概要

概要

科学研究の成果であるデータや論文を公開して、学術関係者のみならず民間企業を含む一般市民が自由にアクセスして利用できるようにするオープンサイエンス政策は、世界的な拡がりを見せている。

日本でも、2021 年 3 月 26 日に閣議決定された『第 6 期科学技術・イノベーション基本計画』[1]では、第 2 章 Society 5.0 の実現に向けた科学技術・イノベーション政策の「2. 知のフロンティアを開拓し価値創造の源泉となる研究力の強化」において、オープンサイエンスとデータ駆動型研究等の推進のために新たな研究システムを構築することが明記されている（p. 58〜）。こうした政策を適切かつ効率的に実現していくためには、まず、現状を把握した上で、学術機関、出版社、学協会、政策担当者、助成機関といった幅広いステークホルダーによる議論が誘発されることが望ましいと考えられる。特に研究者が論文に加えて研究データの共有・公開することによって、データ駆動型研究が促進されるが、その実態は十分に明らかになっていない。

そこで科学技術・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認識を明らかにするために、2016 年[2]と 2018 年[3]に科学技術専門家ネットワークを対象としてウェブ質問紙調査を実施した。3 回目となる本調査では、大学、企業、公的機関・団体に所属する研究者 1,349 名（回答率 70.5%）による回答を分析した結果を 2016/2018 年調査の結果と比較した。調査項目は 2016/2018 年調査を踏襲しつつ、新たにデータ公開に対する評価に関する質問を行った。以下では概要として、(1)データと論文の公開状況、(2)公開データの利用状況と課題、(3)データマネジメントプラン（DMP）の作成状況、(4)データ公開の障壁、(5)データ公開のインセンティブ、(6)研究データ管理（RDM）に対する認識についての結果を示す。

---

[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 2021-10-18).

[2] 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 2021-10-18).

[3] 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 2021-10-18).

(1) データと論文の公開状況

研究のために収集・作成・観測したデジタルデータで、論文など研究成果の根拠となるもの（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%、論文を OA にした経験を有する回答者（以下、「論文の OA 率」）は 80.1%であった（図 1）。

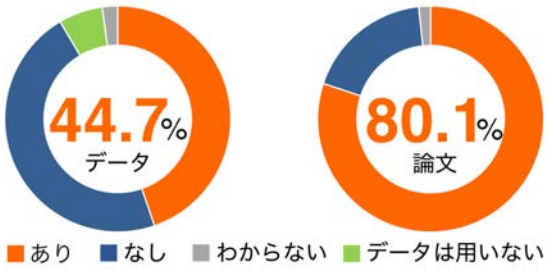

図 1　データと論文の公開経験（n=1,268）

2016 年調査におけるデータ公開率は 51.0%、2018 年調査は 51.9%であり、2 年前と比較して 7.2 ポイント減少していた。論文の OA 率は 2016 年調査が 70.9%、2018 年調査は 78.0%であり、2 年前と比較して 2.1 ポイント増加した。

図 2 に示すように、分野別のデータ公開率は、地球科学（70.2%）から工学（27.7%）まで 42.5 ポイントの差がみられた。全体の公開率は低下していたものの、地球科学と数学は増加していた。

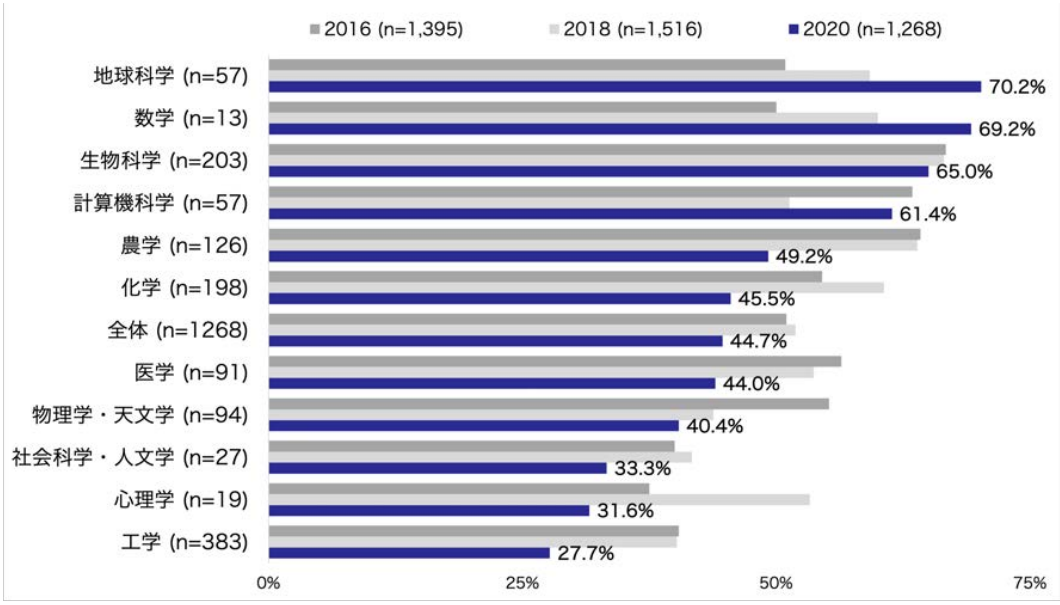

図 2　分野別データ公開率（2016/18/20 年）

データの公開方法は、「論文の補足資料」（54.1%）、「個人や研究室のウェブサイトへの掲載」（31.4%）の順に選択率が高かった（図 3）。オープンサイエンス政策や学術雑誌のデータ共有ポリシーで推奨、あるいは想定されている永続性のあるリポジトリによる公開

は「特定分野のリポジトリ」(29.1%)が 2018 年度から 10.5 ポイント増加していた一方で、「学術機関のリポジトリ」(25.4%) は 2018 年度から 1.7 ポイント減少していた。「学術系 SNS」[4] (10.1%)、「コード共有サービス」[5] (9.7%)、「データ共有サービス」[6] (5.3%) の選択率はそれほど高くないものの、2016/2018 年調査と比較すると増加していた (図 6)。

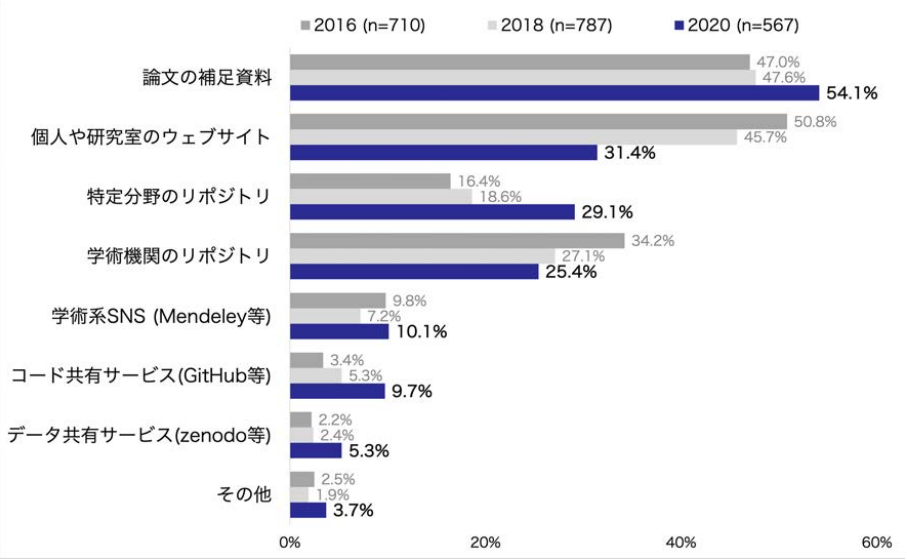

図 3　データの公開方法（2016/18/20 年：複数回答）

データの公開理由は 1 位が「論文を投稿した雑誌のポリシーだから」(54.1%)、2 位が「研究成果を広く認知してもらいたいから」(52.6%) であり、2016/2018 年調査とは 1 位と 2 位の順位が逆転していた。論文の公開理由は 1 位が「論文を投稿した雑誌が OA だから」(75.8%)、2 位が「研究成果を広く認知してもらいたいから」(57.6%) であり、2016/2018 年調査と同様であった。データ、論文のいずれも雑誌のポリシーが主たる理由であった。

データの未公開理由は 1 位が「論文を投稿した雑誌のポリシーではないから」(38.3%)、2 位が「業績にならないから」(26.6%) であった。論文の未公開理由は 1 位が「資金がないから」(57.6%)、2 位が「論文を投稿したい雑誌が OA ではないから」(40.3%) であり、2016/2018 年調査とは 1 位と 2 位の順位が逆転していたものの、この 2 つの理由が突出しているという点は同様であった。

データの未公開理由を尋ねた後に、その問題が解決された場合のデータの公開意思を尋ねた結果、「はい」は 29.4%、「いいえ」は 29.2%、「わからない」は 41.4%であった (図 4)。全体的にデータ公開に対して慎重な姿勢は継続しているものの、やや「はい」の比率が増え、「わからない」の比率が減少している傾向がみられた。

[4] 質問紙では、学術系 SNS の例として Mendeley（Elsevier）と ResearchGate を示した。
[5] 同様に、ソースコードを共有できる GitHub を示した。
[6] 同様に、無料でデータを公開できる figshare と zenodo（CERN）を示した。

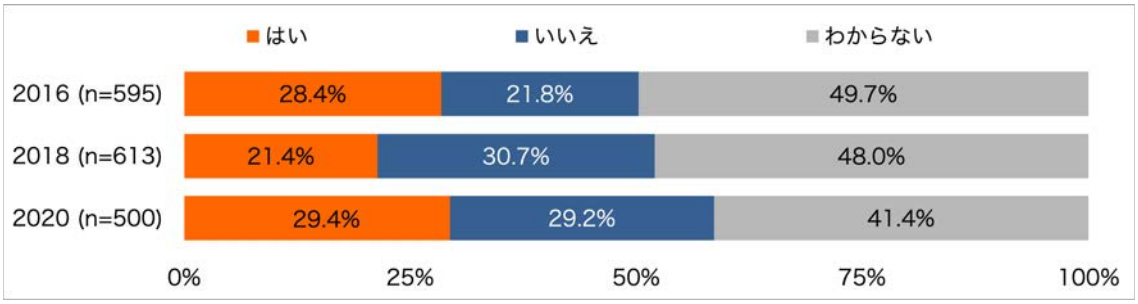

図 4　未公開理由が解決した場合のデータの公開意思（2016/18/20 年）

(2) 公開データの利用状況と課題

公開データの入手経験がある回答者は 69.7%であった。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（図 5）。分野別のデータ入手経験はデータ公開経験及びデータ共有（提供）経験と正の相関がみられた。すなわち、よくデータを入手している分野は、よくデータを公開している傾向や、よくデータを共有している傾向がみられた。

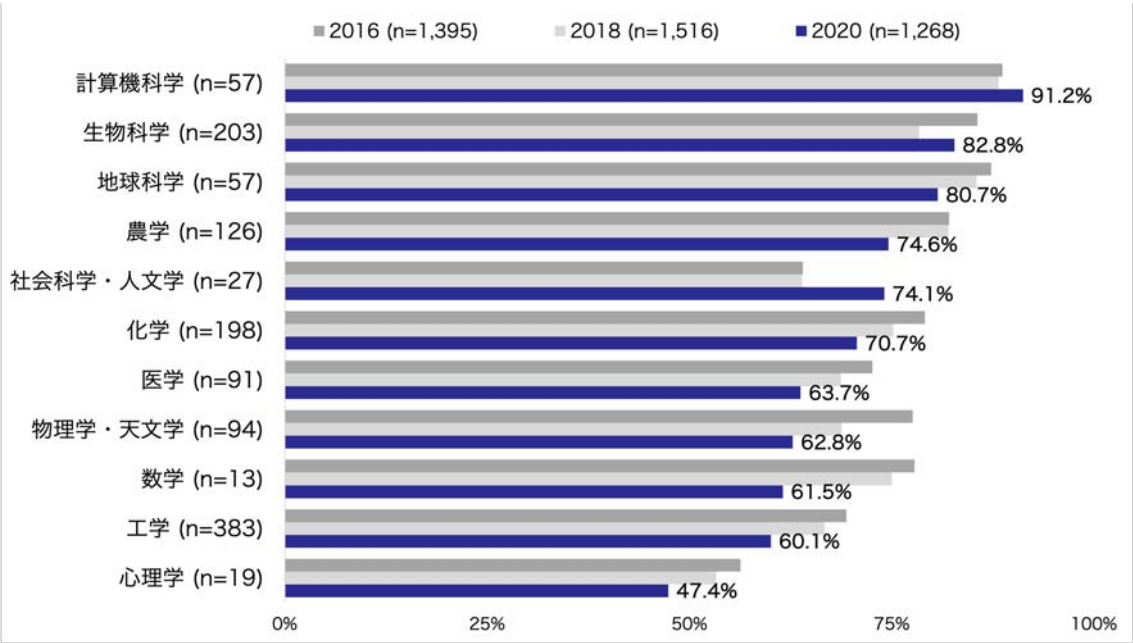

図 5　分野別公開データの入手経験（2016/18/20 年）

公開データの入手方法のうち、最も選択率が高かったのは「論文の補足資料」(59.4%)、次いで「学術機関のリポジトリ」(57.6%)、「個人や研究室のウェブサイト」(38.9%) であった（図 6)。2016/2018 年調査においては、1 位が「個人や研究室のウェブサイト」、2 位が「論文の補足資料」、3 位が「学術機関のリポジトリ」であったため変化がみられた。また「特定分野のリポジトリ」、「学術系 SNS (Mendeley 等)」「コード共有サービス (GitHub 等)」の割合が増加していた。

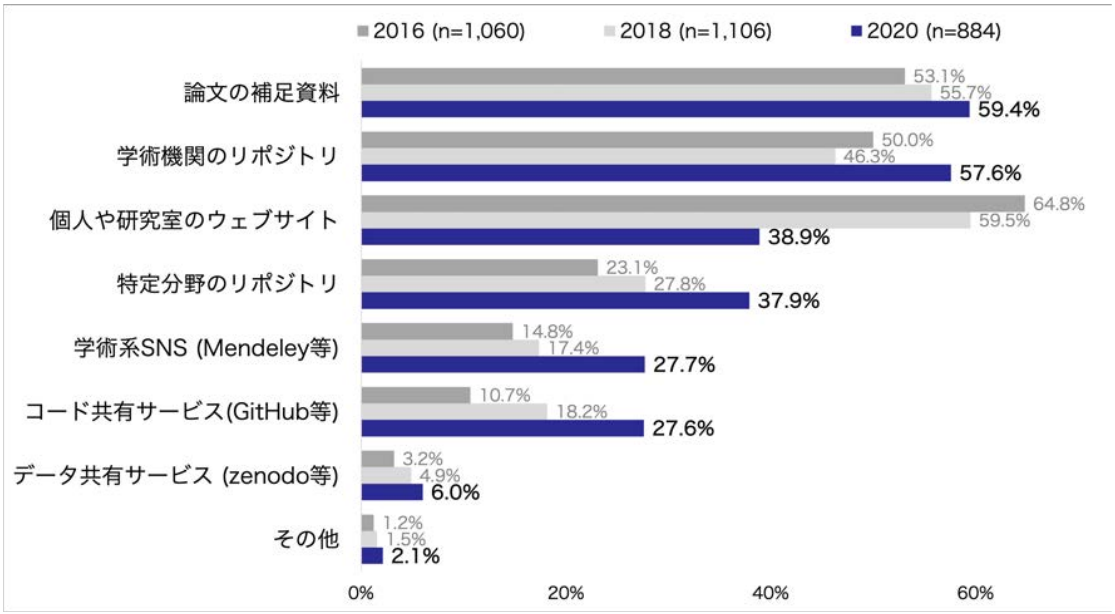

図 6　公開データの入手方法（2016/18/20 年：複数回答）

公開データの入手経験がある回答者 884 名のうち、32.0%が何らかの障壁があると回答していた。具体的な項目を確認すると、1 位は「データごとに品質が異なる」(47.7%)、2 位は「データごとにフォーマットが異なる」(45.9%)、3 位は「利用条件（営利利用が可能かどうかなど）がよくわからない」(31.8%）であった（表 9)。

表 1　データ入手の障壁の経年変化（2016/18/20 年）

順位	2016 年	2018 年	2020 年
1 位	利用料金	データの品質が異なる	データの品質が異なる
2 位	利用者登録	データのフォーマット	データのフォーマット
3 位	利用条件がわからない	利用条件がわからない	利用条件がわからない
4 位	データの品質が異なる	著作者情報がわからない	著作者情報がわからない
5 位	データのフォーマット	利用料金	データの解釈・再利用
6 位	アクセス方法	利用者登録	アクセス方法
7 位	著作者情報がわからない	データの解釈・再利用	利用料金
8 位	データの解釈・再利用	アクセス方法	利用者登録
9 位	入手までの時間	入手までの時間	入手までの時間
10 位	最新データの入手	最新データの入手	最新データの入手

2016/2018 年調査と比較すると、2016 年調査ではデータの入手時点における問題（利用料金や利用者登録が必要であること）の選択率が高かったが、2018 年調査、本調査ではこれらの順位が徐々に低下しており、データの再利用における問題の順位が徐々に高まっていた。

(3) データマネジメントプラン（DMP）の作成状況

データマネジメントプラン（Data Management Plan, 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。

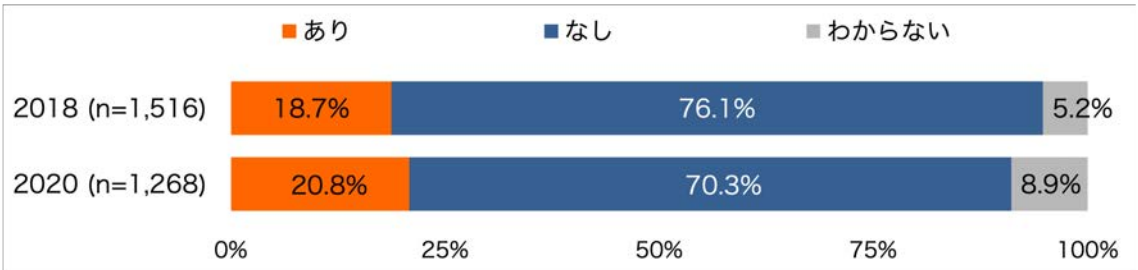

図 7　DMP 作成経験の経年変化（2018/20 年）

例示した DMP のうち選択率が最も高かったのは「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ための DMP」（30.3%）（図 8）。作成理由のうち、選択率が最も高かったのは「助成機関が要求しているから」であり、JST は 9.5 ポイント、日本医療研究開発機構（AMED）は 6.7 ポイント、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。

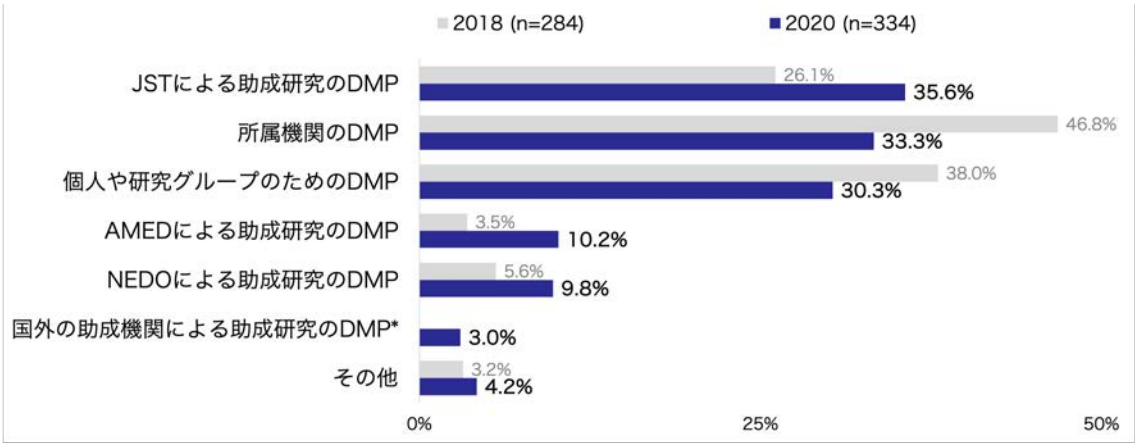

図 8　作成経験がある DMP（2018/20 年：複数回答）

DMP を作成していない理由の 1 位は「DMP を知らなかったから」であり、2018 年調査と同様であった。現状では DMP の認知度がそれほど高くないことが示唆されたものの、前回調査と比べると 3.7 ポイント減少していた。2 位は「所属機関から要求されていないから」（39.3%）、3 位は「助成機関から要求されていないから」（33.1%）であった。

(4) 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、資源の充足度や懸念の強さを尋ねた。全体的に不十分であるという認識をもつ回答者が多く、人材、時間、資金については「不十分」とする回答者が過半数であった（図 9)。「わからない」とする回答者の比率も高く、人材は 12.0%、時間は 13.1%、資金は 15.0%であった。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」と「やや不十分」の合計選択率が 5 割を下回っており、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。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%であり、認知度が高くないことがうかがえた。

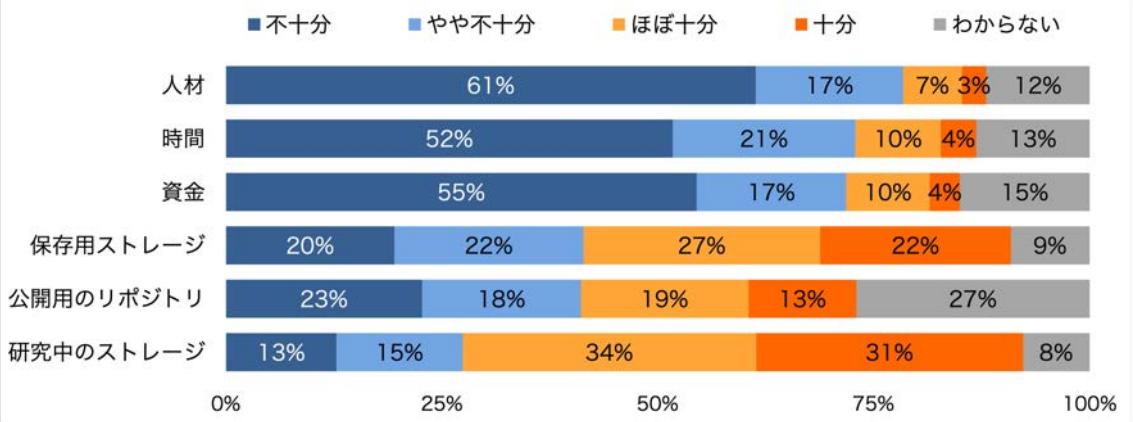

図 9　データの整備や公開に関する資源の充足度（n=1,188）

カレントデータの公開に関する問題のうち、最も懸念が強かった項目は、「引用せずに利用される可能性」（「問題である」と「やや問題である」の合計 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。「研究の誤りを発見する可能性」（同 19.4%）は懸念をもつ回答者が比較的少なく、「問題ではない」（35.3%）と「あまり問題ではない」（37.5%）を選択する回答者の比率が高かった（図 10)。

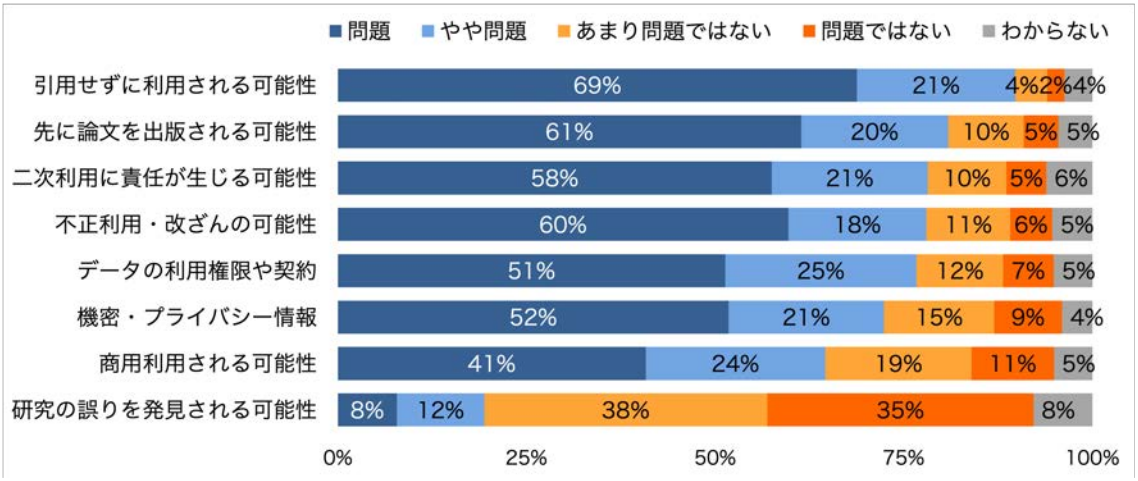

図 10　カレントデータの公開に関する懸念（n=1,188）

2016 年調査や先行研究によってデータ公開に対する懸念が強いことが明らかにされてきた。そこで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 年調査からデータ公開経験を有する回答者に自由記述で尋ねることとした。その結果、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。表 2 に示すように、最も多かったのは 2018 年調査と同様、公開したデータに対する「問い合わせ等への対応」であった。次いで「誤用された」、「引用せずに利用された」、「データの権利に関する問題」の順であった。

表 2　データ公開によるデメリット（2018/20 年）

2018 年調査	人数	2020 年調査	人数
問い合わせ等への対応	17	問い合わせ等への対応	6
引用せずに利用された	14	誤用された	5
先取権の喪失	9	引用せずに利用された	4
誤用された	7	データの権利に関する問題	4
更新のコストがかかる	2	懸念が生じた	3
徒労感	2	ミスを発見した	2
不正アクセス	1	公開に手間がかかる	2
商用利用された	1	先取権の喪失	2
研究者以外に利用された	1	長期公開・保存のコストがかかる	2
		不正アクセス	2
		情報流出	1
合計	54	合計	33

(5) データ公開のインセンティブ

データ公開のインセンティブを明らかにするために、2018 年調査からデータ公開経験を

有する回答者に自由記述で尋ねることとした。その結果、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。表 12 に 7 項目に分類した集計結果を 2018 年調査の結果とともに示す。なお、複数の内容を含むコメントはそれぞれカウントしたため、合計 152 件となった。最も多かったのは 2018 年調査と同様に「研究上の利点」(35.5%) であり、次いで「研究・データ・研究者のビジビリティ向上」（25.8%)、「科学・分野の進展」（17.8%）であった。

表 3　データ公開によって得られた良い結果（2018/20 年）

内容	2018 年		2020 年	
	件数	比率	件数	比率
研究上の利点	104	40.6%	54	35.5%
研究･データ･研究者のビジビリティ向上	66	25.8%	41	27.0%
科学・分野の進展	27	10.5%	27	17.8%
人とのつながり	26	10.2%	14	9.2%
評価	11	4.3%	7	4.6%
個人的な利点	9	3.5%	6	3.9%
その他	13	5.1%	3	2.0%
合計	256	100.0%	152	100.0%

また、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 11)。図 10 に示したように「引用されずに利用される可能性」は最も重要な懸念であったのと同時に、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 年調査でも同様の結果であった。

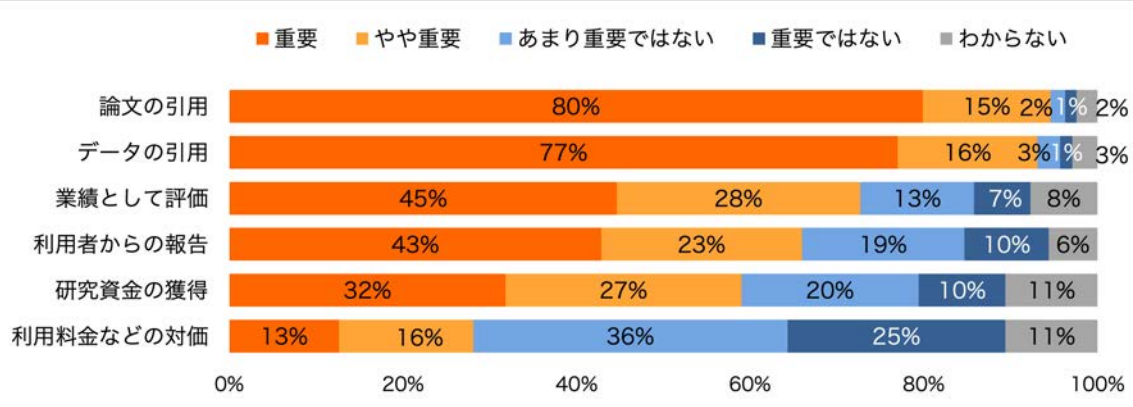

図 11　データ公開によるインセンティブの重要度（n=1,180）

データ公開に対する評価の状況を明らかにするために、回答者自身が所属するコミュニティや機関、及び回答者自身が評価しているかどうかを尋ねた。両者を比較するために、回答から「その他」と「指導する研究者がいない」を除いて再集計した結果を示す（図 12）。

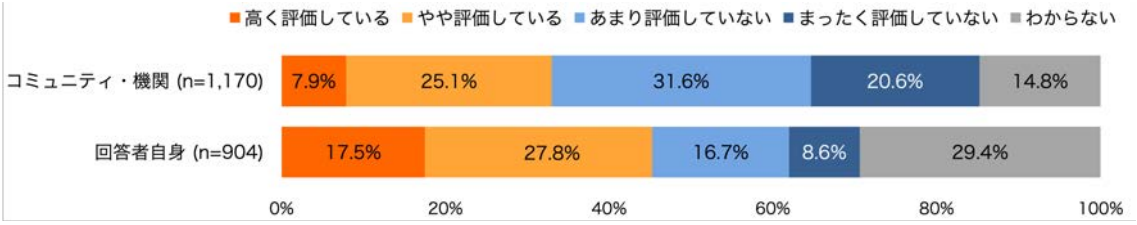

図 12　データ公開に対するコミュニティ・機関と回答者自身の評価

コミュニティや機関よりも回答者自身の方が、評価している回答者の比率が高く、評価していない回答者の比率が低かった。「わからない」とする回答者の比率も高く、回答者自身については 29.4%であった。データを公開することが評価につながるようになるのかどうか、今後も継続的に調査していきたい。

(6) 研究データ管理（RDM）に対する認識

研究データ管理（Research Data Management, RDM）に関するリテラシーへの関心を把握するために、データの整備や公開に関する項目を挙げて複数選択方式で尋ねたところ、回答者の 87.5%が何らかの項目を選択していた。2016/2018 年調査から引き続き、「適切なデータ形式」、「知的財産権やライセンス」、「データの安全な管理方法」の選択率が高かった（図 13）。

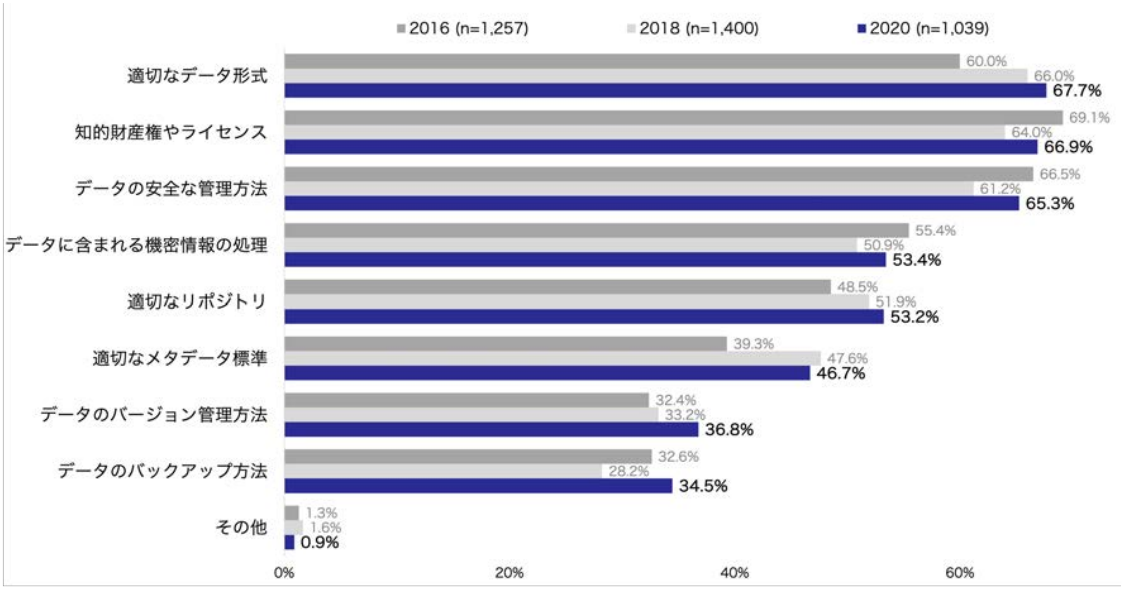

図 13　研究データ管理（RDM）に関して知りたい項目（2016/18/20 年：複数回答）

RDM プロセスを、回答者自身や共同研究者にかわって図書館員やデータキュレーターに依頼したいと考えるかどうかを尋ねたところ、回答者の 41.1%が依頼したいと考えていた。具体的に依頼したいプロセスとしては、「適切なデータ形式への変換」、「適切なリポジトリの選択」、「適切なライセンスの選択」の選択率が高かった（図 14）。

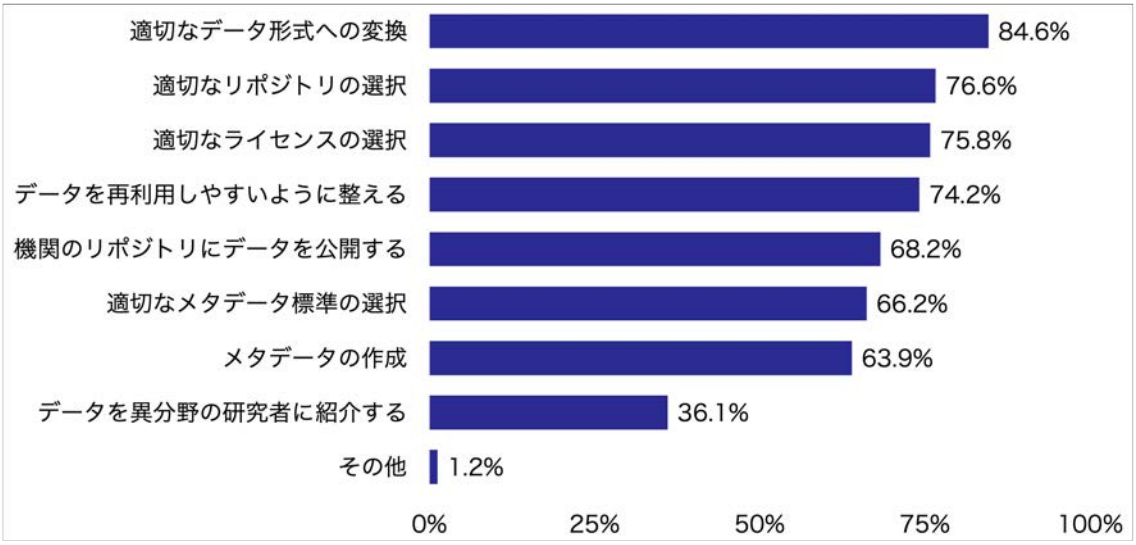

図 14　研究データ管理（RDM）に関して依頼したい項目（n=488：複数回答）

本調査によって、日本の研究者によるオープンサイエンスの実践状況や認識、課題とその経年的な変化を明らかにした。今後も継続的な調査を実施することによって、日本のオープンサイエンスがどのように変化していくのかを明らかにしていきたい。