 言及原稿数

情報学分野を直接的には含有しない bio/medRxiv においても github 言及原稿の割合がそれぞれ 2 割，1 割程度となっている点は興味深い。そこで，原稿に割り付けられた分野の単位での把握も試みた。なお，bio/medRxiv では，規定の分野からひとつを選択する形式のため，分野別数の総計は全数に一致する。結果を表 17 及び表 18 に示す。

生物学系の bioRxiv について，表 17 をみると，ゲノミクスやバイオインフォマティクスで特に割合が大きいことわかる。その一方で，その他の分野でも 1 割程度の言及原稿が見られており，濃淡はあるものの，広く活用されている様子が観察できる。

医学系の medRxiv については表 18 の通り。こちらも，遺伝子関係や医療情報学の割合は相対的に高いものの，多くの分野で活用が進んでいることがわかる。

表 17: bioRxiv の github 言及原稿数（分野別）

Category	分野（機械翻訳）	総数	github 言及原稿数	割合
Animal Behavior and Cognition	動物行動と認知	613	89	14.5%
Biochemistry	生化学	1,774	127	7.2%
Bioengineering	生物工学	1,451	157	10.8%
Bioinformatics	バイオインフォマティクス	3,237	1,404	43.4%
Biophysics	生物物理学	1,872	230	12.3%
Cancer Biology	がん生物学	1,789	270	15.1%
Cell Biology	細胞生物学	2,456	222	9.0%
Developmental Biology	発生生物学	1,171	167	14.3%
Ecology	生態学	1,576	267	16.9%
Evolutionary Biology	進化生物学	1,663	535	32.2%
Genetics	遺伝学	1,076	287	26.7%
Genomics	ゲノミクス	1,594	738	46.3%
Immunology	免疫学	1,695	168	9.9%
Microbiology	微生物学	3,665	713	19.5%
Molecular Biology	分子生物学	1,725	244	14.1%
Neuroscience	神経科学	7,151	1,301	18.2%
Oncology	腫瘍学	1	0	0.0%
Paleontology	古生物学	54	7	13.0%
Pathology	病理学	290	31	10.7%
Pharmacology	薬理学	472	24	5.1%
Physiology	生理学	761	67	8.8%
Plant Biology	植物生物学	1,431	227	15.9%
Scientific Communication	科学コミュニケーション	139	22	15.8%
Synthetic Biology	合成生物学	429	65	15.2%
Systems Biology	システム生物学	775	215	27.7%
Zoology	動物学	234	21	9.0%
N/A	N/A	1	0	0.0%
Total	総計	39,095	7,598	19.4%

表 18: medRxiv の github 言及原稿数（分野別）

Category	分野（機械翻訳）	総数	github 言及原稿数	割合
Addiction Medicine	依存症医学	62	6	9.7%
Allergy and Immunology	アレルギー学と免疫学	85	7	8.2%
Anesthesia	麻酔学	33	0	0.0%
Cardiovascular Medicine	心血管医学	843	40	4.7%
Dentistry and Oral Medicine	歯科学と口腔医学	54	2	3.7%
Dermatology	皮膚科	46	5	10.9%
Emergency Medicine	救急医学	69	2	2.9%
Endocrinology	内分泌学	195	19	9.7%
Epidemiology	疫学	1,241	239	19.3%
Forensic Medicine	法医学	3	0	0.0%
Gastroenterology	消化器病学	144	19	13.2%
Genetic and Genomic Medicine	遺伝医学とゲノム医学	812	226	27.8%
Geriatric Medicine	老年医学	82	4	4.9%
Health Economics	医療経済学	106	17	16.0%
Health Informatics	医療情報学	459	109	23.7%
Health Policy	医療政策	127	8	6.3%
Health Systems and Quality Improvement	医療システムと品質改善	198	4	2.0%
Hematology	血液学	75	11	14.7%
HIV/AIDS	HIV/AIDS	184	9	4.9%
Infectious Diseases (except HIV/AIDS)	感染症（HIV/AIDSを除く）	1,037	133	12.8%
Intensive Care and Critical Care Medicine	集中治療医学	91	8	8.8%
Medical Education	医学教育	85	1	1.2%
Medical Ethics	医療倫理学	22	5	22.7%
Nephrology	腎臓学	84	6	7.1%
Neurology	神経学	820	112	13.7%
Nursing	看護学	49	0	0.0%
Nutrition	栄養学	123	6	4.9%
Obstetrics and Gynecology	産科と婦人科	150	6	4.0%
Occupational and Environmental Health	労働衛生と環境衛生	95	5	5.3%
Oncology	腫瘍学	360	45	12.5%
Ophthalmology	眼科学	116	12	10.3%
Orthopedics	整形外科学	58	3	5.2%
Otolaryngology	耳鼻咽喉科学	43	1	2.3%
Pain Medicine	痛み医学	54	2	3.7%
Palliative Medicine	緩和医療	7	0	0.0%
Pathology	病理学	84	15	17.9%
Pediatrics	小児科学	216	15	6.9%
Pharmacology and Therapeutics	薬理学と治療学	86	10	11.6%
Primary Care Research	一次医療研究	87	8	9.2%
Psychiatry and Clinical Psychology	精神医学と臨床心理学	605	83	13.7%
Public and Global Health	公衆衛生と国際保健	932	83	8.9%
Radiology and Imaging	放射線医学と画像診断	285	61	21.4%
Rehabilitation Medicine and Physical Therapy	リハビリテーション医学と理学	178	8	4.5%
Respiratory Medicine	呼吸器内科学	126	10	7.9%
Rheumatology	リウマチ学	58	4	6.9%
Sexual and Reproductive Health	性と生殖の健康	88	4	4.5%
Sports Medicine	スポーツ医学	74	5	6.8%
Surgery	外科学	94	4	4.3%
Toxicology	毒物学	6	1	16.7%
Transplantation	移植医学	38	2	5.3%
Urology	泌尿器科学	34	3	8.8%
N/A	N/A	2	2	100.0%
Total	総計	11,005	1,390	12.6%

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論文や書籍にも付与される 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。

まず，DOI については 2022 年と 2023 年との間で割合に大きな差は見られず，ほぼ横ばい傾向となっている。DOI の増加が始まるのは 2016，2017 年あたりからであり，これを基準に普及が進んだと考えた場合，2023 年時点でもすでに 6 年近くが経過していることになる。特に，arXiv の主たる対象分野のうちのひとつ情報科学（arXiv の分野では cs）は直近数年で，深層学習（ディープラーニング）や大規模言語モデルなど新たな手法が誕生し，多くの研究が行われた。こうした新しい研究の論文には DOI が付与されている確率が高いと期待される。さらに，DOI は論文を中心に付与が始まったこともあり，旧来の「文献引用」という論文執筆の作法の枠組みの中でもちいると利便性が向上する。しかし，その DOI であっても言及原稿割合が 2 割程度に留まるという点は興味深い。また，生物系を主とする bioRxiv，医学系を主とする medRxiv でも同程度の割合で，今回試行した範囲では分野差もある程度小さいことが予想される[18]。

次に，一般的なデータの利用について見る。データ共有のためのプラットフォームである Zenodo, figshare は各アイテムに DOI を割り振るため，参照も基本的に DOI ベースとなり DOI 言及原稿数を大きく上回ることは原理上ありえないが，比較的数の多い Zenodo でも全体の 2% とシェアは少ない。Zenodo の登録アイテム数では 2024 年 5 月 30 日時点で Image 957,013 件，Dataset 326,489 件となっており，「データが少ないために利用が進んでいない」という可能性は低いと考えられる。研究者へのアンケートによって日本におけるデータ公開・利用の状況を調べた [池内 22] では，12% (n=823) の研究者が「データ共有サービス (figshare や zenodo など)」から公開データを取得した経験があると回答している。日本を対象としており地域が限定される上，研究者の単位であり，一度でも入手していれば回答できるため，これを持って「論文においても 12% 程度が Zenodo などから得たデータを使っている」と考えるのは当然誤りである。しかしながら，Zenodo 等からのデータ取得経験のある研究者の割合 12% に対して，言及のある原稿の割合は 2% という結果は低いように思われる。おなじく文献 [池内 22] では利用の障壁として，利用条件が不明確であることや，著者の情報が明確ではないこと，品質が不明なこと，などが挙げられている。仮に他国の研究者も同様の意見を抱いているとして，Zenodo や figshare では少なくとも利用条件は明示できるため，主に著者情報（信頼性）や品質の理由から取得はしたものの，利用には至っていないことが考えられる。

他方で，OSS 利用の代理変数とした github の言及は，arXiv では DOI の 2 割を越え，bioRxiv でも DOI に迫るなど，Zenodo 等で想定する数値や図表などのデータに比べて非常に大きい。特に非情報

[18] なお図 18 では q-bio（定量生物学）, q-fin（金融学） などが 3 割に迫っている。ただし，これらは cs（計算機科学）の 1/30 程度の原稿数で，かつ分野共起から cs（計算機科学）との重複も多いと考えられる点に留意が必要である。

系である bioRxiv や medRxiv でも DOI と近しいレベルで一定のシェアがあることは興味深い。数値や図表などの一般的なデータと異なり，github についての言及原稿が多い理由について考えた場合，例えばソースコードについては，1. 基本的に実際に動作させるものであって，かつ，多くの利用者にとっては任意の入力に対して所望の結果が出てくれば中身はブラックボックスで良いこと。2. 出力結果を通じて品質の確認ができること。3. さらに，著者情報も必ずしも実名を用いないユーザが多いこと。などの要因から複合的に利用が進んでいる可能性が考えられる。

arXiv の 2023 年分原稿において，言語モデルなどの共有サービスを提供する Hugging Face への言及が Zenodo の言及を超えている点も，github と同じく利用することで容易に品質を確認できることが一因と考えられる。

これらから，1. 画像や数値などの「データ」については，論文とは独立した形で公開されつつある一方，原稿中での言及は進んでいないこと。2. OSS については，情報系が多い arXiv では DOI 以上に，生物・医療系の bioRxiv,medRxiv でも DOI と同程度に言及されていること。が，わかった。

ソースコードを通じて，アイデアを実際にすぐ動かせる形で公開する，あるいは，それを活用する，という行為は「ＤＸによる研究活動の変化等」の一端とみることができる。

これらが一般化していくことで，例えば就職などプロモーション活動においても，論文数などに加えてソースコードの公開数やそれらのダウンロード数，スター[19]の数，などが評価の俎上（そじょう）にあがる可能性はある。また，研究プログラムなどにおける評価指標として，従来指標に加えて採用することも考えられる。例えば [池内 22] では，「所属機関がデータを公開することを業績として評価しているか」を聞いているが，ここでは，73% (n=1.237) が「あまり・全く評価していないと思う」と回答しているが，前述の研究プログラムなどにおける評価指標として採用することでデータ公開が促進され，結果として我が国における「ＤＸによる研究活動の変化等」も促進される可能性がある。

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[20]。

---

19) github の機能として搭載されている，利用者からのいわゆる「いいね」に相当するもの。

20) 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。

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 をみても，これらの選択は比較的妥当と考えられるものの，例えば，github に類似するサービスとして bitbucket[21]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 件ずつ DataCite[22]に問い合わせるなどして，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。

21) https://bitbucket.org/
22) https://datacite.org/

参考文献

[MEXT20] MEXT–NISTEP プレプリント調査・検討チーム． プレプリントをめぐる近年の動向及び今後の科学技術行政への示唆, **科学技術・学術審議会 情報委員会, ジャーナル問題検討部会（第７回）**, 2020.

[池内 22] 池内 有為, 林 和弘． 研究データ公開と研究データ管理に関する実態調査 2022：日本におけるオープンサイエンスの現状, *Research Material*, *No.335*, 文部科学省科学技術・学術政策研究所, 2020. DOI: https://doi.org/10.15108/rm335

[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

[林 22] 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, *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

付録 A　URL 含有原稿の状況

本編では，DOI をはじめ特定の文字列を含む URL のみについて，また原稿単位でのカウントを示した。ここでは，単純に検出できた URL の数や，それを原稿当たりでみた数，全原稿に対する URL を含む原稿の割合について表 19 に示す。

表 19 をみると，URL の記載がある原稿の割合は年々増加しており，2020 年には過半数を超える原稿に何らかの URL 記載が認められる。また，2023 年には原稿あたりの URL の数も 7.5 件となっており，これも増加している。

基本的にメジャーな論文誌では DOI の導入が進んでいること，さらに本編で述べたとおり，DOI は現状 2 割程度に留まること，github も 2.5 割程度であること，などを勘案すると，多くは論文やソースコード以外の，何らかの情報を指し示すために使われている可能性が高い。

年	検出URL数 (A)	URL含有原稿数 (B)	平均URL数 (A/B)	全原稿数 (C)	URL含有原稿率 (B/C)
2010	52,596	20,207	2.6	69,957	28.9%
2011	58,847	22,173	2.7	76,574	29.0%
2012	75,235	26,109	2.9	84,603	30.9%
2013	88,185	29,319	3.0	92,641	31.6%
2014	101,447	32,063	3.2	97,517	32.9%
2015	123,663	36,456	3.4	105,280	34.6%
2016	152,775	41,693	3.7	113,380	36.8%
2017	197,581	48,717	4.1	123,523	39.4%
2018	277,342	61,031	4.5	140,616	43.4%
2019	354,532	71,675	4.9	155,866	46.0%
2020	514,721	89,972	5.7	178,329	50.5%
2021	629,564	97,864	6.4	181,630	53.9%
2022	712,070	102,534	6.9	185,692	55.2%
2023	891,615	119,620	7.5	208,492	57.4%

表 19: URL 含有原稿の推移

調査資料-342

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 2024

2024 年 09 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
小柴 等・林 和弘

〒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 2024

Nov. 2024

KOSHIBA Hitoshi,　HAYASHI Kazuhiro
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/rm342

https://www.nistep.go.jp