

# Data journals: incentivizing data access and documentation within the scholarly communication system

Data journals provide strong incentives for data creators to verify, document and disseminate their data. They also bring data access and documentation into the mainstream of scholarly communication, rewarding data creators through existing mechanisms of peer-reviewed publication and citation tracking. These same advantages are not generally associated with data repositories, or with conventional journals' data-sharing mandates. This article describes the unique advantages of data journals. It also examines the data journal landscape, presenting the characteristics of 13 data journals in the fields of biology, environmental science, chemistry, medicine and health sciences. These journals vary considerably in size, scope, publisher characteristics, length of data reports, data hosting policies, time from submission to first decision, article processing charges, bibliographic index coverage and citation impact. They are similar, however, in their peer review criteria, their open access license terms and the characteristics of their editorial boards.

## Keywords

data journals; data repositories; incentives; open access; open data

## Introduction

The benefits of free, unmediated access to research data are widely acknowledged, especially in the life sciences. Despite mandates from both funding agencies and publishers, however, open data initiatives have been only partially successful. Previous research suggests that this can be attributed to a lack of incentives for data creators, who are often expected to expend considerable effort without receiving meaningful rewards for their work. Data creators who have documented their procedures in detail, made their data user-friendly, and met data archives' strict submission requirements, will often receive nothing more than an acknowledgment, which counts for little within the framework of research funding, promotion and tenure.

The first section of this paper, 'Incentivizing data access', shows how data journals – those that publish data reports rather than conventional articles – provide strong incentives for data creators to thoroughly verify, document, review and disseminate their work (i.e. to document and publish their data in accordance with open data principles). Unlike conventional journals with data-sharing mandates, data journals reward authors who share their data. Unlike data archives, data journals bring access and documentation into the mainstream of scholarly communication through conventional practices such as authorship, publication and citation.

The second section of the paper, 'Characteristics of data journals', updates an earlier study by Leonardo Candela and associates,<sup>1</sup> providing current information about the data journal landscape: the characteristics and policies of data journals in biology, environmental science, chemistry, medicine, and health sciences. The results may be useful to librarians, to authors in STEM disciplines and to researchers in areas such as scholarly communication, science studies and information science.



WILLIAM H WALTERS

Executive Director  
Mary Alice & Tom  
O'Malley Library  
Manhattan College

'open data initiatives  
have been only  
partially successful'

- 2 The final section, 'Data journals: potential and continuing challenges', summarizes the advantages of data journals – advantages that may grow or diminish with changes in the scholarly communication system. It also describes three continuing difficulties: the need for sustainable data management practices, the fact that the incentives provided by data journals may not always offset the advantages of keeping data private, and the exclusionary effect of open access (OA) article processing charges (APCs).

## Incentivizing data access

The authors of journal articles do not usually engage in *unmediated* data sharing (i.e. sharing through methods that do not require the requester to contact the author or data provider) unless mandates or other mechanisms require or encourage it. Only 37% of natural and social scientists have ever shared their data through a public repository or archive, and just 25% have shared their data through a journal's website.<sup>2</sup> Moreover, *mediated* requests for data, such as e-mail messages sent to authors, are unsuccessful at least one-third of the time.<sup>3</sup> The success rate for mediated requests may be higher in particular fields, however.<sup>4</sup>

'authors of journal articles do not usually engage in *unmediated* data sharing'

## Open data mandates

At least 24 major funding agencies in the life sciences have established policies that require or promote data sharing.<sup>5</sup> Nearly a quarter of health and life science journals have data-sharing policies of one kind or another, and many prominent journals have adopted policies that require authors to provide unmediated access to the data used in their analyses.<sup>6</sup> For example, the Public Library of Science (PLOS) requires authors to make their data publicly available when the paper is submitted.<sup>7</sup>

There are two problems with these mandates, however. First, many researchers fail to comply with data-sharing mandates, even when the submission of data is a nominal requirement. In the BioMed Central (BMC) journals, which require data sharing, full data are immediately available for just 31% of the papers. Sixty per cent include a notation that data are available on request, while 9% appear not to comply with the requirement in any way.<sup>8</sup> Although strict standards for the protection of human subjects may partially account for non-compliance with data-sharing requirements, it is not clear why an author would submit to a journal with a data-sharing mandate while knowing that his or her data could not be shared. Higher rates of unmediated data sharing, from 78% to 86%, have been reported for life science journals other than the BMC journals, but more than a third of the data files are incomplete; they do not provide enough details for replication.<sup>9</sup> In the field of metabolomics, there is only a weak relationship between journals' data-sharing policies and the extent to which data are actually made available.<sup>10</sup> Second, and perhaps more significantly, data-sharing mandates provide no real incentives for compliance, since the reward for publishing in a journal that requires data sharing is no greater than the reward for publishing in a journal that does *not* require data sharing. Given the extensive time and effort that data sharing requires, mandates may even encourage adaptive strategies that are not in the best interests of the scholarly community. For instance, data-sharing requirements at individual journals may encourage authors to send their work to other journals of similar quality and scope that do *not* require data sharing, since sharing involves a cost without a commensurate benefit.<sup>11</sup> Likewise, authors may be less likely to generate or compile original data if the relative benefits of that work (to them) are reduced through mandatory data disclosure. That is, effective data-sharing mandates at conventional journals might increase data availability in the short term, but decrease it in the long term once authors adjust their behavior in accordance with the more widespread availability of data – once they fully understand that it is more cost-effective to use available data than to generate their own. Authors may also postpone their article submissions in order to more fully exploit their data files before handing them over to others, thereby delaying the reporting of potentially important findings.<sup>12</sup>

'data-sharing mandates provide no real incentives for compliance'

3 It is important to realize that even the widespread enforcement of data-sharing mandates is unlikely to change the underlying system of incentives. We can expect data sharing to become more common system-wide only if scholars, promotion committees and funding agencies assign greater credit to the production, documentation and dissemination of data.

### The importance of incentives

Surveys and discussions with researchers in the life sciences suggest several reasons why authors may be reluctant to make their data available on the open web.<sup>13</sup> (See Table 1.) When asked about obstacles to the free, online dissemination of data, survey respondents mentioned the need to keep data private during ongoing projects, inadequate credit for those who create and share data, legal concerns, the possible misinterpretation or misuse of publicly available data and the potential loss of control over valuable intellectual property. Researchers recognize the benefits of open data initiatives, but they also want full credit for their contributions.

‘Researchers recognize the benefits of open data initiatives, but they also want full credit for their contributions’

#### Poor research practices

---

- Absence of a culture of data sharing in the academic field
- Potential for discovery of errors in the data creator’s published analyses
- Inadequate documentation of data-related procedures
- Failure to save and safeguard data, metadata or statistical code
- Loss of data or interpretive expertise due to the retirement or migration of personnel

#### Limited data storage and dissemination mechanisms

---

- Relatively few journals or other outlets devoted to data publication
- Lack of technical expertise in data publishing
- Hardware or software problems
- Obsolete devices and file formats

#### Limited awareness of open data principles

---

- Concern that public disclosure of data will violate legal or ethical norms
- Difficulty dealing with open access licensing terms (e.g. Creative Commons licenses)

#### Ongoing research

---

- Desire to keep data private until the research project is completed

#### Expenditure of effort

---

- Considerable effort required to produce documentation that is unlikely to be needed by the data creator in his or her own research
- Awareness that the expenditure of effort needed to comply with a data request, even if minimal, could be otherwise devoted to activities that bring greater rewards

#### Inadequate credit for data-sharing activities

---

- Reluctance to share valuable data due to a general sense of ownership
- Absence of universal mechanisms, such as authorship and citation, by which data creators can be recognized and rewarded
- Concern that the costs of data dissemination are considerably greater than the individual rewards – that the sharing of data without compensation, and the use of data without credit, are inherently unfair
- Concern that commercial firms will use the data inappropriately or without compensation

#### Potential for misuse of data

---

- Fear that data dissemination will facilitate plagiarism
- Concern that users without an understanding of the data will draw unwarranted or misleading conclusions

Table 1. Reasons authors may be reluctant to share their data

4 Incentives are especially important to scholars working on long-term research projects. James A. Mills et al. surveyed 73 ecologists serving as principal investigators on projects of five to 68 years' duration, reporting that two thirds were unhappy with mandates for public data access and that only 8% were in favor of making their data freely accessible online.<sup>14</sup> Nearly 55% stated that they would avoid publishing in journals that required them to share their data on the open web. Several noted that the data collected for a long-term project may provide the foundation for an entire career of scientific activity. Because no single grant provides funding for the entire duration of a project, the principal investigators must constantly identify new sources of short-term funding for interim projects that make use of the unique data they have collected. Widespread data dissemination 'could lead to a loss of funding opportunities if data for their next project are routinely mined by other researchers'.<sup>15</sup> Consequently, data sharing may reduce the number of long-term projects by decreasing the incentives for undertaking them.

Individual motives and incentives are not mentioned in most open data policies, which focus almost exclusively on the broader societal benefits of data sharing. For instance, the FAIR Data Principles do not address authorship, citation, or credit for the creation or maintenance of data files.<sup>16</sup> After all, the main beneficiaries of data-sharing policies are data *users* rather than producers. 'Researchers' incentives to release their own data may or may not align with their motivations to gain access to the data of others', especially when data submission and data archiving require a considerable investment of time and effort.<sup>17</sup>

### Data reports and data journals

Data journals are those that publish *data reports* on a regular basis, either exclusively or as a primary article type. Each data report describes the data that underlie an empirical paper or a broader research project. Data reports often include greater methodological detail than would normally be found in a research paper – information on the procedures used to generate or compile the data, the population of interest, the sampling methods, the variable names and response codes, difficulties encountered, decisions made, user notes and suggestions for further use of the data. Nearly all data reports are peer-reviewed. Although some present simple descriptive statistics or frequency tables, they do not normally include cross-tabulations, multivariate analyses, or other attempts to describe the relationships among the variables. For any empirical study that draws on the data, three elements – the study itself, the data report and the data file – should provide all the information needed to replicate the analysis.<sup>18</sup>

Although the *Journal of Chemical and Engineering Data* first appeared in 1956, most data journals have been founded within the past five or ten years. Most conform to OA principles and are therefore *open* data journals.

As noted earlier, data-sharing mandates and data repositories seldom provide substantial incentives for authors to make their data openly available. In contrast, data journals do so by adopting a universally accepted mechanism of quality control (peer review), providing authorship credit for journal articles (data reports), and facilitating the indexing and citing of those reports. Perhaps most importantly, they provide both authorship and citation credit, since data reports are peer-reviewed articles that can be readily cited and recognized for their scholarly impact in terms that are widely understood by promotion and tenure committees.<sup>19</sup> Moreover, data journals bring data publishing into the mainstream of scholarly communication, since data reports are authored, published, indexed, cited and used in much the same way as conventional journal articles. After all, initiatives to encourage or require data dissemination are unlikely to be as effective as the systems and incentives that underlie data *publishing*.<sup>20</sup> Through publication, data reports emerge as first class research products that are fully integrated into the scholarly communication system through processes such as authorship, validation (e.g. peer review), dissemination, preservation and citation. Moreover, they can provide valuable information that is often not fully presented within either data sets or conventional research papers.<sup>21</sup>

'Nearly all data reports are peer-reviewed'

'data journals bring data publishing into the mainstream of scholarly communication'

'data reports emerge as first class research products'

- 5 Although the number of data journals has been increasing over the past few years, many researchers are still unaware of them. In a recent survey, only 16% of researchers in the natural and social sciences were able to name one or more data journals. Nonetheless, many respondents were intrigued by the prospect of authoring data reports, and one wrote 'I've never heard of this, but it sounds fantastic'.<sup>22</sup> A similar 2009 survey found that only 9% of meteorologists had heard of *Earth System Science Data*, but 69% said they would use a data journal to find data relevant to their work. Likewise, 67% reported that the prospect of getting authorship credit for their contributions would make them more likely to publish their data.<sup>23</sup>

## Characteristics of data journals

This section examines the characteristics of data journals in the fields of biology, environmental science, chemistry, medicine and health sciences, presenting detailed information on each of the 13 data journals that regularly publish in those areas.

### Methods

I used Google to identify an initial set of journals, following much the same procedure as used by Candela and associates in 2015.<sup>24</sup> Specifically, I searched for the phrases 'data journals', 'open data journals' and 'open access data journals'. I also searched for the names of well-known data journals, since the websites that mention one data journal often mention others as well. Finally, I added all the data journals identified by Candela et al., which resulted in a list of 169 journals.

Further review of the journals' websites revealed that only 19 of the 169 journals (11%) currently publish data reports on a regular basis. Specifically, the 169 journals include:

- 19 'pure' data journals, for which data reports comprise at least half the papers in the journal (Group 1)
- 109 journals that publish data reports but are devoted mainly to other types of contributions (Group 2). On average, data reports account for just 1.6% of the papers in these journals, and there are just three journals for which data reports comprise more than 8% of the published items
- 21 journals that do not actually include data reports as a publication type (Group 3). Some may have been inadvertently described as data journals due to their coverage of data science topics or their strict requirements for dissemination of the data used in empirical papers
- 20 journals that are no longer published, or that no longer publish data reports (Group 4). Very few of these journals were devoted mainly to data reports.

The Appendix lists the journals in each group. The fact that only 19 of the journals are pure data journals (Group 1) is consistent with the findings of Candela et al., who reported that only seven of their 116 journals (6%) were devoted solely to data reports.<sup>25</sup>

'No online list of data journals is comprehensive'

No online list of data journals is comprehensive. Moreover, most of the online lists include journals that are *not* data journals in any real sense. Some journals claim to accept data reports but have never published any. Others publish articles *about* data science rather than data reports. (See the Appendix.)

The variables (journal characteristics) examined in this study are more extensive than those presented by Candela et al. (See Table 2.) However, information is provided only for the 13 Group 1 data journals that publish in the fields of medicine, health, biology, or chemistry. Information was compiled mainly from the websites of publishers and journals, although other sources were consulted.<sup>26</sup>

## General information

---

URL  
 Open access?  
 Year founded  
 Items published, July 2018 through June 2019  
 Percentage of published items that are data reports  
 Subject scope  
 Publisher  
 Publisher information  
 General note

## Characteristics of data reports

---

Term for data reports  
 Typical length of data reports  
 Required or recommended sections of data reports  
 Original or secondary data?  
 Data files hosted on journal's platform or elsewhere?  
 % of data files included in text of report  
 % of data files on journal's website as supplementary files  
 % of data files in external data repository  
 % of data files not found  
 Data hosting note

## Editors and peer review

---

Editor(s) in chief  
 Editorial board  
 Review process  
 Time from submission to first decision  
 Time from acceptance to publication  
 Acceptance rate

## Licenses and article processing charges

---

Creative Commons license(s) for OA data reports  
 Article processing charge (APC) for OA data reports  
 Waivers or reductions of the APC?

## Indexing and citation impact

---

Indexed in BIOSIS?  
 Indexed in PubMed?  
 Indexed in Science Citation Index (SCI)?  
 Indexed in Scopus?  
 SCI Impact Factor percentile  
 Scopus CiteScore percentile

Table 2. Variables for which information was compiled

The concentration of data journals in the life sciences is nothing new. The current percentage – 68%, or 13 of the 19 Group 1 journals – is consistent with the 77% value reported by Candela et al.<sup>27</sup> The number of multidisciplinary data journals appears to have increased over time, however. Candela and associates identified just one multidisciplinary data journal, but this investigation includes three: *Data in Brief*, *Scientific Data* and *Data*.

## Results and discussion

The results are presented in five subsections that correspond to the headings shown in Table 2. For the full results and associated notes, see Supplementary Table 1. (Details may be found in the data accessibility statement at the end of this article.)

### General information

Table 3 shows general information on the 13 data journals included in the investigation. Three of the 13 journals, including the two oldest, are not actually *open* data journals, since they require a subscription for access. The other ten journals, all OA, are relatively new, with founding dates no earlier than 2013.

Data journal	OA?	Founded	Items <sup>a</sup>	Data reports <sup>b</sup>	Subject scope	Publisher
<i>Data in Brief</i>	Yes	2014	1,520	100%	All subjects	Elsevier
<i>Scientific Data</i>	Yes	2014	274	90%	Natural sciences	Springer Nature
<i>IUCrData</i>	Yes	2016	181	100%	Crystallography & related fields	International Union of Crystallography
<i>Data</i>	Yes	2016	145	50%	Natural scis., some social scis.	Multidisciplinary Digital Publishing Institute
<i>Earth System Science Data (ESSD)</i>	Yes	2009	130	55%	Earth system sciences	Copernicus Publications
<i>Biodiversity Data Journal</i>	Yes	2013	80	74%	Biodiversity science	Pensoft
<i>Geoscience Data Journal</i>	Yes	2013	18	87%	Geosciences	Wiley, Royal Meteorological Society
<i>Journal of Open Psychology Data</i>	Yes	2013	4	97%	Psychology	Ubiquity Press
<i>Open Data Journal for Agricultural Research</i>	Yes	2016	2	100%	Agriculture & food (in) security	Several universities and research foundations
<i>Open Health Data</i>	Yes	2013	2	100%	Health & medicine	Ubiquity Press
<i>Journal of Chemical &amp; Engineering Data</i>	No	1956	569	97%	Materials science	American Chemical Society
<i>Chemical Data Collections</i>	No	2016	155	100%	Chemistry	Elsevier
<i>Journal of Physical &amp; Chemical Reference Data</i>	No	1972	14	99%	Physical sciences	American Institute of Physics, with NIST

Table 3. Data journals included in the investigation

a. Items published, July 2018 through June 2019. Includes data reports, conventional research articles, and other items such as editorials.

b. Percentage of published items that are data reports.

The three multidisciplinary journals – *Data*, *Data in Brief* and *Scientific Data* – are larger than most of the others. In fact, *Data in Brief* publishes about as many data reports as the other 12 journals combined. In contrast, the three smallest data journals each publish fewer than five data reports per year.

Ten of the 13 journals are devoted exclusively to data reports, except for occasional editorials or feature articles. Data reports account for at least 87% of the items published in those journals. However, three of the journals – *Biodiversity Data Journal*, *Data* and *Earth System Science Data* – routinely publish items other than data reports, such as reviews or empirical articles on data science topics.

Nine of the 13 journals are published by commercial publishers, and Elsevier accounts for more than half the data reports that appear each year. Notably, non-profit organizations publish two of the three journals that are accessible only to subscribers.

Although Candela et al. list BMC as a major publisher of data journals,<sup>28</sup> the Appendix shows that data reports account for no more than 3% of the items published in any BMC journal. Data reports are welcome at nearly all the BMC journals, but always as one of several article types.

'the publishers of data journals are notable for their good reputations'

Overall, the publishers of data journals are notable for their good reputations. Publishers such as Elsevier and the American Chemical Society are well known, and Pensoft earned the 2016 Innovator Award of the Scholarly Publishing and Academic Resources Coalition (SPARC).

#### Characteristics of data reports

The 13 journals use 11 different terms for their data reports: data paper (five instances), article (two instances), data descriptor (two instances), data article, data description paper, data in brief, interactive key, research article, single taxon treatment, species conservation profile and taxonomic paper. *Biodiversity Data Journal* is unique in publishing several distinct types of data reports: data papers, interactive keys, single taxon treatments, species conservation profiles and taxonomic papers. The characteristics of data reports vary from one journal to the next. (See Table 4.)

Data journal	Typical length (printed pages) <sup>b</sup>	Percentage of data file(s) <sup>c</sup>				
		Original or secondary data?	Included in text of report <sup>d</sup>	On journal's website as suppl. files	In external data repository <sup>e</sup>	Not found
<i>Data in Brief</i>	6–9	Either	14%	56%	26%	4%
<i>Scientific Data</i>	7–10	Either	0%	0%	100%	0%
<i>IUCrData</i>	6–10	Original	0%	100%	0%	0%
<i>Data</i>	8–16	Either	0%	28%	60%	12%
<i>Earth System Science Data (ESSD)</i>	13–21	Original	0%	0%	100%	0%
<i>Biodiversity Data Journal</i>	12–22	Original	26%	44%	28%	2%
<i>Geoscience Data Journal</i>	8–14	Either	0%	0%	87%	13%
<i>Journal of Open Psychology Data</i>	4–6	Either	0%	0%	100%	0%
<i>Open Data Journal for Agricultural Research</i>	6–9	Either	0%	0%	100%	0%
<i>Open Health Data</i>	4–6	Either	0%	4%	74%	22%
<i>Journal of Chemical &amp; Engineering Data</i>	8–11	Original	60%	38%	0%	2%
<i>Chemical Data Collections</i>	8–14	Original	76%	6%	10%	8%
<i>Journal of Physical &amp; Chemical Reference Data</i>	25–40	Original	98%	0%	0%	2%

Table 4. Characteristics of data reports in each of the 13 journals<sup>a</sup>

- Length and data file statistics are based on the 50 most recent data reports in each journal – or on all the published data reports, for journals with fewer than 50.
- Each page range represents the middle two-thirds of the values (i.e. the median  $\pm$  1 standard deviation, adjusted to account for the natural breaks in the distribution of page lengths).
- If the same data were presented in multiple places, they were counted in the leftmost column: *in text of report* rather than *on journal's website*, and *on journal's website* rather than *in external data repository*. Data sites owned by the journal publisher but separate from the journal were counted as external repositories.
- This category includes most chemical data as well as most image data (photographs, blots, diagnostic images, etc.).
- Includes cases in which the data could be readily located despite an incorrect URL or identifier in the data report.

Although the average length of a data report is nine pages, the typical length ranges from five pages at *Open Health Data* and the *Journal of Open Psychology Data*, to more than 30 pages at the *Journal of Physical and Chemical Reference Data*. These differences in length often represent differences in the number of sections or elements expected by the editors of each journal (e.g. data collection methods, sampling strategies, validation, limitations, unique or innovative characteristics, variables, coding, descriptive statistics, file specifications and user notes).

Six of the 13 journals accept only reports that describe original data – data based on the author's own experimental, observational, computational or statistical work. Seven also accept reports based on secondary data – data compiled from publicly available sources (e.g. archives, documents or websites). With secondary data, the author is expected to have added value through processes such as compilation, standardization or verification.

While all these data journals require authors to make their data freely available without mediation, only two of the 13 have policies that require authors to host their data on the journals' own websites. Five require authors to deposit their data in an external repository and six allow authors to present their data either on the journal's website or elsewhere. Table 4 shows, for each journal, the percentage of data sets that are (a) included in the text of the data report itself, (b) available on the journal's website as supplementary files or (c) hosted in an external repository. As Table 4 reveals, there is no consistency in the practices adopted by the 13 journals, other than a tendency to present chemical and image data within the data report itself. In particular, the three largest data journals – *Data in Brief*, *Scientific Data* and *IUCrData* – have each adopted different approaches to data access. Aside from a few minor discrepancies, the actual data access practices of the journals are consistent with their stated policies. For example, 96% of the data reports published in *Open Health Data* have data files hosted elsewhere, in keeping with the journal's policy, and 4% have data files hosted on the journal's website.

Previous research has revealed high rates of non-compliance with data mandates at conventional journals, from 14% to 69%.<sup>29</sup> For the 13 data journals shown in Table 4, the

9 rate of non-compliance is considerably lower at around 3%.<sup>30</sup> Data journals' lower rate of non-compliance is presumably due to the fact that only authors with a commitment to data archiving (publication) are likely to submit their work to a data journal. In contrast, conventional journals, including those that require data sharing, may attract authors who have no particular interest in making their data accessible.

For the non-compliant data reports – those that did not provide immediate access to the data – broken links were the main difficulty. Specifically, the 565 data reports evaluated for this purpose (i.e. those that appeared most recently in each data journal) include ten with broken links to data repositories (for which the available information was not enough to provide ready access to the data), five with instructions such as 'contact the author for data access', four that include only summary statistics rather than raw data or microdata, three for which a supplementary file is mentioned but not accessible, three for which data access requires registration with the data repository and one for which the data repository includes a data set that is clearly incomplete. Broken links are especially prevalent at *Data* and *Geoscience Data Journal*. They do not necessarily represent non-compliance on the part of the author, however, since they can also result from errors by journal publishers and repository managers.

'access restrictions may sometimes be instituted to protect the privacy or safety of human subjects'

Access restrictions that require data users to identify themselves (i.e. 'contact the author' and 'register with the data repository') were counted as a form of non-compliance, since they are contrary to the spirit of OA; they provide an opportunity for authors and repository administrators to deny particular data requests. We should keep in mind, however, that access restrictions may sometimes be instituted to protect the privacy or safety of human subjects. At *Open Health Data*, the most clinically oriented of the 13 data journals, all five instances of non-compliance can be traced to access restrictions that require data users to identify themselves.

In 2015 most data journals required authors to submit their data to third-party archives, since 'maintaining a 24/7 operational data repository service requires investments in specialized computing, software resources, and skilled technical staff'.<sup>31</sup> Although this is understandable, a system that relies on multiple agencies and technologies is inherently less stable than one in which responsibility is clearly delineated.<sup>32</sup> With external (third-party) data deposit, at least three actors are involved in every transaction – every attempt to deposit, evaluate, revise, verify or maintain the data. For instance, there is no mechanism by which the modification of a data file on PhysioNet necessarily leads to a change in the data report published by *Scientific Data*. Moreover, limited evidence suggests that non-compliance rates are lower when authors are required to post their data on the journal's platform, either within the report or as a supplementary file. The point-biserial correlation,<sup>33</sup>  $r_{pb}$ , between compliance rate and data policy is 0.30 when the policies are coded 1 (data are included within the report or on the journal's website), -1 (data are hosted in an external repository), or 0 (either option is acceptable).

All 13 data journals are willing to publish data that have not (yet) been used in a conventional research paper. However, the editors of *Data in Brief* recommend that authors first publish the research that draws on their data, then cite that research in the subsequent data report. This practice gives authors the exclusive use of their data, at least for a time, and ensures that the data are of proven utility. In contrast, the editors of *Data* recommend that authors first publish a data report, then cite the data report in their research.<sup>34</sup>

#### **Editors and peer review**

The editors of the 13 journals are almost all at well-known universities or research institutes, such as Harvard University, Oxford University, MIT, the University of Copenhagen, the University of Melbourne, Uppsala University and the National Institute of Standards and Technology (NIST). The editorial boards vary in size from eight to 258 members (median = 20). Apart from the *Journal of Physical & Chemical Reference Data*, which has a strong US focus, no journal is dominated editorially by a single institution or country. All have broad international representation.

10 Candela et al. reported in 2015 that nearly all data journals use conventional peer review, in which:

- the review process is intended to both evaluate papers and improve them through revision
- at least two anonymous reviewers are selected by the editors
- the reviewers' comments are the primary factor in the editors' decision
- the reviews are not made available to readers of the journal
- there is no provision for post-publication review.<sup>35</sup>

More recent evidence (Table 5) shows that conventional peer review is still the norm. Nonetheless, the peer review criteria used by data journals do account for the journals'

---

**Data in Brief**

Six criteria: Is the data format in alignment with existing standards? Are the protocol/references for generating data sufficiently explained? Is the data description complete and is data well-documented? Do the authors adequately explain the data's utility? Are the data potentially reusable? Does the article adhere to the template?

---

**Scientific Data**

Each paper is reviewed by one data standards expert and at least one subject expert based on 'the technical quality of the procedures used to generate the data, the reuse value of the resulting datasets and their alignment with existing community standards, and the completeness of the data description. [Acceptance] is not based on the perceived impact or novelty of the findings'.

---

**IUCrData**

Single-blind review by at least two reviewers. Papers not accepted after two rounds of revision will not be published.

---

**Data**

Each paper is evaluated by at least two reviewers. Reviewers may choose to sign their reviews. Authors may choose to include the reviewers' reports as supplementary materials.

---

**Earth System Science Data (ESSD)**

Papers that meet the standards of an initial rapid review are posted to the journal's website. Readers are invited to submit reviews or comments, and the editors' decision accounts for both the solicited reviews and any additional remarks. If the paper is accepted, it is published with the referees' comments (anonymous or attributed), the readers' comments (attributed), and the authors' replies.

---

**Biodiversity Data Journal**

After initial editorial review, each paper is sent to two or three nominated reviewers, who are expected to submit their comments within ten days; and to several panel reviewers, who may choose whether to comment. Authors' revisions are expected within one week, although extensions may be granted. Most revised papers are re-evaluated by the editors, although some are sent for another round of review.

---

**Geoscience Data Journal**

The review process evaluates the data report (completeness, appropriateness of methods, uniqueness, applicability and utility of the data), the metadata (completeness and quality) and the data (accessibility and usability).

---

**Journal of Open Psychology Data**

The review criteria include content, structure and argument, figures/tables, formatting and language.

---

**Open Data Journal for Agricultural Research**

No information provided.

---

**Open Health Data**

The review process evaluates the data report (description of methods, appropriateness of methods, ability to replicate methods, correctness of data description, extent to which reuse of the data is addressed, accessibility of the data) and the data (appropriateness of data repository, accessibility and licensing, file formats, labeling and user notes, study participants' privacy, inclusion of software or other necessary supplements).

---

**Journal of Chemical and Engineering Data**

'Articles should present a significant amount of experimental or computational data on properties of systems of technological or theoretical interest that are not available in the original literature, that have lower uncertainty than those published, or that help resolve conflicts in previously published values.'

---

**Chemical Data Collections**

Six criteria: Is the data format in alignment with existing standards? Are the protocol/references for generating data sufficiently explained? Is the data description complete and data well-documented? Do the authors adequately explain the data's utility? Are the data potentially reusable? Does the article adhere to the template?

---

**Journal of Physical and Chemical Reference Data**

No information provided.

Table 5. Notes on review processes

11 unique characteristics. For example, *Open Health Data* asks reviewers to consider several criteria that apply to both the data and the data report: correctness of data description, level of methodological detail, appropriateness of methods, ability to replicate the data, extent to which reuse of the data is addressed, accessibility of the data, protection of study participants' privacy, appropriateness of data repository, accessibility, licensing, use of non-proprietary file formats, labeling, user notes and inclusion of software or other materials needed to make use of the data. Overall, the review criteria adopted by data journals correspond closely to data users' expectations. Data users want reliable, well-documented data collection and processing methods, adequate metadata that allow for replication, technical details that inspire confidence in the quality of the data, and data files and notes that can be understood without assistance.<sup>36</sup>

Just one of the 13 journals, *Scientific Data*, has adopted the 'soundness-only' standard used by some OA journals such as *PLOS ONE* and *Scientific Reports*. This standard is meant to ensure that scientifically rigorous work is not excluded due to its presumed lack of novelty, significance or expected impact, and to avoid the publication bias that results when only statistically or theoretically significant work appears in the literature. However, there is evidence that reviewers consider perceived importance even when instructed not to do so,<sup>37</sup> and the distinction between conventional and 'soundness-only' peer review may be less meaningful for data reports than for conventional research articles.

A few of the 13 journals have adopted innovative procedures while maintaining conventional peer review standards. *Scientific Data* sends each paper to at least one subject expert and at least one data standards expert. *IUCrData* accepts or rejects each paper after no more than two rounds of revision. In a somewhat greater departure from the norm, *Biodiversity Data Journal* solicits reviews from both regular reviewers (who agree to review the paper) and panel reviewers (who may or may not choose to comment). Reviewers' comments are expected within ten days, and authors are expected to complete their revisions ten days later, although extensions may be granted. Likewise, papers that meet rapid review standards at *Earth System Science Data* are posted to the journal's website. Readers are asked to submit their comments and the editors' decision accounts for both the solicited reviews and any additional remarks. If the paper is accepted, it is published with the referees' reviews (anonymous or attributed), the readers' comments (attributed) and the authors' replies.

For the ten journals with available data, the median time from submission to first decision is 52 days. Two of the ten have median review times of 30 days or less (*Data*: 17 days; *IUCrData*: 24 days), five of 35 to 60 days, and three of 130 days or more (*Journal of Chemical and Engineering Data*: 132 days; *Scientific Data*: 165 days; *Geoscience Data Journal*: 167 days). There is far less variation in the usual time from acceptance to publication, the median being 18 days, with values of 38 days or fewer for all but *Geoscience Data Journal*.<sup>38</sup> (See Supplementary Table 1 – details may be found in the data accessibility statement at the end of this article.) Unfortunately, acceptance rate data were available for only three of the journals: *Chemical Data Collections* (37%), *Data in Brief* (39%) and *IUCrData* (83%).

#### **Licenses and APCs**

As noted earlier, ten of the 13 data journals are *open* data journals that comply with the Berlin Declaration on Open Access.<sup>39</sup> OA principles are represented fully in the Creative Commons CC BY licenses adopted by each of the data journals for which information is available. The CC BY license allows others to redistribute, modify and build upon the data report (and the accompanying data) as long as they credit the author/creator of the original work.<sup>40</sup> Six of the data journals – *Data in Brief*, *Scientific Data*, the *Journal of Open Psychology Data*, *Open Health Data*, the *Journal of Chemical and Engineering Data* and *Chemical Data Collections* – also offer other licensing options that (for example) limit the creation of derivative works, restrict commercial use, or limit redistribution and use in the first 12 months after publication.

Data journal	APC <sup>a</sup>	Waivers and reductions
<i>Data in Brief</i>	\$600	Possible, especially for authors in developing countries
<i>Scientific Data</i>	\$1,690	Automatic for authors in developing countries; possible for others
<i>IUCrData</i>	\$200	Possible, especially for authors in developing countries
<i>Data</i>	\$1,020	Possible, especially for authors in developing countries and in disciplines with less funding
<i>Earth System Science Data (ESSD)</i>	\$0	Not applicable
<i>Biodiversity Data Journal</i>	\$110–\$510	Automatic for retirees, independent scholars, students, and authors in developing countries
<i>Geoscience Data Journal</i>	\$1,200–\$1,500	Automatic for authors in developing countries
<i>Journal of Open Psychology Data</i>	\$0	Not applicable
<i>Open Data Journal for Agricultural Research</i>	Not stated	Not stated
<i>Open Health Data</i>	\$0	Not applicable
<i>Journal of Chemical &amp; Engineering Data</i>	\$1,250–\$5,000	Automatic for authors in developing countries
<i>Chemical Data Collections</i>	\$500	Possible, especially for authors in developing countries
<i>Journal of Physical &amp; Chemical Reference Data</i>	Not stated	Not stated

Table 6. Article processing charges (APCs) and waiver policies

a. The *Journal of Open Psychology Data* and *Open Health Data* ask for voluntary contributions of \$435 and \$125, respectively.

At the ten open data journals – the first ten shown in Table 6 – the APCs vary dramatically, from no charge at all (three journals) to \$1690 at *Scientific Data*. The average APC is \$574, with no consistent difference between commercial and non-profit journals. These results are in line with those of Candela et al., who reported an average APC of \$523–\$566 in 2019 dollars.<sup>41</sup> Although many authors have grant funds or institutional support to cover these charges, that is not always the case, and data archives/repositories may charge additional fees. Fortunately, most of the 13 journals have generous APC waiver policies for authors in developing countries, and most will also consider granting waivers for other reasons.

#### **Indexing and citation impact**

Data journals bring data dissemination efforts into closer alignment with scholarly norms through peer review, indexing and citation. Although several authors have set forth guidelines for the direct citation of data files,<sup>42</sup> none of those guidelines have been widely adopted. Data files are often used but not cited and data citations, when they do appear, are often inconsistent in format. The inclusion of data reports in bibliographic databases such as BIOSIS, PubMed, Google Scholar, Science Citation Index (SCI) and Scopus provides a way around these difficulties. Data reports can be indexed and cited in the same way as conventional research articles, using the same mechanisms that have proven effective within the broader system of scholarly communication.

'Data reports can be indexed and cited in the same way as conventional research articles'

These advantages will be achieved, however, only if data journals are actually included in the major bibliographic databases. Of the 13 data journals, just three are indexed in BIOSIS. Ten are indexed in PubMed, however, six in SCI, and eight in Scopus. (See Table 7.) Despite the poor coverage of these data journals in BIOSIS and SCI, there are two reasons why the inclusion of data journals in bibliographic databases may provide an incentive for authors to publish there. First, the indexing of Group 1 data journals – those devoted mainly to data reports – appears to have improved substantially in recent years. In 2015, none of the seven Group 1 journals identified by Candela et al. were included in either SCI or Scopus. Since many bibliographic databases are reluctant to include recently founded journals, we might expect better coverage of data journals in the coming years as each builds a multi-year record of publication and scholarly impact. Second, the index coverage rates for all 13 journals, combined, are higher than might be suggested by the entries for the individual journals (Table 7). This is because the journals that publish more data reports are more likely to be indexed in BIOSIS ( $r_{pb} = 0.53$ ) and Scopus ( $r_{pb} = 0.34$ ). Ninety-four per cent of the data reports in these 13 journals are indexed in PubMed, 91% in Scopus, 63% in BIOSIS, and 33% in SCI.<sup>43</sup>

Data journal	BIOSIS	PubMed	SCI	Scopus	IF <sup>a</sup>	CiteScore <sup>b</sup>
<i>Data in Brief</i>	Yes	Yes	—	Yes	—	71
<i>Scientific Data</i>	Yes	Yes	Yes	Yes	87	99
<i>IUCrData</i>	—	Yes	—	—	—	—
<i>Data</i>	—	Yes	—	—	—	—
<i>Earth System Science Data (ESSD)</i>	—	Yes	Yes	Yes	99	99
<i>Biodiversity Data Journal</i>	Yes	Yes	Yes	Yes	26	43
<i>Geoscience Data Journal</i>	—	Yes	Yes	Yes	66	88
<i>Journal of Open Psychology Data</i>	—	—	—	—	—	—
<i>Open Data Journal for Agricultural Research</i>	—	—	—	—	—	—
<i>Open Health Data</i>	—	Yes	—	—	—	—
<i>Journal of Chemical &amp; Engineering Data</i>	—	Yes	Yes	Yes	51	74
<i>Chemical Data Collections</i>	—	—	—	Yes	—	41
<i>Journal of Physical &amp; Chemical Reference Data</i>	—	Yes	Yes	Yes	76	91

Table 7. Bibliographic index coverage and citation impact (Impact Factor and CiteScore percentiles)

- a. IF (Impact Factor) is the average number of times the articles published in the journal over a two-year period (the two years prior to the report year) were cited during the report year, based on SCI data. It is expressed here as a percentile rank among journals in the appropriate subject category.
- b. CiteScore is the average number of times the articles published in the journal over a three-year period (the three years prior to the report year) were cited during the report year, based on Scopus data. It is expressed here as a percentile rank among journals in the appropriate subject category.

Among the 13 journals, inclusion in BIOSIS, PubMed, SCI and Scopus is not generally associated with variables such as founding date, non-profit status, report length or APC. There are a few exceptions, however. SCI is especially likely to index the journals that were founded earlier ( $r_{pb} = 0.51$ ), those that publish longer data reports ( $r_{pb} = 0.56$ ) and those with higher APCs ( $r_{pb} = 0.49$ ). Scopus is especially likely to index the journals that publish longer data reports ( $r_{pb} = 0.47$ ) and those with higher APCs ( $r_{pb} = 0.41$ ).

Only the journals indexed by SCI are assigned Impact Factors (IFs). Likewise, only the journals indexed by Scopus are assigned CiteScores. As Table 7 shows, both indicators reveal the same pattern: two data journals, *Scientific Data* and *Earth System Science Data*, have exceptionally high citation impact; four have higher impact than the average journal in their subject areas and seven have lower impact or are not covered by SCI and Scopus. The omission of a journal from those two databases does not necessarily indicate low impact, however. It may also be due to insufficient data (e.g. a recent founding date), failure to adhere to a regular publication schedule, a high self-citation rate or other factors.<sup>44</sup> The high citation impact of *Scientific Data* is notable, especially since it was among the less cited data journals just a few years ago.<sup>45</sup> Its CiteScore puts it in second place (99th percentile) among the 206 journals in the Scopus 'statistics and probability' category. It also ranks at or above the 94th percentile in five other subject categories. Likewise, *Earth System Science Data* is ranked first of the 182 journals in general earth and planetary sciences. Even *Data in Brief*, somewhat lower in the hierarchy, is ranked 26th of the 90 journals (71st percentile) in the Scopus 'multidisciplinary' category. Despite their recent emergence, nearly half of the 13 data journals have higher citation rates than most of the conventional journals in their subject areas.<sup>46</sup>

## Summary

Of the 169 journals identified as data journals by Candela et al., or in various online sources,<sup>47</sup> only 19 are Group 1 data journals – pure data journals devoted primarily to data reports. The 13 Group 1 journals that publish in the fields of medicine, health, biology or chemistry vary greatly in size, subject scope, publisher characteristics, length of data reports, data hosting policies, time from submission to first decision, APCs, bibliographic index coverage and citation impact. Nonetheless, nearly all are similar in their peer-review criteria, their OA license terms and the characteristics of their editorial boards.

## Data journals: potential and continuing challenges

Data archives and data journals both make data freely accessible online. However, there are several advantages specific to data journals:

- **Quality control**  
Conventional peer review ensures the quality and completeness of both data and documentation, thereby facilitating replication and reuse. The data report format encourages the replication and transparency that are essential to scientific research.<sup>48</sup>
- **Discoverability**  
The indexing of data reports increases their discoverability, thereby encouraging the use and citation of data while also promoting opportunities for collaboration between data creators and other scholars.
- **Incentives for data publishing**  
The article format allows authors and institutions to receive full credit for their data-related work. It also facilitates citation linking from the data report to the studies that have used the data, and vice versa.<sup>49</sup> The established system of authorship and citation credit gives researchers strong, direct incentives to publish their data, and these same incentives may encourage the production and dissemination of new data.<sup>50</sup>
- **Efficiency of effort**  
Data reports reduce the need to include data details in all the papers that use the data; authors may simply refer to the earlier data report.
- **Sustainability**  
Hosting data on the publisher's online platform helps ensure long-term access by reducing data users' dependence on multiple organizations and multiple links.<sup>51</sup>

The benefits of data journals are system-wide. The author who publishes a data report gets a peer-reviewed article, perhaps in a high-impact journal, for work that might otherwise go unacknowledged. Other researchers get a free data resource that has been evaluated more rigorously and described more fully than it might otherwise have been. The author's institution gets an opportunity to raise its profile in rankings that account for publishing productivity or for the citation impact of scientific research. Finally, the publisher gets a journal – perhaps a highly cited journal – and an initial advantage in the data publishing arena.

Of course, authorship credit is a reliable incentive only if scholars acknowledge its value. Recent survey evidence suggests that they do. Natural and social scientists agree that conventional peer-reviewed articles carry more weight than peer-reviewed data reports, which in turn carry more weight than peer-reviewed, stand-alone data files. Most data files are not peer-reviewed, however, and all forms of peer-reviewed work are regarded more highly than other contributions.<sup>52</sup>

Moreover, authorship credit is widely understood and accepted by scholars in a broad range of fields, unlike new forms of acknowledgment such as 'data steward' credit.<sup>53</sup> At the moment, data archives provide no similar incentive – no true authorship credit – since the peer-reviewed article (or the book, in some fields) remains the gold standard by which research outputs are evaluated. Although changes in formal assessment programs such as the Research Excellence Framework might provide greater credit for data archive submissions (and thereby reduce the advantages associated with data reports and data journals), there is currently no sign that changes of this type are anticipated.

Despite the advantages of data journals, three problems remain. The most serious problem, which affects data archives and data journals equally, is the need for sustainable data management practices. A shift in responsibility – from individual authors to stand-alone data archives to data journals, for instance – does not alter the underlying fact that some individual or group must undertake the long-term management of hardware, software and data (e.g. maintenance of links and migration from older to newer file formats).<sup>54</sup> Unfortunately, none

15 of the 13 data journals shown in Table 3 have formal, publicly accessible policies that describe how they will ensure long-term data preservation and access. In that respect, they are similar to conventional economics journals.<sup>55</sup> As the Appendix shows, not all data journals are financially or administratively viable, and many appear to have no ‘insurance’ in the event that they are no longer able to maintain the reports and files entrusted to them.

Second, previous research suggests that the advantages of keeping data private are especially great in fields such as ecology, where long-term, externally funded research projects (up to several decades’ duration) are the norm.<sup>56</sup> Data journals may have limited impact in those subject areas, since the credit associated with a few data articles is unlikely to offset the benefits of maintaining exclusive access to unique research materials for an extended period.

Finally, the exclusionary effect of APCs is no less a problem for open data journals than for other OA journals. Although authors in developing countries can usually obtain APC waivers or reductions, others – independent scholars and students, for instance – may not be able to do so.

#### Data accessibility statement

Supplementary Table 1 is available at <http://doi.org/10.5281/zenodo.3755191>.

#### Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the ‘full list of industry A&As’ link: <http://www.uksg.org/publications#aa>

#### Competing Interests

The author has declared no competing interests.

#### References

- Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani, “Data journals: a survey,” *Journal of the Association for Information Science and Technology* 66, 9 (2015): 1747–1762, DOI: <https://doi.org/10.1002/asi.23358> (accessed 21 April 2020).
- Patrick Andreoli-Versbach and Frank Mueller-Langer, “Open access to data: An ideal professed but not practised,” *Research Policy* 43, 9 (2014): 1621–1633, DOI: <https://doi.org/10.1016/j.respol.2014.04.008> (accessed 21 April 2020); John Ernest Kratz and Carly Strasser, “Researcher Perspectives on Publication and Peer Review of Data,” *PLOS ONE* 10, 0117619 (2015): 1–21, DOI: <https://doi.org/10.1371/journal.pone.0117619> (accessed 21 April 2020).
- William G. Dewald, Jerry G. Thursby, and Richard G. Anderson, “Replication in Empirical Economics: The *Journal of Money, Credit and Banking* Project,” *The American Economic Review* 76, 4 (1986): 587–603, <https://www.jstor.org/stable/1806061> (accessed 21 April 2020); Jelte M. Wicherts, Denny Borsboom, Judith Kats, and Dylan Molenaar, “The Poor Availability of Psychological Research Data for Reanalysis,” *American Psychologist* 61, 7 (2006): 726–728, DOI: <https://doi.org/10.1037/0003-066X.61.7.726>
- James A. Mills et al., “Archiving Primary Data: Solutions for Long-Term Studies,” *Trends in Ecology & Evolution* 30, 10 (2015): 581–589, DOI: <https://doi.org/10.1016/j.tree.2015.07.006> (accessed 21 April 2020).
- “Sharing research data to improve public health: full joint statement by funders of health research,” 2019, Wellcome, <https://wellcome.ac.uk/what-we-do/our-work/sharing-research-data-improve-public-health-full-joint-statement-funders-health> (accessed 21 April 2020).
- Christine L. Borgman, “The conundrum in sharing research data,” *Journal of the American Society for Information Science and Technology* 63, 6 (2012): 1059–1078, DOI: <https://doi.org/10.1002/asi.22634> (accessed 21 April 2020); Eleni Castro et al., “Evaluating and Promoting Open Data Practices in Open Access Journals,” *Journal of Scholarly Publishing* 49, 1 (2017): 66–88, DOI: <https://doi.org/10.3138/jsp.49.1.66> (accessed 21 April 2020).
- Liz Silva, Theo Bloom, Emma Ganley, and Maggie Winker, “PLOS’ New Data Policy: Public Access to Data,” *EveryONE* (blog), *PLOS ONE*, February 24, 2014, <https://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2> (accessed 21 April 2020).
- Giovanni Colavizza et al., The citation advantage of linking publications to research data, 2019, <https://arxiv.org/abs/1907.02565> (accessed 21 April 2020).
- Colavizza et al., The citation advantage; Tom E. Hardwicke et al., “Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*,” *Royal Society Open Science* 5, 180448 (2018): 1–18, DOI: <https://doi.org/10.1098/rsos.180448> (accessed 21 April 2020).

10. Rachel A. Spicer and Christoph Steinbeck, "A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it," *Metabolomics* 14, 1 (2018): 16, DOI: <https://doi.org/10.1007/s11306-017-1309-5> (accessed 21 April 2020).
11. William G. Dewald and Richard G. Anderson, "Replication and Reflection: A Decade at the *Journal of Money, Credit, and Banking*," in *Secrets of Economics Editors*, ed. Michael Szenberg and Lall Ramrattan (Cambridge, MA: MIT Press, 2014), 199–212; Campbell R. Harvey, "Reflections on Editing the *Journal of Finance*, 2006 to 2012," in *Secrets of Economics Editors*, ed. Michael Szenberg and Lall Ramrattan (Cambridge, MA: MIT Press, 2014), 67–81.
12. Frank Mueller-Langer and Patrick Andreoli-Versbach, "Open access to research data: Strategic delay and the ambiguous welfare effects of mandatory data disclosure," *Information Economics and Policy* 42 (2018): 20–34, DOI: <https://doi.org/10.1016/j.infoecopol.2017.05.004> (accessed 21 April 2020).
13. Candela et al., "Data journals"; Mark J. Costello, "Motivating Online Publication of Data," *BioScience* 59, 5 (2009): 418–427, DOI: <https://doi.org/10.1525/bio.2009.59.5.9> (accessed 21 April 2020); Melissa A. Haendel, Nicole A. Vasilevsky, and Jacqueline A. Wirz, "Dealing with Data: A Case Study on Information and Data Management Literacy," *PLOS Biology* 19, 5 (2012): 1–4, DOI: <https://doi.org/10.1371/journal.pbio.1001339> (accessed 21 April 2020); Mills et al., "Archiving Primary Data"; Sébastien Renaut et al., "Data Management, Archiving, and Sharing for Biologists and the Role of Research Institutions in the Technology-Oriented Age," *BioScience* 68, 6 (2018): 400–411, DOI: <https://doi.org/10.1093/biosci/biy038> (accessed 21 April 2020); Jeffrey N. Rouder, "The what, why, and how of born-open data," *Behavior Research Methods* 48, 3 (2016): 1062–1069, DOI: <https://doi.org/10.3758/s13428-015-0630-z> (accessed 21 April 2020); Birgit Schmidt, Birgit Gemeinholzer, and Andrew Treloar, "Open Data in Global Environment Research: The Belmont Forum's Open Data Survey," *PLOS ONE* 11, 0146695 (2016): 1–29, DOI: <https://doi.org/10.1371/journal.pone.0146695> (accessed 21 April 2020); Silva et al., "PLOS' New Data Policy: Public Access to Data."
14. Mills et al., "Archiving Primary Data."
15. Mills et al., "Archiving Primary Data", 585.
16. Mark D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data* 3, 160018 (2016): 1–9, DOI: <https://doi.org/10.1038/sdata.2016.18> (accessed 21 April 2020).
17. Borgman, "The conundrum in sharing research data," 1067; Christine L. Borgman, Andrea Scharnhorst, and Milena S. Golshan, "Digital data archives as knowledge infrastructures: Mediating data sharing and reuse," *Journal of the Association for Information Science and Technology* 70, 8 (2019): 888–904, DOI: <https://doi.org/10.1002/asi.24172> (accessed 21 April 2020); Rouder, "The what, why, and how."
18. Vishwas Chavan and Lyubomir Penev, "The data paper: a mechanism to incentivize data publishing in biodiversity science," *BMC Bioinformatics* 12, 15 (2011): 1–12, DOI: <https://doi.org/10.1186/1471-2105-12-S15-S2> (accessed 21 April 2020); John E. Kratz and Carly Strasser, "Data publication consensus and controversies," *F1000Research* 3, 94 (2014): 1–21, DOI: <https://doi.org/10.12688/f1000research.3979.3> (accessed 21 April 2020).
19. Mueller-Langer and Andreoli-Versbach, "Open access to research data."
20. Mark J. Costello et al., "Biodiversity data should be published, cited, and peer reviewed," *Trends in Ecology & Evolution* 28, 8 (2013): 454–461, DOI: <https://doi.org/10.1016/j.tree.2013.05.002> (accessed 21 April 2020).
21. Candela et al., "Data journals"; Kratz and Strasser, "Data publication consensus"; Kratz and Strasser, "Researcher Perspectives."
22. Kratz and Strasser, "Researcher Perspectives," 14.
23. Sarah Callaghan et al., "Overlay Journals and Data Publishing in the Meteorological Sciences," *Ariadne* 60 (2009), <http://www.ariadne.ac.uk/issue60/callaghan-et-al/> (accessed 21 April 2020).
24. Candela et al., "Data journals."
25. Candela et al., "Data journals."
26. Beall's List of Predatory Journals and Publishers, <https://beallslist.net> (accessed 21 April 2020); "Journal Citation Reports," Clarivate; "Master Journal List," Clarivate <https://mjli.clarivate.com/home> (accessed 21 April 2020); "About the Licenses", Creative Commons, <https://creativecommons.org/share-your-work/licensing-examples> (accessed 21 April 2020); "Journal Finder," Elsevier, <https://journalfinder.elsevier.com> (accessed 21 April 2020); Scopus "Sources," Elsevier, <https://www.scopus.com/sources> (accessed 21 April 2020); "National Library of Medicine Catalog," NCBI, <https://www.ncbi.nlm.nih.gov/nlmcatalog> (accessed 21 April 2020); Judith M. Nixon, "Core Journals in Library and Information Science: Developing a Methodology for Ranking LIS journals," *College & Research Libraries* 75, 1 (2014): 66–90, DOI: <https://doi.org/10.5860/crl12-387> (accessed 21 April 2020); "Journal suggester," Springer Nature, <https://journalsuggester.springer.com> (accessed 21 April 2020); Wikipedia, <https://www.wikipedia.org> (accessed 21 April 2020).
27. Candela et al., "Data journals."
28. Candela et al., "Data journals."
29. Colavizza et al., "The citation advantage"; Hardwicke et al., "Data availability"; Spicer and Steinbeck, "A lost opportunity."
30. This estimate is based on the 'not found' values presented in the table, and on the number of data reports published in each journal.
31. Candela et al., "Data journals", 1755.
32. Gregory W. Lawrence et al., *Risk Management of Digital Information: A File Format Investigation* (Washington, DC: Council on Library and Information Resources, 2000), <http://www.clir.org/pubs/reports/pub93> (accessed 21 April 2020); William H. Walters, "E-Books in Academic Libraries: Challenges for Acquisition and Collection Management," *portal: Libraries and the Academy* 13, 2 (2013): 187–211, DOI: <https://doi.org/10.1353/pla.2013.0012> (accessed 21 April 2020).

33. Peter Y. Chen and Paula M. Popovich, *Correlation: Parametric and Nonparametric Measures* (Thousand Oaks, CA: SAGE, 2002); David C. Howell, *Statistical Methods for Psychology* (Belmont, CA: Wadsworth, 2013); Laerd Statistics, Point-Biserial Correlation Using SPSS Statistics, <https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php> (accessed 21 April 2020); Jason T. Newsom, Point-Biserial Correlation, Phi, & Cramer's V, <http://web.pdx.edu/~newsomj/pa551/lecture15.htm> (accessed 21 April 2020).
34. "About Data," Multidisciplinary Digital Publishing Institute, <https://www.mdpi.com/journal/data/about> (accessed 21 April 2020).
35. Candela et al., "Data journals."
36. Kratz and Strasser, "Researcher Perspectives"; Limor Peer, Ann Green, and Elizabeth Stephenson, "Committing to Data Quality Review," *International Journal of Digital Curation* 9, 1 (2014): 263–291, DOI: <https://doi.org/10.2218/ijdc.v9i1.317> (accessed 21 April 2020).
37. Valerie Spezi et al., "'Let the community decide'? The vision and reality of soundness-only peer review in open-access mega-journals," *Journal of Documentation* 74, 1 (2018): 137–161, DOI: <https://doi.org/10.1108/JD-06-2017-0092> (accessed 21 April 2020).
38. These review and publication times were calculated from the dates shown on each website for the 50 most recent data reports. If the necessary dates were not available for individual reports, the values were estimated from statements on the journals' websites or elsewhere (e.g., Elsevier Journal Finder and Springer Nature Journal suggester).
39. "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," *Max-Planck-Gesellschaft zur Förderung der Wissenschaften*, 2003, <https://openaccess.mpg.de/Berlin-Declaration> (accessed 21 April 2020).
40. Creative Commons Licenses; Iain Hrynaszkiewicz and Matthew J. Cockerill, "Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals," *BMC Research Notes* 5, 494 (2012): 1–12, DOI: <https://doi.org/10.1186/1756-0500-5-494> (accessed 21 April 2020).
41. Candela et al., "Data journals."
42. "Joint Declaration of Data Citation Principles," Data Citation Synthesis Group, 2014, ed. M. Martone, <https://www.force11.org/datacitationprinciples> (accessed 21 April 2020); Bryan Lawrence et al., "Citation and Peer Review of Data: Moving Towards Formal Data Publication," *International Journal of Digital Curation* 2, 6 (2011): 4–37, DOI: <https://doi.org/10.2218/ijdc.v6i2.205> (accessed 21 April 2020); Gianmaria Silvello, "Learning to cite framework: How to automatically construct citations for hierarchical data," *Journal of the Association for Information Science and Technology* 68, 6 (2017): 1505–1524, DOI: <https://doi.org/10.1002/asi.23774> (accessed 21 April 2020); Gianmaria Silvello, "Theory and practice of data citation," *Journal of the Association for Information Science and Technology* 69, 1 (2018): 6–20, DOI: <https://doi.org/10.1002/asi.23917> (accessed 21 April 2020).
43. These estimates are based on data for journal size and index coverage. See Tables 3 and 7.
44. William H. Walters, "Citation-Based Journal Rankings: Key Questions, Metrics, and Data Sources," *IEEE Access* 5 (2017): 22036–22053, DOI: <https://doi.org/10.1109/ACCESS.2017.2761400> (accessed 21 April 2020).
45. David Stuart, "Data bibliometrics: metrics before norms," *Online Information Review* 41, 3 (2017): 428–435, DOI: <https://doi.org/10.1108/OIR-01-2017-0008> (accessed 21 April 2020).
46. Although Impact Factors, CiteScores, and related metrics do provide a rough guide to scholarly impact, it is important to recognize that all such measures have significant flaws, that impact may vary for reasons unrelated to "quality", and that citation-based metrics and subjective ratings of journals are only modestly related. See, for example, Peter Haddawy, Saeed-Ul Hassan, Awais Asghar, and Sarah Amin, "A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality," *Journal of Informetrics* 10, 1 (2016): 162–173, DOI: <https://doi.org/10.1016/j.joi.2015.12.005> (accessed 21 April 2020); San Francisco Declaration on Research Assessment (DORA), 2012, <https://sfidora.org/read/> (accessed 21 April 2020); William H. Walters, "Do subjective journal ratings represent whole journals or typical articles? Unweighted or weighted citation impact?" *Journal of Informetrics* 11, 3 (2017): 730–744, DOI: <https://doi.org/10.1016/j.joi.2017.05.001> (accessed 21 April 2020); William H. Walters and Susanne Markgren, "Do faculty journal selections correspond to objective indicators of citation impact? Results for 20 academic departments at Manhattan College," *Scientometrics* 118, 1 (2019): 321–337, DOI: <https://doi.org/10.1007/s11192-018-2972-7> (accessed 21 April 2020).
47. Candela et al., "Data journals."
48. Thu-Mai Christian, Sophia Lafferty-Hess, William G. Jacoby, and Thomas M. Carsey, "Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals," 2018, <https://osf.io/preprints/socarxiv/cfdb> (accessed 21 April 2020); Dewald and Anderson, "Replication and Reflection"; Dewald et al., "Replication in Empirical Economics"; Jan H. Höfler, "Replication and Transparency in Economic Research," *Perspectives* (blog), Institute for New Economic Thinking, 2015, <https://www.ineteconomics.org/perspectives/blog/replication-and-transparency> (accessed 21 April 2020); Arthur Lupia and Colin Elman, "Openness in Political Science: Data Access and Research Transparency," *PS: Political Science & Politics* 47, 1 (2014): 19–42, DOI: <https://doi.org/10.1017/S1049096513001716> (accessed 21 April 2020).
49. Although agencies such as Crossref and DataCite have worked to strengthen the mechanisms that underlie data citation, many (perhaps most) data citations conform to neither standard, and individual disciplines have established their own standards that are sometimes inconsistent with those of Crossref and DataCite. Moreover, there is currently no mechanism by which conventionally archived data files can be included in citation indexes such as Science Citation Index and Scopus.
50. Despite the emergence of new evaluation methods such as post-publication review, conventional peer review remains the only mechanism of quality control that is widely accepted by scholars in all disciplines. See, for example, Melissa Blankstein and Christine Wolff-Eisenberg, *Ithaka S+R US Faculty Survey 2018* (New York: Ithaka S+R, 2019), <https://doi.org/10.18665/sr.311199> (accessed 21 April 2020); Christine Wolff-Eisenberg, Alisa B. Rod, and Roger C. Schonfeld, *UK Survey of Academics 2015* (New York: Ithaka S+R, 2016), <https://doi.org/10.18665/sr.282736> (accessed 21 April 2020); and the many discipline-specific surveys conducted by Ithaka S+R.
51. Walters, "E-Books in Academic Libraries."
52. Blankstein and Wolff-Eisenberg, *Ithaka S+R US Faculty Survey*; Kratz and Strasser, "Researcher Perspectives"; Wolff-Eisenberg, Rod, and Schonfeld, *UK Survey of Academics*.

53. Barbara E. Bierer, Mercè Crosas, and Heather H. Pierce, "Data Authorship as an Incentive to Data Sharing," *New England Journal of Medicine* 376, 17 (2017): 1684–1687, DOI: <https://doi.org/10.1056/NEJMs1616595> (accessed 30 April 2020).
54. Ana Sofia Figueiredo, "Data Sharing: Convert Challenges into Opportunities," *Frontiers in Public Health* 5, 327 (2017): 1–6, DOI: <https://doi.org/10.3389/fpubh.2017.00327> (accessed 21 April 2020); Lawrence et al., "Citation and Peer Review."
55. Courtney R. Butler and Brett D. Currier, "You Can't Replicate What You Can't Find: Data Preservation Policies in Economic Journals" (paper presented at the 2017 Annual Conference of the International Association for Social Science Information Services & Technology, Lawrence, Kansas, May 2017), <https://osf.io/preprints/lissa/hf3ds>.
56. Mills et al., "Archiving Primary Data."

## Appendix

Journals identified as data journals by one or more websites or by Candela et al., 'Data Journals':

### **Group 1: 'pure' data journals, for which data reports comprise at least half the papers in the journal**

Asterisks indicate the 13 data journals included in the analysis – those that publish in the fields of biology, environmental science, chemistry, medicine, and health sciences.

*Atomic Data and Nuclear Data Tables, Biodiversity Data Journal\**, *Chemical Data Collections\**, *Data in Brief\**, *Data\**, *Earth System Science Data\**, *Geoscience Data Journal\**, *IUCrData\**, *Journal of Astronomical Data*, *Journal of Chemical and Engineering Data\**, *Journal of Open Archaeology Data*, *Journal of Open Humanities Data*, *Journal of Open Psychology Data\**, *Journal of Physical and Chemical Reference Data\**, *Nuclear Data Sheets*, *Open Data Journal for Agricultural Research\**, *Open Health Data\**, *Research Data Journal for the Humanities and Social Sciences*, *Scientific Data\**

### **Group 2: journals that publish data reports but are devoted mainly to other types of contributions**

For each journal, the value in parentheses is the percentage of items in the journal that are data reports, for the period since data reports first appeared in the journal. On average, data reports account for just 1.6% of the papers in these journals and there are just three journals for which data reports comprise more than 8% of the content: *GigaScience* (30), the *International Journal of Food Contamination* (15), and the *Journal of Economics and Statistics* (24). The *Open Journal of Bioresources* (0) publishes bioresource papers, which may include data reports. In practice, however, all of the bioresource papers are descriptions of biobanks – collections of biological samples such as tissue, plasma and DNA.

*Annals of Forest Science* (3), *Biology of Sex Differences* (1), *BioRisk* (0), *BMC Anesthesiology* (0), *BMC Biochemistry* (1), *BMC Bioinformatics* (1), *BMC Biotechnology* (0), *BMC Cancer* (0), *BMC Cardiovascular Disorders* (0), *BMC Chemistry* (0), *BMC Complementary & Alternative Medicine* (0), *BMC Dermatology* (2), *BMC Developmental Biology* (1), *BMC Ecology* (1), *BMC Emergency Medicine* (0), *BMC Endocrine Disorders* (0), *BMC Evolutionary Biology* (0), *BMC Family Practice* (0), *BMC Gastroenterology* (1), *BMC Genetics* (0), *BMC Genomics* (2), *BMC Geriatrics* (0), *BMC Health Services Research* (0), *BMC Hematology* (0), *BMC Immunology* (1), *BMC Infectious Diseases* (0), *BMC International Health and Human Rights* (3), *BMC Medical Education* (0), *BMC Medical Ethics* (0), *BMC Medical Genetics* (0), *BMC Medical Genomics* (1), *BMC Medical Imaging* (0), *BMC Medical Informatics & Decision Making* (0), *BMC Medical Research Methodology* (2), *BMC Medicine* (0), *BMC Microbiology* (0), *BMC Molecular and Cell Biology* (1), *BMC Molecular Biology* (0), *BMC Musculoskeletal Disorders* (0), *BMC Nephrology* (0), *BMC Neurology* (0), *BMC Neuroscience* (0), *BMC Nursing* (0), *BMC Ophthalmology* (1), *BMC Oral Health* (0), *BMC Palliative Care* (0), *BMC Pediatrics* (0), *BMC Pharmacology & Toxicology* (0), *BMC Plant Biology* (0), *BMC Pregnancy & Childbirth* (0), *BMC Psychiatry* (0), *BMC Psychology* (0), *BMC Public Health* (0), *BMC*

*Pulmonary Medicine* (0), *BMC Research Notes* (3), *BMC Sports Science, Medicine and Rehabilitation* (0), *BMC Structural Biology* (1), *BMC Surgery* (0), *BMC Systems Biology* (1), *BMC Urology* (0), *BMC Women's Health* (0), *Botanical Studies* (0), *Cell & Bioscience* (0), *Chiropractic & Manual Therapies* (1), *Comparative Cytogenetics* (1), *Data Science Journal* (6), *Deutsche Entomologische Zeitschrift* (0), *Earthquake Spectra* (0), *Ecological Research* (8), *Ecology* (7), *EvoDevo* (0), *F1000 Research* (2), *Genome Medicine* (1), *Geochemistry, Geophysics, Geosystems* (0), *GigaScience* (30), *Health and Justice* (0), *Human Genomics* (3), *International Journal of Food Contamination* (15), *International Journal of Robotics Research* (3), *International Journal of Spatial Data Infrastructures Research* (6), *Irish Veterinary Journal* (1), *Italian Botanist* (0), *Journal of Applied Volcanology* (2), *Journal of Biomedical Semantics* (5), *Journal of Cheminformatics* (4), *Journal of Economics and Statistics* (24), *Journal of Environmental Quality* (2), *Journal of Hymenoptera Research* (1), *Journal of Occupational Medicine and Toxicology* (0), *Journal of the International Society of Sports Nutrition* (0), *Movement Ecology* (0), *MycoKeys* (1), *Nature Biotechnology* (0), *Nature Conservation* (3), *NeoBiota* (2), *Neuroinformatics* (2), *Nota Lepidopterologica* (1), *One Ecosystem* (5), *Open Journal of Bioresources* (0), *PhytoKeys* (3), *Plant & Cell Physiology* (4), *Plant Methods* (1), *Poverty & Public Policy* (0), *Rio* (1), *Substance Abuse Treatment, Prevention, and Policy* (0), *Subterranean Biology* (1), *Theoretical Biology and Medical Modelling* (0), *ZooKeys* (2), *Zoosystematics and Evolution* (0).

### **Group 3: journals that do not actually include data reports as a publication type**

Several of these journals, including *Ecological Applications*, *Ecological Monographs*, *Geoscientific Model Development* and *PLOS ONE*, require public dissemination of the data used in empirical papers. The *Journal of Open Research Software* and the *Journal of Statistical Software* publish software reports but not data reports. *Internet Archaeology* is distinctive because it showcases the graphic artifacts of early internet culture, but the materials presented in *Internet Archaeology* are not really data reports.

*Applied Vegetation Science*, *BMC Biology*, *BMC Veterinary Research*, *British Journal of Educational Technology*, *Bulletin of the Ecological Society of America*, *Diagnostic Pathology*, *Ecological Applications*, *Ecological Monographs*, *Ecosphere*, *Environmental Microbiome*, *Frontiers in Cellular and Infection Microbiology*, *Frontiers in Ecology and the Environment*, *Geoscientific Model Development*, *Health Information Science and Systems*, *In Silico Pharmacology*, *International Economics*, *Internet Archaeology*, *Journal of Open Research Software*, *Journal of Statistical Software*, *Journal of Vegetation Science*, *PLOS ONE*.

### **Group 4: Journals that are no longer published, or that no longer publish data reports**

For each journal, the value in parentheses is the year of the final issue, or of the final issue that included data reports. Very few of these journals were devoted mainly to data reports. *Acta Crystallographica Section E* formerly published data reports, but those reports now appear in *IUCrData*. *Biomedical Data Journal* has an active website and may be accepting submissions, but it has not published a paper since 2015. *Ecological Archives* formerly hosted the data reports of the Ecological Society of America, but those data reports are now published within the Society's other journals. *European Data Watch* has been incorporated into the *Journal of Economics and Statistics*. *Genomics Data* has been incorporated into *Data in Brief*. *Open Network Biology* was announced as a new BioMed Central journal in 2011, but the publisher's site and PubMed provide no evidence that it was ever published.

*Acta Crystallographica Section E* (2016), *Applied Informatics* (2018), *Aquatic Biosystems* (2015), *Biomedical Data Journal* (2015), *BMC Biophysics* (2019), *BMC Clinical Pathology* (2019), *BMC Ear, Nose and Throat Disorders* (2019), *BMC Medical Physics* (2015), *BMC Obesity* (2019), *BMC Physiology* (2019), *Dataset Papers in Science* (2017), *Earth Perspectives* (2017), *Ecological Archives* (2016), *European Data Watch* (2016), *Genomics Data* (2018), *Journal of Clinical Bioinformatics* (2015), *Journal of Systems Chemistry* (2015), *Microbial Informatics and Experimentation* (2014), *Open Network Biology* (2011), *SpringerPlus* (2016).

**Article copyright: © 2020 William H Walters. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use and distribution provided the original author and source are credited.**



William H Walters  
Executive Director  
Mary Alice & Tom O'Malley Library  
Manhattan College  
4513 Manhattan College Parkway  
Riverdale, NY 10471, US  
E-mail: [william.walters@manhattan.edu](mailto:william.walters@manhattan.edu)  
ORCID ID: <https://orcid.org/0000-0001-9490-4032>

To cite this article:

Walters W H, "Data journals: incentivizing data access and documentation within the scholarly communication system", *Insights*, 2020, 33: 18, 1–20; DOI: <https://doi.org/10.1629/uksg.510>

Submitted on 28 February 2020

Accepted on 21 April 2020

Published on 10 June 2020

Published by UKSG in association with Ubiquity Press.