２Ｆ１９

arXiv, bioRxiv に掲載されたプレプリントの分析

○ 林 和弘（NISTEP）, 小柴 等（NISTEP）

1　はじめに

研究成果を発信・共有し，そして研究コミュニティ内での評価する営みは，研究における標準的な活動であり，近年では多くの分野で査読済みの学術ジャーナル論文（以降，学術論文とする）がそのコミュニケーションを媒介している。一方，研究分野によっては，学術論文だけでなく，その査読前段階の草稿であるプレプリントを活用したコミュニケーションが進展してきた。そして，近年の急速な ICT の進展に伴うオープンサイエンスの潮流によってこうした動きが加速していたところ，COVID-19 流行下にあってさらに拍車がかかり，医学を中心により幅広い分野でプレプリントの活用に注目が集まっている。しかしながら，研究活動におけるプレプリントの役割や位置付け，そしてその存在感に関する定量的なエビデンスは少なく，これまで研究者や政策関係者の間でもごく限定的にしか語られてこなかった。

著者らは，プレプリントの活用が研究者のコミュニケーションや研究活動をどのように変えているかの実態を把握することを目的に，また，政策的には学術ジャーナルに掲載される学術論文の量（論文数）と被引用数に基づく質に関する調査研究を補完することを目的とし，一定以上の歴史と掲載数を有するプレプリントサーバ（プレプリントの公開・共有サービス）に投稿されたプレプリントに着目した分析を行ってきた [文科 20, 林 20a, 林 20b, 林 21, 小柴 21]。分析対象は，1991 年から運用している物理系のプレプリントサーバーである arXiv と，2010 年代に入って進展している生物科学系の bioRxiv に着目し，原著論文との関係，プレプリントの引用などの観点から，その特徴および分野別特性を分析した。また，bioRxiv については，プレプリントとその後ジャーナルに掲載された論文との差異についても比較を行った。本講演ではこれらを総括して，プレプリントの現状と動向について述べる。

2　arXiv のプレプリント分析

arXiv は物理学分野を中心に 1991 年から運用を開始してた最古かつ最大手のプレプリントサーバであり，近年では人工知能など情報系の分野でも活用が進んでいる。この arXiv 上にある約 160 万本（2020 年 1 月時点）を分析した結果を分野別件数推移（図 1），DOI 付与率推移（図 2），分野別の DOI 付与割合（図 3），分野別の DOI 付与までの期間（図 4），ならびに分野別の被引用数（図 5）に示す。

なお，arXiv では CS_DL（計算機学分野：デジタル図書）のような粒度で 153 の分野が存在する。ここでは筆者らが更に 8 分野でまとめた独自の分野分類をもちいている。詳細は別稿 [林 20a] に示す。

arXiv においては，掲載数においては 2010 年代の情報系の伸びが大きいこと，DOI の付与は，時間の経過とともに 2/3 程度にとどまること，また，DOI の付与割合は分野によって大きく異なることがわかる。arXiv は論文管理に独自の ID を用いており，DOI が記載されているものは当該原稿に関して arXiv 外で取得したものであることを意味する。そこでこの DOI の付与について，プレプリント公開後に査読付きのジャーナル等，既存の出版物として発行されたとみなしてその割合（ジャーナル掲載率）を調べると，このジャーナル掲載率は 2/3 程度であり，全体の 1/3 のプレプリントはプレプリントのままであることが示唆される。この 1/3 のプレプリント群は，これまで学術論文としては表に出てこなかった研究成果であるために，プレプリント群の分析は，学術論文群の分析とは違った結果になる可能性がある。その一方，査読に通らなかった可能性やフェイク情報の可能性を含めて，その価値付けには慎重を要する。

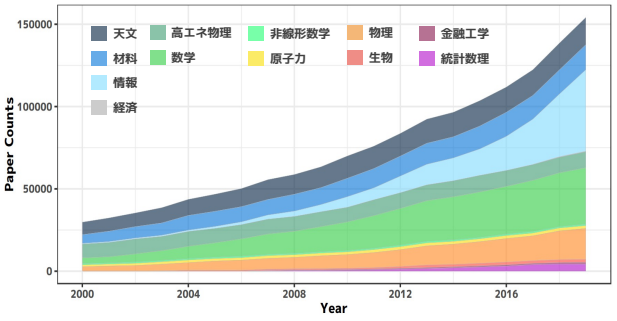

図 1: 分野別件数推移 (arXiv)

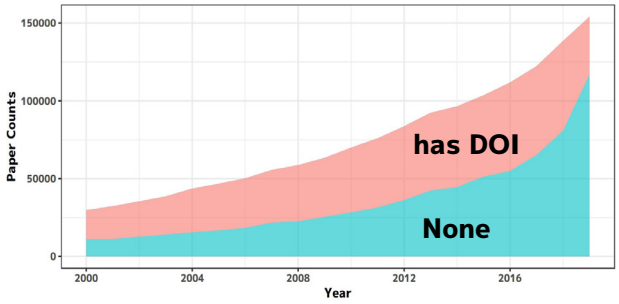

図 2: DOI 付与率推移 (arXiv)

次に，プレプリントが学術論文になる過程において，DOI の付与率，付与までの期間や被引用において分野別に大きな違いが出ていることは，分野別の研究成果共有活動の差を示唆する。特に情報系では，深層学習等の競争が非常に激しい分野において，プレプリントを公開し，プレプリントを引用して研究活動を行っていると言われているが，そのことがデータとしても示されている。COVID-19 によって，迅速な研究成果の共有が求められる中，また，プレプリントの公開によって，先取権の一定の確保が可能という研究者に対するメリットと合わせて，プレプリントが今後より受け入れられていく可能性を示唆する。

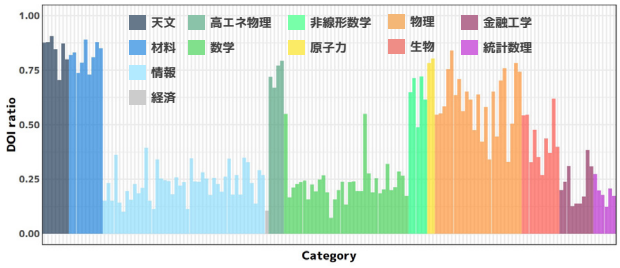

図 3: 分野別 DOI 割合 (arXiv)

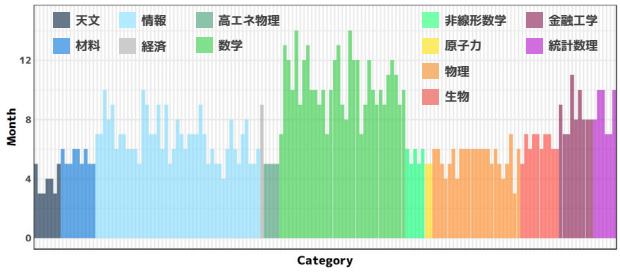

図 4: 分野別 DOI 付与期間 (arXiv)

3　bioRxiv のプレプリント分析

続いて，生物系（バイオ）分野を主なターゲットとして 2013 年に開始した bioRxiv に掲載された，約 12 万件

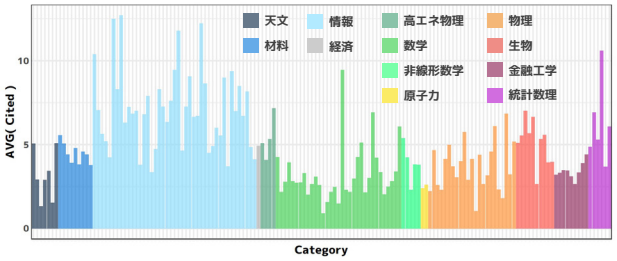

図 5: 分野別平均被引用数 (arXiv)

のプレプリント（2021 年 4 月時点）について，arXiv と同様に分析した結果を，分野別件数推移（図 6），DOI 付与率推移（図 7），分野別の DOI 付与割合（図 8），分野別の DOI 付与されるまでの期間（図 9），ならびに分野別の被引用数（図 10）で示す。

bioRxiv については各原稿を DOI で管理しているが，さらに，ジャーナル等他の媒体における DOI を付与できるようになっていることから，図 7 などの DOI 付与率はこの他媒体の DOI 情報を用いて算出した.

これらの図を見ると，例えば bioRxiv 内の掲載数においては，神経科学 (Neuroscience) と微生物学 (Microbiology) の伸びが相対的に顕著であることが分かる。その一方，分野別のジャーナル掲載率や，DOI の付与期間にはほとんど差がない。また，被引用数においても，ゲノミクス（Genomics) やバイオインフォマティクス（Bioinfomatics）に特徴が見られるものの，arXiv に比較して大きな差は見いだされなかった [林 21]。

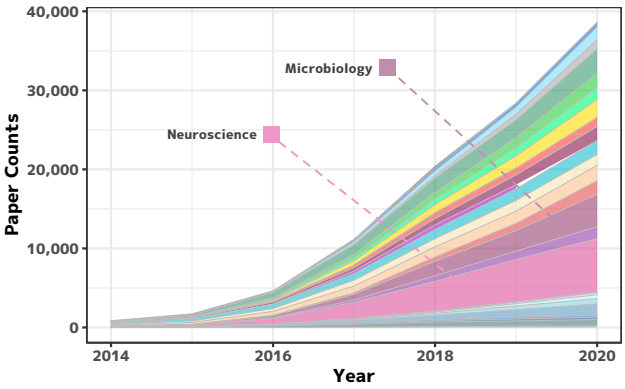

図 6: 分野別件数推移 (bioRxiv)

arXiv は 30 年にわたる歴史の中で，数理学を中心としながらも分野が拡張し，また，長年の運用の中で研究者コミニティのプレプリントに対する扱いが変わってきた経緯を持つ。対して，bioRxiv はバイオを中心として予め決められた分野の中で運用を行っているため，まだ 10 年未満であることも含めて，分野別の差がでにくいことが示唆される。bioRxiv でも今後年数を重ねることで

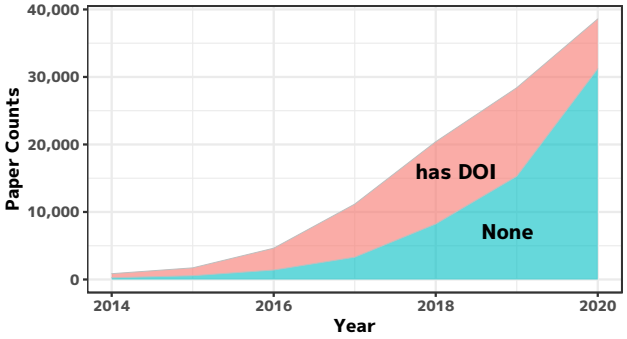

図 7: Journal DOI 付与率推移 (bioRxiv)

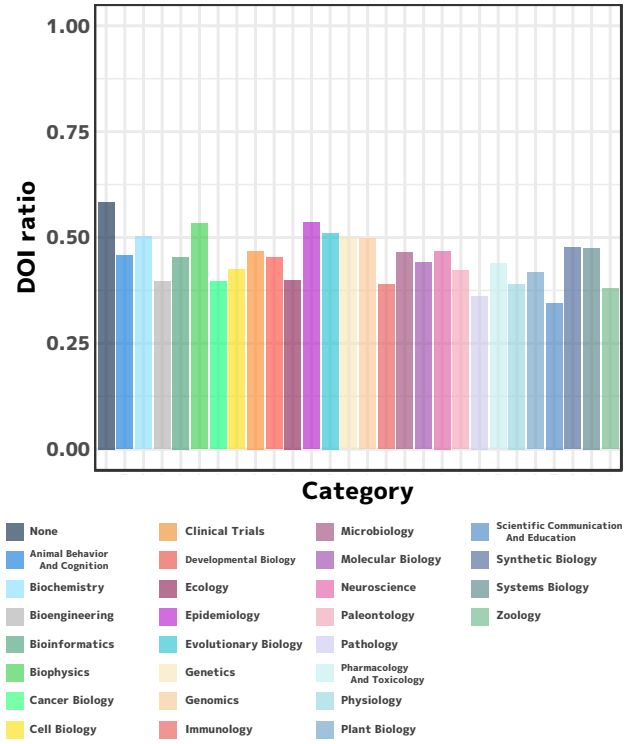

図 8: 分野別 Journal DOI 割合 (bioRxiv)

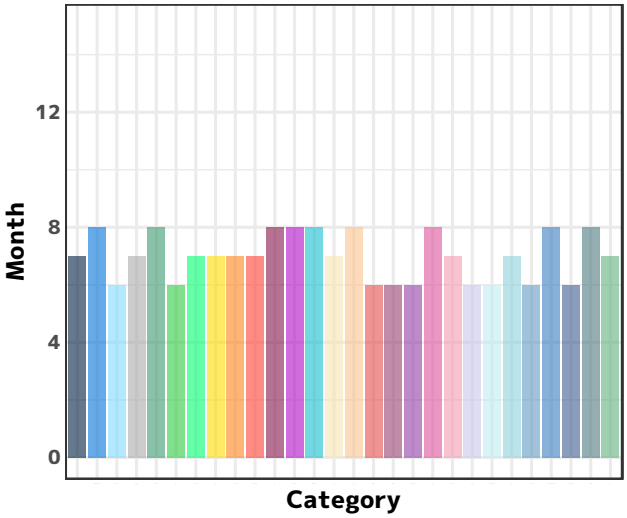

図 9: 分野別 DOI 付与期間 (bioRxiv)

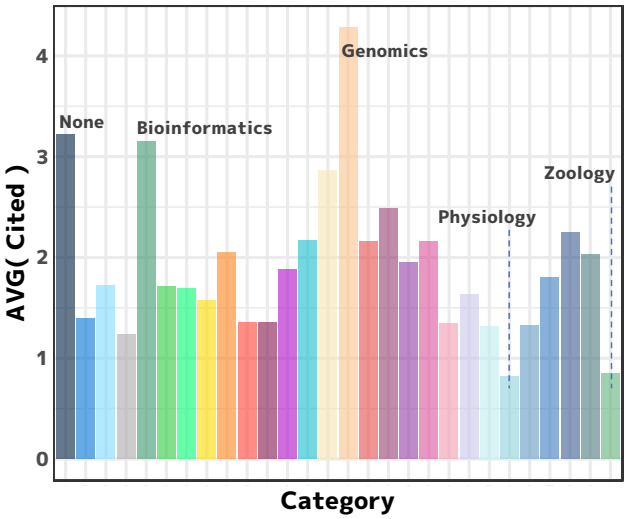

図 10: 分野別平均被引用数 (bioRxiv)

arXiv と同様の経緯をたどって，プレプリントの取り扱いにも分野別の特徴が出るかに注目したい。いずれにせよ，学術ジャーナル同様プレプリントにおいても，分野別の特徴が出うることを念頭に，プレプリントサーバー全体の調査と，分野別の調査を目的に応じて使い分ける必要がある。

4 bioRxiv のプレプリントと掲載 OA 学術論文との比較

bioRxiv のプレプリントについては，その後 OA 学術論文となったものの全文 XML を一定量（7,985 件）確保することができたので，プレプリントと (OA) 学術論文の比較を行った（図 11）。その結果，参考文献数や単語数などの外形的な基準や，簡単な文書類似度から両者の差分を明らかにしようとした試行の範囲では，プレプリントと学術論文（図 11 の A と B），学術論文になったプレプリントとそうではないプレプリントの間（同 A と C）で明確な違いを見いだすことはできなかった [小柴 21]。

プレプリントから学術論文として発行するメリットとして，査読による内容の向上が考えられるが，そのことは，外形的な変化という形では明らかにすることはできなかった。今回の限定的な調査をもって査読の価値付けが低いということは言えない。一方で，プレプリントについて少なくとも論文の体裁を満たさないようなものは多くなく，プレプリントであるからと言って必ずしも質が低い，信頼性が疑わしいようなものが多いとはいえないことも示唆される。引き続き，内容の解釈等を含めて査読によって変わる論文の価値を定量的に表現することの検討が求められる。なお，bioRxiv から OA 学術論文

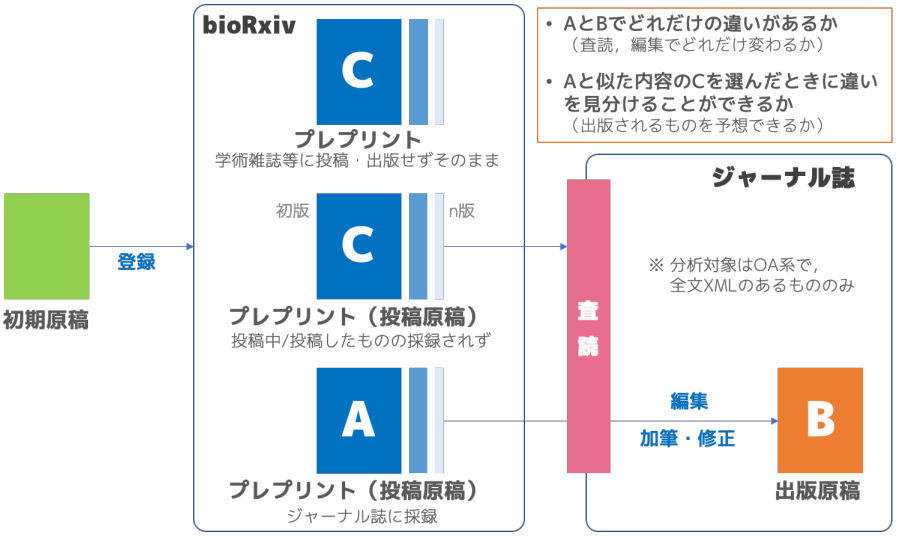

図 11: bioRxiv のプレプリントと学術論文の比較

以外の学術論文になったものと元のプレプリントとの比較や，arXiv のプレプリントにおいて同様の調査を行うには，プレプリント全文データと，掲載論文全文の入手と比較のための加工が難しいか相当の手間がかかる[*1]。しかしながら，プレプリントの研究活動への影響を一定の定量性をもって推し進めるために，今後の課題として検討を続ける。

5 まとめ

プレプリントの活用が研究者のコミュニケーションや研究活動をどのように変えているかの実態を把握することを目的に，また，政策的には，学術ジャーナルに掲載される学術論文の量（論文数）と被引用数に基づく質に関する調査研究を補完することを目的に，原著論文の草稿であるプレプリントに着目した分析を行った。

1991 年から運用している物理系のプレプリントサーバである arXiv と，2010 年代に入って進展している生物科学系の bioRxiv に着目し，原著論文との関係，プレプリントの引用などの観点から，その特徴および分野別特性を分析した。分野の粒度やカバー範囲に違いがあるため，一概には比較が難しいものの，arXiv ではジャーナル掲載率（DOI 付与率）や DOI が付与されるまでの期間に大きな差があるのに対し，bioRxiv ではそれらが相対的に小さいなど，分野の特徴が得られた。

また，bioRxiv については，プレプリントとその後ジャーナルに掲載された論文との差異についても比較を行い，自然言語処理で単純に比較した場合においては内容面での差は小さく，図表や文字数など外形的基準では，プレプリントか学術論文かを判別することは困難であることが分かった。

以上の結果から，プレプリントサーバのプレプリントを分析することで，学術論文の交換とは異なる研究者のコミュニケーションが分野ごとに存在していることが，一定の定量性をもって示された。引き続き，他の分野のプレプリントサーバの調査を含む，定量的なアプローチによる分析を行い，かつ，分野ごとの定性的な調査分析と合わせて，プレプリントの活用が研究者のコミュニケーションや研究活動をどのように変えるか，そして，その活動が研究評価にどのような影響を与えるかについて明らかにしたい。

[*1] bioRxiv や多くの OA ジャーナルでは JATS(Journal Article Tag Suite)-XML 形式で全文データを持つが，arXiv は JATS 策定以前からの歴史とも相まって TEX 形式のデータをもつ。

参考文献

[文科 20] 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. **文部科学省 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 配布資料**資料 1-別添, (2020). https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

[小柴 21] 小柴 等，林 和弘: プレプリントとジャーナル論文の差異: bioRxiv を用いた試行, *NISTEP DISCUSSION PAPER*, No.200, (2021). https://doi.org/10.15108/dp200

[林 20a] 林 和弘，他: arXiv に着目したプレプリントの分析, *NISTEP DISCUSSION PAPER*, No.187, (2020). https://doi.org/10.15108/dp187

[林 20b] 林 和弘：MedRxiv, ChemRxiv にみるプレプリントファーストへの変化の兆しと オープンサイエンス時代の研究論文. *STI Horizon 2020* **春号**, Vol.6, No.1, (2020). https://doi.org/10.15108/stih.00205

[林 21] 林 和弘，他: bioRxiv に着目したプレプリントの分析, *NISTEP DISCUSSION PAPER*, No.197, (2021). https://doi.org/10.15108/dp197