A Survey of Researchers' Needs and Priorities for Data Sharing

RESEARCH PAPER

IAIN HRYNASZKIEWICZ
JAMES HARNEY
LAUREN CADWALLADER
*Author affiliations can be found in the back matter of this article

ABSTRACT

One of the ways in which the publisher PLOS supports open science is via a stringent data availability policy established in 2014. Despite this policy, and more data sharing policies being introduced by other organizations, best practices for data sharing are adopted by a minority of researchers in their publications. Problems with effective research data sharing persist and these problems have been quantified by previous research as a lack of time, resources, incentives, and/or skills to share data.

In this study we built on this research by investigating the importance of tasks associated with data sharing, and researchers' satisfaction with their ability to complete these tasks. By investigating these factors we aimed to better understand opportunities for new or improved solutions for sharing data.

In May-June 2020 we surveyed researchers from Europe and North America to rate tasks associated with data sharing on (i) their importance and (ii) their satisfaction with their ability to complete them. We received 617 completed responses. We calculated mean importance and satisfaction scores to highlight potential opportunities for new solutions to and compare different cohorts.

Tasks relating to research impact, funder compliance, and credit had the highest importance scores. 52% of respondents reuse research data but the average satisfaction score for obtaining data for reuse was relatively low. Tasks associated with sharing data were rated somewhat important and respondents were reasonably well satisfied in their ability to accomplish them. Notably, this included tasks associated with best data sharing practice, such as use of data repositories. However, the most common method for sharing data was in fact via supplemental files with articles, which is not considered to be best practice.

We presume that researchers are unlikely to seek *new* solutions to a problem or task that they are satisfied in their *ability* to accomplish, even if many do not attempt this task. This implies there are few opportunities for *new solutions or tools* to meet these researcher needs. Publishers can likely meet these needs for data sharing by working to seamlessly integrate *existing* solutions that reduce the effort or behaviour change involved in some tasks, and focusing on advocacy and education around the benefits of sharing data.

CORRESPONDING AUTHOR:
Iain Hrynaszkiewicz
PLOS, Carlyle House, Carlyle Road, Cambridge CB4 3DN, UK
ihrynaszkiewicz@plos.org

KEYWORDS:
Open research; data sharing; publishing practices; survey results; open science; reuse of research data

TO CITE THIS ARTICLE:
Hrynaszkiewicz, I, Harney, J and Cadwallader, L. 2021. A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, 20: 31, pp. 1–16. DOI: *https://doi.org/10.5334/dsj-2021-031*

There may however be opportunities – unmet researcher needs – in relation to better supporting data reuse, which could be met in part by strengthening data sharing policies of journals and publishers, and improving the discoverability of data associated with published articles.

DOI: 10.5334/dsj-2021-031

INTRODUCTION

PLOS introduced a strong data availability policy in 2014 requiring all authors make the research data that support their results publicly available without restriction, with rare exceptions. The availability of research data supporting scholarly publications is increasing, slowly (Colavizza et al. 2020), and since 2015 many journals and publishers have also introduced journal data sharing policies (Hrynaszkiewicz et al. 2020). Policies contribute to increased availability of research data but new solutions may be needed to further accelerate best practice in data sharing – in compliance with the Findable, Accessible, Interoperable and Reusable (FAIR) principles (Wilkinson et al. 2016).

Aspects of researchers' experiences and attitudes about sharing research data are, relative to other aspects of open science, well-studied. A meta-synthesis of 45 qualitative studies of researchers' practices and perceptions about data sharing found that researchers lack time, resources, and skills to effectively share their data in public repositories (Perrier et al. 2020). In previous research, a lack of suitable infrastructure for data sharing is commonly cited as a barrier to the availability of research data along with a lack of incentives (Science et al. 2019).

Considering the results of previous studies (Allagnat et al. 2019; Borghi & Van Gulick 2018; Eynden et al. 2016; Federer et al. 2018; Houtkoop et al. 2018; Kratz et al. 2015; Open Data 2017; Rathi et al. 2012; Schmidt et al. 2016; Science et al. 2017; 2018; Stuart et al. 2018; Tenopir et al. 2011; 2015; 2018; 2020), beyond infrastructure, researchers' concerns about misuse and scooping (lost publication opportunities) are amongst the most common concerns about, and barriers to, data sharing. In previous research, these concerns are followed in their frequency by, more practical, concerns about copyright and licensing (ownership) and the time and effort required to make research data openly available. The mounting evidence of the common concerns about data sharing, considered in isolation, might suggest a substantial need for new solutions to address these concerns, but the importance of these problems to researchers and researchers' ability to easily solve them is less clear.

In principle, there are numerous solutions available for the problems represented in these findings – from repositories, institutional research support, training programmes, to journal policies and procedures. Yet in practice — to give one example of FAIR data practice — only about a fifth of researchers use repositories to share data when published articles are analysed (Colavizza et al. 2020). Most authors, including PLOS authors, who share research data publicly choose to share data as supporting information files with their published articles (Stuart et al. 2018).

Conversations with PLOS authors have suggested researchers favour the convenience of sharing data as supplemental files (supporting information) with their papers, and, consistent with some surveys, view journals and publishers as a trustworthy steward of their research data (Science et al. 2019). While common problems with data sharing have been repeatedly identified, there is little evidence on how important each of these problems are, and if and how well existing tools, products, and services help to solve these problems.

Adapting an approach to user research and surveying rooted in Jobs To Be Done theory (Christensen et al. 2016), we sought to understand how important different tasks ("factors") associated with data sharing are to researchers, and how satisfied researchers are with their ability to complete that task ("factor"). Part of our motivation for this research was exploring opportunities for new products, partnerships or services that support better data sharing practices by researchers, in particular PLOS authors. We also believed this approach would illuminate the likelihood that researchers will adopt new solutions, providing insight not available from previous studies.

We hypothesised that researchers had unmet needs, which would be represented by factors rated as important but unsatisfied, in tasks relating to:

- Preparing, managing, publishing and understanding reuse of their research data

- Compliance with the data sharing policies of funding agencies, institutions and journals
- Their ability to obtain and access other researchers' data for reuse

METHODS

RECRUITMENT

Our recruitment plan utilized a wide range of channels to reach researchers. This strategy leveraged (a) direct email campaigns, (b) promoted Facebook and Twitter posts, (c) a post on the PLOS Blog, and (d) emails to industry contacts who distributed the survey on our behalf. URL variables were assigned to track the efficacy of each recruitment channel.

Participation was incentivized with 3 random prize draws, each with a $200 prize. The prize draw was managed via a separate survey to maintain anonymity, and 559 of 728 eligible participants entered the prize draw.

The survey received 617 completed responses although 1477 people responded to some of the survey.

The effectiveness of these recruiting methods varied widely. Of the participants who completed the survey, nearly 80% were recruited via direct email campaigns, with the overwhelming majority of these coming in response to a dedicated message about the survey.

Given the importance of our direct email campaigns, it is unsurprising that our cohort was composed largely of former PLOS authors, accounting for 82% of the users who completed the survey.

SURVEY INSTRUMENT

The individual factors associated with data sharing tasks were recast into outcome statements, which represent a researcher's hypothetical success measures for completing a task. Statements were identified by considering the policies, procedures and tasks associated with various aspects of research data management and publishing, the solutions and support currently available to researchers to complete these tasks, and were further informed by conversations with researchers to develop the survey. The outcome statements were constructed with a standard syntax, to ensure that they could be usefully compared to each other.

Typically, a desired outcome statement is composed of a direction of change, a metric of change, an object of change, and an optional context. For example,

> "Spend less time creating a data availability statement"

Where, "spend less" is the direction, "time" is the metric, and "data availability statement" is the object. The context in this case is provided in the associated survey question phrase, "when submitting your research for peer review".

In some cases we have forgone the direction when the context is sufficient to define the goal the researcher is trying to achieve such as "My research data has its own Digital Object Identifier (DOI)" (context from associated question: "when preserving or archiving your data"). While the resulting statements diverge from standard practice, it should not impact how well they can be tested in survey work, or their usefulness in identifying researcher needs.

Using these statements, we constructed a survey in SurveyGizmo, now known as Alchemer, which measured how important a researcher thought the task was, and their level of satisfaction with being able to complete it. The survey was tested by individuals not involved with the study, and who have scientific backgrounds, before deployment to ensure it is understandable.

Importance was measured using a five-point *unipolar* scale, which was later mapped to a value from 0 to 100:

1. Not at all important: 0
2. Slightly important: 25
3. Moderately important: 50
4. Very important: 75
5. Extremely important: 100

Since **satisfaction** can be expressed in *negative* terms, it needs a different scale that can account for this. Therefore, satisfaction was measured using a seven-point *bipolar* scale, also mapped to a value from 0 to 100:

1. Completely dissatisfied: 0
2. Mostly dissatisfied: 16.7
3. Somewhat dissatisfied: 33.3
4. Neither satisfied nor dissatisfied: 50
5. Somewhat satisfied: 66.7
6. Mostly satisfied: 83.3
7. Completely satisfied: 100

DATA ANALYSIS PROCESS

We are most likely to see a need for a new solution when we identify user needs that are both important and underserved, measured by importance and satisfaction scores. We can see the relationship of importance and satisfaction by mapping the mean importance and satisfaction scores for each factor on a scatter plot, with importance on the `y-axis`, and satisfaction on the `x-axis` (***Figure 1***). On each axis, a neutral response is mapped to 50. Viewed in this way, the factors that map to the upper-left quadrant indicate opportunities for new solutions, as they are generally regarded by researchers as both important and underserved.

Figure 1 In quadrants defined by the relationship of importance and satisfaction the best opportunities for new solutions exist where there are both important and underserved needs.

Ethical considerations

We did not obtain approval from a research ethics committee as the research was considered to be low risk and we did not collect sensitive information about the participants. All data were collected anonymously. Participants were informed that their participation in this survey was completely voluntary, and that they were free to withdraw from the study at any time until they submitted their response. Answers will never be associated with individual participants and the results will only be analyzed in aggregate. The data collection procedures and survey tool were compliant with the General Data Protection Regulation 2016/679.

RESULTS

SURVEY PARTICIPANTS

The survey received 617 completed responses. The distribution of respondents by discipline and career stage is very similar when comparing the completed and whole cohort (completed and partial responses). Respondents self-identified their career stage in question 5 of the survey. Responses from the 617 individuals who have completed the survey —those who answered all questions in the survey— are used in our analysis (***Table 1***).

Table 1 Survey respondent demographics.

	TOTAL		COMPLETE	
	(n = 1477)		(n = 617)	
Career Stage				
Early-Career	471	31.9%	278	45.1%
Mid-Career	403	27.3%	223	36.1%
Late-Career	261	17.7%	112	18.2%
(blank)	342	23.2%	4	0.6%
Discipline				
Biology and Life Sciences	449	30.4%	247	40.0%
Earth Sciences	27	1.8%	13	2.1%
Ecology and Environmental Sciences	103	7.0%	56	9.1%
Engineering and Technology	31	2.1%	18	2.9%
Medicine and Health Sciences	308	20.9%	148	24.0%
Other – Please Specify	65	4.4%	39	6.3%
Physical Sciences	26	1.8%	16	2.6%
Social Sciences	144	9.7%	80	13.0%
(blank)	324	21.9%	0	0.0%
Location				
Europe	337	22.8%	132	21.4%
North America	934	63.2%	485	78.6%
Other – Please Specify	82	5.6%	0	0.0%
(blank)	124	8.4%	0	0.0%

Over half of the respondents were from Biology and Life Sciences or Medicine and Health Sciences disciplines. The cohorts from Physical Sciences, Engineering and Technology, Earth Sciences are small (with a maximum of 18 completed responses) (***Figure 2***). The largest proportion of survey respondents self-identified as Early Career researchers (45%), followed by Mid-Career researchers (36%) and Late-Career (18%). The majority of survey respondents were from North America (79%).

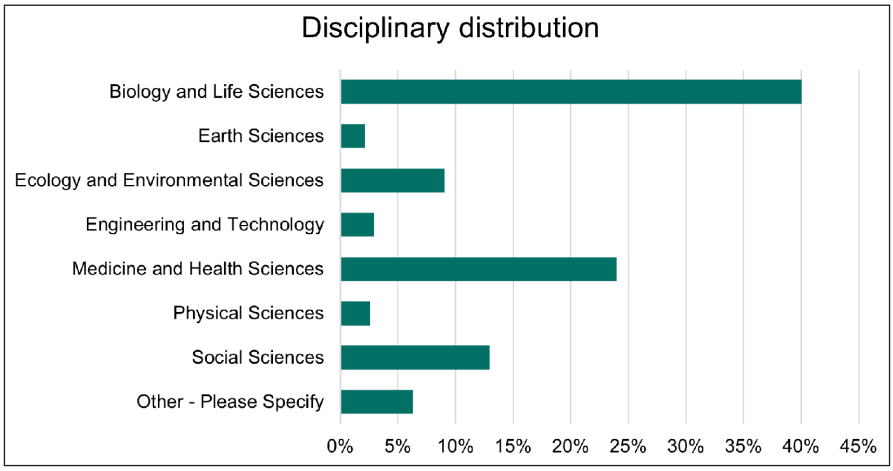

Figure 2 The most common discipline of respondents who completed the survey was Biology and Life Sciences, followed by Medicine and Health Science.

CURRENT DATA SHARING PRACTICES

Data sharing approaches

Respondents were asked to select all the methods of data sharing that they had previously used. Sharing data as supplemental files alongside a research paper was the most common method for all career levels (67%), followed by deposition in a public repository (59%) and sharing privately on request (49%). Only 10% of respondents reported that they had never shared their research data – the largest proportion of whom (42%) work in Medicine and Health Science disciplines. Sharing data privately, upon request was more common for more experienced researchers (***Figure 3***).

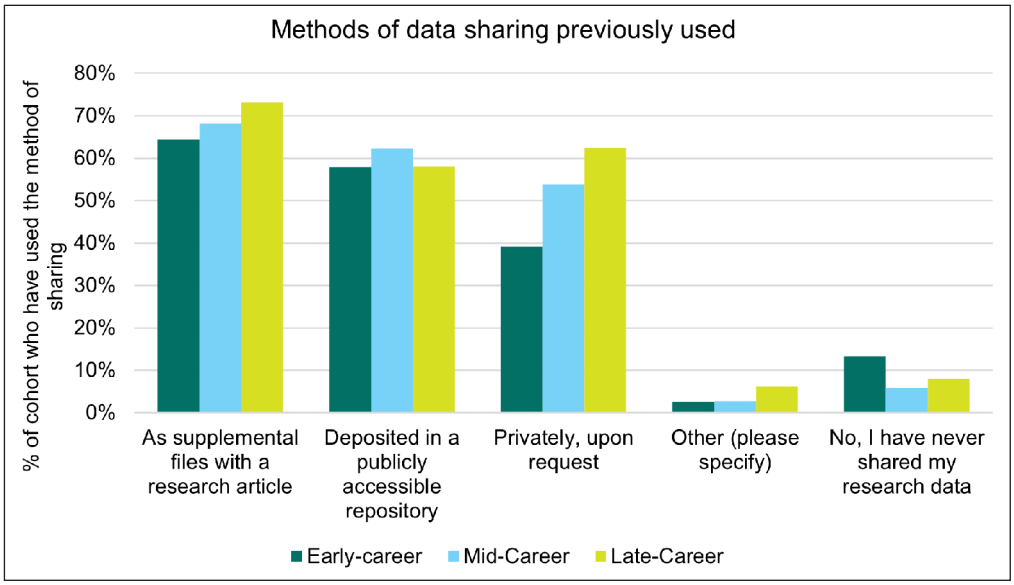

Figure 3 The most common method for sharing research data in the past is as supplemental files.

Prevalence of data reuse

Respondents were also asked if they have ever reused someone else's data. 52% responded 'yes' and 48% 'no'. These proportions are very similar when segmenting for career stage cohorts, with the yes/no split being 51%/49% for early-career, 52%/48% for mid-career and 53%/47% for late-career.

IMPORTANCE AND SATISFACTION SCORES OF DATA SHARING TASKS

Respondents were asked to rate the importance and their satisfaction for 36 factors related to sharing or reusing data. These answers have been turned into importance and satisfaction scores (see Methods section for details). A score of 0 indicates that researchers do not find the factor at all important or they are completely dissatisfied with their ability to carry out the task. A score of 100 indicates that they regard the factor as of the highest importance or they are completely satisfied with their ability to undertake the task. The mean importance scores ranged from 37.8 to 85.0 and the mean satisfaction scores ranged from 41.4 to 69.1. Tasks related to data sharing and reuse have been grouped according to the section of the research lifecycle that they primarily fall in (***Table 2***). The following groupings were used in our analysis: data preparation, policy requirements, data publishing, and data reuse.

Data preparation

This stage includes time preparing data for sharing, such as organising files, deciding which datasets to share, describing the data and preparing usage rights statements. Overall, these factors scored in the lower to middle range of importance and mid to high satisfaction when considering all of the factors presented in the survey.

Policy requirements

The factors concerning policy requirements include policies from funders, institutions and journals related to data sharing and data management. In terms of importance, four of the factors rank very highly (between 62.6 and 73.8), three of which are policy compliance factors

	n	IMPORTANCE			SATISFACTION		
		mean	± stdev	CI	mean	± stdev	CI
Data Preparation							
Spend less time organizing my data files	617	57.9	28.1	2.2	60.8	25.0	2.0
Spend less time deciding which datasets to share	617	37.8	31.1	2.5	65.7	24.0	1.9
Spend less time describing my research data	617	47.0	28.2	2.2	63.8	21.9	1.7
Prepare usage rights statement outlining conditions of use and acknowledgment	617	54.7	31.1	2.5	52.5	25.3	2.0
Policy Requirements							
Spend less time preparing Data Management Plan(s)	617	48.5	28.0	2.2	58.5	24.9	2.0
Comply with journal policies on data sharing	617	69.5	27.4	2.2	68.1	25.3	2.0
Comply with funder policies on data sharing	617	73.8	27.9	2.2	69.1	24.2	1.9
Comply with institutional policies on data sharing	617	67.1	30.1	2.4	68.5	24.9	2.0
Meet funder requirements for data management plans	617	62.6	29.8	2.4	64.7	23.2	1.8
Ensure funder knows my Data Management Plan has been followed	617	52.7	29.9	2.4	61.4	22.8	1.8
Data Publishing							
Get help determining which datasets I have permission to share	617	47.8	33.4	2.6	59.1	26.7	2.1
Spend less time finding a repository for my data	617	44.0	31.4	2.5	61.4	27.7	2.2
Ability to place an embargo on my data	617	44.0	34.2	2.7	60.1	24.5	1.9
Spend less time describing my supplemental files	617	43.4	29.8	2.4	60.0	23.0	1.8
Ability to upload my data along with my article	617	54.8	31.3	2.5	59.0	24.2	1.9
Spend less time creating a Data Availability Statement	617	44.7	28.9	2.3	55.2	22.7	1.8
Ability to create a Data Availability Statement that includes links to my research data files	617	51.8	29.9	2.4	53.6	23.1	1.8
Ability to create a Data Availability Statement that includes a description of each of my research data files	617	47.2	28.3	2.2	52.8	22.1	1.7
Spend less time uploading my data files	617	45.5	31.0	2.4	58.3	24.2	1.9
Choose an appropriate license for my data	617	54.4	31.6	2.5	51.8	25.7	2.0
Increase the discoverability of my research data	617	64.8	30.9	2.4	51.0	23.1	1.8
My research data has its own Digital Object Identifier (DOI)	617	59.2	33.4	2.6	58.8	26.6	2.1
Reuse of my data							
Understand who is using my data set	617	54.3	32.0	2.5	46.5	26.9	2.1

(Contd.)

Table 2 Mean scores, standard deviations, 95% confidence interval for the mean, and number of responses for each factor for both importance and satisfaction.

	n	IMPORTANCE			SATISFACTION		
		mean	± stdev	CI	mean	± stdev	CI
Ability to control who can use my data	617	39.1	35.1	2.8	53.8	26.2	2.1
Trust the researchers who request my data	617	58.3	33.4	2.6	55.3	24.2	1.9
Increase my co-authorship opportunities	617	53.5	33.7	2.7	53.4	24.3	1.9
Increase the likelihood that my research papers are cited	617	70.0	27.6	2.2	52.5	21.3	1.7
Increase the likelihood that my research benefits science	617	85.0	21.8	1.7	54.9	21.2	1.7
Ability to track downloads of my research data	617	51.7	30.2	2.4	49.5	23.3	1.8
Ability to track citations of my research data	617	65.4	28.2	2.2	54.2	25.5	2.0
Reuse of other researchers' data							
Spend less time searching for articles with reusable datasets	318	61.4	33.1	3.7	41.4	23.5	2.6
Determine how many other researchers are sharing data with their publications	318	45.3	33.4	3.7	42.7	22.0	2.4
Find articles with data that are available on request	318	47.8	32.3	3.6	45.2	23.6	2.6
Spend less time making individual requests for datasets	318	58.9	33.2	3.7	43.0	23.6	2.6
Determine how many papers in a journal have data that is publicly available	318	45.3	35.6	3.9	43.6	23.0	2.5
Determine which researchers are sharing data with their publications	263	50.2	34.0	4.1	44.7	23.9	2.9

and the other meeting funder requirements for data management plans. The other two factors in this group fall within the mid-range of scores for all factors surveyed. These factors scored towards the higher end of all the factors for mean satisfaction, with the three factors about compliance scoring the highest mean satisfaction scores when all factors are considered.

Data publishing

Factors around ensuring data are discoverable, citable and licensed correctly scored higher in importance than factors more related to the process of publishing data, e.g. spending less time uploading files. None of the factors scored towards the extremes of the range of importance scores for all factors. Respondents were generally satisfied with all factors when considered alongside all the factors surveyed.

Data reuse

Reuse of my data

Factors within this group were spread across a range of importance scores. Factors such as, "increase the likelihood that my research benefits science" and "increase the likelihood that my research papers are cited" are amongst the highest rated of all factors for importance (85.0 and 70.0 respectively). Conversely, "ability to control who can use my data" is one of the lowest scoring factors in terms of importance (39.1). The factors in the group have similar mean satisfaction scores compared to other groups, ranging from 46.5 to 55.3, with only two factors scoring less than 50. The lowest satisfaction score belongs to "understand who is using my data set" (mean = 46.5, 95% CI = 2.1). This factor also scores 54.3 for mean importance, making it important yet underserved.

Reuse of other researchers' data

Factors in this group scored in the lower and middle of all the importance scores, ranging from 45.3 to 61.4. The most important factors in this group are the two associated with spending less time finding and getting hold of others' data. All of these factors had low satisfaction scores compared to the rest of the survey, ranging from 41.4 to 45.2. For two of the factors – "Spend less time searching for articles with reusable datasets" and "Spend less time making individual requests for datasets" these were regarded as both important and not satisfied meaning that these are both significantly underserved factors, as the 95% confidence intervals for the importance scores are above 50 and the satisfaction scores are below 50 ("Spend less time searching for articles with reusable datasets" mean importance = 61.4, 95% CI = 3.7 and mean satisfaction = 41.3, 95% CI = 2.6; "Spend less time making individual requests for datasets" mean importance = 58.9, 95% CI = 3.7 and mean satisfaction = 43.0, 95% CI = 2.6).

CAREER STAGE AND DISCIPLINARY DIFFERENCES

Career stage

Early career researchers gave the highest importance scores to 22 out of the 36 factors surveyed when compared to mid- and late-career respondents. This difference was only statistically significant for 10 of these factors when a two-tailed t-test was used to compare the cohorts ($p < .05$). Late-career researchers gave the highest average scores for 6 factors and mid-career researchers gave the highest score for 7, but none of these differences were statistically significant when compared to the next-highest score. The mean score for one factor was the same for both early and mid career researchers who scored it higher than late career researchers. Differences in mean importance scores between the 3 career-based cohorts ranged from 1.7 to 25.5 between highest and lowest score for each factor.

Satisfaction scores were less variable by career stage. Although late-career researchers were on average more satisfied with their ability to complete the tasks, the scores between career stages were more similar in comparison to the importance scores, with the maximum difference being 11.1. Fewer statistically significant differences between the cohorts were seen with satisfaction scores using t-tests but again, no clear trends emerged.

One notable difference between the early and mid-career cohorts versus the late career cohort was that the factors "Spend less time making individual requests for datasets" and "Spend less time searching for articles with reusable datasets" both fell into the important and underserved segment for early and mid-career researchers when taking the 95% confidence intervals into account but did not for late career researchers (***Table 3***).

Discipline

One notable difference between the disciplines surveyed was seen in the answers provided by those who identified as researchers in physical sciences. This group scored more factors as both important and not satisfied than the other disciplinary groups. Researchers from the 'Social Science' group scored fewer factors as not satisfied than the other disciplinary cohorts. There was consensus across the disciplines that "Increase the likelihood that my research benefits science" was the most important factor. In all disciplines, factors relating to the reuse of other researchers' data received low satisfaction scores.

DISCUSSION AND CONCLUSION

LIMITATIONS

The survey focused on researchers from North America and Europe, a high proportion of whom have published with PLOS. This scope limitation was to help ensure a sufficiently large sample of researchers in certain regions to draw meaningful conclusions. Further, the survey was written only in English, and we assumed that this impacts response rates in some regions.

Some of the disciplinary samples (Earth Sciences, Engineering, and Physical Sciences) were too small to be considered representative of the corresponding research community. The high proportion of PLOS authors could impact our results, as their experience with data sharing requirements – PLOS's strong data availability policy relative to most other publishers – may differ from the non-PLOS cohort.

	n			MEAN IMPORTANCE			MEAN SATISFACTION		
	EARLY	MID	LATE	EARLY	MID	LATE	EARLY	MID	LATE
Data Preparation									
Spend less time organizing my data files	278	223	112	57.7	58.6	56.9	60.9	59.8	62.9
Spend less time deciding which datasets to share	278	223	112	38.8	38.8	33.3	65.3	64.4	68.7
Spend less time describing my research data	278	223	112	46.6	49.1	44.9	63.0	63.4	66.4
Prepare usage rights statement outlining conditions of use and acknowledgment	278	223	112	55.8	53.8	54.5	52.6	51.2	54.5
Policy Requirements									
Spend less time preparing Data Management Plan(s)	278	223	112	47.7	49.7	47.5	58.4	58.1	59.5
Comply with journal policies on data sharing	278	223	112	69.5	70.6	66.5	68.5	68.1	67.4
Comply with funder policies on data sharing	278	223	112	73.8	74.2	71.9	68.7	68.4	71.9
Comply with institutional policies on data sharing	278	223	112	**70.6**	64.9	61.8	69.1	66.7	71.0
Meet funder requirements for data management plans	278	223	112	63.4	61.8	60.7	63.8	65.3	65.3
Ensure funder knows my Data Management Plan has been followed	278	223	112	53.7	52.4	50.0	60.7	61.9	62.5
Data Publishing									
Get help determining which datasets I have permission to share	278	223	112	50.4	47.0	43.8	55.5	59.9	**66.5**
Spend less time finding a repository for my data	278	223	112	46.9	42.3	40.0	60.6	63.2	60.6
Ability to place an embargo on my data	278	223	112	39.7	48.9	45.5	59.8	60.3	60.3
Spend less time describing my supplemental files	278	223	112	42.4	43.9	44.9	60.7	59.7	59.5
Ability to upload my data along with my article	278	223	112	**59.5**	52.0	48.4	59.1	59.6	58.3
Spend less time creating a Data Availability Statement	278	223	112	45.6	42.9	45.8	56.1	54.6	54.0
Ability to create a Data Availability Statement that includes links to my research data files	278	223	112	**56.8**	49.7	43.1	54.4	52.7	53.9
Ability to create a Data Availability Statement that includes a description of each of my research data files	278	223	112	50.3	46.0	41.3	53.6	52.4	51.5
Spend less time uploading my data files	278	223	112	46.0	44.3	46.4	59.2	58.6	56.0
Choose an appropriate license for my data	278	223	112	57.6	54.1	47.3	51.9	51.2	53.0
Increase the discoverability of my research data	278	223	112	**70.3**	64.0	53.3	50.6	51.0	52.5

Table 3 Mean importance and satisfaction scores by career stage for each factor. The factor where the highest score is statistically significantly different based on t-tests ($p < .05$) to both of the other cohorts are marked in bold.

(Contd.)

	n			MEAN IMPORTANCE			MEAN SATISFACTION		
	EARLY	MID	LATE	EARLY	MID	LATE	EARLY	MID	LATE
My research data has its own Digital Object Identifier (DOI)	278	223	112	63.2	60.4	46.9	59.1	60.5	55.1
Reuse of my data									
Understand who is using my data set	278	223	112	53.5	54.1	56.5	47.3	45.7	46.1
Ability to control who can use my data	278	223	112	35.8	40.9	43.8	54.8	53.5	51.8
Trust the researchers who request my data	278	223	112	58.6	56.7	61.4	56.4	55.6	51.9
Increase my co-authorship opportunities	278	223	112	**63.0**	49.8	37.5	50.1	53.5	**60.9**
Increase the likelihood that my research papers are cited	278	223	112	**75.7**	69.1	58.0	50.4	**54.9**	52.7
Increase the likelihood that my research benefits science	278	223	112	**88.1**	83.4	81.0	53.4	56.3	56.6
Ability to track downloads of my research data	278	223	112	53.9	51.1	48.2	51.1	48.2	48.2
Ability to track citations of my research data	278	223	112	**69.0**	62.6	61.8	56.2	52.1	53.7
Reuse of other researchers' data									
Spend less time searching for articles with reusable datasets	141	115	59	**67.9**	59.6	50.8	38.4	43.5	44.9
Determine how many other researchers are sharing data with their publications	141	115	59	48.6	47.2	34.7	41.5	43.0	44.3
Find articles with data that are available on request	141	115	59	51.2	48.3	39.4	43.5	47.4	44.4
Spend less time making individual requests for datasets	141	115	59	**66.7**	58.0	42.8	40.1	43.6	48.3
Determine how many papers in a journal have data that is publicly available	141	115	59	50.9	44.8	35.2	42.7	42.9	46.6
Determine which researchers are sharing data with their publications	117	98	46	52.8	53.1	39.7	43.7	44.4	47.8

The factors that the survey asked about do not cover all aspects of data sharing, as they were derived in part from our assumptions about which tasks researchers might find problematic and where there might be opportunities for new solutions. They were also intended to be discipline-agnostic. For example, issues specific to certain types of data, such as sensitive data, are not included. There may be important and underserved needs around data sharing that were not tested in our survey. For example, we did not include factors relating to technical problems or the quality of the data that is being shared.

IMPACT OF CAREER STAGE

There is a general tendency from Early- to Mid- to Late-Career researchers, and researchers who have published more articles, of declining importance scores, although the differences are mostly not statistically significant ($p < 0.5$). More experienced researchers and authors rate the importance of the majority of factors lower on average, although there are fewer differences with levels of satisfaction with existing tools, as satisfaction numbers remain more stable across these segments. This suggests that ECRs regard these factors as more important, as opposed

to having not yet mastered the tools needed to effectively share and reuse data. However, the factors with the greatest mean importance scores when comparing early and late career researchers can be considered those that are more likely to be relevant to junior researchers, for example "Increase my co-authorship opportunities" and "Spend less time making individual requests for datasets" both of which have p value < 0.001.

COMPARISON WITH PREVIOUS RESEARCH

Multiple surveys have quantified how common researchers' problems or concerns are with data sharing (Allagnat et al. 2019; Borghi et al. 2018; Lucraft et al. 2019; Science et al. 2017; 2018; 2019; Tenopir et al. 2011; Wiley Open Science Researcher Survey 2016). Our findings suggest that while many factors (problems) associated with sharing research data are important to researchers, on average, researchers are reasonably satisfied with their ability to share data, from their perspective. Overall our findings are additive to previous research, providing additional context as to why solutions such as data repositories are still used by a minority of researchers, despite data repositories, ostensibly, being available for most types of research data. If researchers are generally satisfied with their ability to complete a task associated with data sharing, this suggests that researchers will be unlikely to be motivated to seek (new) solutions to that problem, no matter how common it is.

For example, our finding that the 'ability to control who can use my dataset' was slightly important (39.1 importance) extends previous findings that researchers' concerns relating to misuse of their data is very common (Science et al. 2019). If this concern is common yet not very important to the average researcher, then it may be viewed as "low stakes", and not a motivator for action. The ability to 'trust the researchers who request my data' may also relate to potential for misuse of researchers' data but was rated as moderately important, not viewed as "low stakes", (58.3) as was 'choose an appropriate license for my data' (54.4). Both of these factors, associated with reuse of researchers' own data, were somewhat satisfied, however.

The highest average score for importance was found for 'Increase the likelihood that my research benefits science' (85.0 importance; 54.9 satisfaction). This supports previous research, where increasing the benefit to science or society of research is commonly amongst the top reasons or motivations for sharing research data (Science et al. 2019). The second most important factor overall related to compliance with funder policies on data sharing (73.8 importance; 69.1 satisfaction), ranking it more highly than in previous research exploring factors that motivate data sharing (Science et al. 2019). The third most important factor related to increasing the likelihood that researchers' papers are cited (70.0 importance; 52.5 satisfaction). This reputational factor, citations, is consistent with increased impact of research and desire for greater credit (recognition) for data sharing found by previous research. It may also offer opportunities for promotion of the potential benefits of sharing research data by tool and service providers. Sharing research data is associated with increased citations to researchers' papers (Colavizza et al. 2020; Piwowar et al. 2013).

POTENTIAL OPPORTUNITIES RELATING TO DATA REUSE

While most of the factors we assessed appear to be reasonably well satisfied from the researchers' perspective, a small number of factors suggest potential opportunities for new or better solutions. Around half of survey respondents indicated that they have reused research data in the past, consistent with findings from other surveys (Science et al. 2018; 2019). These factors feature in the upper left quadrant (***Figure 4***) — albeit moderately so — and relate to reuse of other researchers' data:

- Spend less time searching for articles with reusable datasets (61.4 importance; 41.4 satisfaction)
- Spend less time making individual requests for datasets (58.9 importance; 43.0 satisfaction)

Both these factors are relevant to scholarly publishers, who can influence the accessibility and availability of research data associated with publications — with research data policies (Vines et al. 2013) and associated workflows. Making research data available in repositories that enable compliance with the FAIR Data principles, and creating prominent and visible links to those data in journal articles, might be a simple solution to the first factor.

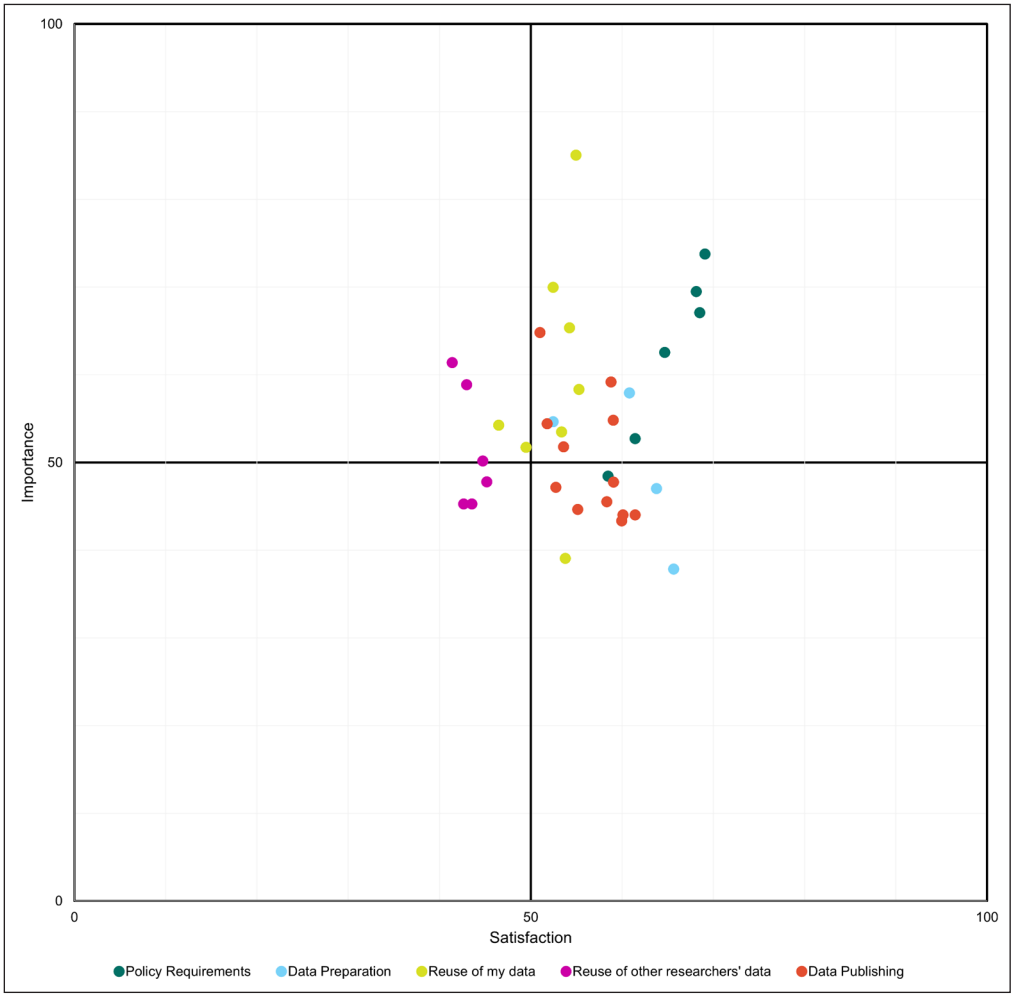

Figure 4 Respondents were on average satisfied with their ability to complete the majority of tasks associated with Data Preparation, Data Publishing and Reuse of their own data but dissatisfied with their ability to complete tasks associated with Reuse of other researchers' data. Tasks associated with meeting policy requirements are important and satisfied. 95% Confidence intervals for the mean values ranged from 1.7 to 4.1 for importance scores and 1.7 to 2.9 for satisfaction scores.

Researchers' dissatisfaction with obtaining research data from individual requests to other researchers has policy implications for journals and publishers who wish to further support open science and open research. While many journals now have policies on sharing research data, and many peer-reviewed papers include statements about the availability of data supporting publications, many of those statements state that data are "available on [reasonable] request". Multiple studies (Rowhani-Farid et al. 2016; Savage & Vickers 2009; Vanpaemel et al. 2015; Wicherts et al. 2006) have found that researchers have been unable to obtain data supporting publications when those data are 'available on request', consistent with the dissatisfaction amongst our survey respondents. Since PLOS introduced its data availability policy in 2014, "data available on request" has not been permitted when publishing in PLOS journals. Publishers can potentially meet this data reuse need by strengthening their policies on data sharing – requiring all data supporting publications to be publicly available unless legal or ethical restrictions apply, and working to eliminate "data available on request" as an acceptable policy. And, in such cases where data must be available under restricted access, requiring information on conditions and procedures for data access and reuse.

OPPORTUNITIES TO BETTER SUPPORT DATA SHARING

The relative unimportance of some factors associated with best data publishing practice, such as deposition of data in repositories, suggests the need for more advocacy to researchers and education of the benefits, or for data repositories to be more integrated with the traditional publishing experience in such a way that researchers do not need to change their behaviour in order to use them. Amongst PLOS authors and the survey respondents, the most common method for sharing data is via supplemental (supporting information) files with their publications. More than half our survey respondents indicated they had shared data in a repository in the past but when published articles are analysed, data repositories are used by around a quarter of authors publishing with PLOS. At PLOS this proportion has been slowly growing each year, from 18% of authors in 2015 (Colavizza et al. 2020).

While part of our motivation for this research was to explore opportunities for new products or services to support researchers in sharing, discovering and managing research data, the results imply that, amongst the PLOS author community in particular, researcher needs and better support for FAIR data can likely be met by working with existing solutions. This includes tactics such as more closely partnering with established data repositories and improving the linking of research data and publications, as well as maintaining, or enhancing where appropriate, stringent journal data sharing policies.

DATA ACCESSIBILITY STATEMENT

Both the survey questions used and the resulting anonymised data are available from figshare at *https://doi.org/10.6084/m9.figshare.13858763* (Harney et al. 2021).

ACKNOWLEDGEMENTS

The authors thank Samira Vijghen for support in developing and deploying the survey instrument and contributions to data analysis. The authors also thank Dan Morgan at PLOS for his comments on the draft of this manuscript.

COMPETING INTERESTS

All authors are employees of PLOS.

AUTHOR CONTRIBUTIONS

Study conception: IH, JH

Survey design: IH, JH

Data analysis: JH, LC

Manuscript preparation and approval for submission: IH, LC, JH

Dataset curation for public release: LC

AUTHOR AFFILIATIONS

Iain Hrynaszkiewicz *orcid.org/0000-0002-9673-5559*
PLOS, Carlyle House, Carlyle Road, Cambridge CB4 3DN, UK

James Harney *orcid.org/0000-0003-4339-5105*
PLOS, 1265 Battery Street, Suite 200, San Francisco, CA 94111, US

Lauren Cadwallader *orcid.org/0000-0002-7571-3502*
PLOS, Carlyle House, Carlyle Road, Cambridge CB4 3DN, UK

REFERENCES

Allagnat, L, Allin, K, Baynes, G, Hrynaszkiewicz, I and **Lucraft, M.** 2019. Challenges and Opportunities for Data Sharing in Japan. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.7999451.v1*

Borghi, JA and **Van Gulick, AE.** 2018. Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLoS ONE*, 13: e0200562. DOI: *https://doi.org/10.1371/journal.pone.0200562*

Christensen, CM, Hall, T, Dillon, K and **Duncan, DS.** 2016. *Competing Against Luck: The Story Of Innovation And Customer Choice*.

Colavizza, G, Hrynaszkiewicz, I, Staden, I, Whitaker, K and **McGillivray, B.** 2020. The citation advantage of linking publications to research data. *PLoS ONE*, 15(4): e0230416. DOI: *https://doi.org/10.1371/journal.pone.0230416*

Eynden, VVD, Knight, G, Vlad, A, Radler, B, Tenopir, C, Leon, D, et al. 2016. Survey of Wellcome researchers and their attitudes to open research. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.4055448.v1*

Federer, LM, Belter, CW, Joubert, DJ, Livinski, A, Lu, Y-L, Snyders, LN, et al. 2018. Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLoS ONE*, 13: e0194768. DOI: *https://doi.org/10.1371/journal.pone.0194768*

Harney, J, Hrynaszkiewicz, I and **Cadwallader, L.** 2021. Data from: A survey of researchers' needs and priorities for data sharing. *figshare*. DOI: *https://doi.org/10.31219/osf.io/njr5u*

Houtkoop, BL, Chambers, C, Macleod, M, Bishop, DVM, Nichols, TE and **Wagenmakers, E-J.** 2018. Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1: 251524591775188. DOI: *https://doi.org/10.1177/2515245917751886*

Hrynaszkiewicz, I, Simons, N, Hussain, A, Grant, R and **Goudie, S.** 2020. Developing a research data policy framework for all journals and publishers. *Data Sci J*, 19: 5. DOI: *https://doi.org/10.5334/dsj-2020-005*

Kratz, JE and **Strasser, C.** 2015. Researcher perspectives on publication and peer review of data. *PLoS ONE*, 10: e0117619. DOI: *https://doi.org/10.1371/journal.pone.0117619*

Lucraft, M, Baynes, G, Allin, K, Hrynaszkiewicz, I and **Khodiyar, V.** 2019. Five Essential Factors for Data Sharing. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.7807949.v2*

Open Data: the researcher perspective – survey and case studies [Internet]. 4 Apr 2017 [cited 15 Nov 2018]. Available: *https://data.mendeley.com/datasets/bwrnfb4bvh/1.*

Perrier, L, Blondal, E and **MacDonald, H.** 2020. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PLoS ONE*, 15: e0229182. DOI: *https://doi.org/10.1371/journal.pone.0229182*

Piwowar, HA and **Vision, TJ.** 2013. Data reuse and the open data citation advantage. *PeerJ*, 1: e175. DOI: *https://doi.org/10.7717/peerj.175*

Rathi, V, Dzara, K, Gross, CP, Hrynaszkiewicz, I, Joffe, S, Krumholz, HM, et al. 2012. Clinical trial data sharing among trialists: a cross-sectional survey. *BMJ*, 345: e7570. DOI: *https://doi.org/10.1136/bmj.e7570*

Rowhani-Farid, A and **Barnett, AG.** 2016. Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open*, 6: e011784. DOI: *https://doi.org/10.1136/bmjopen-2016-011784*

Savage, CJ and **Vickers, AJ.** 2009. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, 4: e7078. DOI: *https://doi.org/10.1371/journal.pone.0007078*

Schmidt, B, Gemeinholzer, B and **Treloar, A.** 2016. Open data in global environmental research: the belmont forum's open data survey. *PLoS ONE*, 11: e0146695. DOI: *https://doi.org/10.1371/journal.pone.0146695*

Science, D, Fane, B, Ayris, P, Hahnel, M, Hrynaszkiewicz, I, Baynes, G, et al. 2019. The State of Open Data Report 2019. *Digital Science*. DOI: *https://doi.org/10.6084/m9.figshare.9980783.v1*

Science, D, Hahnel, M, Fane, B, Treadway, J, Baynes, G, Wilkinson, R, et al. 2018. The State of Open Data Report 2018.

Science, D, Hahnel, M, Treadway, J, Fane, B, Kiley, R, Peters, D, et al. 2017. The State of Open Data Report 2017.

Stuart, D, Baynes, G, Hrynaszkiewicz, I, Allin, K, Penny, D, Lucraft, M, et al. 2018. Whitepaper: Practical challenges for researchers in data sharing [Internet]. Available: *https://figshare.com/articles/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011.*

Tenopir, C, Allard, S, Douglass, K, Aydinoglu, AU, Wu, L, Read, E, et al. 2011. Data sharing by scientists: practices and perceptions. *PLoS ONE*, 6: e21101. DOI: *https://doi.org/10.1371/journal.pone.0021101*

Tenopir, C, Christian, L, Allard, S and **Borycz, J.** 2018. Research data sharing: practices and attitudes of geophysicists. *Earth and Space Science*, 5: 891–902. DOI: *https://doi.org/10.1029/2018EA000461*

Tenopir, C, Dalton, ED, Allard, S, Frame, M, Pjesivac, I, Birch, B, et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE*, 10: e0134826. DOI: *https://doi.org/10.1371/journal.pone.0134826*

Tenopir, C, Rice, NM, Allard, S, Baird, L, Borycz, J, Christian, L, et al. 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS ONE*, 15: e0229003. DOI: *https://doi.org/10.1371/journal.pone.0229003*

Vanpaemel, W, Vermorgen, M, Deriemaecker, L and **Storms, G.** 2015. Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra*, 1. DOI: *https://doi.org/10.1525/collabra.13*

Vines, TH, Andrew, RL, Bock, DG, Franklin, MT, Gilbert, KJ, Kane, NC, et al. 2013. Mandated data archiving greatly improves access to research data. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*. fj.12-218164-. Available: *http://www.fasebj.org/content/early/2013/01/07/fj.12-218164.*

Wicherts, JM, Borsboom, D, Kats, J and **Molenaar, D.** 2006. The poor availability of psychological research data for reanalysis. *Am Psychol*, 61: 726–728. DOI: *https://doi.org/10.1037/0003-066X.61.7.726*

Wiley Open Science Researcher Survey 2016 [Internet]. [cited 15 Nov 2018]. Available: *https://figshare.com/articles/Wiley_Open_Science_Researcher_Survey_2016/4748332/2.*

Wilkinson, MD, Dumontier, M, Aalbersberg, IJJ, Appleton, G, Axton, M, Baak, A, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3: 160018. DOI: *https://doi.org/10.1038/sdata.2016.18*

TO CITE THIS ARTICLE:
Hrynaszkiewicz, I, Harney, J and Cadwallader, L. 2021. A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, 20: 31, pp. 1–16. DOI: *https://doi.org/10.5334/dsj-2021-031*

Submitted: 25 February 2021
Accepted: 22 September 2021
Published: 13 October 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.