１Ｃ０７

「知」の生産活動におけるプレプリントの意義と役割 ―arXiv からのエビデンス―

○ 林 和弘（NISTEP），依田 洸（文部科学省），小柴 等（NISTEP），岡村 圭祐（文部科学省）

1　はじめに

近年のオープンサイエンスの潮流にあって，研究成果の発信・共有，そして研究コミュニティ内での評価に係る営みの顕著な割合を，研究分野によってはジャーナル論文だけでなく，その査読前段階のプレプリントが担うようになっている．近年の急速な ICT の進展，そして今般の COVID-19 流行下にあって，こうした動きにはさらに拍車がかかっている．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，そして頓に高まるその存在感に関する定量的なエビデンスは，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．本稿では，プレプリントサーバーの先駆けである「arXiv」に掲載された様々な分野の論文に関するビッグデータをもとに，現代の知の生産活動におけるプレプリントの意義と役割について計量書誌学的アプローチにより試行的に検証した結果を報告する．あわせて，研究評価の在り方等の観点を含め，今後の科学技術政策上の示唆についても議論したい．

2　背景

研究者は他の研究者との交流にあたって研究論文の執筆・発表を行う．多くの場合，研究論文の主な発表舞台である学術ジャーナル（論文誌）には，その出版プロセスにおいて，各学問分野に精通した研究者や専門家が事前に論文内容の評価を行う査読（ピアレビュー）プロセスが組み込まれており，これはジャーナル側にとっては質担保の観点からのスクリーニングとして機能する．査読者のコメントを踏まえて，ジャーナル編集部から著者に対して論文の不十分な点についての改善等を要求するプロセスが続くため，場合によっては論文投稿から最終的な掲載・出版までに年単位の時間を要することも珍しくない．このように，査読付きジャーナル論文では投稿から出版までの期間が比較的長くなるが，そうしたジャーナル側での査読を受けていないステータスで，著者による投稿と同時に近いタイミングで公開される論文原稿が「プレプリント」である．一般的なユーザーは金銭的な負担なくプレプリントの投稿・閲覧が可能であり，近年，そのユーザー層は急激に拡大してきた [林 20a]．プレプリントを通じた研究成果の共有様式は，近年のオープンサイエンスの潮流にあって，いまや多くの研究分野においてその研究活動を支える重要な要素となっている．

そうしたプレプリントを公開・管理するプレプリントサーバーの先駆けとなったのが，1991 年 8 月に運用開始された「arXiv（アーカイブ）」[*1]である [Ginsparg16]．後述のとおり，ライフサイエンス系や医学系の分野でプレプリント様式が取り入れられるようになったのはつい近年のことだが，物理学，数学，計算機科学等の分野では，arXiv が既にこの 30 年近くの間，研究コミュニティの活動にとって欠かせないプラットフォームを提供してきた．先行研究や最新の研究動向の把握，参考文献の収集，論文の執筆，ジャーナル掲載に先立つ先行的な公開・成果共有，成果に関する先取権の獲得，研究コミュニティからのフィードバックを踏まえた出版前の随時アップデートに至るまで，基本的な研究活動のおおよその部分がプレプリントを通じた研究成果の共有様式で十分に完結するとの指摘もあるほど，分野によってはその活用が進んでいる．近年では特に人工知能など情報分野での活用が急速に進んでおり [林 20b]，こうしたプレプリントの意義と役割，そして分野特性を踏まえた留意点等については多くの識者により語られてきた．

一般論として，ジャーナル論文とプレプリントとは，その公開までの期間の長短や論文の「質」保証等の観点から互いに相補的な機能が期待されるものであり，その相補性の程度も分野によって大きく異なるのが現状である．例えば物理学分野では，多くの場合，研究論文の著者は，論文原稿をまずプレプリントとして arXiv に投稿しつつ，同時に（少し時間をおいて）ジャーナルにも平行して投稿することで，最終的には両方のメリットを享

[*1] https://arxiv.org/

受しようとすることが多い．そうした研究者の行動原理を踏まえ，いまや多くのジャーナルがプレプリントを経ての論文投稿やその論文中でのプレプリントの引用を認めている[*2]．

こうした中，プレプリントを通じた研究成果の共有様式は近年，その対象分野が大きく拡大し，この数年でも数多くのプレプリントサーバーが立ち上がってきた．このうち，特に医学系やバイオ系のプレプリントは，昨今の新型コロナウイルス感染症 (COVID-19) の感染拡大抑制の及び治療法確立に向けた研究論文のオープン化と公表迅速化の流れを受け，研究者の研究活動に大きな影響を与え（cf. [池内 20a]），社会的にも大きな注目を集めてきた．例えば 2019 年に運用開始したばかりの医学系のプレプリントサーバー「medRxiv」では，COVID-19 に関連するものに限定しても，2020 年 1 月中旬から 5 月上旬までの約 5 か月間で 3 千件近くの投稿が行われている [小柴 20]．こうしたケースとともに「プレプリント」という研究成果の共有様式が社会的にも広く認知されていく中で，情報の信頼性や伝達の正確性といった観点で，これまでにない学術情報と社会との関わり方の難しさも浮き彫りとなってきた．COVID-19 研究関係では，特に研究論文としての「質」の観点が問題視されてきたが，これらの点については例えば文献 [池内 20a] を参照されたい．

プレプリントをめぐる動向は，アカデミアと社会とのより良い共創関係を目指す科学技術・イノベーション政策上も，固有の価値ある有用な情報源となり得る．例えば，研究評価に当たっては，いわゆる Top 10% 論文（各論文について）や インパクトファクター（各ジャーナルについて）等の定量的指標がこれまで広く使われてきたが，そこではプレプリントの関与する各種の活動状況は反映されてこなかった．ジャーナル論文と合わせてプレプリントの情報までをいかに活用していくべきかについて今後検討を進めていくに当たっては，以下の問：

プレプリントを通じた研究成果の共有様式は，現代の研究活動（論文引用による新たな知の生産活動）において，実際にどれほどの存在感・インパクトを担っているのか

に対する答について，定量的な裏付けを持って把握しておくことが EBPM (Evidence-Based Policy Making) の前提として欠かせない．他方で，この問に対する具体的な答は（世界的にも）見当たらず，プレプリント・サーバーに眠る情報を有用な情報源として活用し切れていないのが現状である．そこで本稿では，この問に向き合い一定の回答を与えるべく，まずは最も歴史があり投稿数も多いプレプリントサーバーである arXiv を対象に，そのデータを活用して試行的に分析した結果を概観的に報告する[*3]．

3 計量書誌学的アプローチに基づく分析

被引用回数ベースの定量的指標は，特に自然科学系の分野において研究論文の「質」を測る上での有用な媒介変数としてしばしば用いられ，知の生産活動における一種のインパクト指標と見なされてきた [孫 20]．本稿では，ジャーナル論文に留まらずプレプリントまでを含んだ形での被引用状況の分析を通じて，現代の知の生産活動の全体像及びその中でプレプリントの担っているインパクトを一部なりとも定量的に描き出すことを目指す．その際，被引用回数ベースの新たな指標を提案し，これをもとに分野横断的な議論を試みる．

3.1 分析手法

あるプレプリントに着目したとき，それがプレプリントサーバに投稿されてから何らかの媒体で出版されるまで，つまり DOI が付与されるまでの期間と，DOI が付与されてから現在（データ取得時点：2020 年初め）に至るまでの期間とを，それぞれ「プレプリント期」,「ジャーナル論文期」と呼称することにする[*4]．また，同じくあるプレプリントに着目したとき，それがデータ取得時点において獲得している総被引用回数のうち，それがプレプリント期にある文献からの引用なのか，あるいはジャーナル論文期にある文献からの引用なのかを区別して分析する[*5]．図 1 にその概念図を示す．その上で，本稿の関心対象とする新指標は以下の三つである：

指標 1　総被引用回数に占めるプレプリント期に獲得した被引用回数の割合：$\alpha\,(\%) = \frac{A+B}{A+B+C+D}$

[*2] 例えば，http://transpose-publishing.github.io/, https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy 等を参照．

[*3] さらに詳細な解説については，追加的な分析結果等とも合わせて別稿 [岡村 20] に譲る．

[*4] ジャーナル論文以外にも DOI の付与された文献も存在するほか，ジャーナル論文として出版されていても DOI 付与のない文献も存在し得るが，ここでは便宜上，DOI 付与をもってジャーナル論文としての出版ステータスの代理変数とする．

[*5] 引用は arXiv 外からのものを当然含んでおり，そこには会議の予稿集など DOI の付与されない文献も含まれる．ここでは，それらも含めて「プレプリント期」にカウントしている点には留意を要する．

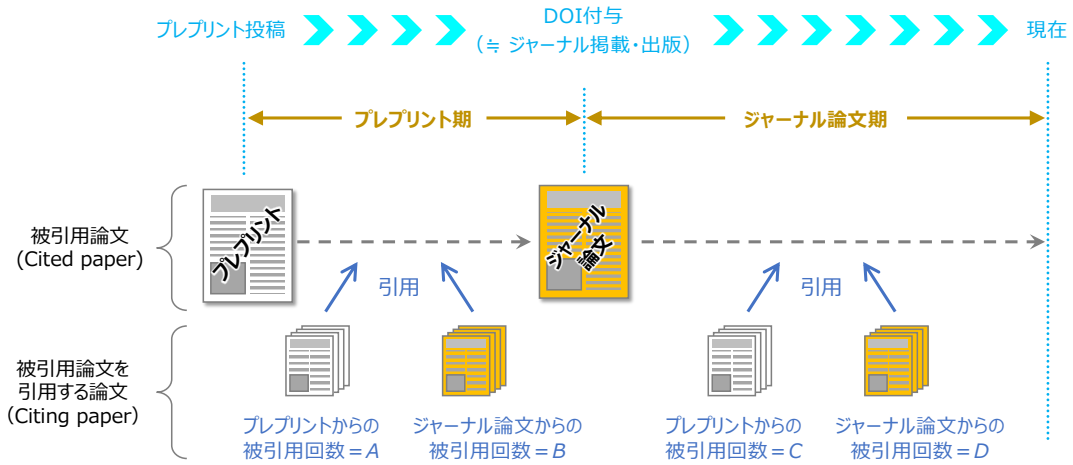

図 1: 被引用回数のカウントに関する整理

指標 2　総被引用回数に占めるプレプリントから獲得した被引用回数の割合：$\beta\,(\%) = \frac{A+C}{A+B+C+D}$

指標 3　総被引用回数に占めるプレプリントの関与する被引用回数の割合：$\gamma\,(\%) = \frac{A+B+C}{A+B+C+D}$

指標 1 は，論文が被引用を獲得するにあたってプレプリント期がいかに重要な稼ぎ時であるかを表すものと言える．また，指標 2 は，論文にとってプレプリントがいかに重要な被引用の獲得源であるかを表すものと言える．そして指標 3 にはこれらの両観点が含まれており，最も広い意味でプレプリントのインパクトを捕捉できる指標となっている．これらの指標設定のもとで分析を行った結果，指標 1 の値が α %，指標 2 の値が β %，指標 3 の値が γ % と算出されたなら，平均的には以下のとおり結論付ける（推定する）ことができる：

1. ある文献が獲得する被引用回数のうち α % は未出版の時期に獲得している．
2. ある文献が獲得する被引用回数のうち β % は未出版の文献から獲得している．
3. もしプレプリントという様式が存在しなければ（つまり，ジャーナル論文がジャーナル論文を引用するという形でしか文献引用がなされないのであれば），潜在的に獲得可能であった被引用回数のうち γ % 分を失うことになる．言い換えれば，文献引用を通じた知の生産活動の γ % はプレプリントなしには成立していない．

本稿第 3.3 節では実際の arXiv データについて α–γ 値を算出した結果をもとに議論を行う。

3.2　データ

本稿における分析で使用するデータセットは文献 [林 20b] で使用されたものと同じであり，データの取得法等についての詳細はそちらを参照されたい．arXiv データとしては，2020 年 1 月 21 日時点で収集可能なものを全収集しており，2020 年 1 月 17 日までに投稿された計 1,622,763 件のプレプリント情報を使用している．被引用回数についてのデータは 2020 年 1 月 24 日から 2 月 7 日までの期間で Semantic Scholar API を通じて取得している．いずれも，本稿における分析に当たっては「年」を時間に関する最小粒度として扱う．ここで，プレプリントの投稿年と DOI 付与年とが同じ場合，当該年に獲得した被引用回数は，本稿ではプレプリント期に獲得されたものと整理して分析を行う[*6]．被引用回数は，Semantic Scholar 側で推定・同定されており，プレプリントがその後ジャーナル論文に採録され，その形で引用された場合でも同じプレプリント ID において一括管理されている．分野カテゴリーの分類としては，arXiv で使用されている 153 分野分類に基づき，arXiv の統計情報サイト[*7] で利用されているものと同じ粒度で表 1 のとおり 6 分野に大括り化したものを採用した[*8]．

[*6] これをジャーナル論文期に獲得されたものと整理して分析を行った結果との比較については別稿 [岡村 20] に譲るが，いずれにしろ定性的な結論は本稿で得られるものと変わらない．

[*7] `https://arxiv.org/help/stats/2019_by_area/index`

[*8] arXiv では一つのプレプリントに複数の分野カテゴリーを設定（cross-listing）できるため，今回の分析で使用した arXiv でも，プレプリントによっては複数の分野カテゴリー属性を持つこともあるが，その場合は個々の分野に一つずつカウントしている．

表 1: 6 分野カテゴリー（arXiv 分類の大括り化）

分類カテゴリー	arXiv 上の分類
天体物理学	Astrophysics (astro-ph*)
物性物理学	Condensed Matter Physics (cond-mat*)
計算機科学	Computer Science (cs*)
高エネルギー物理学	High Energy Physics (hep-*)
数学	Mathematics (math + math-ph)
その他の物理学	physics + nucl + gr-qc + quant-ph + nlin

3.3 分析結果及び考察

まず arXiv への新規プレプリント投稿数の分野毎の経年推移を図 2 に示す．先の arXiv の統計情報サイトでも指摘されているとおり，近年の AI 研究進展を反映して計算機科学分野カテゴリーの新規投稿数が急増している．2000 年代後半までは 6 分野の中でも最低位にあったものが，2016 年以降は他分野を大きく引き離す勢いで伸びている様子が見て取れ，2019 年に至っては新規投稿数が約 7 万報を記録している．数学分野も 4 万報強の新規投稿数をマークしており，高度情報化社会にあっての近年の数学研究の盛り上がりを象徴していると言える．

図 3 はプレプリント期（DOI が付与されていない期間）の長さの分布を分野毎に表したものであるが，分野間で顕著な傾向差があることが見て取れる．物理学カテゴリーの分野群では，プレプリント投稿から概ね 2 年間ほどで過半数のプレプリントに対して DOI が付与されているのに対し，計算機科学や数学では投稿後初めの数年間への集中やその後の減衰の度合いが緩やかである．これらの分野では，そもそも DOI が付されないままに現在に至るプレプリントの割合が大きい（計算機科学：68.5%，数学：77.2%）．そのようなプレプリントが毎年積み上がっていくために，図 3 に見るとおり，これらの分野については経過年数に対して heavy-tailed なヒストグラムとなっている[*9]．今後の分析や考察において，被引用回数にまつわる各種指標について分野間比較を見ていくに当たっては，図 3 に見られる時間スケールの違いを念頭に置いて解釈していくことが重要となる．

図 4 は総被引用回数に占めるプレプリント期に獲得した被引用回数の割合（指標 1：α 値）の経年推移を分野毎に表したものである．この指標は先のとおり，端的に言うなら，論文が被引用を獲得するにあたってプレプリント期がいかに重要な稼ぎ時であるかを表すものである．まず全体に着目すれば，計算機科学・数学とそれ以外とで傾向が大きく異なる[*10]．計算機科学・数学の場合は α 値が高く，特に数学分野の場合には期間を通じて平均的に 60% から 70% ほどの被引用価値がプレプリント期に集中していることがわかる．計算機科学分野についても高い値を維持しているが，2006 年頃（ちょうど第 3 次 AI ブーム：ディープラーニング時代が始まった時期）を境に以下のとおり傾向に差が見られる．2006 年以降に投稿されたプレプリント（図の左半分）は，平均的に見たとき，時間の経過に比例して α 値が下がる傾向にあ

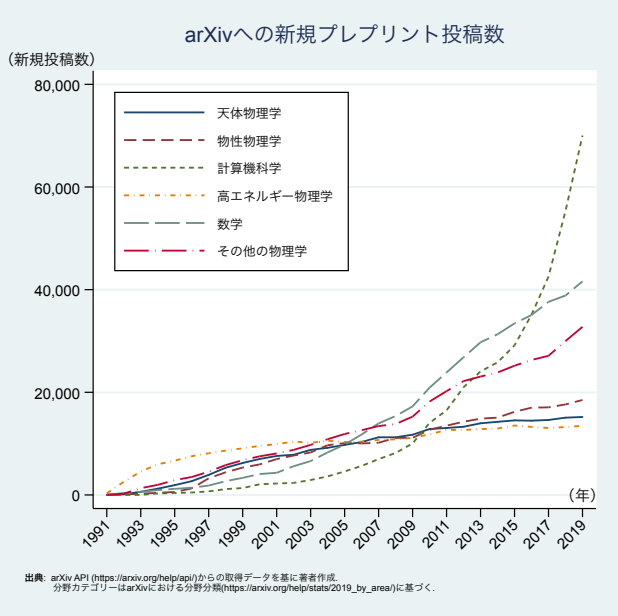

図 2: arXiv への新規プレプリント投稿数

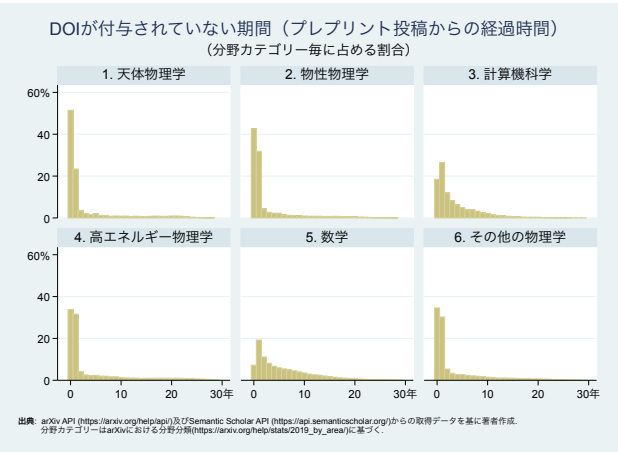

図 3: 「プレプリント期」の長さ

[*9] 数学分野では他分野と比べ DOI 付与までに比較的長い時間を要することは文献 [林 20b] でも指摘されている．

[*10] 用語の定義上，投稿後に DOI が付与されないプレプリントについては，全期間をプレプリント期として扱っていること，また，分野ごとに DOI 付与率は大きく異なり，特に計算機科学分野は DOI 付与率が低いとの指摘 [林 20b] もあることから，解釈に当たってはこれらの点に注意を要する．

り，このことは計算機科学分野のプレプリントがジャーナル論文として出版された後，10 数年ほどの長期にわたって引用され続ける傾向にあることを示している．これに対し，2007 年以前に投稿された同分野のプレプリント（図の右半分）は全体的に右上がりとなっており，特に 1990 年台（arXiv 立上げ後初めの 10 年間）では α 値は 100% に近い．他方で，物理学系の 4 分野については，一定程度の年数（5 年程度）が経過した後は，α 値は概ね 15%〜30% で推移しており，経過年数の長い（古い）論文ほどその値は上昇する傾向にある．分野によって多少の差はあれ，いずれの分野でも α 値が一定の高さを持って推移している様子は，それだけ被引用回数を稼ぐ上でプレプリント期が重要な役割を担っているということ，言い換えれば，新たな知を生み出していく上でプレプリント様式が本質的に重要な役割を果たしていることの表れと言える．

図 5 は総被引用回数に占めるプレプリントから獲得した被引用回数の割合（指標 2：β 値）の経年推移を分野毎に表したものである．これは，現在（2020 年初頭）までに文献が稼いだ被引用回数のうち，現在に至るまで DOI が付与されずに残っているプレプリントからの被引用回数の割合を示したものである．面白いことに，一定期間（10 年間ほど）以上経過した後は，多少の幅こそあれ，6 分野カテゴリー間でそれほど相違なく，β 値は概ね 25%〜40% の範囲に収まって推移している．その間，古い論文ほど β 値は上昇傾向にある．

最後に，総被引用回数に占めるプレプリントの関与する被引用回数の割合，すなわち，プレプリント期に獲得したあるいはプレプリントから獲得した被引用回数の割合（指標 3：γ 値）について見る．紙面の制約上，グラフは省略するが，その定義上も明らかなとおり，概ね α 値（図 4）と β 値（図 5）の傾向を併せ持つものとなる．γ 値の比較的安定する期間（プレプリント投稿からの経過時間が 5 年以上 20 年未満）について平均値をとれば，物理系の 4 分野ではいずれも 4 割前後，計算機科学分野では約 7 割，数学分野では約 8 割という結果になる．したがって，分析や解釈上の諸条件や制約には十分に留意する必要があるものの，端的かつ標語的にまとめるなら：

論文引用を通じた「知の生産活動」の顕著な割合—物理系の分野では約 *4* 割，計算機科学・数学分野では約 *7*〜*8* 割—はプレプリントなしには成立していない

と言えることになる．

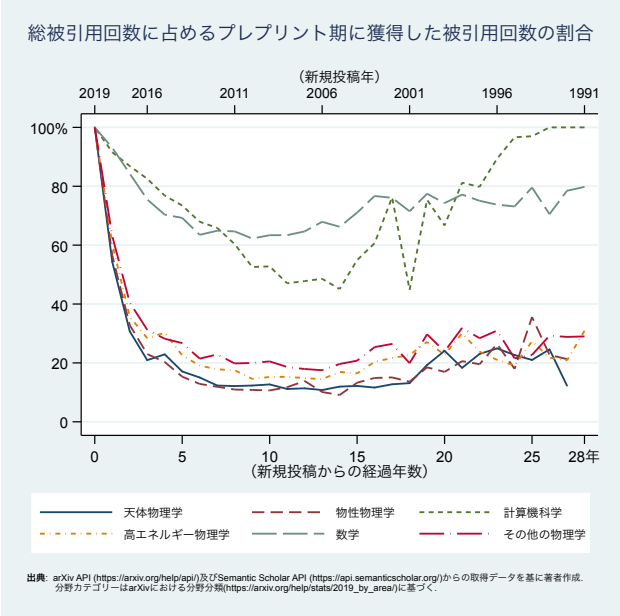

図 4: 総被引用回数に占めるプレプリント期に獲得した被引用回数の割合（α 値）

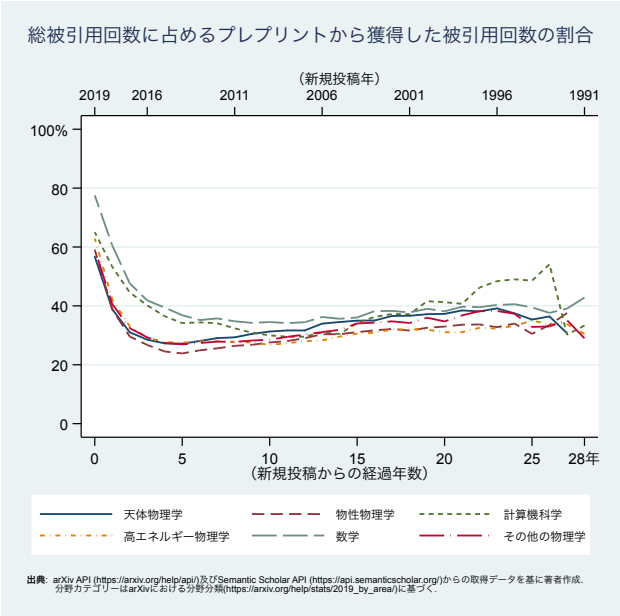

図 5: 総被引用回数に占めるプレプリントから獲得した被引用回数の割合（β 値）

4　今後の科学技術政策への示唆

学術情報の社会への発信に際してよく課題として挙げられる情報の正確さ・わかりやすさの観点，そして論文（プレプリント）そのものの「質」の問題については稿を改めて論じる [岡村 20] こととし，ここでは特に研究評価の観点から今後の科学技術政策への示唆を論じたい．

これまでの研究評価では，各種の政府目標等にもあ

るとおり，いわゆる Top 10% 論文やインパクトファクターといった定量的指標が幅広く使われてきた．しかしながら，これらはいずれも査読付きジャーナルやそれに掲載された論文を対象として定義・算出されているものであり，本稿で詳細分析の対象としたプレプリントまでを含むものではなかった．また，こうしたジャーナル論文ベースの指標をもとにした分析や評価は，計算機科学（情報科学）[*11]等の一部の分野では特別なバイアスがかかりがちであった．今回の調査・分析を通じて，プレプリントがいまや現代の知の生産活動にとって欠かせない機能を担っていること，そしてその存在感や重要性は年々増していることが，今回提案された新しい指標を通じても実際に見て取れたことは，研究評価の在り方を含む今後の科学技術行政を考える上で示唆に富む．プレプリントの役割や意義について研究者や政策関係者が半ば肌感覚で抱いてきた印象が，今回の調査によって定量的な裏付けを持って検証されたことも意義深い．例えば，政府の科学技術基本計画においても標榜する Society 5.0 を描く上で政策的にも大きな注目と期待を集めている人工知能研究を含む計算機科学系の分野について，ジャーナル論文ベースの指標からはなかなかつかみきれない研究動向やインパクトに関する情報やエビデンスが，こうしてプレプリントのデータから新たに得られたことは，今後の科学技術政策を EBPM の観点から支えていく上でも有用である．

加えて，世界的潮流として，研究費申請や研究実績報告等に際しても，最新の研究動向の説明やそれをもとにした研究者個人の今後の研究計画の提案等を行う際，プレプリント情報を活用していくことが効果的である場面は今後益々増えていくと考えられる．世界各国の代表的なファンディング・エージェンシー（FA）・研究支援団体やその研究費事業等における研究評価（申請書，中間・事後評価書等）に際してのプレプリントの扱いについては，FA 毎に対処方針が異なっているのが現状である[*12]．今後，我が国においても，プレプリントを通じた活動状況を研究評価にどのように反映していくかについて，一定の考え方を整理する必要が生じてくるだろう．

今般，NISTEP において日本国内の研究機関に在籍する研究者等を対象としたプレプリント関連のアンケート調査 [池内 20b] が行われているが，その結果からも，プレプリントの利用実績や定着度は分野によって大きく異なることが示唆されている．今後の科学技術政策への反映に当たっては，こうした分野特性にも十分に留意した上でさらに検討が深められていく必要があるだろう．

本稿で紹介した今回の成果が，現状では数少ない（特に定量的な）エビデンス・ベースの一つとして今後の科学技術政策に活かされていくことを期待したい．また，研究者や政策関係者のみならず，広くアカデミア，行政，政治，民間，そして科学ジャーナリズムを含む関係者間でさらなる調査・分析や議論が行われる際の検討材料として活かされていくなら幸いである．

謝辞

今回の調査・分析にあたり御知見・御助言をいただいた野崎 光昭氏（高エネルギー加速器研究機構; KEK），引原 隆士氏（京都大学），武田 英明氏（国立情報学研究所; NII）に御礼申し上げる．

参考文献

[Ginsparg16] Ginsparg, Paul：Preprint Déjà Vu. *The EMBO Journal,* Vol.35, No.24, pp.2620–2625, Oct 2016. https://doi.org/10.15252/embj.201695531

[池内 20a] 池内 有為： オープンサイエンスの効果と課題―新型コロナウイルスおよび COVID-19 に関する学術界の動向． 情報の科学と技術， Vol.70, No.3, pp.140–143, Mar 2020. https://doi.org/10.18919/jkg.70.3_140

[池内 20b] 池内 有為, 林 和弘：日本の研究者によるプレプリントの利活用状況と認識（仮題）． *NISTEP Discussion Paper. (To appear)*

[岡村 20] 岡村 圭祐，依田 洸，林 和弘，小柴 等：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（仮題）． *MEXT Evidence Paper. (To appear)*

[小柴 20] 小柴 等，林 和弘，伊藤 裕子： COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析． *NISTEP Discussion Paper,* No.186, Jun 2020. http://doi.org/10.15108/dp186

[住井 19] 住井 英二郎：「情報系」の業績評価について ―「若手」研究者の視点から―. 日本学術会議科学者委員会研究評価分科会 公開シンポジウム「研究評価の客観化と多様化をめざして–分野別研究評価の現状と課題」, 2019. http://www.scj.go.jp/ja/event/pdf2/190524-5.pdf

[孫 20] 孫 媛： 引用に基づく学術研究のインパクト評価． 情報の科学と技術， Vol.70, No.5, pp.255–260, May 2020. https://doi.org/10.18919/jkg.70.5_255

[林 20a] 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,* Vol.6, No.1, Mar 2020. https://doi.org/10.15108/stih.00205

[林 20b] 林 和弘，小柴 等： arXiv に着目したプレプリントの分析． *NISTEP Discussion Paper,* No.187, Aug 2020. http://doi.org/10.15108/dp187

[*11] 研究評価に際して，いわゆるトップカンファレンスでの採択実績が重視され，ジャーナル論文そのものは比較的重視されないとの指摘もある [住井 19].

[*12] 例えば，https://asapbio.org/funder-policies/ を参照.