DISCUSSION PAPER No.187

arXiv に着目したプレプリントの分析

Analysis of preprints on arXiv

2020 年 8 月

文部科学省　科学技術・学術政策研究所

林 和弘　　小柴 等

【執筆者】

林 和弘　　文部科学省科学技術・学術政策研究所　科学技術予測センター　上席研究官

小柴 等　　文部科学省科学技術・学術政策研究所　第2調査研究グループ　上席研究官

【Authors】

HAYASHI Kazuhiro　Science and Technology Foresight Center,
National Institute of Science and Technology Policy (NISTEP), MEXT

KOSHIBA Hitoshi　2nd Policy-Oriented Research Group,
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 paper.

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, 文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/dp187

HAYASHI Kazuhiro, KOSHIBA Hitoshi "Analysis of preprints on arXiv," *NISTEP DISCUSSION PAPER*, No.187, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/dp187

arXiv に着目したプレプリントの分析

文部科学省　科学技術・学術政策研究所
林 和弘，小柴 等

要旨

定量的なデータに裏打ちされたエビデンスに基づく科学技術政策形成が求められる中，学術ジャーナルに掲載される原著論文の量（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試行分析を行った。

プレプリントとしては，もっとも歴史が長いプレプリントサーバである arXiv 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，arXiv の特徴および分野別特性を分析した。

その結果，プレプリント公開から，原著論文になるまでの期間に分野による差が見られることや，情報系を中心に，必ずしも原著論文を出口としないプレプリントが多数掲載されていることなどが分かった。

また，Award 情報との一定の紐付けも可能であり，我が国のファンデイング政策の効果を多少観察できる可能性や，被引用数の分析によって，分野ごとのプレプリント利用スタイルが大きく異なる可能性も見出した。

Analysis of preprints on arXiv

HAYASHI Kazuhiro, KOSHIBA Hitoshi
National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

In order to complement the evidence-based science and technology policy formation backed by quantitative data, we conducted a pilot analysis focusing on preprints, which are drafts of original papers, for the purpose of complementing research on the quality of original papers based on the quantity (number of papers) and the number of citations of original papers published in academic journals. We focus on arXiv, which is the oldest preprint server, and analyze its characteristics and field characteristics in terms of its relationship to the original paper and citation of preprints.

The results show that there is a difference in the time period between the release of preprints and the publication of the original paper depending on the discipline, and that there are many preprints published, mainly in the information field, that do not necessarily exit from the original paper. We also found that there is a possibility that we can observe the effect of the Japanese research funding policy, and that the style of preprints usage differs greatly depending on the analysis of the number of citations.

By analyzing the set of preprints while taking into account the characteristics of these preprints, we were able to show the possibility of complementary analysis of aspects of science and technology that cannot be determined from the previous analysis of papers and citations alone. It is hoped that the addition of analysis of preprint servers and other preprints from other disciplines will allow us to generalize the preprint analysis by comparing the preprints' persistence and interdisciplinarity.

目次

本文

1　序論

我が国の科学技術イノベーション（以下，STI とする）政策立案において，エビデンスに基づく政策立案 (以下，EBPM：Evidence-based Policy Making) 機能の強化が求められている．「第 5 期科学技術基本計画 (平成 28 年 1 月 22 日閣議決定)」においても，エビデンスに基づく政策立案等を推進することとされ，内閣府などにおいてもエビデンス等を整備する関連する取組が進められてきた．

その中でも研究力については，計量書誌学や科学計量学を基礎とした学術ジャーナルの査読を通った原著論文（以下，原著論文とする）に着目した定量的な分析と，政策づくりへの反映が試みられている．例えば，科学技術・学術政策研究所（以下 NISTEP とする）においては，サイエンスマップ，国別ベンチマーキング，大学ベンチマーキング，などの調査分析を行い，その内容が STI 政策づくりの一助となっている．一方，原著論文の分析においては，研究成果が生まれてから，査読・編集・出版を経て公開されるまでのタイムラグが含まれるため，全体の傾向（trends）をレビューすることには向いているが，新興領域を早く押さえることは構造的に難しい．また，原著論文においては原則それぞれの領域で確立された科学的判断基準によって査読が行われるため，学際領域や融合領域，あるいは全く新しい概念の論文が不利になりやすい．

ここで，データジャーナルなども含めた研究プロセスにおける研究データの共有や公開に着目して，より早期の把握を行うことも考えられる．こうした動向はオープンサイエンスの一環として近年大きな注目を浴びており，その将来性は大きく期待されるものの，原著論文とは違って，データの粒度，形式，流通形態が標準化されていないために，現状では研究力を測る分析に原著論文と同じレベルで適用することが難しい．

そこで本報では，論文原稿の草稿であり，査読による選別もされていない「プレプリント」に着目し，その分析を行うことで，原著論文を基とした研究力の分析に対して相補的に新しい知見を得ることができるのではないかと考えた．

本調査研究は，プレプリントサーバに登載されたプレプリントの分析について，プレプリント・サーバとしての嚆矢とされる arXiv に着目して試行的に行うと共に，科学技術政策への示唆を得ることを目的に行った．

2　プレプリント，プレプリントサーバと arXiv

2.1　プレプリント，プレプリントサーバ と arXiv の概要

プレプリントとは，主に査読付きジャーナルに投稿する前の草稿原稿のことを指す．したがってプレプリントは論文の体裁は満たしているものの，査読済みでも出版されたものでもないという位置づけである[1)]．プレプリントを研究者仲間に事前に共有して意見を求めることは従来より分野を問わず広く行われる情報共有活動であった [1] が，1990 年代に入って Web が登場すると，このプレプリントを Web に掲載して誰でも読めるようにするプレプリントサーバが物理系分野で登場し，新しい知見の迅速な共有とより多くのフィードバックを得ることができるようになった．

現在，プレプリントサーバにも様々なものが存在する [1] が，その代表例としては arXiv（アーカイブ）が挙げられる．arXiv はプレプリントサーバの嚆矢で，1991 年にロスアラモス国立研究所の Paul Ginsparg 氏 によって開設されたものである．現在は米国コーネル大学を中心として世界各国の協力のもとに運営され，特に物理・数学・情報系の分野でメジャーかつ最も歴史が長いプレプリントサーバとなっている．

2.2　arXiv の活用と課題

2020 年現在，arXiv は物理・数学・情報系の分野におけるプレプリントサーバの代名詞ともなっている．この主要収録分野のうち，物理・数学系においては，査読付きジャーナルに投稿する前か同時にプレプリントを公開する慣習が広がっている．

また，arXiv には情報系のプレプリントも多数収録されている．情報系は，「進展速度が速い研究領域のために出版までに時間がかかる原著論文よりもトップカンファレンスの予稿などが重視される」と言われて [2, 3] おり，情報系分野の動向把握等を迅速に行う上で特に有用と考えられる．その他，原著論文を対象とした いわゆる論文データベース を用いた書誌情報分析との差分としては，1．論文投稿に先立って登録されるというプレプリントサーバの性質上，投稿から論文誌採録までの査読および出版期間についての情報を得られる可能性が高いこと，2．また論文誌への掲載が行われていないと目されるものと，掲載が行われたものとの比較ができること，など（一部の論文について）論文誌掲載前後の比較が可能であるということが想定される．したがって，上手く活用すれば既存の論文データベースを用いた書誌情報分析を補完するものとして機能する可能性が高い．

例えば文献 [4] においても複数の研究者へのインタビュー結果として“ **特に，プレプリントについては，雑誌への論文投稿後，査読が完了するまでにプレプリントサーバにアップロードすることが急速に一般化しており，いち早く情報を取得できる場として有効であるとの意見が，複数の研究者から得られている．** ”と，記載されている．

しかしながら，これまで arXiv 上の論文に関するまとまった実態調査結果等は見当たらない．そこで，本論文では arXiv に投稿された論文の全体的な傾向や，これらのデータを使ったいくつかの分析を試み，arXiv の状況を明らかにするとともに，研究分野の動向把握等への活用可能性など科学技術・学術政策への寄与について検討する．

[1)] 公開することで，より多くの読者の目に触れることになり，これにより読者全員が査読者であるという見方もあるが，一般的にジャーナル論文でいう意味での査読とは異なる．

3　分析の手法

分析の手法について以下にまとめる．

まず arXiv の論文書誌データについては，arXiv が提供する API [2] を通じて収集する．ここでは，arXiv 上の投稿分野と投稿日時の範囲を指定することで，作業時点で収集可能な全ての論文の書誌データを収集する．なお，今回は論文そのものについては収集を行わない．

書誌データの項目としては例えば以下が挙げられる．

- arXiv の ID
- タイトル
- 概要
- 著者名
- 投稿先分野
- 初版投稿日
- 最終更新日
- DOI

この他にも項目が存在するが，詳細は arXiv API の User Manual [3] を参照いただきたい．

ここで，AI2 [4] が提供する Semantic Scholar [5] では，arXiv の引用文献情報（任意の arXiv の論文を引用している文献の情報）を得ることができる．そこで，Semantic Scholar の API [6] を通じて，arXiv の論文ごとに被引用先のデータ（タイトル，雑誌名，DOI 等）も収集する．なお今回は対象論文数が膨大なため，収集には約 2 週間の期間がかかっており，序盤に収集したものと，終盤に収集したものとで約 2 週間分期間が異なる．また，上記収集期間における被引用数であるため，将来に向かって数が増えていく可能性が高い．結果の分析においては，これらの点について若干の注意を要する．

次に，DOI が付与されているものについては，Crossref[7]が提供する Crossref REST API [8] を用いると，雑誌名や公開日などの情報を得ることができる．そこで，DOI が付与されている arXiv の論文については，この API を通じて，掲載雑誌名等のデータも収集する．

以上の収集したデータについて，分野，期間，DOI の有無，などの軸で計量することで，arXiv の状況を明らかにするとともに，研究分野の動向把握等への活用可能性について検討する．

---

[2] https://arxiv.org/help/api
[3] https://arxiv.org/help/api/user-manual
[4] Allen Institute for AI
[5] https://www.semanticscholar.org/
[6] http://api.semanticscholar.org/
[7] https://www.crossref.org/
[8] https://github.com/CrossRef/rest-api-doc

4　結果

4.1　データ件数等

2020 年 1 月 21 日時点で収集可能なものを全収集し，以下の通りとなった.

データ総数　1,622,763 件
期間　1986 年 4 月 25 日 〜 2020 年 1 月 17 日

arXiv のサービス開始は 1991 年であるため，1986 年 4 月 25 日に登録というデータは矛盾するが，ここでは収集データの内容をそのまま採用した.

4.2　arXiv の分野について

arXiv には独自の分野割りがなされおており，2020 年 1 月 21 日時点では，以下の 8 分類と，それらを更に細分化した 153 の小分類が設定されている．

1. Physics
2. Mathematics
3. Computer Science
4. Quantitative Biology
5. Quantitative Finance
6. Statistics
7. Electrical Engineering and Systems Science
8. Economics

なお，投稿時には上記の小分類の中から少なくとも一つの分野を選択する[9)]．

投稿数等に応じて 8 分類それぞれに属する小分類の数にも偏りがあるため，本論文では，小分類（分野）をベースに独自に再整理し，以下の表 1〜10 に示す 12 分野（大分野）を設定している．

表 1　分野分類 (1/10)

大分野	分野	説明（原文）	説明（機械翻訳）
astro-ph	astro-ph	Astrophysics	天体物理学
astro-ph	astro-ph.CO	Cosmology and Nongalactic Astrophysics	宇宙論と非銀河天体物理学
astro-ph	astro-ph.EP	Earth and Planetary Astrophysics	地球惑星天体物理学
astro-ph	astro-ph.GA	Astrophysics of Galaxies	銀河の天体物理学
astro-ph	astro-ph.HE	High Energy Astrophysical Phenomena	高エネルギー天体物理現象
astro-ph	astro-ph.IM	Instrumentation and Methods for Astrophysics	天体物理学の計測と方法
astro-ph	astro-ph.SR	Solar and Stellar Astrophysics	太陽および星の天体物理学
astro-ph	gr-qc	General Relativity and Quantum Cosmology	一般相対性理論と量子宇宙論
cond-mat	cond-mat.dis-nn	Disordered Systems and Neural Networks	無秩序システムとニューラルネットワーク
cond-mat	cond-mat.mes-hall	Mesoscale and Nanoscale Physics	メソスケールおよびナノスケールの物理学
cond-mat	cond-mat.mtrl-sci	Materials Science	材料科学
cond-mat	cond-mat.other	Other Condensed Matter	その他の凝縮物質
cond-mat	cond-mat.quant-gas	Quantum Gases	量子ガス
cond-mat	cond-mat.soft	Soft Condensed Matter	ソフト凝縮物質
cond-mat	cond-mat.stat-mech	Statistical Mechanics	統計力学
cond-mat	cond-mat.str-el	Strongly Correlated Electrons	強く相関した電子
cond-mat	cond-mat.supr-con	Superconductivity	超伝導

9) したがって，複数の分野を割り当てることもできる

表 2　分野分類 (2/10)

大分野	分野	説明（原文）	説明（機械翻訳）
cs	cs.AI	Artificial Intelligence	人工知能
cs	cs.AR	Hardware Architecture	ハードウェアアーキテクチャ
cs	cs.CC	Computational Complexity	計算の複雑さ
cs	cs.CE	Computational Engineering, Finance, and Science	計算工学、金融、科学
cs	cs.CG	Computational Geometry	計算幾何学
cs	cs.CL	Computation and Language	計算と言語
cs	cs.CR	Cryptography and Security	暗号化とセキュリティ
cs	cs.CV	Computer Vision and Pattern Recognition	コンピュータビジョンとパターン認識
cs	cs.CY	Computers and Society	コンピューターと社会
cs	cs.DB	Databases	データベース
cs	cs.DC	Distributed, Parallel, and Cluster Computing	分散、並列、およびクラスターコンピューティング
cs	cs.DL	Digital Libraries	デジタル図書館
cs	cs.DM	Discrete Mathematics	離散数学
cs	cs.DS	Data Structures and Algorithms	データ構造とアルゴリズム
cs	cs.ET	Emerging Technologies	新技術
cs	cs.FL	Formal Languages and Automata Theory	形式言語とオートマトン理論
cs	cs.GL	General Literature	一般文学

表 3　分野分類 (3/10)

大分野	分野	説明（原文）	説明（機械翻訳）
cs	cs.GR	Graphics	グラフィックス
cs	cs.GT	Computer Science and Game Theory	コンピュータサイエンスとゲーム理論
cs	cs.HC	Human-Computer Interaction	人間とコンピューターの相互作用
cs	cs.IR	Information Retrieval	情報検索
cs	cs.IT	Information Theory	情報理論
cs	cs.LG	Learning	学習
cs	cs.LO	Logic in Computer Science	コンピュータサイエンスのロジック
cs	cs.MA	Multiagent Systems	マルチエージェントシステム
cs	cs.MM	Multimedia	マルチメディア
cs	cs.MS	Mathematical Software	数学ソフトウェア
cs	cs.NA	Numerical Analysis	数値解析
cs	cs.NE	Neural and Evolutionary Computing	ニューラルおよび進化コンピューティング
cs	cs.NI	Networking and Internet Architecture	NWとインターネットのアーキテクチャ
cs	cs.OH	Other Computer Science	その他のコンピューターサイエンス
cs	cs.OS	Operating Systems	オペレーティングシステム
cs	cs.PF	Performance	性能
cs	cs.PL	Programming Languages	プログラミング言語

表 4　分野分類 (4/10)

大分野	分野	説明（原文）	説明（機械翻訳）
cs	cs.RO	Robotics	ロボティクス
cs	cs.SC	Symbolic Computation	シンボリック計算
cs	cs.SD	Sound	音
cs	cs.SE	Software Engineering	ソフトウェア工学
cs	cs.SI	Social and Information Networks	ソーシャルおよび情報ネットワーク
cs	cs.SY	Systems and Control	システムと制御
cs	eess.AS	Audio and Speech Processing	オーディオおよび音声処理
cs	eess.IV	Image and Video Processing	画像およびビデオ処理
cs	eess.SP	Signal Processing	信号処理
econ	econ.EM	Econometrics	計量経済学
hep	hep-ex	High Energy Physics - Experiment	高エネルギー物理学-実験
hep	hep-lat	High Energy Physics - Lattice	高エネルギー物理学-格子
hep	hep-ph	High Energy Physics - Phenomenology	高エネルギー物理学-現象学
hep	hep-th	High Energy Physics - Theory	高エネルギー物理学-理論

表 5　分野分類 (5/10)

大分野	分野	説明（原文）	説明（機械翻訳）
math	math-ph	Mathematical Physics	数理物理学
math	math.AC	Commutative Algebra	可換代数
math	math.AG	Algebraic Geometry	代数幾何学
math	math.AP	Analysis of PDEs	PDEの分析
math	math.AT	Algebraic Topology	代数トポロジー
math	math.CA	Classical Analysis and ODEs	古典分析とODE
math	math.CO	Combinatorics	組み合わせ論
math	math.CT	Category Theory	カテゴリー理論
math	math.CV	Complex Variables	複雑な変数
math	math.DG	Differential Geometry	微分幾何学
math	math.DS	Dynamical Systems	動的システム
math	math.FA	Functional Analysis	機能的解析
math	math.GM	General Mathematics	一般数学
math	math.GN	General Topology	一般的なトポロジ
math	math.GR	Group Theory	群論
math	math.GT	Geometric Topology	幾何学的トポロジー
math	math.HO	History and Overview	歴史と概要

表 6　分野分類 (6/10)

大分野	分野	説明（原文）	説明（機械翻訳）
math	math.IT	Information Theory	情報理論
math	math.KT	K-Theory and Homology	K理論とホモロジー
math	math.LO	Logic	論理
math	math.MG	Metric Geometry	メトリックジオメトリ
math	math.MP	Mathematical Physics	数理物理学
math	math.NA	Numerical Analysis	数値解析
math	math.NT	Number Theory	数論
math	math.OA	Operator Algebras	演算子代数
math	math.OC	Optimization and Control	最適化と制御
math	math.PR	Probability	確率
math	math.QA	Quantum Algebra	量子代数
math	math.RA	Rings and Algebras	環と代数
math	math.RT	Representation Theory	表現論
math	math.SG	Symplectic Geometry	シンプレクティックジオメトリ
math	math.SP	Spectral Theory	スペクトル理論
math	math.ST	Statistics Theory	統計理論

表 7　分野分類 (7/10)

大分野	分野	説明（原文）	説明（機械翻訳）
nlin	nlin.AO	Adaptation and Self-Organizing Systems	適応と自己組織化システム
nlin	nlin.CD	Chaotic Dynamics	カオスダイナミクス
nlin	nlin.CG	Cellular Automata and Lattice Gases	セルオートマトンと格子ガス
nlin	nlin.PS	Pattern Formation and Solitons	パターン形成とソリトン
nlin	nlin.SI	Exactly Solvable and Integrable Systems	厳密に可解で統合可能なシステム
nucl	nucl-ex	Nuclear Experiment	核実験
nucl	nucl-th	Nuclear Theory	核理論

表 8　分野分類 (8/10)

大分野	分野	説明（原文）	説明（機械翻訳）
physics	physics.acc-ph	Accelerator Physics	加速器の物理
physics	physics.ao-ph	Atmospheric and Oceanic Physics	大気海洋物理学
physics	physics.app-ph	Applied Physics	応用物理学
physics	physics.atm-clus	Atomic and Molecular Clusters	原子および分子クラスター
physics	physics.atom-ph	Atomic Physics	原子物理学
physics	physics.bio-ph	Biological Physics	生物物理学
physics	physics.chem-ph	Chemical Physics	化学物理学
physics	physics.class-ph	Classical Physics	古典物理学
physics	physics.comp-ph	Computational Physics	計算物理学
physics	physics.data-an	Data Analysis, Statistics and Probability	データ分析、統計および確率
physics	physics.ed-ph	Physics Education	物理教育
physics	physics.flu-dyn	Fluid Dynamics	流体力学
physics	physics.gen-ph	General Physics	一般物理学
physics	physics.geo-ph	Geophysics	地球物理学
physics	physics.hist-ph	History and Philosophy of Physics	物理学の歴史と哲学
physics	physics.ins-det	Instrumentation and Detectors	計装と検出器
physics	physics.med-ph	Medical Physics	医学物理学

表 9　分野分類 (9/10)

大分野	分野	説明（原文）	説明（機械翻訳）
physics	physics.optics	Optics	光学
physics	physics.plasm-ph	Plasma Physics	プラズマ物理学
physics	physics.pop-ph	Popular Physics	人気の物理学
physics	physics.soc-ph	Physics and Society	物理学と社会
physics	physics.space-ph	Space Physics	宇宙物理学
physics	quant-ph	Quantum Physics	量子物理学
q-bio	q-bio.BM	Biomolecules	生体分子
q-bio	q-bio.CB	Cell Behavior	セルの挙動
q-bio	q-bio.GN	Genomics	ゲノミクス
q-bio	q-bio.MN	Molecular Networks	分子ネットワーク
q-bio	q-bio.NC	Neurons and Cognition	ニューロンと認知
q-bio	q-bio.OT	Other Quantitative Biology	その他の定量生物学
q-bio	q-bio.PE	Populations and Evolution	人口と進化
q-bio	q-bio.QM	Quantitative Methods	定量的な方法
q-bio	q-bio.SC	Subcellular Processes	細胞内プロセス
q-bio	q-bio.TO	Tissues and Organs	組織と臓器

表 10　分野分類 (10/10)

大分野	分野	説明（原文）	説明（機械翻訳）
q-fin	q-fin.CP	Computational Finance	計算ファイナンス
q-fin	q-fin.EC	Economics	経済
q-fin	q-fin.GN	General Finance	一般金融
q-fin	q-fin.MF	Mathematical Finance	数理ファイナンス
q-fin	q-fin.PM	Portfolio Management	ポートフォリオ管理
q-fin	q-fin.PR	Pricing of Securities	証券の価格
q-fin	q-fin.RM	Risk Management	危機管理
q-fin	q-fin.ST	Statistical Finance	統計ファイナンス
q-fin	q-fin.TR	Trading and Market Microstructure	取引と市場の微細構造
stat	stat.AP	Applications	応用物理学
stat	stat.CO	Computation	計算
stat	stat.ME	Methodology	方法論
stat	stat.ML	Machine Learning	機械学習
stat	stat.OT	Other Statistics	その他の統計
stat	stat.TH	Statistics Theory	統計理論

4.3　データ登録数推移

図 1 および図 2 に，2000 年からの年ごとの論文投稿数を示す．順調に数が伸びているため累積に見えるが，実際にはその年ごとの投稿数である．

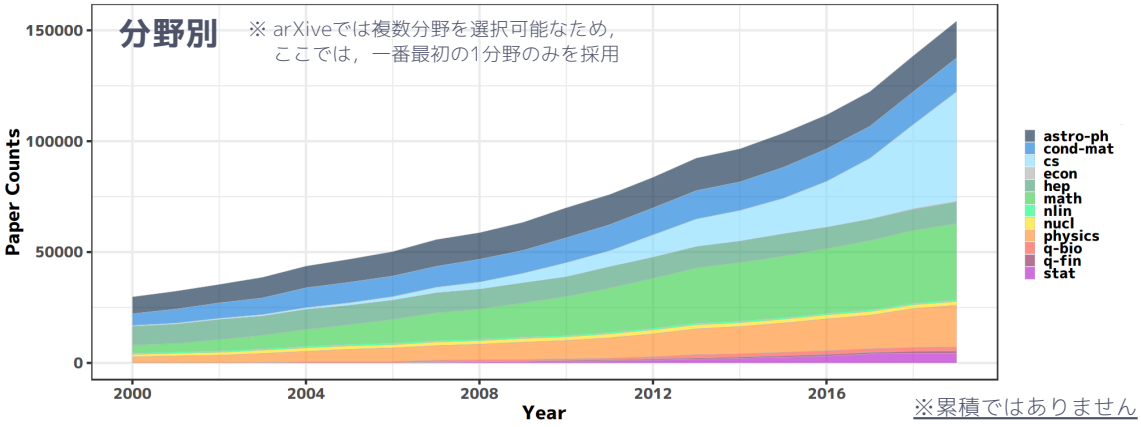

図 1　年・分野別の投稿数推移

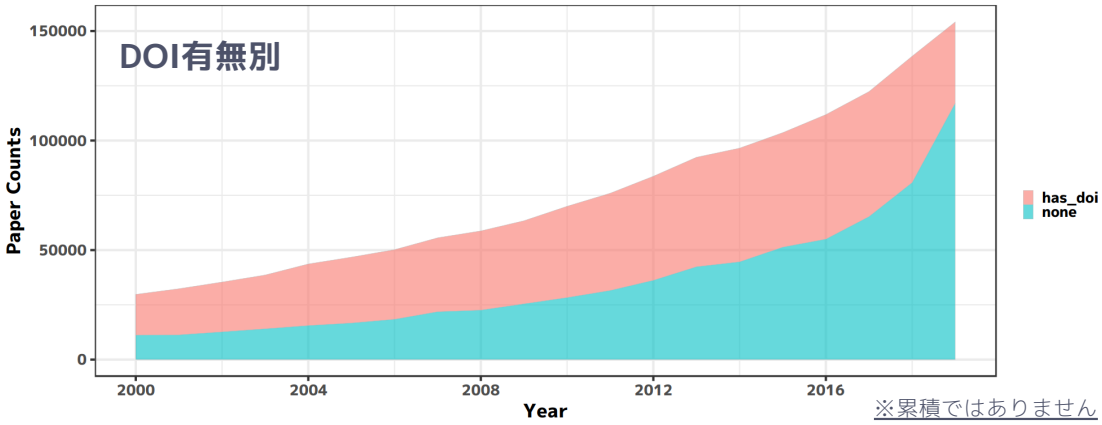

図 2　年・DOI 有無別の投稿数推移

図 1 の分野については，書誌情報上で最初に出現する 1 分野のみを採用しているため，数値はその年の投稿論文数に一致する．図 1 をみると，2015 年頃から情報系 (cs) の投稿数が急激に増加している傾向が読み取れる．また，数学系 (math) については 2000 年から順調に数を伸ばしている傾向が読み取れる．

同じデータを DOI(Digital Object Identifier) の有無で塗り分けた図 2 を見ると，2016 年と 2018 年の辺りで傾向が変化する様子が読み取れる．ここから，単純には arXiv 投稿後 3 年と 1 年の辺りで，DOI が付与されはじめる可能性が推測できる．

現状において，arXiv の論文に DOI が付与されるということは，基本的には何らかの雑誌[10)]に掲載されたこととほぼ同義と捉えられることから，すなわち，投稿から採録されるまでに 1 年から 3 年の幅があることが推定できる．この点については後で詳細に見ていくこととする．

[10)] 論文誌の他に，プロシーディングスなども含む.

4.4　DOI 中の Award 情報

前節で DOI が付与された arXiv 論文数の推移を確認した．ところで，Crossref を通じて得ることのできる DOI 情報の中には Award（研究助成）の情報も含まれる．

そこで，DOI 付き論文のうち，Award 情報が付与されたものの件数および，その中に “Japan” の文字列を含むものが何件程度あるかを調査した．また比較のために China, Chinese とのマッチを意図して “Chin” の文字列を含むものが何件程度あるかを調査した．

結果を以下に示す．

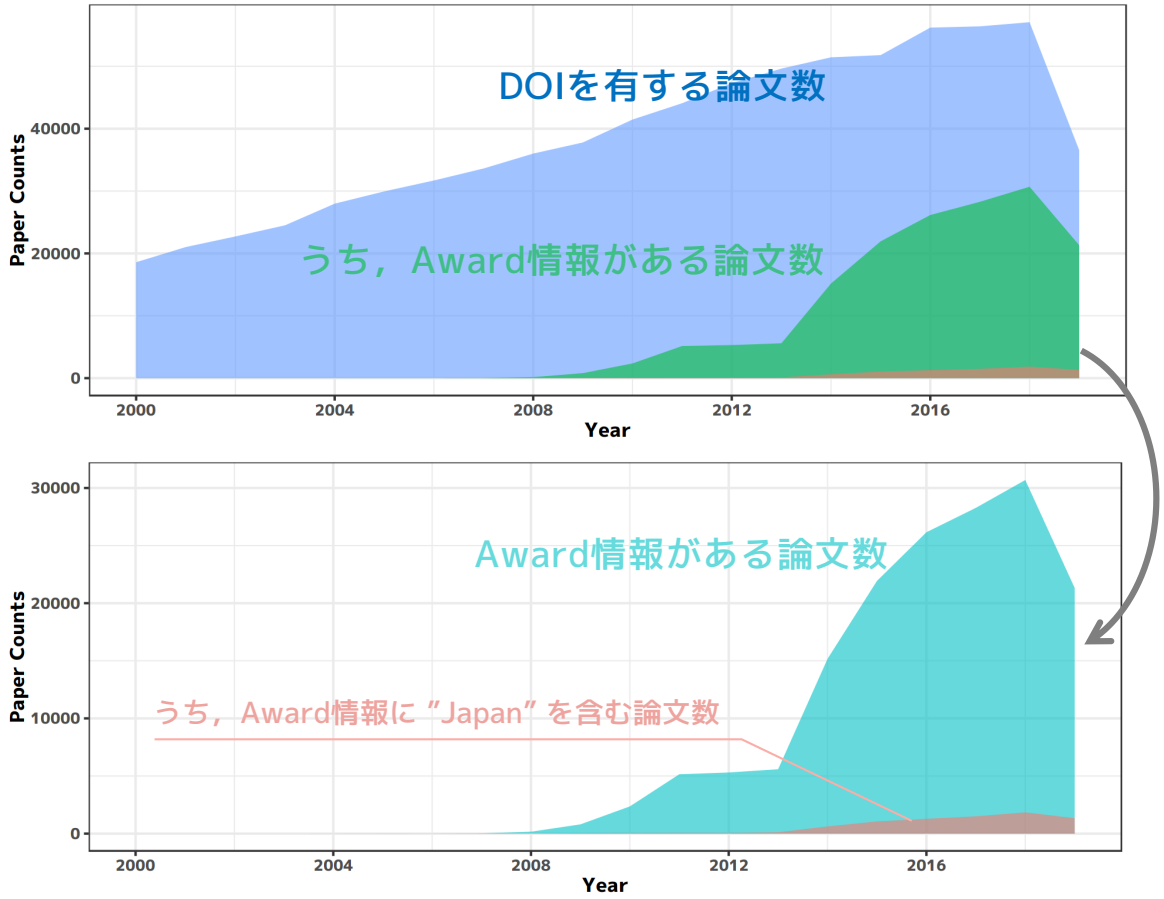

図 3　DOI 付き論文のうち Award 情報を有するものの数

図 3 をみると，2008 年頃から Award 情報が検出されはじめ，2013 年頃から急激に増加しているように見受けられる．これはおそらく，DOI 側が 2008 年頃から項目を追加し，2013 年頃に周知をはじめるなど，なんらかの施策を打った結果と推測される．従って，Award 情報を分析する場合，推察される施策変更から少し期間を空けた，2015 年くらいからを用いると，安定した結果が得られる可能性が高い．

次に図 4 をみると，我が国からの研究助成を得て実施されたと推測される論文も，2015 年以降

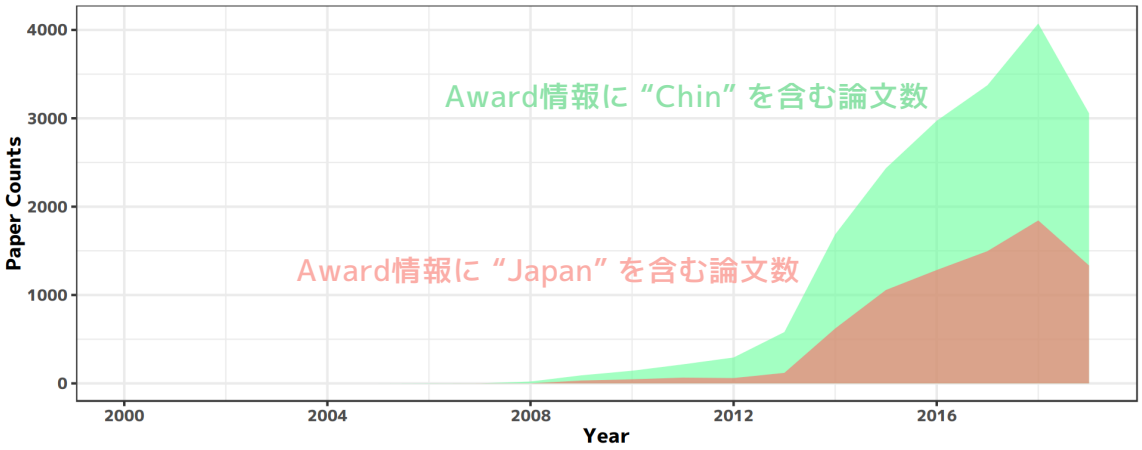

図 4　Award 情報に “Japan”, “Chin” の文字を含むものの数

の直近 5 年間では年あたり千件以上登録されている様子が観察できる．中国の論文数の伸びについては各所で指摘されているところではあるが，“Chin” の文字を含むものと比較しても一定の割合を有していると言え，arXiv を起点とした論文分析においても，我が国のファンデイング政策の効果を多少観察できる可能性が見いだせた．

4.5　分野と DOI の関係性

先に述べたとおり，論文に DOI が付与されるということは，基本的には何らかの雑誌に掲載されたこととほぼ同義と捉えられる．したがって，分野毎にどの程度 DOI が付与されているかを観察すれば，分野毎の論文誌採択率に相当するものを推定できると考えた．

そこで，分野単位で DOI の付与率などを算出した．結果を図 5 に示す．

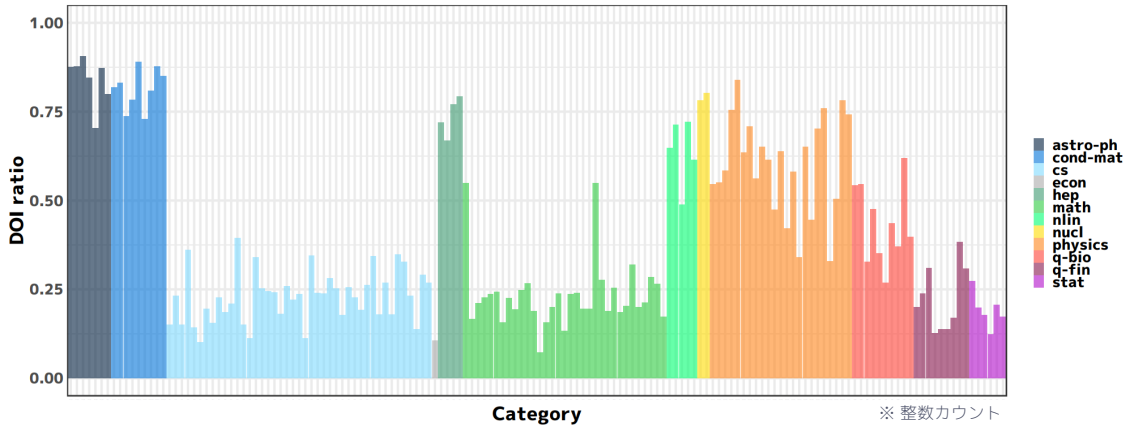

図 5　分野毎の DOI 付与率

図 5 では図 2 の傾向を参考に，期間を 2014 年から 2018 年の 5 年間に限定した上で DOI 付与率を示している．

図 5 を見ると，分野毎に DOI の付与率が大きく異なる様子が見て取れる．

ここで，既に述べたとおり DOI が付与されているということは，基本的には何らかの雑誌に掲載されていると考えて良かった．他方，**DOI が付与されない場合**には以下のような複数の状態が考えられる．

1. 論文誌に投稿中・査読中
2. 論文誌に投稿したがリジェクトされた
3. そもそも arXiv への掲載に留めており，論文誌に投稿していない
4. 論文誌に採録されたが論文誌側に DOI がない
5. 論文誌以外（例えば会議録，書籍）のメディアで発行されたが DOI が付与されていない
6. 論文誌や他のメディアに採録され DOI も付与されたが，arXiv の情報更新を忘れている[11)]

したがって，図 5 に示した結果から DOI 付与率が低いことのみをもって，ある分野の採択率が厳しい，ある分野では論文誌に投稿しない，といった推定を行うことは現時点では難しい．

[11)] arXiv は出版社と連携して DOI を自動的に更新する機能も有する（cf. `https://arxiv.org/help/bib_feed`）が，当然，一定の設定が必要であり全ての論文が自動的に紐付けられるわけではない．arXiv も自動更新に加えて，手動更新のための手段も提供している（cf. `https://arxiv.org/help/jref`）．

しかしながら，分野によって大きな違いがあることは明らかと言える．特に，情報 (cs) および数学 (math) は物理系に比べて DOI 付与率が低く，物理系 (astro-ph, cond-mat, hep など) が 75% 程度の付与率に対して，25% 程度の付与率となっている．

続いて，arXiv に登録されてから，DOI の公開日までの期間を分野毎に調べた．結果を図 6 に示す．

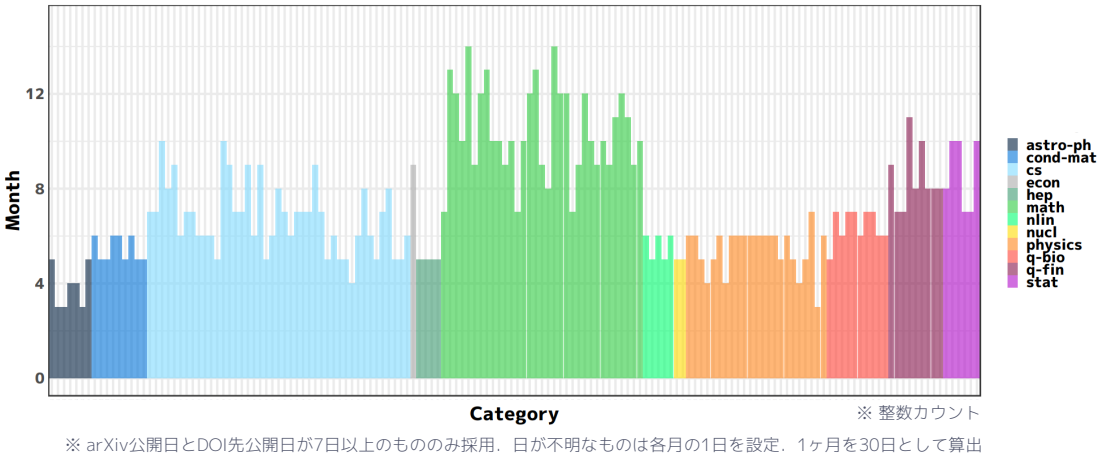

図 6　分野毎の DOI 付与までの期間

図 6 は図 5 とは期間を変え，2000 年〜2017 年投稿 の 18 年分を採用した．これは，期間の算出に際して十分な幅の “窓” を設ける，すなわち DOI が付与されるまでの期間が長かったプレプリントもできるだけ拾うことを意図したものである．その他の詳細については図中の補足文を参照されたい．期間が異なることから，図 5 の結果との単純な比較は困難である点に留意を要する．

その上で図 6 をみると，ここでも分野によって大きな違いがあることが観察できる．特に数学系は平均的に期間が長く 10 ヶ月から 12 ヶ月程度の時間を要しているように見受けられる．その他は概ね半年（6 ヶ月）もあれば DOI が付与されているようである．DOI 付与率が比較的高かった天文物理学系 (astro-ph) はおおよそ 3 ヶ月と，採録までの期間が短い点も興味深い．また，今回調査した物理・数学・情報系分野では査読期間がおおむね半年，査読期間が長い場合でも平均的には 1 年以内に収まる，と言った知見も興味深い．

最後に，分野毎に全期間を通じて具体的にどういった雑誌に掲載されているかを調査した．結果を表 11,12 に示す．

なお表 11,12 は各分野 100 件以上の掲載があるものに絞って Top5 を示しているため，一部の分野は表に含まれていない．また，投稿先が 5 件に満たない場合がある．

表 11　分野別投稿先雑誌名 Top5 (1/2)

1986-2020

ctg	title	count
astro-ph	The Astrophysical Journal	66168
astro-ph	Monthly Notices of the Royal Astronomical Society	46747
astro-ph	Physical Review D	34640
astro-ph	Astronomy & Astrophysics	29896
astro-ph	Journal of Cosmology and Astroparticle Physics	9880
cond-mat	Physical Review B	74769
cond-mat	Physical Review Letters	34033
cond-mat	Physical Review E	20297
cond-mat	Physical Review A	11216
cond-mat	Applied Physics Letters	6801
cs	Electronic Proceedings in Theoretical Computer Science	3983
cs	IEEE Transactions on Signal Processing	1143
cs	IEEE Transactions on Information Theory	1060
cs	Logical Methods in Computer Science	583
cs	IEEE Transactions on Wireless Communications	488
hep	Physical Review D	65614
hep	Journal of High Energy Physics	33701
hep	Physics Letters B	27796
hep	Nuclear Physics B	14027
hep	Physical Review Letters	10340
math	Journal of Mathematical Physics	7674
math	Communications in Mathematical Physics	6842
math	Journal of Physics A: Mathematical and Theoretical	6132
math	Journal of Statistical Physics	2962
math	Journal of High Energy Physics	2757

表 12　分野別投稿先雑誌名 Top5 (2/2)

1986-2020

ctg	title	count
nlin	Physical Review E	4411
nlin	Physical Review Letters	1942
nlin	Journal of Physics A: Mathematical and Theoretical	940
nlin	Journal of Physics A: Mathematical and General	815
nlin	Physics Letters A	805
nucl	Physical Review C	14503
nucl	Physical Review D	4835
nucl	Nuclear Physics A	4555
nucl	Physics Letters B	4094
nucl	Physical Review Letters	3130
physics	Physical Review A	27012
physics	Physical Review Letters	14807
physics	Physical Review E	8539
physics	Physical Review B	5882
physics	New Journal of Physics	3941
q-bio	Physical Review E	1836
q-bio	Physical Review Letters	411
q-bio	PLoS ONE	317
q-bio	The Journal of Chemical Physics	269
q-bio	PLoS Computational Biology	231
q-fin	Physica A: Statistical Mechanics and its Applications	647
q-fin	Physical Review E	119
stat	The Annals of Statistics	1385
stat	The Annals of Applied Statistics	897
stat	Bernoulli	524
stat	Statistical Science	411
stat	IEEE Transactions on Signal Processing	208

4.6　分野と被引用の関係性

ここまで arXiv に投稿された論文がどの程度の期間を経て，どのような雑誌に掲載されているかを分析した．

続いて arXiv に掲載された論文が その後の論文にどのような影響を与えたか，具体的には arXiv に掲載された個々の論文がどの程度引用されたか，について見る．結果を図 7 に示す．

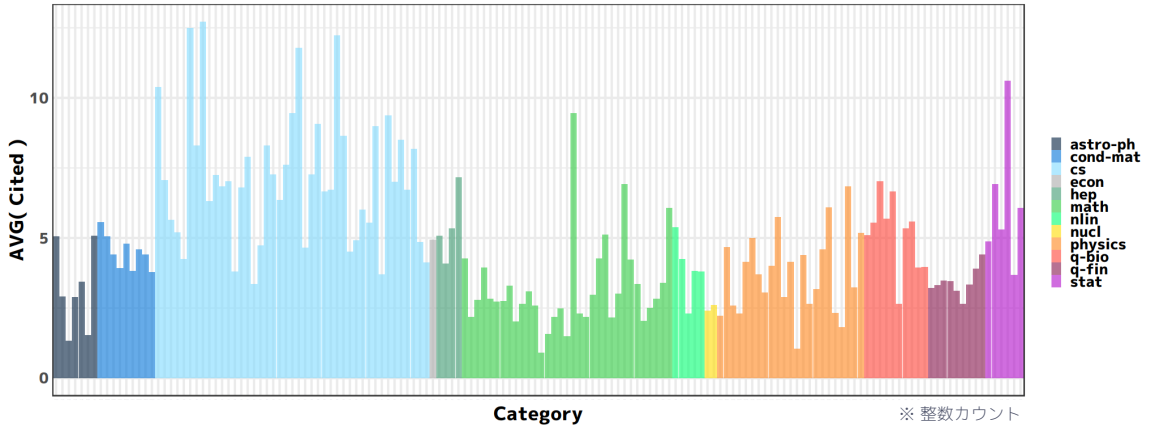

図 7　分野と被引用の関係性

図 7 は 2014 年〜2018 年 の 5 年間に arXiv に投稿された論文を対象に，被引用情報を取得し，分野毎に件数を示したものである．

図 7 をみると，ここでも分野毎に傾向が大きく異なる．たとえば天文物理 (astro-ph) は引用回数が低く，情報系 (cs) の引用回数が大きい．

単純にこれらの結果を閲覧すると，「天文物理については DOI 付与率が高いため，あえて arXiv の論文を引用せず DOI のついた，すなわち論文誌に掲載された論文を引用している」という可能性も推察されるが，Semantic Scholar では，arXiv に投稿されたプレプリントそのものだけでなく，DOI が付与された原著論文等の被引用についても，すべて合算して提示するため，これらの結果はそのまま “各分野における引用のされやすさ” と，想定して差し支えない[12)]．

関連して分野に関係なく，同期間での引用数 Top15 を表 13 に示す．

表 13 をみると上位は情報系が占めている．論文は基本的に深層学習を行う上での基礎的なテクニックに関するものが多い印象で，中には深層学習や機械学習に際して活用されるプログラミング言語のライブラリに関するものもある．プログラミング言語のライブラリは実務上極めて有用である一方，その成果をいわゆる原著論文として雑誌に掲載することは科学的な新規性の観点から難易度が高いと考えられ，arXiv への掲載という選択には情報系特有のパブリシティ（出版・PR）戦略も伺える．

12) 逆に，プレプリントのみの純粋な被引用数ではない点には留意する必要がある．

表 13　高被引用論文 Top15

	aid	date	category	title	cite
1	1502.03167v3	2015-02	cs.LG	Batch Normalization: Accelerating Deep Network Training by Reducing Inter	9999
2	1409.4842v1	2014-09	cs.CV	Going Deeper with Convolutions	9998
3	1201.0490v4	2012-01	cs.LG|cs.MS	Scikit-learn: Machine Learning in Python	9997
4	1310.4546v1	2013-10	cs.CL|cs.LG|stat.ML	Distributed Representations of Words and Phrases and their Compositionali	9997
5	1409.1556v6	2014-09	cs.CV	Very Deep Convolutional Networks for Large-Scale Image Recognition	9996
6	1412.6980v9	2014-12	cs.LG	Adam: A Method for Stochastic Optimization	9996
7	1512.03385v1	2015-12	cs.CV	Deep Residual Learning for Image Recognition	9996
8	1409.0575v3	2014-09	cs.CV|I.4.8; I.5.2	ImageNet Large Scale Visual Recognition Challenge	9994
9	1506.01497v3	2015-06	cs.CV	Faster R-CNN: Towards Real-Time Object Detection with Region Proposal N	9994
10	1301.3781v3	2013-01	cs.CL	Efficient Estimation of Word Representations in Vector Space	8977
11	1408.5093v1	2014-06	cs.CV|cs.LG|cs.NE	Caffe: Convolutional Architecture for Fast Feature Embedding	8977
12	1409.0473v7	2014-09	cs.CL|cs.LG|cs.NE|sta	Neural Machine Translation by Jointly Learning to Align and Translate	8727
13	1406.5823v1	2014-06	stat.CO	Fitting Linear Mixed-Effects Models using lme4	8708
14	1311.2524v5	2013-11	cs.CV	Rich feature hierarchies for accurate object detection and semantic segmer	8145
15	1505.04597v1	2015-05	cs.CV	U-Net: Convolutional Networks for Biomedical Image Segmentation	7797

次に被引用数 0 件から 100 件までの頻度分布を図 8 に示す.

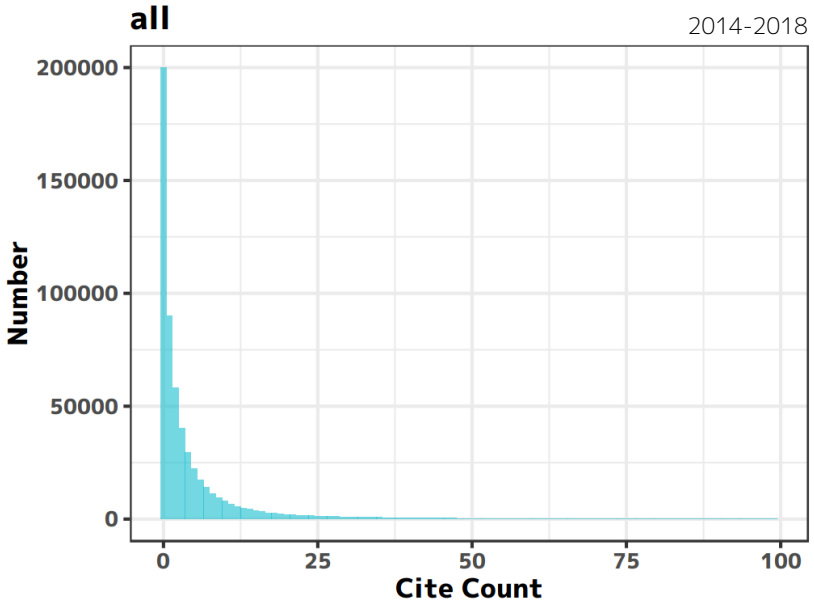

図 8　被引用件数と頻度

計量書誌学では「ロトカの法則」など「べき分布」に従うようなデータが見られるが，ここでも「べき分布」に類する分布形状が観察でき，多くの論文は被引用が 0 件である一方，表 13 に示したように数年で 1 万件近い論文で引用されるようなものも存在することが分かる.

5　考察

5.1　分野毎の差異について

今回の試行からは，arXiv 登録から DOI 付与までの期間や，DOI 付与率などの指標について，分野間で大きな違いが見られた．特に天文物理学系の分野と情報系の分野とで違いが顕著に見られた．

ただし，これらの結果の読み取りには若干の注意も必要と考えられる．

たとえば，天文物理学では登録から DOI 付与まで概ね 3 ヶ月となっている．一般的な査読・出版プロセスを考えると，1. 投稿を受け付けてから査読者を選定して査読を依頼，2. 査読結果を編集委員会で集約して採否を確定，（条件付き採録の場合は一旦投稿者に戻して一定期間内でのリバイスを依頼，リバイスがあった場合は再査読し再度査読者の確認と委員会での再査読,）3. 採録決定の場合，出版社に回して文章校正，4. 著者による校正確認を経て論文誌掲載，となる．近年ではオンライン出版も増加しており，紙媒体への印刷を伴わず，かつ，論文単位での順次公開も可能となっているため，出版までの期間は短縮されていると期待できる．それでも一般的にこれらの全プロセスを 3 ヶ月以内で完了することは難易度が高いと推定される．天文物理学系の高い DOI 付与率と期間の短さを併せて考えると，草稿が完成して論文誌に投稿するのとほぼ同時に登録するのではなく，採録がほぼ決まった時点で arXiv に登録している可能性も排除できない．

このように，分野毎の文化の違いが行動の差異を生んでいる可能性もあり，結果の活用には留意する必要がある．

その留意をおいた上で，出版社や雑誌とは独立した同一のプラットフォーム上で，複数分野を比較できること，またその結果として（差異の要因がどこにあるかは一旦慮外に置いた上で）差異が確認できたことは興味深い．

また，数ヶ月〜半年，場合により 1 年程度とは言え，論文分析の先行指標として活用できる可能性も見いだせた．

5.2　政策への活用について

文献 [4] では，研究分野把握を目的とした研究者へのインタビュー結果として“ **査読がされていないこともありプレプリントとしてアップロードされた論文は，雑誌に掲載された査読付き論文と比べて質的に不十分なものも多く，当該分野に専門的知識を持たない政策担当者が利用するのは難しいのではないかとの意見もあった．** ”といったことを記載している．引用件数等を活用することで，ある程度質をはかることも可能とは考えられるが，その場合，従来のジャーナル論文を対象とした書誌情報分析同様に，速報性の点が犠牲になる．

たとえば医療系のプレプリントサーバ bioRxiv [13]を活用した Rχivist [14]ではダウンロード数や，マイクロ SNS での言及数を用いて論文をランキングし，提供している．こうしたオルトメトリクス的な指標を用いることは可能と考えられ，前述の文献 [4] においても，“**プレプリントのみならず従来の論文誌にも言えることであるが，従来の被引用数のみならずオルトメトリクスにより各論文の注目度を可視化する指標も利用されるようになってきているとの意見が得られており，こうしたものの活用も検討していくべきである．**”との記述が見られる．

また，arXiv には必ずしも「論文誌」に投稿する論文だけが掲載されているわけではない可能性が高く，論文書誌情報だけでは評価が困難と言われる情報系分野についても，同一のプラットフォーム上で比較できる可能性が示された．たとえば，表 11 を見ると cs（情報系）のトップは「Electric Proceedings in Theoretical Computer Science」で，「Journal」や「Transaction」ではなく「Proceedings」である．つまり，論文誌用の論文だけではなく，国際会議用の論文（予稿）も arXiv に掲載されていることが分かる．

5.3 留意点

分野の差異の読み取りに関して既に留意点を述べたが，その他の留意点についてまとめる．

arXiv は登録に際して内容については学術論文としての体裁や整合性など最低限の確認に留めて，基本的には掲載が行われる．つまり一般的な意味での査読を経たものではない．従って玉石混淆の度合いは相対的に高いと考えられ，一概に “arXiv に掲載されている論文なので信頼できそうだ・信頼できそうにない” とは言えない．また，arXiv のプレプリント群の分析で得られる傾向の取り扱いにも留意が必要であり，例えば，原著論文群の分析で得られる傾向と同列に捉えることはできない．

また，DOI の付与についても，近年ではプレデトリー・ジャーナルなどの問題も生じてきており，DOI がついた原著論文だから信頼できるものかどうかについても留意が必要である．従って，質をそろえて分析したい場合には，分野毎のトップジャーナルやトップカンファレンスをいくつか定義して，“トップジャーナル採択率” のような原著論文の掲載元を限定する形で分析する必要がある．

その他，arXiv は様々な分野を横断しているとは言え，物理・数学・情報系の論文が多く投稿されるプレプリントサーバである．ジャーナル論文掲載までいかずトップカンファレンス採録でも成果としてカウントされたり，研究サイクルや進展速度が速いため論文書誌情報には向かないとされる情報系 [2, 3] などが追えるメリットや，出版社・論文誌に固定されないメリットもあるものの，バイアスがあることは想定され，また，bioRxiv など他分野に強みを持つ他のプレプリントサーバと一概に比較はできない．

分析結果を読み取る上ではそれらの点について，十分に留意する必要がある．

---

[13] https://biorxiv.org/

[14] https://rxivist.org/

6 まとめ

COVID-19 による緊急性の高いグローバルな社会問題解決に向けて，研究成果の共有がこれまでにない速さで行われており，オープンサイエンスで予察されていた研究成果のあり方や研究そのものの変容が現実のものとなりつつある．その変容を駆動するものとして，研究データに加えて，プレプリントは迅速な成果の共有を担うメディアとしても広く注目されている．NISTEP では COVID-19 に関連するプレプリントの試行的分析 [7] を公表している．

一方で，プレプリント自体は分野によっては 30 年近く運用され，プレプリントサーバに掲載された研究内容の質については懐疑的な見方も多く，その評価のあり方が問われてもいる．

今回プレプリントサーバの嚆矢である arXiv に着目した分析を行ったことで，プレプリント公開から原著論文になるまでの期間に分野による差が見られることや，情報系を中心に，必ずしも原著論文を出口としないプレプリントが多数掲載されていることなどが分かった．また，Award 情報との一定の紐付けも可能であり，我が国のファンディング政策の効果を多少観察できる可能性や，被引用数の分析によって，分野ごとのプレプリント利用スタイルが大きく異なる可能性も見出した．

これらのプレプリントの特性を踏まえながらプレプリントの集合を解析することで，これまでの論文と被引用数の解析からだけではわからない，科学技術の側面を補完的に分析できる可能性を示すことができた．

今後他の分野のプレプリントサーバなどの分析を加えることで，プレプリントの定着具合や，分野間比較を行い，プレプリント分析の一般化を行うことが望まれる．

謝辞

考察・分析の一部において，角田英之氏，岡村圭祐氏にご助言頂いた．記して感謝する．

参考文献

[1] 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, Mar 2020. https://doi.org/10.15108/stih.00205

[2] 鷲尾 隆： 一流国際会議発表のための研究戦略とは?. **人工知能学会誌**, Vol.23, No.3, pp.362–366, May 2008. http://id.nii.ac.jp/1004/00006982/

[3] 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

[4] 文部科学省：「海外の最新科学技術動向に係る新興・融合領域に関する調査分析業務」業務成果報告書. **平成** *30* **年度科学技術調査資料作成委託事業**, 2019. https://www.mext.go.jp/a_menu/kagaku/kihon/1404334.htm

[5] Arthur, David and Vassilvitskii, Sergei： K-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* pp,1027–1035, 2007. http://dl.acm.org/citation.cfm?id=1283383.1283494

[6] Leland McInnes, John Healy and James Melville： UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint,* 2018. arXiv:1802.03426

[7] 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 Discuttion Paper,* No.186, June 2020. http://doi.org/10.15108/dp186

DISCUSSION PAPER No.187

arXiv に着目したプレプリントの分析

2020 年 08 月

文部科学省 科学技術・学術政策研究所
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: 03-3581-2391 FAX: 03-3503-3996

Analysis of preprints on arXiv

Aug 2020

HAYASHI Kazuhiro, KOSHIBA Hitoshi
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

https://doi.org/10.15108/dp187

https://www.nistep.go.jp