



研究力の新たな側面

オープンデータ, オープンアクセス,
プレプリント, RDX に関する調査

第14回政策研究レビューセミナー
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長
林 和弘
2022年2月18日(金)

背景

- **定量的な研究評価は未だ発展途上**
 - ◆ 研究“インパクト”計量は論文と被引用数を中心に一定の成果
 - ◆ 評価にどう活かすかは未だに議論の途上（分野別）
- **定量的な研究評価を推し進める技術が、研究のあり方、研究成果自体も変えつつある状況**
 - ◆ デジタル化、ネットワーク化、オープン化は、既存のインパクト計量を効率化し、さらにその先の非連続変化を生み出しつつある
 - 例：研究データに注目した成果公開のあり方、プレプリントによる査読の意義の再確認
- **オープンサイエンス政策は包括的な科学，社会，科学と社会の変容を見越し，現在特に研究データに着目**
 - ◆ 第6期科学技術・イノベーション基本計画，G7:イノベーションの新たな基盤や産業の創出

今回発表の概要

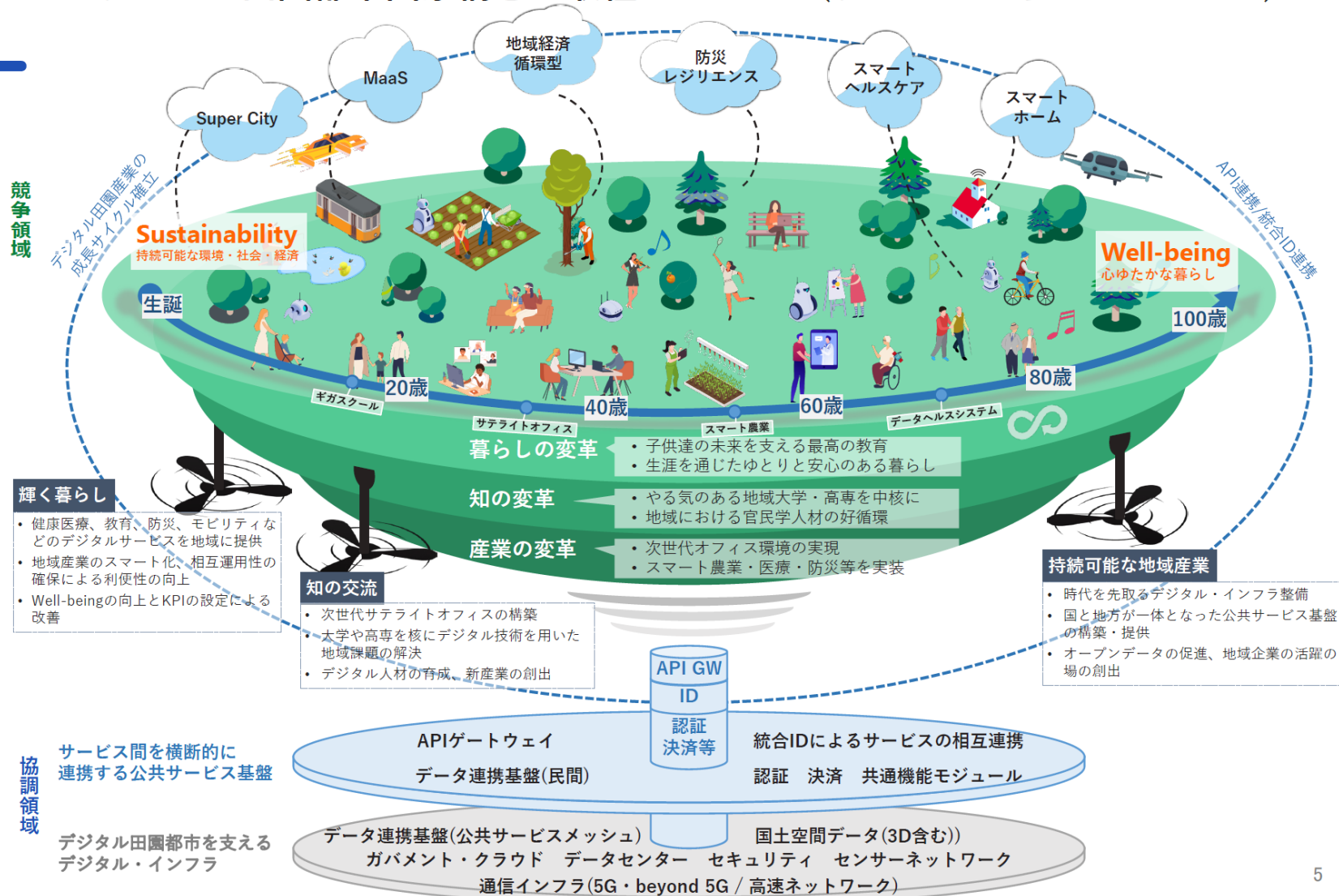
- **研究データに着目した調査**
 - ◆ 研究データの共有や利活用を推進するためのモニタリング
 - ◆ ボトルネックの洗い出しによる施策へのつなぎ
- **プレプリントに着目した調査**
 - ◆ プレプリントがどれだけ受容されているか（2020年度レビューセミナー）
 - ◆ プレプリントの集合を分析することで、研究活動の変容を掴むことができるか

A 論文と研究データの公開に関する実態調査2020

B arXiv, bioRxiv に掲載されたプレプリントの分析

研究DXの調査に対する明確なフレームがあるわけではなく、むしろ、研究DXの調査フレームをデザインするための調査

デジタル田園都市国家構想の取組イメージ（デジタルからのアプローチ）



G7 Research Compact



G7 Research Compact

As Open Societies with democratic values we believe in academic freedom. The freedom to pursue intellectual enquiry and to innovate allows us to make progress on shared issues and drive forward the frontiers of knowledge and discovery for the benefit of the entire world. We recognise that research and innovation are fundamentally global endeavours. Nations, citizens, institutions, and businesses have made huge strides forward, not otherwise possible, through open research collaboration across borders. Working together we will use our position as leading science nations to collaborate on global challenges, increase the transparency and integrity of research, and facilitate data free flow with trust to drive innovation and advance knowledge.

Shared Values

The global response to COVID-19 has demonstrated the progress that arises from long-term collaboration which puts science at the heart of prevention, preparedness, response, recovery and resilience. This progress requires sustained investment in research and supporting infrastructure, including in basic research and high-risk, high-reward undertakings. As our nations and communities start to recover from the pandemic and build resilience for future shocks, we will continue to work with our research and business communities to remove barriers to the open and rapid sharing of knowledge, data and tools, to the greatest extent possible, recognising the importance of research security in particular in cutting-edge fields, and to promote open science and increase open, safe and transparent dissemination of science to citizens, and to strive to minimise technology-related risk.

We can only tackle the greatest challenges that we face and will face over coming decades – such as climate change, pandemics and biodiversity loss – through transparent, open and agile research collaboration. We must bring the widest possible range of resources, expertise and perspectives to bear on solutions which will benefit people across the globe.

We commit to promoting international research cooperation and the conditions of freedom, independence, openness, reciprocity and transparency under which it flourishes. Our governments have the right and responsibility to effectively ensure the security and integrity of the research ecosystem, in partnership with the research community, preventing the theft, misuse and inappropriate exploitation of our intellectual property and personal data, and other forms of misconduct.

We are committed to developing a strong, diverse and resilient science and research community which is inclusive of all groups, as recognised by the Working Group on Financing Science for Inclusive Growth. It is important to deepen participation of underserved, underrepresented and marginalised communities and expand their participation in the research and innovation ecosystem. Inclusion will enhance the strength of our research base and increase momentum on dismantling the social, legal, and regulatory barriers limiting participation, and complementing our G7 gender equality goals by tackling gender gaps. Principles and practices of inclusive growth distribute the benefits of science among diverse communities and regions across the G7 and beyond.

Open Science WG

UNESCO



UNESCO
"Building peace in the minds of men and women"

IN BRIEF WHAT WE DO WHERE WE WORK PARTNERS JOIN US RESOURCES

Home > Open Science > UNESCO Recommendation on Open Science

Open Science

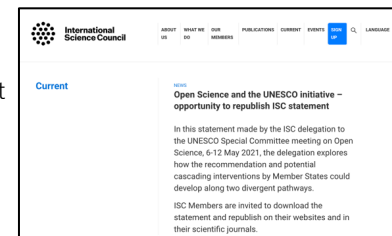
193 Countries

UNESCO Recommendation on Open Science

At the 40th session of UNESCO's General Conference, 193 Member States tasked the Organization with the development of an international standard-setting instrument on Open Science in the form of a UNESCO Recommendation on Open Science to be adopted by Member States in 2021.

ISC

(International Science Council)



International Science Council

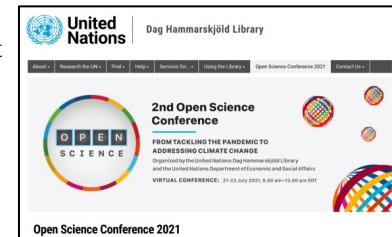
ABOUT WHAT WE DO OUR MEMBERS PUBLICATIONS CURRENT EVENTS CL LANGUAGE

Current

Open Science and the UNESCO Initiative – opportunity to republish ISC statement

In this statement made by the ISC delegation to the UNESCO Special Committee meeting on Open Science, 6-12 May 2021, the delegation explores how the recommendation and potential cascading interventions by Member States could develop along two divergent pathways. ISC Members are invited to download the statement and republish on their websites and in their scientific journals.

UN



United Nations
Dag Hammarskjöld Library

Open Science Conference 2021

2nd Open Science Conference

FROM TACKLING THE PANDEMIC TO ADDRESSING CLIMATE CHANGE
Organized by the Dag Hammarskjöld Library and the United Nations Department of Economic and Social Affairs

VIRTUAL CONFERENCE: 21-23 July 2021, 9.00 am-12.00 pm EDT

Open Science Conference 2021

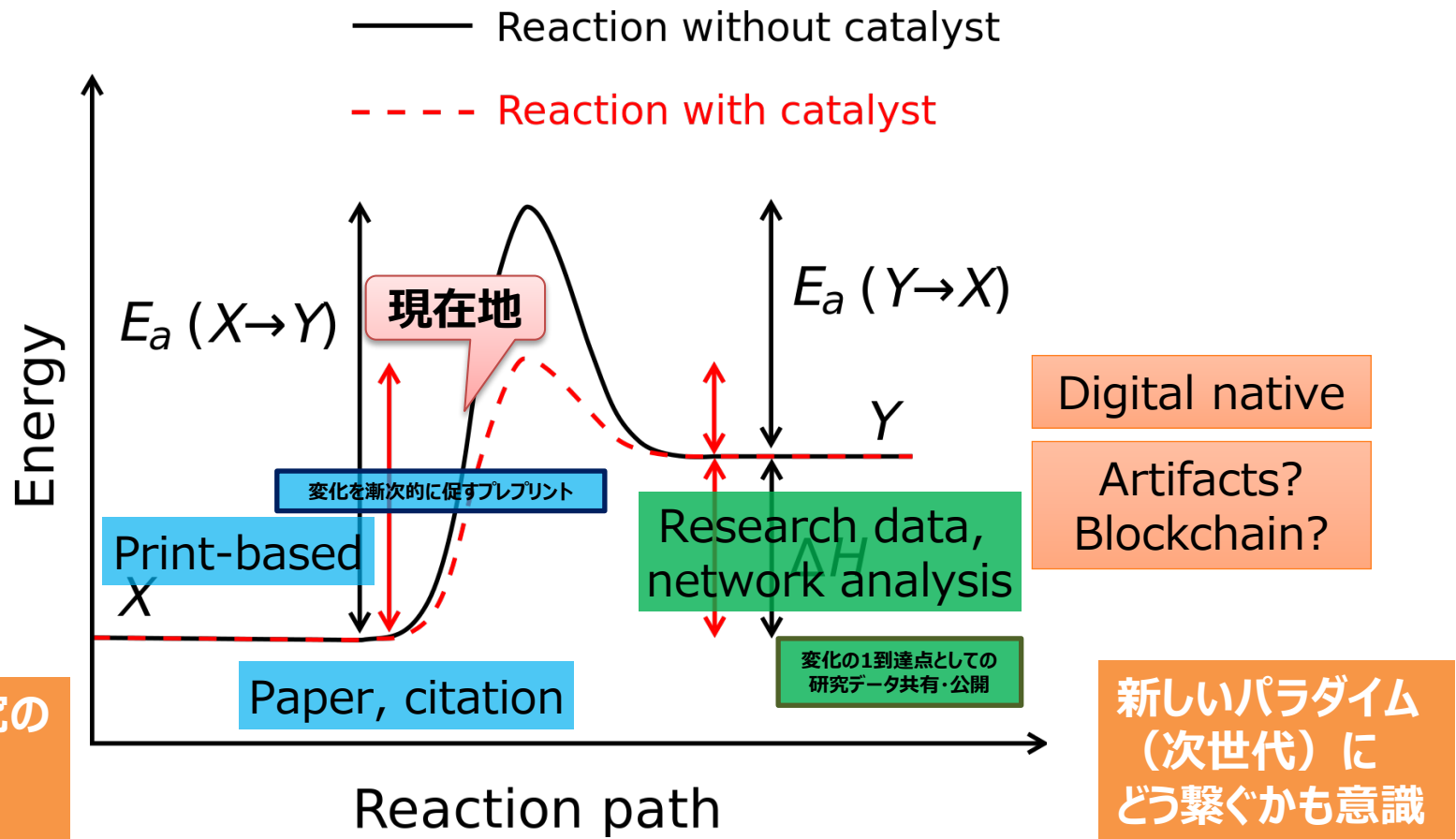
support

support

新しいイノベーション基盤に

<https://www.mofa.go.jp/mofaj/files/100200013.pdf>
<https://en.unesco.org/science-sustainable-future/open-science/recommendation>
<https://council.science/current/news/open-science-and-the-unesco-initiative/>
<https://www.un.org/en/library/OS21>

オープンサイエンスが目指す位相の変化と遷移状態



これまでの研究の
エコシステムは
依然重要

(5-10年の計)

- プレプリントを調査することで漸次的な変化を捉え
- 研究データ共有・公開を調査することで、非連続な変化を捉える

(100年の計)



A 論文と研究データの公開に関する 実態調査2020 ：オープンサイエンス支援体制の構築に向けて

- 池内有為, 林 和弘, 他研究データ公開と論文のオープンアクセスに関する実態調査2020, NISTEP RESEARCH MATERIAL, No.316, (2021).
- <https://doi.org/10.15108/rm316>

■ 調査対象

- ◆ 科学技術専門家ネットワーク：1,914名
- ◆ 有効回答：1,349名（回答率70.5%）

同様の調査を
2016年，2018年にも実施

■ 調査期間

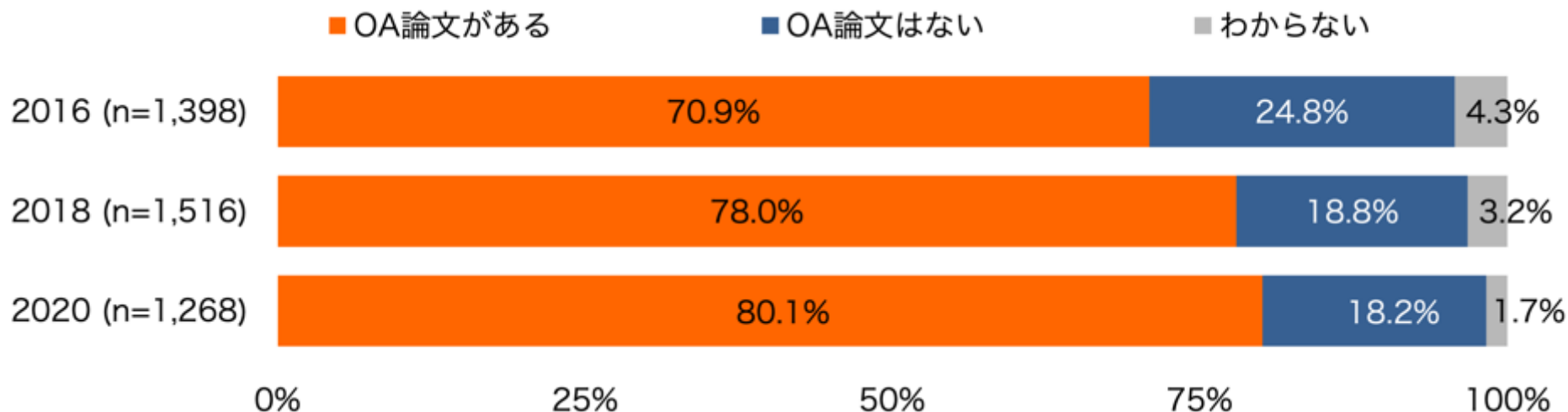
- ◆ 2020年11月5日～11月17日

■ 調査項目

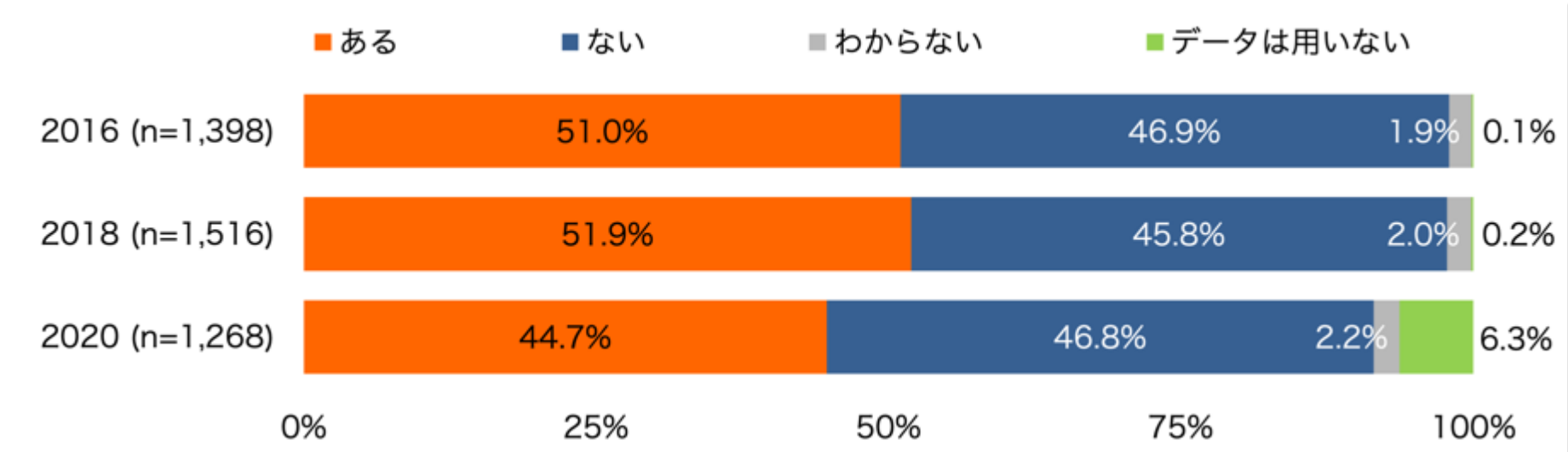
- ◆ 論文と研究データの公開状況
- ◆ 研究データの提供（共有）状況
- ◆ 公開データの利用状況と課題
- ◆ データマネジメントプラン（DMP）の作成状況
- ◆ 研究データ公開の障壁
- ◆ 研究データ公開のインセンティブ
- ◆ 研究データの整備・公開・保存と支援の可能性
- ◆ 研究成果公開の展望・問題点等（自由記述）

属性による分析

- 年齢層
- 所属機関
- 分野



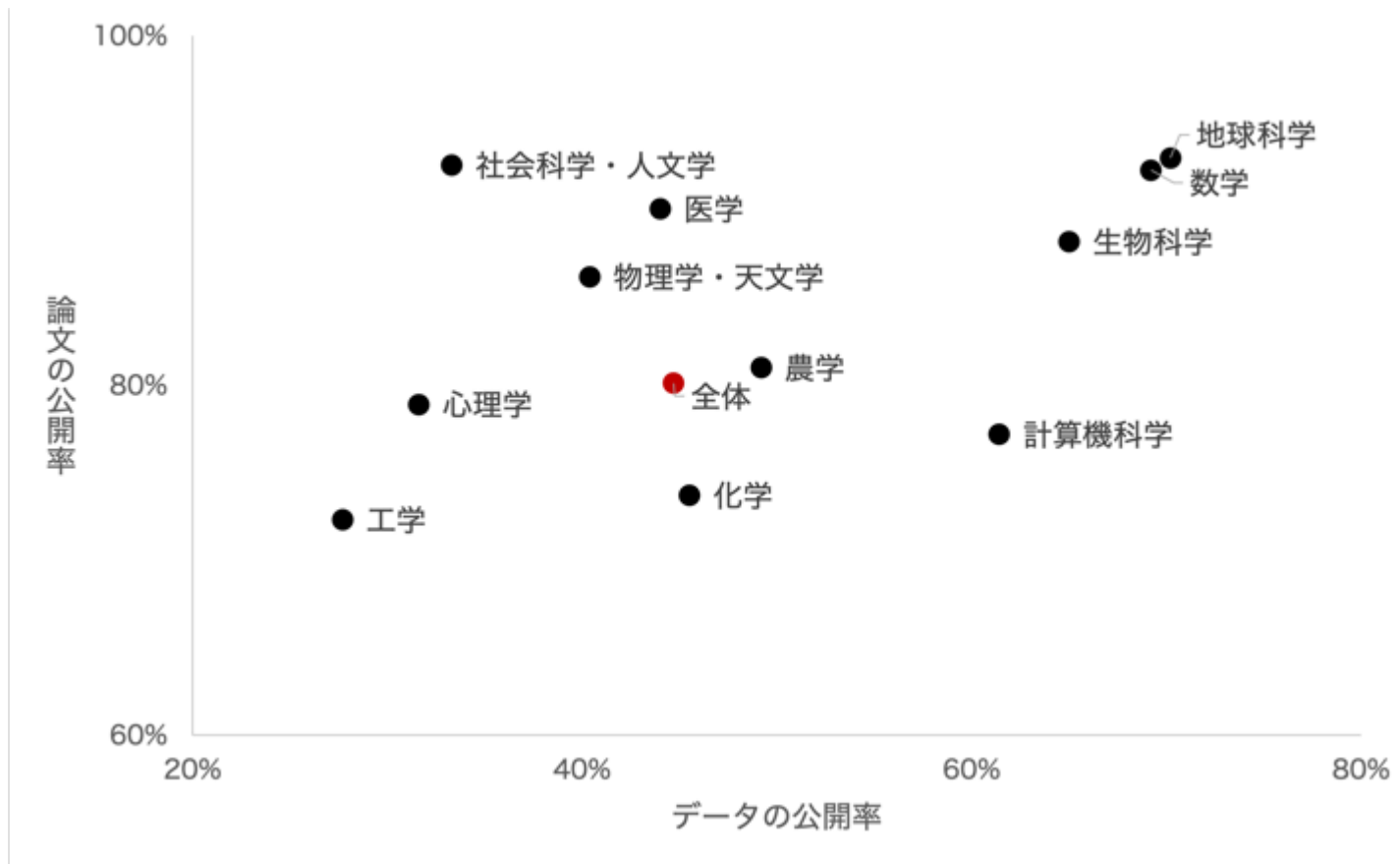
- オープンアクセス（OA）経験をもつ回答者は増加傾向。
- 公開方法は「OA誌への投稿」（77.1%）, 「雑誌が論文をOAにした」（32.6%）, 「所属機関のリポジトリ」（30.7%）, 「雑誌のOAオプション」（30.6%）など。【複数回答】
- 公開理由は「投稿した雑誌がOA」（75.8%）, 「研究成果を広く認知してもらいたい」（57.6%）に集中。【複数回答】
- 未公開理由は「資金がないから」（57.6%）, 「投稿したい雑誌がOAではない」（40.3%）に集中。【複数回答】



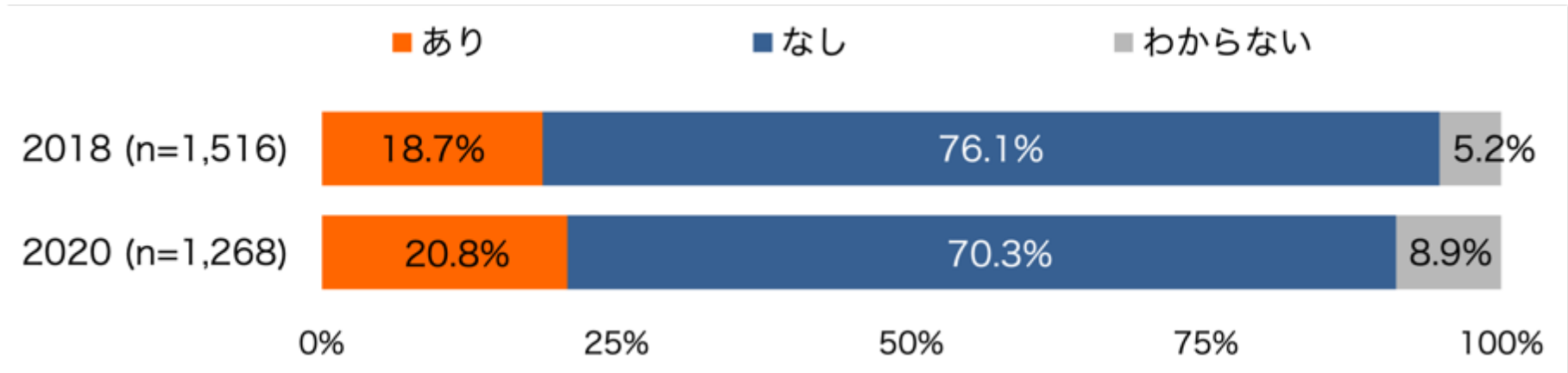
■ 2016/2018年調査と比較して、公開率がやや低下

■ 低下の要因

- ◆ アンケートシステムの変更に伴う質問方法が変更されたことによるもの
- ◆ 回答者の揺れ（同一回答者の、同一システム・同一質問の2016/2018年調査でも矛盾あり、データ公開以外の項目も矛盾あり）

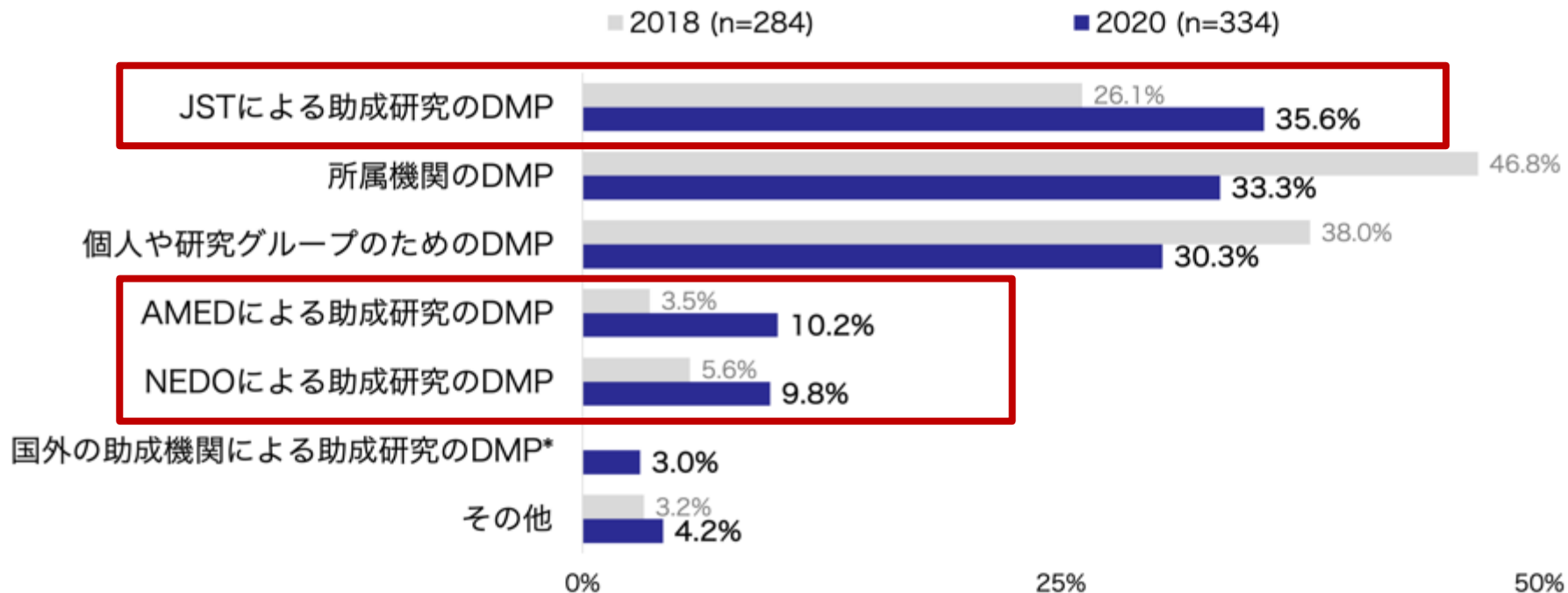


- データ公開経験とOA経験に有意な相関はなし
- いずれも高いのは地球科学・数学，いずれも低いのは工学
- OA率が高いのは社会科学・人文学，データ公開率が高いのはCS



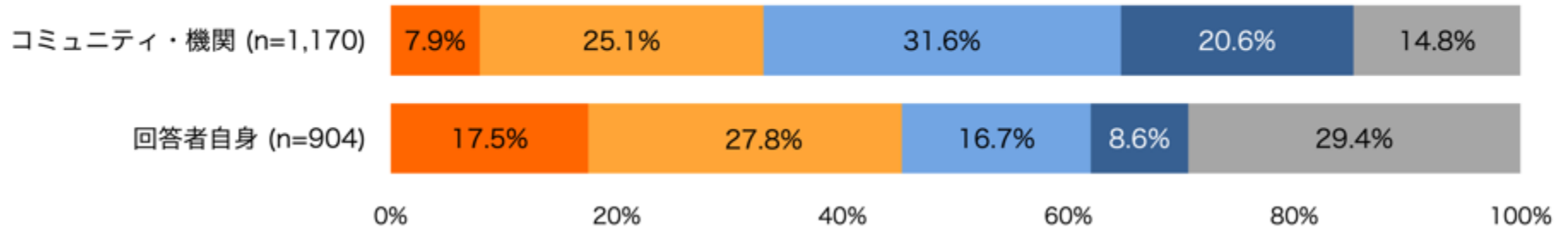
- 作成経験をもつ研究者の比率がわずかに増加
- 「わからない」の選択率が増加

作成経験があるDMP（複数回答）



- 助成機関（JST, AMED, NEDO）の選択率は全て増加
- 作成理由の1位は「助成機関が要求しているから」（n=351, 51.5%）
 - ◆ → 助成機関の要求がDMP作成を牽引
- 未作成理由の1位は「DMPを知らなかったから」（n=891, 52.0%）

■ 高く評価している ■ やや評価している ■ あまり評価していない ■ まったく評価していない ■ わからない



- コミュニティや機関よりも，回答者自身の方が評価していると認識している傾向
- 「わからない」の比率も高い
- データを公開していない理由の2位は「業績にならないから」（26.6%）

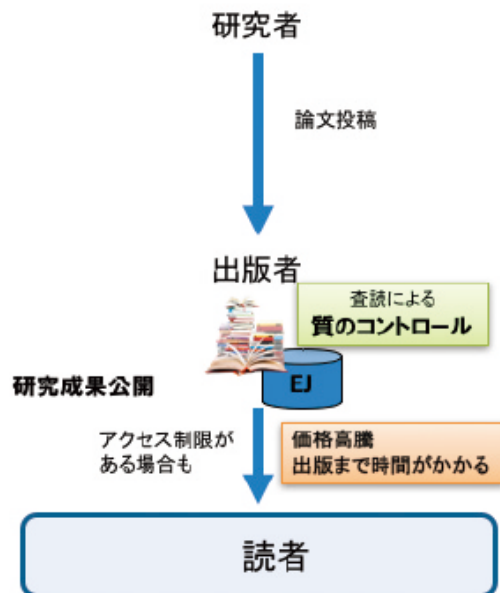


B arXiv, bioRxiv に掲載された プレプリントの分析

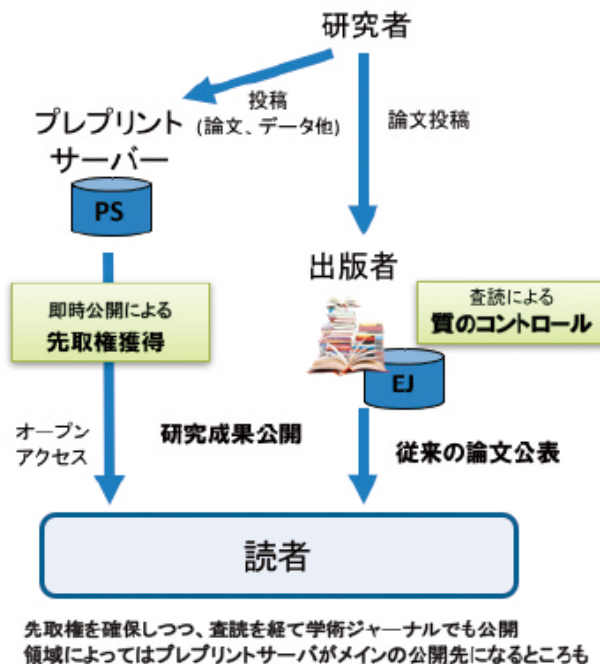
- 林 和弘, 小柴 等: bioRxiv に着目したプレプリントの分析, NISTEP DISCUSSION PAPER, No.197, (2021). <https://doi.org/10.15108/dp197>
- 小柴 等, 林 和弘: プレプリントとジャーナル論文の差異: bioRxiv を用いた試行, NISTEP DISCUSSION PAPER, No.200, (2021). <https://doi.org/10.15108/dp200>

プレプリントサーバー

従来の仕組み



プレプリントサーバ(PS)の活用



プレプリント（査読前の草稿）
の共有自体は紙媒体の頃
からあった

成果公開のゲームチェン
ジャーとして、COVID-19
以前に注目し、政策向け
メディアにて報告

学術情報流通のオープン化がもたらすオープンサイエンスに向けた成果公開プロセスと共有の変革

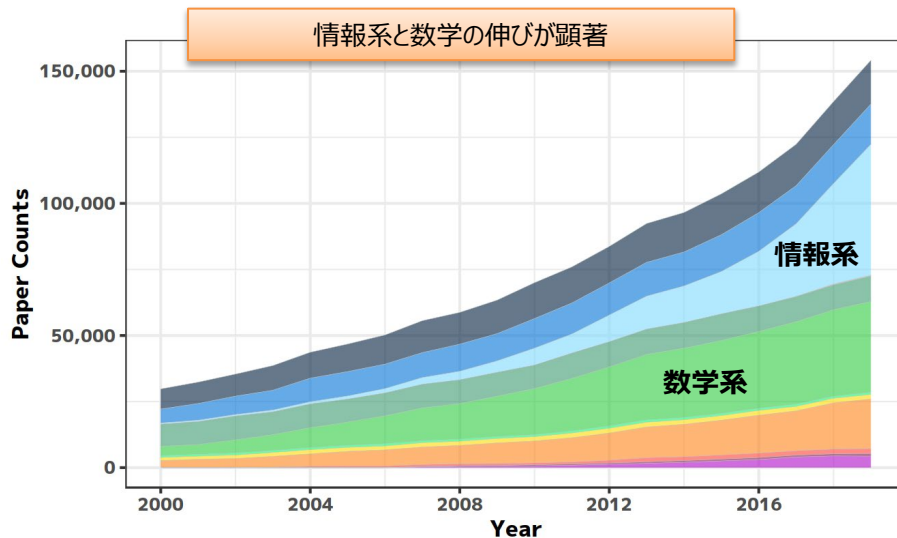
<http://doi.org/10.15108/stih.00092> (2017)



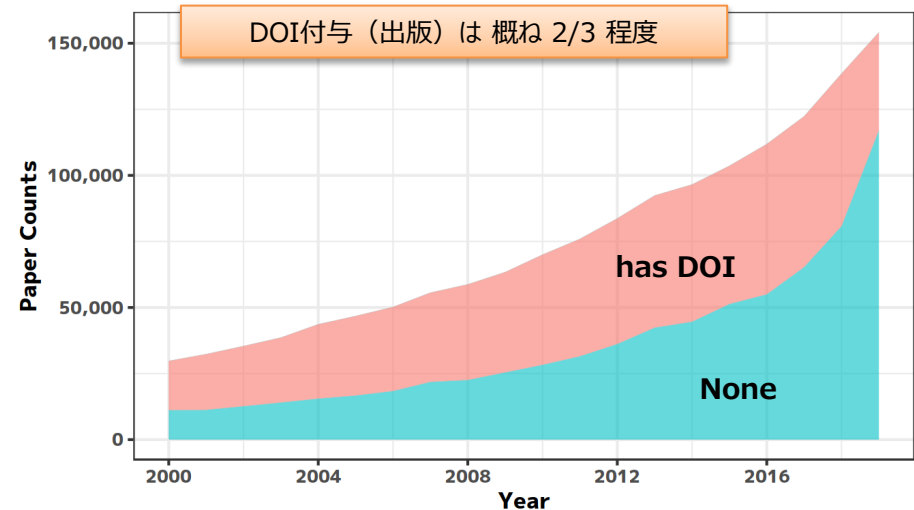
主要プレプリントサーバの コンテンツ分析

■ 物理学分野で 1990年に開始し、物理・情報系を中心とした最大・最古のサーバ arXivに掲載された、約 162 万件のプレプリント（2020年1月時点）分析

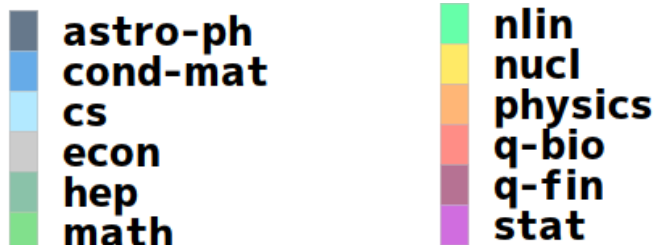
- ◆ arXiv については分析時点でDOIを有しておらず、原稿が公開され別媒体でDOIを取得した際に、それを記載する方式であったため、DOIの有無をジャーナル掲載の代理変数に設定

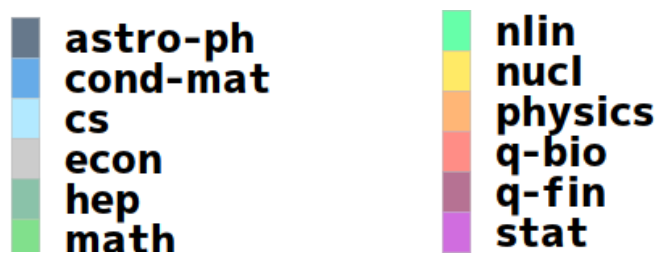
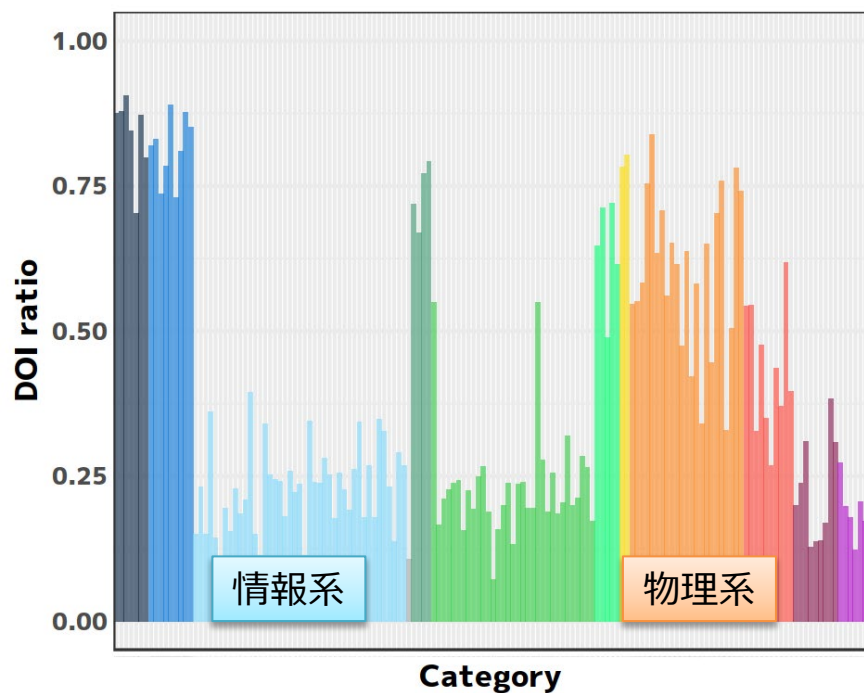


分野別件数推移 (arXiv)



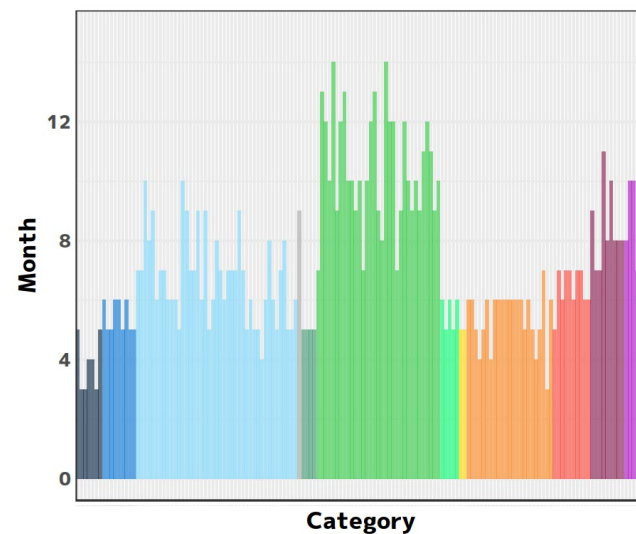
Journal DOI 付与率推移 (arXiv)



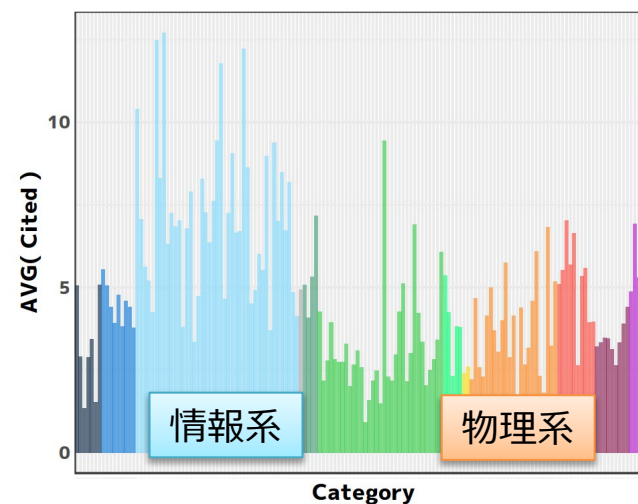


分野別 Journal DOI 割合 (arXiv)

分野による差が大きい



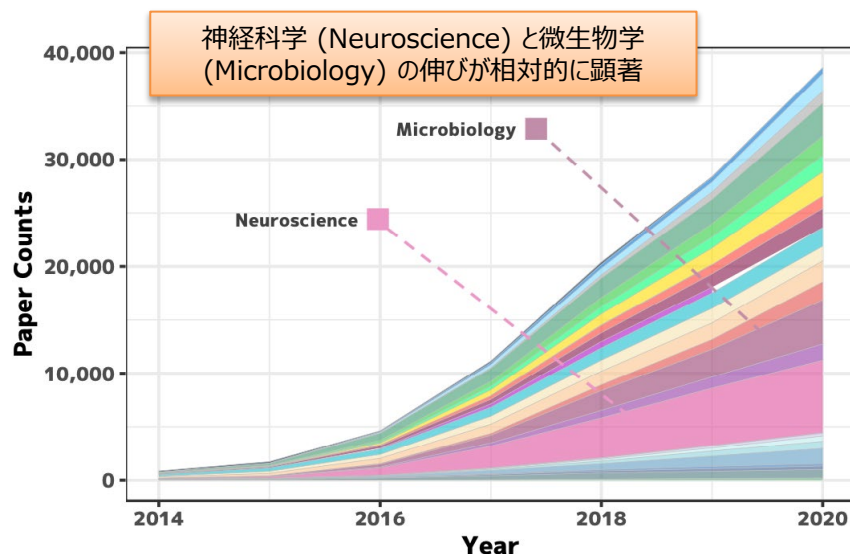
分野別 DOI 付与期間 (arXiv)



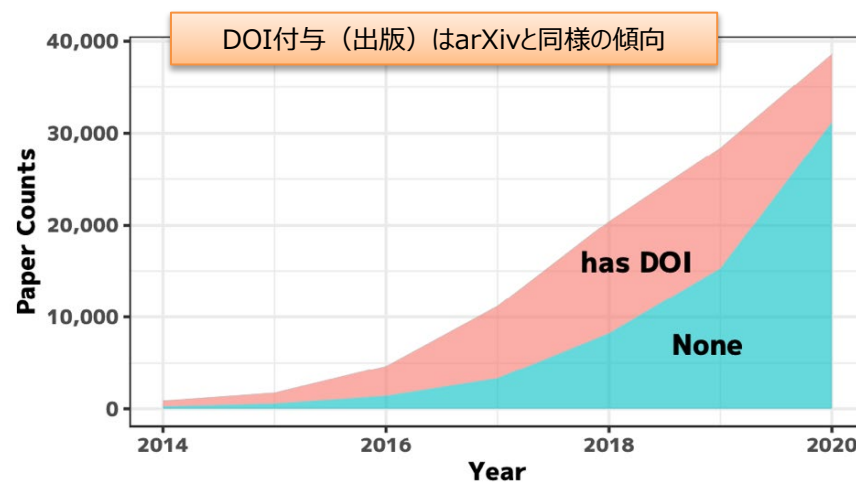
分野別平均被引用数 (arXiv)

■ 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 万件のプレプリント（2021年4月時点）分析

- ◆ 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 を付与できることから、DOI 付与率はこの他媒体の DOI 情報を用いて算出



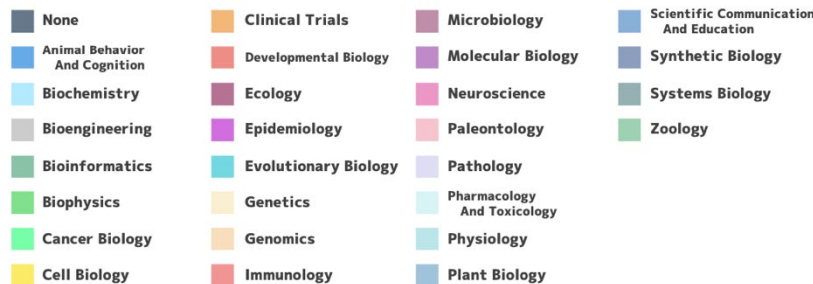
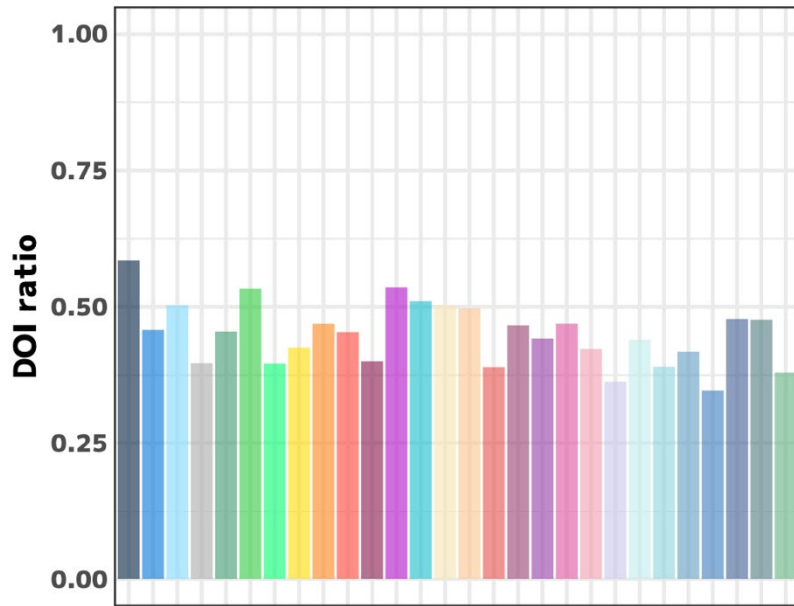
分野別件数推移 (bioRxiv)



Journal DOI 付与率推移 (bioRxiv)

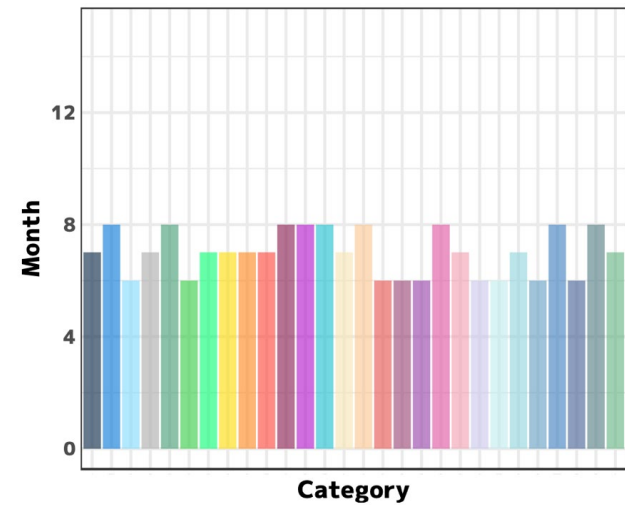
*Noneにも bioRxiv の DOI は付いている



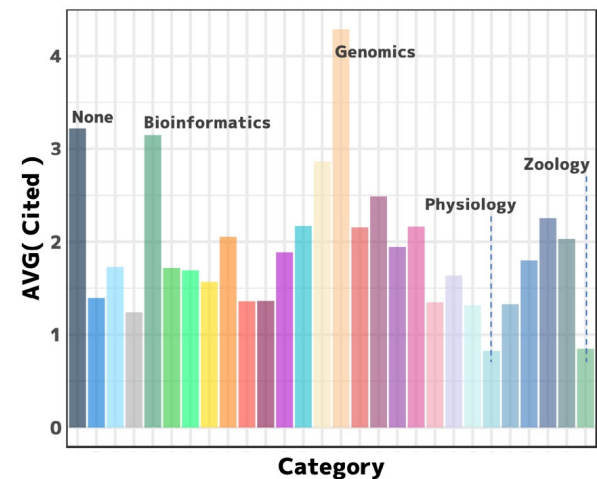


分野別 Journal DOI 割合 (bioRxiv)

分野による差が無い/認めづらい



分野別 DOI 付与期間 (bioRxiv)



分野別平均被引用数 (bioRxiv)

査読でどれだけ変わるのか、出版されるものを予測できるか

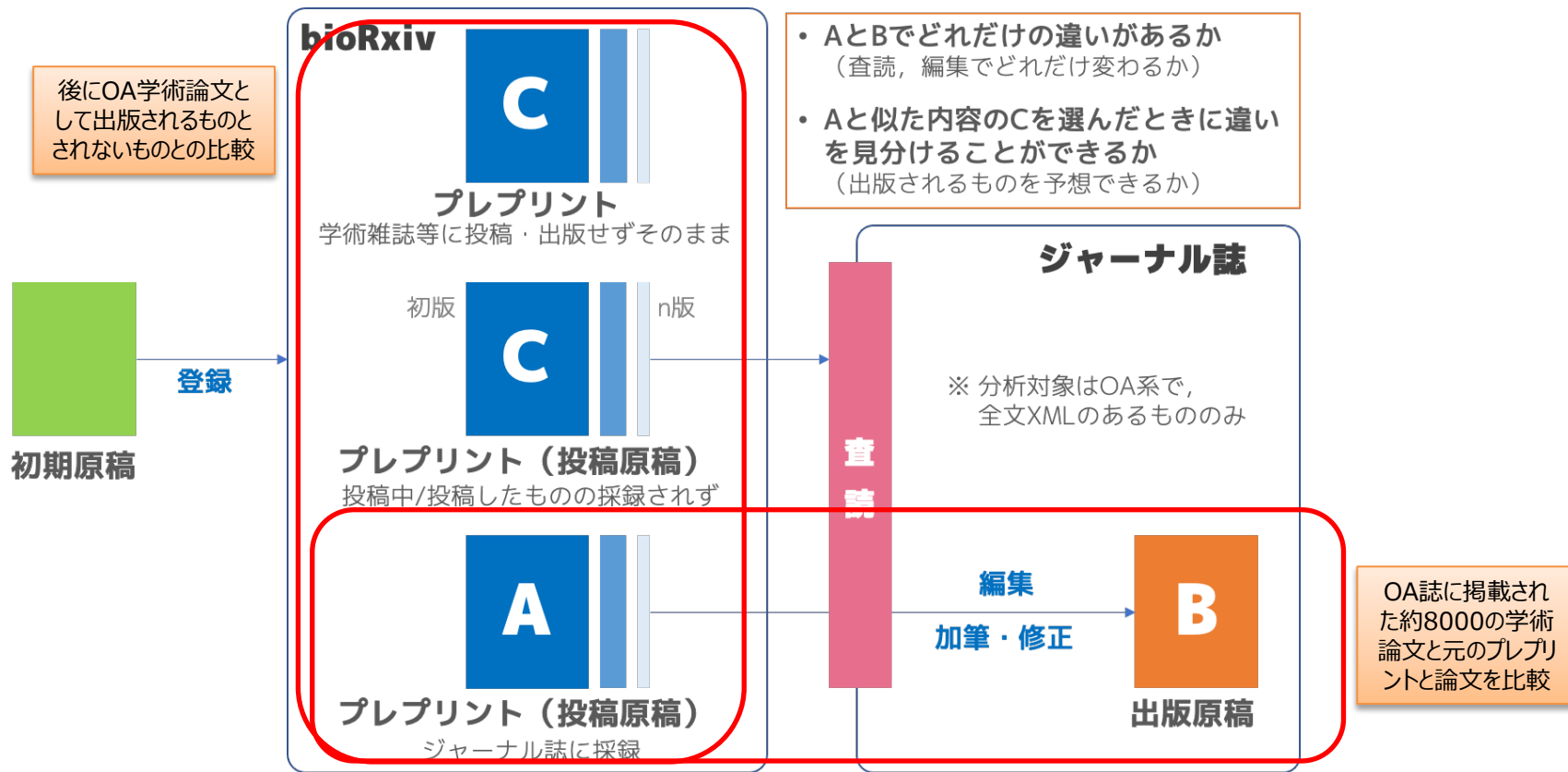
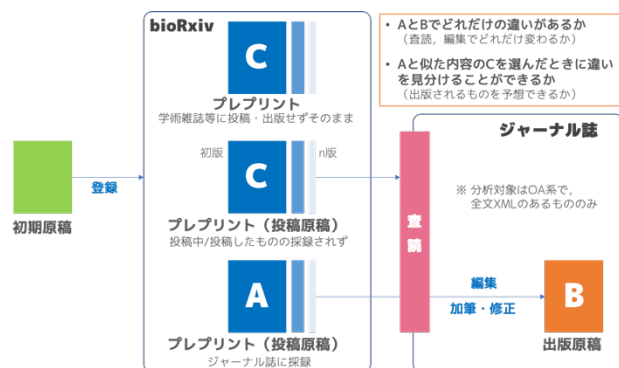


図 11: bioRxiv のプレプリントと学术论文の比較

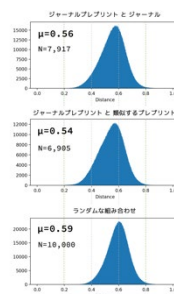
- bioRxiv のプレプリントについて、その後 OA 学术论文となったものの全文XML を一定量 (7,985 件) 確保することができたので、プレプリントと (OA) 学术论文の比較を行った。

- 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，明確な違いを見いだすことはできなかった。

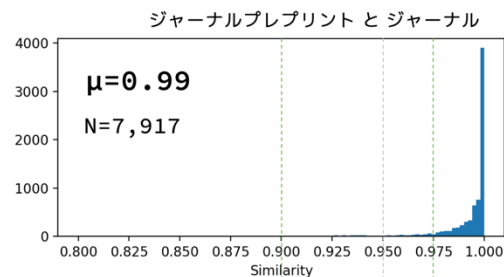
- ◆ プレプリントと学術論文（図のAとB），
- ◆ 学術論文になったプレプリントとそうではないプレプリントの間（同AとC）



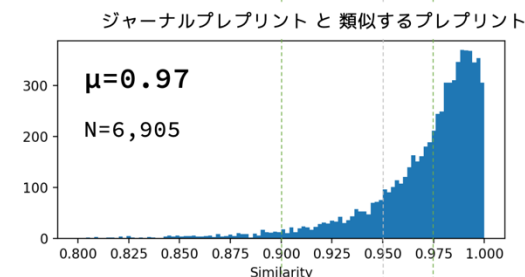
- ・ AとBでどれだけの違いがあるか（査読、編集でどれだけ変わるか）
- ・ Aと似た内容のCを選んだときに違いを見分けることができるか（出版されるものを予想できるか）



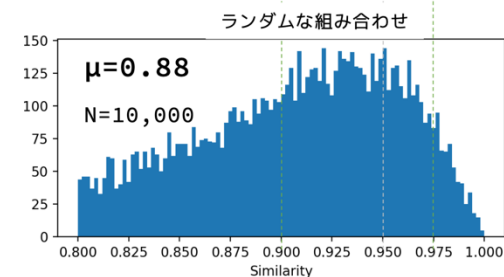
類似度の閾値を0.3に



査読の前後
でほとんど変
わらない



査読に通る
ものと通らな
いものや投
稿しないもの
との差も小さ
い



全部が似て
いるということ
では無いこと
の確認

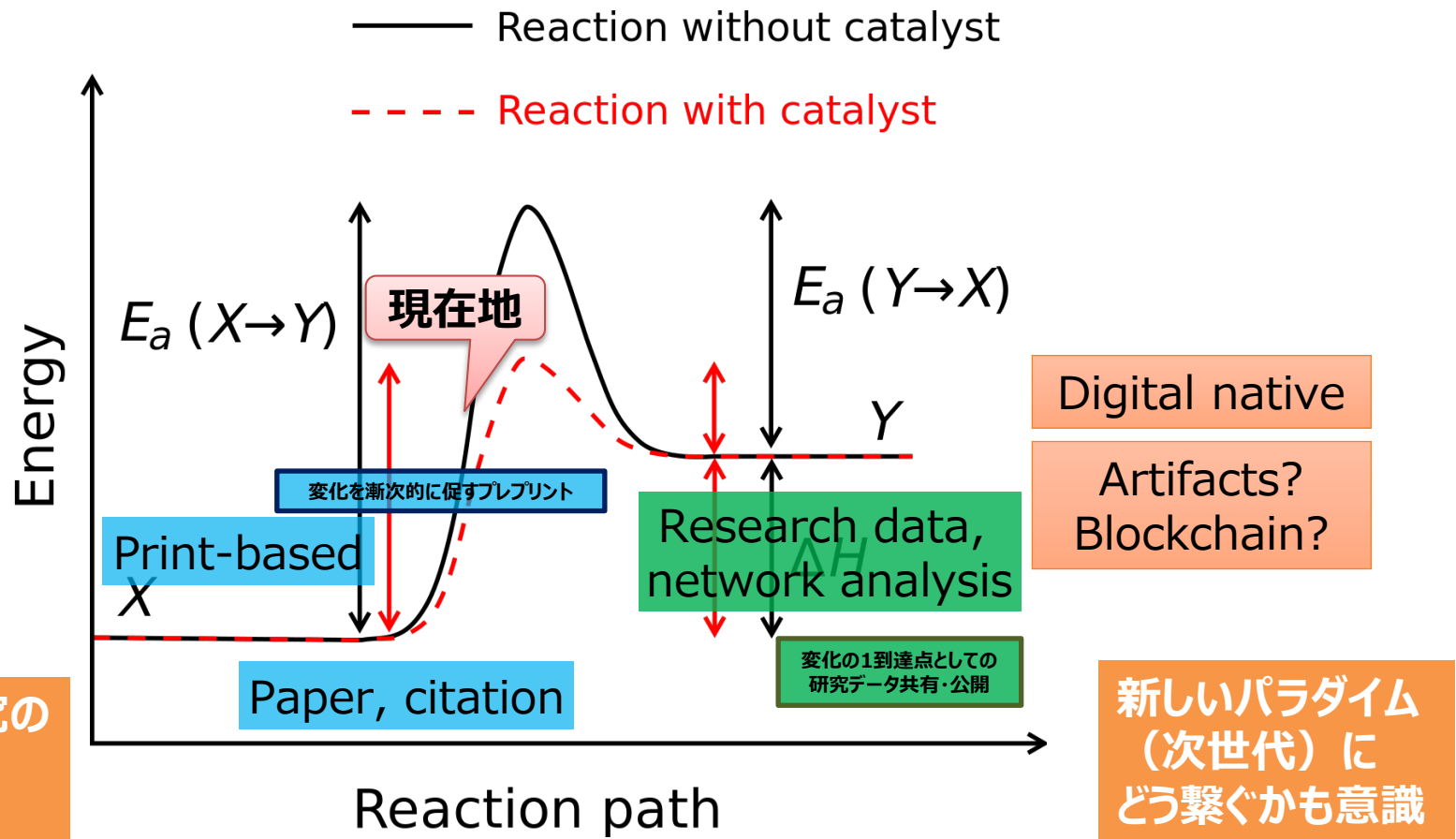
特徴量：著者数，参考文献数，段落数，単語数，章タイトル
内容の類似度

1. A,B2つの文書の全段落間で類似度を検出する（分散表現）
2. A側の各段落から見て最も高い類似度を持つB側の段落をペアにする
3. A,Bを入れ替えて2の処理を行い，B側からみたペアを作成する
4. 2，3でできたペアを全体として，類似度が一定以上のもののみを2つの文書の共通ペアとする。
5. Jaccard*係数を計算し，2つの文書A,B間の類似度とする

*集合（文書）の類似度を計る指標 集合XXと集合YYに対して $|X \cap Y| / |X \cup Y|$

<https://doi.org/10.15108/dp200>

オープンサイエンスが目指す位相の変化と遷移状態



これまでの研究の
エコシステムは
依然重要

(5-10年の計)

- プレプリントを調査することで漸次的な変化を捉え
- 研究データ共有・公開を調査することで、非連続な変化を捉える

(100年の計)

- 研究力を定量的に計る手法の進展と共に，研究や研究成果そのもののあり方が変化しており，日本のオープンサイエンス政策は，研究データの利活用に軸足を置いた施策を打ち出している。これに応じて新たな研究力を計る手法が生み出される可能性
- この潮流を踏まえ，NISTEPでは，研究論文（OA），研究データ公開のモニタリング調査を実施
 - ◆ 論文のOA率は経年的に増加して8割に達したものの，資金（APC）等の問題が残存
 - ◆ データの公開率は5割に満たない。DMPは，助成機関の要求により作成率が2割に到達
 - ◆ データ公開に対する評価は十分に行われておらず，「わからない」とする比率も高い

（今回の発表以外の知見）

 - データ公開には分野による差がみられ，OAの公開率とは傾向が異なる。現時点では，雑誌の要求により論文の補足資料として公開するケースが多い。大学等のデータポリシー策定により，今後データ公開が進む可能性
 - データの整備・公開・保存のための資源は，特に人材・時間・資金に不足感が強く，公開用のリポジトリは認知度が低い
 - データ公開に対する懸念は依然として強く，特に引用されないことが問題視されているものの，実際にはそれほど問題が起きていない可能性
 - データ公開によるインセンティブとして論文やデータの引用が重視されているが，実際にはデータを公開しても引用に結びついていない可能性
 - データの整備・公開・保存を第三者に依頼したいとする回答者は4割強
 - ◆ 2022年度に第4回の調査を実施予定

- より研究者の実活動に身近である，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も実施
 - ◆ 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
 - ◆ 2010年代に入って進展している生物科学系の bioRxiv
 - ◆ 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
 - プレプリントを中心として研究活動が行われている分野の把握（情報系）
 - arXiv とbioRxivの違い、arXiv内の分野別の特徴を把握
 - ◆ 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
 - 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い
 - 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判明
 - ◆ 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
 - ◆ 今後は，分野の拡大や定性的アプローチと組み合わせることによって，プレプリントがもたらす研究，研究成果共有の変容を把握し，研究力の向上や研究評価への応用を検討