

レポート

研究データの公開と論文のオープンアクセスに関する
実態調査 2020

ーオープンサイエンスとデータ駆動型研究の推進に向けた課題ー

データ解析政策研究室 客員研究官 池内 有為、室長 林 和弘

概 要

科学技術・学術政策研究所（NISTEP）は、日本の研究者によるデータと論文の公開状況や課題を明らかにするために、2016年、2018年に続き、2020年にウェブ質問紙調査を実施した。回答対象は様々な分野・所属・職種・年齢で構成される約2,000名の科学技術専門家ネットワークであり、大学、企業、公的機関・団体に所属する研究者1,349名から回答を得た（回答率70.5%）。

現在、研究活動を行っている1,268名のうち、データについては44.7%、論文については80.1%が公開経験を有していた。データの公開率は分野による差が大きく、最大の地球科学で70.2%、最小の工学で27.7%であった。また、データの提供経験については71.2%の回答者が、公開データの入手経験については69.7%の回答者が経験を有しており、分野別のデータの公開・共有・入手経験には有意な相関がみられた。“データ公開に必要な資源の充足度”は経年的にやや改善されていたものの全体的に低く、人材については78.5%が、時間については72.8%が、資金については71.8%が、「不足」又は「やや不足」していると認識していた。データを公開することに対する懸念も依然として強く、引用せずに利用される可能性を89.8%が、先に論文を出版される可能性を80.9%が「問題」又は「やや問題」とであると認識していた。助成機関等が要求するデータマネジメントプラン（DMP）の作成経験がある回答者は20.8%であった。データ形式の変換等、研究データ管理（RDM）を図書館員やデータキュレーターに依頼したいと考える回答者は41.1%であった。

キーワード：オープンサイエンス，研究データ公開，オープンアクセス（OA），
データマネジメントプラン（DMP），研究データ管理（RDM）

1. はじめに

科学研究の成果であるデータや論文を公開し、国や分野を超えた活用を促進するオープンサイエンス政策が加速している。2021年3月に閣議決定された「第6期科学技術・イノベーション基本計画」¹⁾では、オープンサイエンスとデータ駆動型研究の推進が打ち出され、具体的な取組として、「データの共有・活用については、研究の現場において、高品質な研究データが取得され、これら研究データの横断的検索を可能にするプラットフォームの下で、自由な研究と多様性を尊重しつつ、オープン・アンド・クローズ戦略に基づいた研究データの管理・利活用を進める環境を整備する。特にデータの信頼性が確保される仕組み

が不可欠となる」(p.59)等が明記されている。

オープンサイエンスやデータ駆動型研究の適切な効率な推進に当たっては、現状を把握した上で、根拠に基づく政策立案を行うことが重要であると考えられる。科学技術・学術政策研究所（NISTEP）は、2020年11月に日本の研究者を対象として、オープンサイエンスに関するウェブ質問紙調査を実施した²⁾。これは2016年と2018年に実施した調査^{3,4)}（以下、それぞれ「2016年調査」、「2018年調査」、両者をまとめて「2016/2018年調査」という）の後続調査である。調査対象は、大学、企業、公的機関・団体に所属する研究者や専門家、技術者等によって構成される約2,000名の科学技術専門家ネットワークである。科学技術専門家ネットワークの構成員は

毎年一部入替えがあるため、パネル調査ではない点に御留意いただきたい。

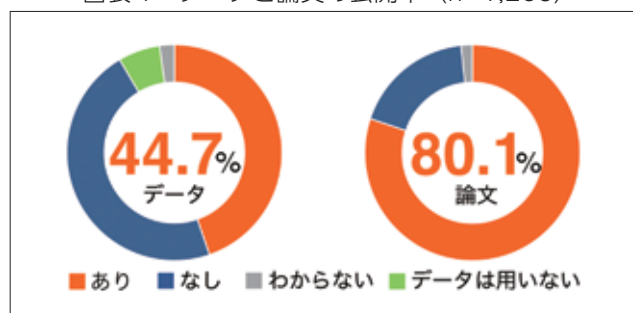
本稿では、1,349 名（回答率 70.5%）による有効回答のうち、現在研究活動を行っている 1,268 名の回答を分析した結果から、(1) データと論文の公開状況、(2) データ公開の障壁、(3) データ公開のインセンティブ、(4) データマネジメントプラン（DMP）^{注1}の作成状況、(5) 研究データ管理（RDM）の依頼意思について報告する。調査の詳細については、調査資料として公開される報告書を参照いただきたい。

2. データと論文の公開状況

2020 年調査において研究のために収集・作成・観測したデジタルデータ（以下、「データ」）を公開した経験を有する回答者（以下、データ公開率）は 44.7%、論文のオープンアクセス（以下、「OA」）公開経験がある回答者（以下、論文の OA 率）は 80.1%であった（図表 1、以下「n」は回答者数を示す）。データ公開率は、2016 年は 51.0%、2018 年は 51.9%であったため、減少傾向がみられた。減少した原因を明らかにするために、3 回の調査に共通する回答者（267 名）の回答を確認したところ 72 名の回答に矛盾がみられた。つまり、過去の調査では後述するデータ公開方法で「公開した経験がある」と考えていたものの、取り下げた回答者が一定数存在すると考えられる。また、2016/2018 年調査の時点と比較して、2020 年には「データ公開」や「オープンサイエンス」に対する理解が深まり、より正確な回答が得られるようになった結果、データ公開率が低下した可能性もあると考えられる^{注2}。一方、論文の OA 率は、2016 年は 70.8%、2018 年は 78.0%であり、増加傾向がみられた。

分野別にデータ公開率を集計すると、地球科学（70.2%）、数学（69.2%）、生物科学（65.0%）の順

図表 1 データと論文の公開率（n=1,268）

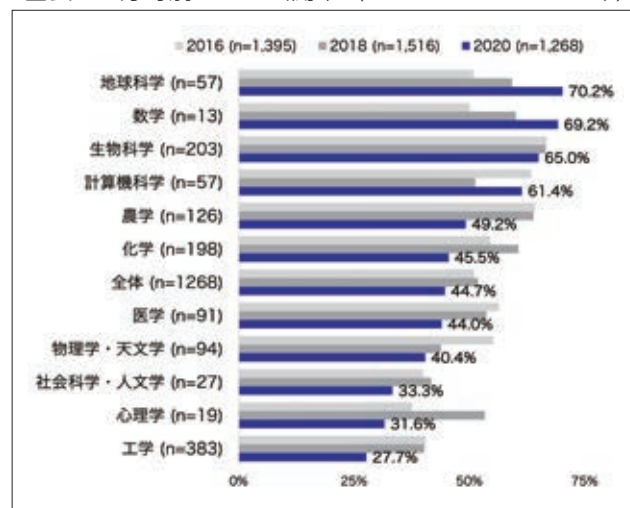


に公開率が高く、工学（27.7%）、心理学（31.6%）、社会科学・人文学（33.3%）は比較的低かった（図表 2）。分野によって公開率に差がみられたのは、2016/2018 年調査と同様であった。

データと同様に、論文の OA 率にも分野による差がみられた。そこで、分野を単位としてデータと論文の OA 率について相関を調べたが、両者に有意な相関はみられなかった（図表 3）。つまり、データ公開率が高い（低い）分野は論文の OA 率も高い（低い）とはいえなかった。データと論文の公開率がいずれも高かったのは地球科学と数学、いずれも低かったのは工学、論文の OA 率と比較してデータ公開率が高かったのは計算機科学、逆にデータ公開率に対して論文の OA 率が高かったのは社会科学・人文学等であった。

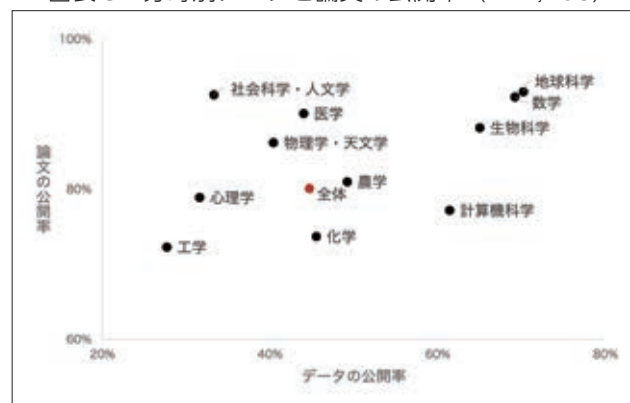
なお、データの提供（共有）経験は 71.2% の回答者が有しており、分野別にみると地球科学（93.0%）

図表 2 分野別データ公開率（2016/2018/2020 年）



※分野ごとの「n」及び肩付きの「%」はそれぞれ 2020 年調査の値を示す。

図表 3 分野別データと論文の公開率（n=1,268）



注 1 研究のために収集・作成するデータをどのように管理するか、取扱いや整備・保存・公開に関する計画を記した書類を指す。

注 2 このほか、単純な回答ミスやアンケートシステムの変更による影響も考えられる。

から心理学 (57.9%) まで 35.1 ポイントの差がみられた。分野別のデータ提供経験はデータ公開経験と正の相関がみられ、よくデータを提供している分野はよくデータを公開している、そして、よくデータを公開している分野はよくデータを提供しているという傾向がみられた。また、公開データの入手経験は 69.7% の回答者が有しており、分野別にみると、計算機科学 (91.2%) から心理学 (47.4%) まで 43.8 ポイントの差がみられた。分野別のデータ入手経験はデータ公開経験及びデータ提供経験と正の相関がみられ、よくデータを入手している分野は、同様によくデータを公開している傾向や、よくデータを提供している傾向があることが明らかになった。

データの公開方法を複数選択方式で尋ねた結果、最も選択率が高かった項目は「論文の補足資料 (54.1%)」、次いで「個人や研究室のウェブサイト (31.4%)」であった (図表 4)。2016 年調査では「個人や研究室のウェブサイト」の選択率が最も高かったが、2018 年調査で逆転し、2020 年調査では論文補足資料と研究室等ウェブサイトの間には 22.7 ポイントの差がみられた。2020 年に Hrynaszkiewicz 氏らが北米と欧州の研究者を対象として実施した調査においても、最もよく用いられているデータ共有の方法は「論文の補足資料 (67%)」であった⁵⁾。本調査では、2016/2018 年調査と比較して「特定分野のリポジトリ (29.1%)」の選択率が増加していた一方で、「学術機関のリポジトリ (25.4%)」はやや減少していた。研究者にとって身近な存在であり、かつ、永続性をもつ学術機関のリポジトリの整備が期待される。

データの公開経験を有する回答者 567 名を対象として、公開理由を複数選択方式で尋ねた。最も選択率が高かったのは「論文を投稿した雑誌のポリシー

(投稿規定) だから (54.1%)」であった (図表 5)。2020 年の調査時点では、学術雑誌のデータ公開要求に応じて補足資料によるデータ公開がよく行われていると考えられる。「オープンデータに貢献したいから (17.5%)」と「分野・コミュニティの規範だから (13.6%)」は、わずかながら経年的に増加していた。

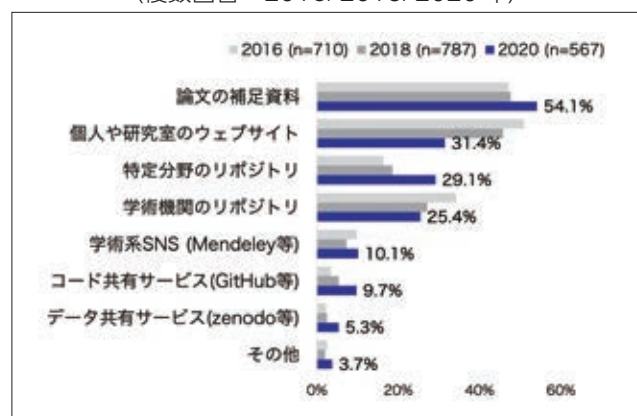
3. データ公開の障壁

データの公開の障壁を明らかにするために、データ公開経験の有無にかかわらず、研究にデータを用いる回答者全員を対象として、“論文などの成果を発表済みの、最近の主要な研究 1 件のために収集・作成・観測したデータ (以下、「カレントデータ」)” を管理し、公開することを想定させた上で^{注3}、資源の充足度や懸念の強さを尋ねた。まず、データ公開に関する資源を 6 項目挙げて、それぞれの充足度を「不十分」から「十分」までの 4 件法で尋ねた。結果を図表 6 に示す。

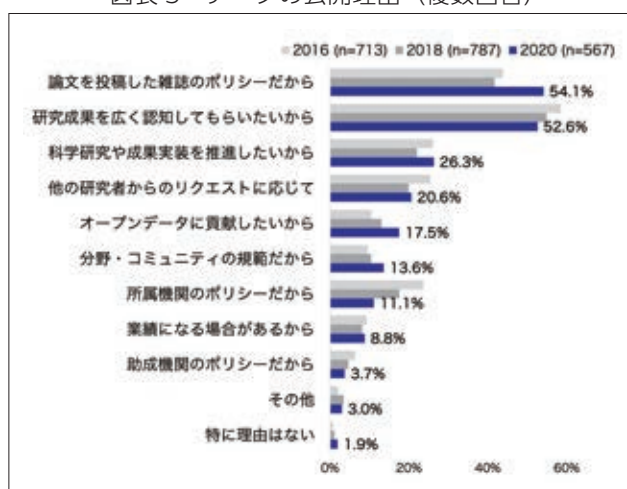
人材、時間、資金は「不十分」とする回答者が過半数であり、「十分」又は「ほぼ十分」とする回答者は 15% 未満であった。全体的に「わからない」とする回答者の比率も高く、特に公開用のリポジトリについては「わからない」とした回答者が 27.0% と、認知度が低いことがうかがえた。保存用ストレージ、公開用のリポジトリ、研究中のストレージについては、「不十分」又は「やや不十分」とした回答者が 50% を下回っており、2016/2018 年調査と比較して、やや不足感が低減している傾向がみられた。

続いてデータを公開する場合の懸念について 8 項目を挙げて、「問題」から「問題ではない」までの 4

図表 4 データの公開方法
(複数回答：2016/2018/2020 年)

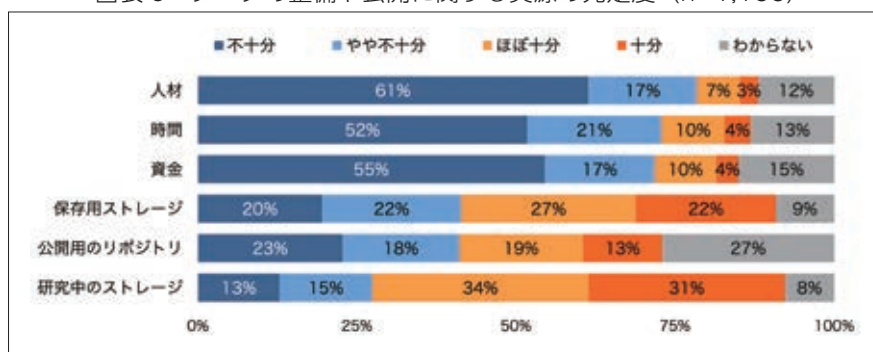


図表 5 データの公開理由 (複数回答)

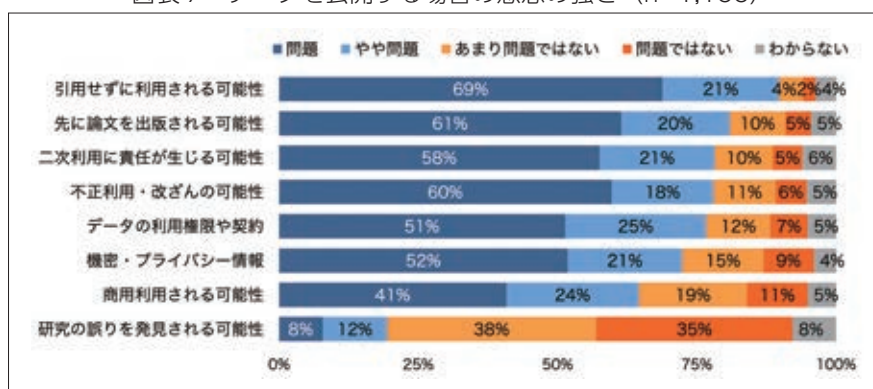


注3 研究によって扱うデータの量や種類が異なる場合があると予想されるため、カレントデータを公開することを想定して回答していた。

図表 6 データの整備や公開に関する資源の充足度 (n=1,188)



図表 7 データを公開する場合の懸念の強さ (n=1,188)



件法で尋ねた。結果を図表 7 に示す。

全体的に懸念が強く、特に「引用せずに利用される可能性」は「問題」と「やや問題」の合計が 89.8% であった。次いで「公開したデータを使って自分より先に論文を出版される可能性 (同 80.9%)」、「二次利用に関して責任が生じる可能性 (同 78.2%)」、「不正利用・改ざんの可能性 (同 78.0%)」、「データの利用権限や契約 (同 76.6%)」の順に懸念が強かった。「研究の誤りを発見される可能性 (同 19.4%)」だけは懸念をもつ回答者が比較的少なく、2016/2018 年調査と同様の傾向が確認された。

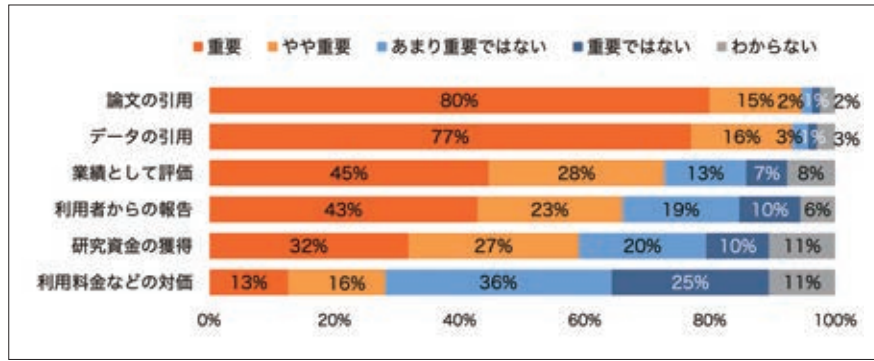
研究者が懸念しているような問題が実際に起きているのかどうかを明らかにするために、2018 年調査からデータの公開経験を有する回答者に自由記述で尋ねた。本調査では、データ公開経験を有する回答者 567 名のうち 33 名 (5.8%) が具体的な問題について記述していた。内容を筆者らでカテゴライズしたところ、最も多かった回答は 2018 年調査と同様に、公開したデータに対する「問い合わせ等への対応 (6 名)」であった。次いで「誤用された (5 名)」、「引用せずに利用された (4 名)」、「データの権利に関する問題 (4 名)」の順であった。

4. データ公開のインセンティブ

研究にデータを用いている回答者を対象として、データ公開によって得られるインセンティブの重要度を 4 件法で尋ねた。最も重要であると考えられていたのは「データに紐づいた論文の引用 (「重要」と「やや重要」の合計 94.5%)」、次いで「データの引用 (論文と同様に、参考文献リストにデータ作成者やデータ名、識別子などを記載する) (同 93.0%)」であった (図表 8)。図表 7 に示したように「引用せずに利用される可能性」は最も重要な懸念であったと同時に、論文やデータを引用されることはデータ公開のインセンティブとして重要視されていた。これは 2018 年調査と同様の結果である。

データを公開することによって、実際にどのようなインセンティブが得られているのかを明らかにするために、2018 年調査からデータ公開経験を有する回答者にインセンティブの内容について自由記述方式で尋ねている。その結果、2020 年調査ではデータ公開経験を有する回答者 567 名のうち 130 名 (22.9%) が具体的な事柄について記述していた。最も多かった回答は 2018 年調査と同様に、共同研究の契機や研究の進展といった「研究上の利点 (54 名)」であり、次いで「研究・データ・研究者のビジビリティ向上 (41 名)」、「科学・分野の進展 (27

図表 8 データ公開のインセンティブの重要性 (n=1,180)



名)」であった。研究者が重視している引用を挙げた回答者は2名にとどまったが、データを公開している論文は引用が増加することを示す複数の研究がある⁶⁾。また、データ自体の引用が将来的に研究業績として評価され、研究費の獲得につながることも期待されている⁷⁾。

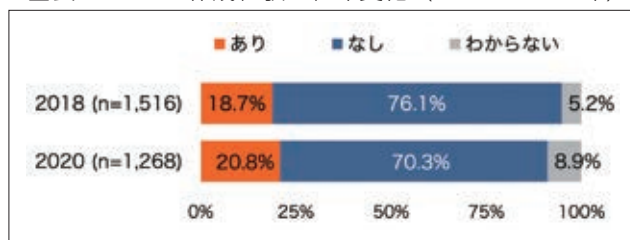
5. データマネジメントプラン (DMP) の作成状況

国立研究開発法人科学技術振興機構 (JST) を始め、複数の助成機関が DMP の作成を求めようになったことを契機として、2018 年調査から DMP の作成状況についての質問も取り入れた。その結果、2020 年調査では DMP の作成経験を有する回答者は 20.8% であり、2018 年調査から 2.1 ポイント増

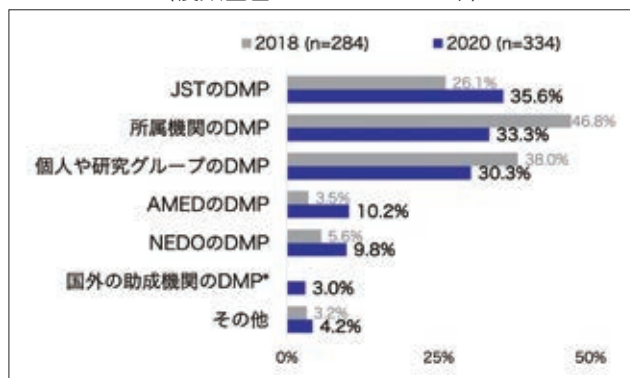
加していた。また、「わからない」を選択した回答者は 8.9% であり、3.7 ポイント増加していた (図表 9)。

DMP 作成の目的・要求元として例示した機関等のうち選択率が最も高かった回答は「科学技術振興機構 (JST) (35.6%)」、2 位は「所属機関の DMP」(33.3%)、3 位は「個人や研究グループの DMP (30.3%)」であった (図表 10)。助成機関に着目すると、JST は 9.5 ポイント、国立研究開発法人日本医療研究開発機構 (AMED) は 6.7 ポイント、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) は 4.2 ポイント増加していた。DMP を作成した理由の 1 位は「助成機関が要求しているから (n=351, 51.5%)」であり、作成していない理由で選択率の 1 位は「DMP を知らなかったから (n=891, 52.0%)」であった。DMP の作成経験を有する研究者の比率や認知度はそれほど高くないものの、助成機関の要求により作成する研究者が少しずつ増加していると考えられる。

図表 9 DMP 作成経験の経年変化 (2018/2020 年)



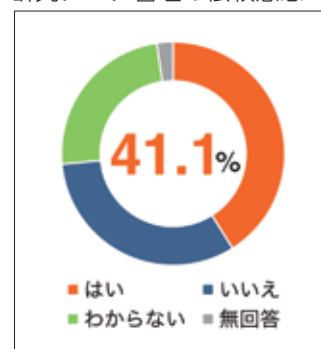
図表 10 作成経験がある DMP (複数回答：2018/2020 年)



6. 研究データ管理 (RDM) の支援

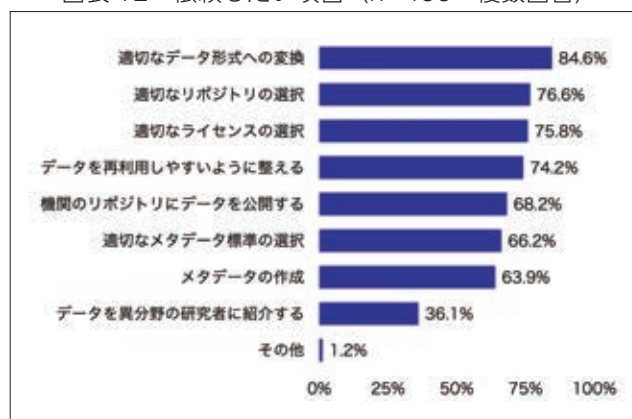
RDM、すなわちデータの整備や公開、保存プロセスを図書館員やデータキュレーターに依頼したいかどうかを尋ねた結果、依頼したいと考えている回答者は 41.1% であった (図表 11)。

図表 11 研究データ管理の依頼意思 (n=1,188)



依頼したい項目を複数選択方式で尋ねた結果、最も選択率が高かったのは「適切なデータ形式への変換 (84.6%)」であった (図表 12)。2016/2018 年調査では、これらのプロセスのうち専門性が必要であると考えられる項目を複数選択方式で尋ねた。その結果、2016/2018 年調査ともに 1 位は「適切なデータ形式への変換」、2 位は「データを再利用しやすいように整える」であり、選択率が低かったのは 7 位「機関リポジトリによるデータ公開」、6 位「データを異分野の研究者に紹介する」であった。つまり、回答者の多くが依頼したいと考えているのは、比較的専門性が高いと認識されているプロセスであることがわかった。

図表 12 依頼したい項目 (n=488：複数回答)



7. おわりに

日本の研究者によるデータ公開は、学術雑誌のポリシーに応じる形で論文の補足資料を中心に実践されているものの、分野による差があり、論文の OA 率ほど増加していないことが明らかになった。データ公開の課題として、まず、人材、時間、資金の不足感が強かった。また、データ公開に対する懸念も依然として強く、特にデータを公開した場合に引用せずに利用される可能性を問題視している一方で、データ公開のインセンティブとして、データに紐づいた論文やデータそのものの引用が重視されていた。データの管理や公開に関わる DMP は助成機関による要求により、作成経験を有する研究者がわずかながら増えていた。また、回答者の 40% 以上は研究データ管理の支援を依頼したいと考えていることがわかった。本調査の結果が、学術機関、出版社、学協会、政策担当者、助成機関といったステークホルダーによる議論に資することを期待している。

謝辞

調査及びプレテストに御協力賜った皆様に心よりお礼申し上げます。また、2021 年 4 月から大森悠生氏 (筑波大学大学院) に RA (リサーチアシスタント) として集計や分析等に御尽力いただいた。厚くお礼申し上げます。

参考文献・資料

- 1) 内閣府. 第 6 期科学技術・イノベーション基本計画. 2021.
<https://www8.cao.go.jp/cstp/kihonkeikaku/index6.html>, (accessed 2021-10-23).
- 2) 池内有為, 林和弘. 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. 文部科学省科学技術・学術政策研究所, 2021, NISTEP RESEARCH MATERIAL No.316, 100p. <https://doi.org/10.15108/rm316>, (accessed 2021-11-30).
- 3) 池内有為, 林和弘, 赤池伸一. 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. 文部科学省科学技術・学術政策研究所, 2017, NISTEP RESEARCH MATERIAL No.268, 108p.
<https://doi.org/10.15108/rm268>, (accessed 2021-10-23).
- 4) 池内有為, 林和弘. 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 2018. 文部科学省科学技術・学術政策研究所, 2020, NISTEP RESEARCH MATERIAL No.289, 96p. <https://doi.org/10.15108/rm289>, (accessed 2021-10-23).
- 5) Hrynaszkiewicz, I.; Harney, J.; Cadwallader, L. A survey of researchers' needs and priorities for data sharing. *Data Science Journal*, 2021, vol. 20, no. 1, p.31. <https://doi.org/10.5334/dsj-2021-031>, (accessed 2021-10-23).
- 6) Colavizza, G.; Hrynaszkiewicz, I.; Staden, I.; Whitaker, K.; McGillivray, B. The citation advantage of linking publications to research data. *PLOS ONE*, 2020, vol. 15, no. 4, e0230416.
<https://doi.org/10.1371/journal.pone.0230416>, (accessed 2021-10-23).
- 7) 能勢正仁, 池内有為. データ引用を研究活動の新たな常識に: 研究データ活用協議会 (RDUF) リサーチデータサイテーション小委員会の活動. *カレントアウェアネス*. 2020, no. 345, p. 2-4.
<https://doi.org/10.11501/11546850>, (accessed 2021-10-23).