

A peer-reviewed version of this preprint was published in PeerJ on 25 April 2017.

[View the peer-reviewed version](https://doi.org/10.7717/peerj.3208) (peerj.com/articles/3208), which is the preferred citable publication unless you specifically need to cite this preprint.

Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. 2017. Reproducible and reusable research: are journal data sharing policies meeting the mark? PeerJ 5:e3208 <https://doi.org/10.7717/peerj.3208>

Reproducible and reusable research: Are journal data sharing policies meeting the mark?

Nicole A Vasilevsky^{1,2}, Jessica Minnier³, Melissa A Haendel^{1,2}, Robin E Champieux^{Corresp. 1}

¹ Library, Oregon Health & Science University, Portland, OR, United States

² Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, United States

³ OHSU-PSU School of Public Health, Oregon Health & Science University, Portland, OR, United States

Corresponding Author: Robin E Champieux

Email address: champieu@ohsu.edu

Background. There is wide agreement in the biomedical research community that research data sharing is a primary ingredient for ensuring that science is more transparent and reproducible. Publishers could play an important role in facilitating and enforcing data sharing; however, many journals have not yet implemented data sharing policies and the requirements vary widely across journals. This study set out to analyze the pervasiveness and quality of data sharing policies in the biomedical literature. **Methods.** The online author's instructions and editorial policies for 318 biomedical journals were manually reviewed to analyze the journal's data sharing requirements and characteristics. The data sharing policies were ranked using a rubric to determine if data sharing was required, recommended, required only for omics data, or not addressed at all. The data sharing method and licensing recommendations were examined, as well any mention of reproducibility or similar concepts. The data was analyzed for patterns relating to publishing volume, Journal Impact Factor, and the publishing model (open access or subscription) of each journal. **Results.** 11.9% of journals analyzed explicitly stated that data sharing was required as a condition of publication. 9.1% of journals required data sharing, but did not state that it would affect publication decisions. 23.3% of journals had a statement encouraging authors to share their data but did not require it. There was no mention of data sharing in 31.8% of journals. Impact factors were significantly higher for journals with the strongest data sharing policies compared to all other data sharing mark categories. Open access journals were not more likely to require data sharing than subscription journals. **Discussion.** Our study confirmed earlier investigations which observed that only a minority of biomedical journals require data sharing, and a significant association between higher Impact Factors and journals with a data sharing requirement. Moreover, while 65.7% of the journals in our study that required data sharing addressed the concept of reproducibility, as with earlier investigations, we found that most data

sharing policies did not provide specific guidance on the practices that ensure data is maximally available and reusable.

1 Reproducible and reusable research: Are journal data sharing
2 policies meeting the mark?

3 Nicole A. Vasilevsky^{1,2}, Jessica Minnier³, Melissa A. Haendel^{1,2}, Robin Champieux¹

4

5 ¹ Library, Oregon Health & Science University, Portland, OR, USA

6 ² Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science
7 University, Portland, OR, USA

8 ³ OHSU-PSU School of Public Health, Oregon Health & Science University, Portland, OR, USA

9

10 Corresponding Author:

11 Robin Champieux¹

12 3181 SW Sam Jackson Park Road, LIB, Portland, OR 97239, USA

13 Email address: champieu@ohsu.edu

14

15

16 **Abstract**

17

18 **Background.** There is wide agreement in the biomedical research community that research data
19 sharing is a primary ingredient for ensuring that science is more transparent and reproducible.
20 Publishers could play an important role in facilitating and enforcing data sharing; however, many
21 journals have not yet implemented data sharing policies and the requirements vary widely across
22 journals. This study set out to analyze the pervasiveness and quality of data sharing policies in
23 the biomedical literature.

24

25 **Methods.** The online author's instructions and editorial policies for 318 biomedical journals were
26 manually reviewed to analyze the journal's data sharing requirements and characteristics. The
27 data sharing policies were ranked using a rubric to determine if data sharing was required,
28 recommended, required only for omics data, or not addressed at all. The data sharing method and
29 licensing recommendations were examined, as well any mention of reproducibility or similar
30 concepts. The data was analyzed for patterns relating to publishing volume, Journal Impact
31 Factor, and the publishing model (open access or subscription) of each journal.

32

33 **Results.** 11.9% of journals analyzed explicitly stated that data sharing was required as a condition
34 of publication. 9.1% of journals required data sharing, but did not state that it would affect
35 publication decisions. 23.3% of journals had a statement encouraging authors to share their data
36 but did not require it. There was no mention of data sharing in 31.8% of journals. Impact factors
37 were significantly higher for journals with the strongest data sharing policies compared to all
38 other data sharing mark categories. Open access journals were not more likely to require data
39 sharing than subscription journals.

40

41 **Discussion.** Our study confirmed earlier investigations which observed that only a minority of
42 biomedical journals require data sharing, and a significant association between higher Impact
43 Factors and journals with a data sharing requirement. Moreover, while 65.7% of the journals in
44 our study that required data sharing addressed the concept of reproducibility, as with earlier
45 investigations, we found that most data sharing policies did not provide specific guidance on the
46 practices that ensure data is maximally available and reusable.

47

48

49 Introduction

50 Over the last several years, the importance and benefits of research data sharing have been
51 emphasized by many communities, including professional societies, funders, policy makers, and
52 publishers [1–5]. Several rationales underpin the arguments for better access to and the curation
53 of research data [6]. While the factors contributing to the poor reproducibility of biomedical
54 research are varied and complex, and even the meaning of reproducible research is fraught, data
55 availability is regarded as one necessary component for the assessment of replication and
56 validation studies [7]. If raw data are made available, others have the opportunity to replicate or
57 correct earlier findings and, ostensibly, influence the pace and efficiency of future research
58 endeavors. Researchers can ask new questions of existing data, and data can be combined and
59 curated in ways that further its value and scholarship [6]. As Fischer and Zigmond argue, the
60 great advances in science depend not only on the contributions of many individual researchers,
61 but also their willingness to share the products on their work [8].

62 The benefits described above have motivated many of the organizations that support research to
63 require that data be made publicly available. Since 2011, the National Science Foundation (NSF)
64 has required applicants to submit a data management plan documenting how investigators will
65 conform to the NSF’s expectation that primary data and research resources will be shared with
66 other researchers [9]. The White House Office of Science and Technology Policy issued a
67 memorandum in 2013 directing agencies to make plans for ensuring public access to federally
68 funded research results, including data [2]. In 2014, the National Institutes of Health (NIH)
69 implemented a strong data sharing policy for large-scale human and non-human genomic data
70 [10]. Additionally, the European Research Council’s Open Access Guidelines include and
71 support public access to research data, and open is the default for all data generated via its
72 Horizon 2020 program [11].

73 However, data sharing and its long-term stewardship involve an array of activities, participants,
74 and technologies, especially if discovery, reuse, and preservation are to be ensured [12].
75 Moreover, despite a belief in the importance of access to other’s data for their own work, many
76 scientists do not consistently share their data, reporting a variety of barriers and disincentives
77 [13]. Roadblocks to sharing include insufficient time, a lack of funding, fear of scrutiny or
78 misinterpretation, a deficit of requirements, attribution concerns, competition, difficulty
79 navigating infrastructure options, and a paucity of data sharing related rewards [14–16]. For
80 quality data sharing to become the norm, broad systemic change and solutions are needed.

81 Journal publication is the current and primary mode of sharing scientific research. While
82 arguably problematic, it has the most influence on an individual’s credibility and success [8]. As
83 Lin and Strasser write, journals and publishers occupy an important “leverage point in the
84 research process”, and are key to affecting the changes needed to realize data sharing as a
85 “fundamental practice” of scholarly communication [17]. There has been significant support for
86 and progress toward this end. At a joint workshop held at the NIH in June 2014, editors from 30
87 basic and preclinical science journals met to discuss how to enhance reproducible, robust, and

88 transparent science. As an outcome, they produced the "Principles and Guidelines for Reporting
89 Preclinical Research", which included the recommendation that journals require that all of the
90 data supporting a paper's conclusion be made available as part of the review process and upon
91 publication, that datasets be deposited to public repositories, and that datasets be bi-directionally
92 linked to published articles in a way that ensures attribution" [1]. In 2013, Nature journals
93 implemented a 18 point reporting checklist for life science articles. It included required data and
94 code availability statements, and a strong recommendation for data sharing via public
95 repositories [18]. Additionally, many large and influential journals and publishers have
96 implemented data sharing requirements, including Science, Nature, the Public Library of Science
97 (PLOS), and the Royal Society [19–22].

98 Given these developments, and the influence of journal publishing on scientific communication
99 and researcher success, we sought to investigate the prevalence and characteristics of journal
100 data sharing policies within the biomedical research literature. The study was designed to
101 determine the pervasiveness and quality of data sharing policies as reflected in editorial policies
102 and the instructions to authors. We chose to focus our analysis on the biomedical literature
103 because of the intense attention data availability and its relationship to issues of reproducibility
104 and discovery have received, and on account of our own roles as and work with biomedical
105 researchers.

106 **Materials & Methods**

107 We evaluated the data sharing policies of journals that were included in Thomson Reuter's
108 InCites 2013 Journal Citations Reports (JCR) [23] classified within the following World of
109 Science schema categories: Biochemistry and Molecular Biology, Biology, Cell Biology,
110 Crystallography, Developmental Biology, Biomedical Engineering, Immunology, Medical
111 Informatics, Microbiology, Microscopy, Multidisciplinary Sciences, and Neurosciences. These
112 categories were selected to capture the journals publishing the majority of peer-reviewed
113 biomedical research. The original data pull included 1,166 journals, collectively publishing
114 213,449 articles. We filtered this list to the journals in the top quartiles by Impact Factor (IF) or
115 number of articles published 2013. Additionally, the list was manually reviewed to exclude short
116 report and review journals, and titles determined to be outside the fields of basic medical science
117 or clinical research. The final study set included 318 journals, which published 130,330 articles
118 in 2013. The study set represented 27% of the original Journal Citation Report list and 61% of
119 the original citable articles. Prior to our analysis, the 2014 Journal Citations Reports was
120 released. While we did not use the 2014 data to alter the journals in the study set, we did employ
121 data from both reports in our analyses. In our data pull from JCR, we included the journal title,
122 International Standard Serial Number (ISSN), the total citable items for 2013 and 2014, the total
123 citations to the journal for 2013 and 2014, the Impact Factors for 2013 and 2014, and the
124 publisher. Table 1 reports the number (and percentage) of journals across Impact Factors, and
125 Table 2 reports the number of citable items per journal.

126 We manually reviewed each journal's online author instructions and editorial policies between
127 February 2016 and June 2016. Because we were specifically interested in the information being
128 communicated to manuscript submitting authors about data sharing requirements, we did not
129 consider more peripheral sources of information, such as footnoted links to additional web pages,
130 unless authors were specifically instructed to review this information in order to understand or
131 comply with a journal's data sharing policy. We ranked the journals' data sharing policies using
132 a rubric adapted from Stodden, Guo, and Ma, which we updated to differentiate those policies
133 that exclusively addressed structural (e.g. proteomic) or genomic data sharing [24] (Table 3).
134 Additionally, we examined the policies to determine the recommended data sharing method (e.g.
135 a public repository or journal hosted), if data copyright or licensing recommendations were
136 mentioned, the inclusion of instructions on how long the data should be made available, and if
137 the policy noted reproducibility or analogous concepts. Finally, each journal was classified as
138 either open access or subscription-based on its inclusion in the Directory of Open Access
139 Journals database (Table 4).

140

141 *Statistical methods*

142 Continuous variables are summarized with medians and interquartile ranges (IQRs) denoting the
143 25th and 75th percentiles. Categorical variables are summarized with counts and percentages.
144 The variables IF and total citable items are not normally distributed (Shapiro Wilk's Test p-
145 values < 0.001), so medians are presented instead of means, and nonparametric methods are used
146 for statistical tests.

147

148 The association of IF with 6-level data sharing mark (DSM) was tested with a nonparametric
149 Kruskal-Wallis one-way analysis of variance (ANOVA) of IF in 2013 and 2014 with DSM as a
150 grouping factor. Post-hoc pairwise two-sample Wilcoxon tests were used to determine whether
151 the median IF for journals differ between the two level data sharing policy (required vs. not
152 required) categories. P-values from the Wilcoxon tests were adjusted for multiple comparisons
153 with the Holm procedure.

154

155 Pearson's chi-square test was used to test the association of data sharing policy (two levels:
156 required vs not required) and open access status. Fisher's Exact Test was used to test the
157 association of the 6-level DSM with open access status. Fisher's Test was used as opposed to
158 Chi-square test due to the low number of open access journals within some DSM categories. To
159 examine the association of open access status and data sharing weighted by publishing volume
160 we examined the number of citable items in each category and tested for the association of open
161 access and data sharing with Pearson's chi-square test.

162

163 All statistical analyses were performed with R version 3.2.1 [25]. All code and data to reproduce
164 these results can be found on GitHub (<https://github.com/OHSU-Ontology-Development-Group/DataSharingPolicies>).

165

166 Results

167 Of the 318 journals examined, 38 (11.9%) required data sharing as a condition of publication and
168 29 (9.1%) required data sharing, but made no explicit statement regarding the effect on
169 publication and editorial decisions. 74 (23.2%) journals explicitly encouraged or addressed data
170 sharing, but did not require it. And, 47 (14.8%) journals only addressed data sharing for
171 proteomic, genomic data, or other specific omics data (Figure 1 and Table 5).

172

173 In order to understand the potential influence of the policies on the published literature, we also
174 evaluated the distribution of publication volume by each data sharing mark. In 2013, the total
175 number of citable items (papers) in the studied journals was 130,330. In 2014, the total number
176 of citable items was 131,107. The median number of citable items per journal was 243.0 and
177 237.5, respectively (Table 5).

178

179 Table 5 shows the 2013 and 2014 publishing volume in citable items for each data sharing mark.
180 While it is likely that some of the journals in the study implemented or revised their data sharing
181 policies after 2014, the publishing volume data is current enough to provide an insight into the
182 potential influence of existing journal data sharing policies on the published literature.

183 While only 21% of the journals in the study required data sharing (DSM 1 and 2), these journals
184 published 42.1% of the citable items in 2013 and 2014 (23.6% and 24.9% of the citable items in
185 2013, 2014 after removing PLoS One) (Table 5).

186 The median 2013 journal IF for journals with the strongest data sharing policies (DSM 1) was
187 8.2; whereas, the median 2013 IF for journals with no mention of data sharing was 3.5. Figure 2
188 shows the median IF for each DSM category by report year. The IF was also analyzed by
189 collapsing the DSM into two categories: Required (DSM 1, 2) and Not Required (DSM 3, 4, 5,
190 6). The median 2013 IF for the journals that required data sharing was 6.8, and the median 2013
191 IF for the journals that did not require data sharing was 4.0.

192 Impact Factor is significantly associated with the six category data sharing marks (Kruskal-
193 Wallis rank sum test, 5 df, $p < 0.001$, 2013 and 2014). Examining pairwise differences between
194 DSM categories, we see that journals with DSM 1 have significantly higher IF than journals with
195 DSM 3, 4, 5, or 6 (Wilcoxon test, $p < 0.001$, < 0.001 , 0.04, < 0.001 ; 2013 data, 2014 similar).
196 Journals with DSM 2 have significantly higher IF than journals with DSM 3, 4, or 6 (Wilcoxon
197 test, $p = 0.034$, 0.0072, 0.0033; 2013 data, 2014 similar). Journals with DSM 5 have significantly
198 higher IF than journals with DSM 3, 4, and 6 (Wilcoxon test, $p = 0.0022$, < 0.001 , < 0.001 ; 2013
199 data, 2014 similar). In general, IF is not significantly different between DSM 1 and 2 and
200 between DSM 2 and 5, reflecting the similar IF for journals with explicit data sharing
201 requirements, either full or partial sharing. After collapsing DSM into two categories, required
202 (DSM 1, 2) and not required (DSM 3, 4, 5, 6), we still see a highly significant increase in IF for

203 journals with required data sharing (Wilcoxon Rank Sum Test, $p < 0.001$, 2013 and 2014 data)
204 (Figure 2).

205 Table 6 shows the count of subscription and open access journals for each DSM category, and
206 the count and percentage of subscription and open access journals for each DSM category. The
207 Fisher's Exact Test result, which yielded a p-value of 0.07, showed no significant association
208 between the DSM and a journal's access model. We also tested this association by collapsing the
209 DSM into two categories, required (DSM 1, 2) and not required (DSM 3, 4, 5, 6), and using a
210 Chi-square test. Again, no significant association was found (Chi-square Test, $df=1$, $p = 0.62$).
211 Both results suggest that journals with a data sharing requirement are not more likely to be open
212 access than journals without a data sharing requirement; nor are open access journals more likely
213 to have a data sharing requirements than subscription journals.

214 Although there was no significant association between open access and DSM at the journal level,
215 we observed a highly significant association at the citable item level (Chi-square Test, $df=1$, $p <$
216 $2e-16$). That is, a citable item that is open access is much more likely to be published in a journal
217 with a data sharing requirement (DSM 1 or 2). The proportion of open access journals that
218 require data sharing is much larger than the proportion of subscription journals (64.3% vs
219 11.3%). The very small p-value is partially due to the large number of total citable articles
220 studied and also due to the large proportion of open access citable items in PLoS One. However,
221 even with PLoS One removed from the analysis, an open access article is still more likely to
222 have been published in a journal with a data sharing requirement and the proportion of open
223 access journals versus subscription journals that require data sharing is 16.0% vs 11.3% (Chi-
224 square Test, $df=1$, $p < 2e-16$).

225 As illustrated in Figure 3, excluding those journals with no mention of data sharing (DSM 6),
226 57.6% (125) of the journals in the data set recommended data sharing via a public repository,
227 20.7% (45) recommended sharing via a journal hosted method, 1.8% (4) recommend sharing by
228 reader request to authors, 5.1% (11) state multiple equally recommended methods and 14.8%
229 (32) do not specify.

230
231 Of the journals requiring data sharing (DSM 1 or 2), 85% (57) recommend data sharing via a
232 public repository. Of the journals that recommended data sharing via a journal hosted method,
233 the majority, 88.8% (40), did not specify any size limitations.

234
235 Only 7.3% (16) journals that addressed data sharing (DSM 1,2,3, 4, and 5) explicitly mentioned
236 copyright or licensing considerations. Even for those journals that required data sharing (DSM 1
237 or 2), only 16.4% (11) mentioned copyright or licensing; however, these journals published
238 31.9% of the citable items in 2013 of the journals that addressed data sharing. Only 2 journals in
239 the entire data set addressed how long the data should be retained.

240

241 In light of its frequently used justification, we also coded the data sharing policies for a mention
242 of scientific reproducibility or analogous concepts. Reproducibility or similar language was
243 mentioned by 16.9% (54) of the total studied journals. Of the journals requiring data sharing
244 (DSM 1 or 2), 65.5% (44) mentioned the concept of reproducibility.

245 **Discussion**

246 Publishers have an influential role to play in promoting, facilitating, and enforcing data sharing
247 [12,17]. However, only a minority of the journals analyzed for this this study required data
248 sharing. While the capacity of the existing policies is more promising if considered from the
249 perspective of publishing volume, our results were consistent with other examinations of data
250 sharing policies [26–28]. Like Piwowar and Chapman [26], we found that a large proportion of
251 the journals we examined (40%) required the deposition of omics data to specific repositories.
252 Less frequent and more varied, however, were requirements that addressed data in general. The
253 higher prevalence of omics data sharing requirements we observed may be due to the more
254 mature guidelines, reporting standards, and centralized repositories for omics data types [26,29–
255 31]. The further development and implementation of well communicated best practices and
256 resources for general data types, could be a means for increasing the prevalence and strength of
257 journal data sharing requirements and ensuring compliance [17].

258

259 While a problematic and often abused proxy for quality, the IF is closely associated with a
260 journal's prestige [32]. It influences publication decisions and the perceived significance of
261 individual papers [32]. Because of its impact on scholarly communication, it is noteworthy that
262 there was a significantly higher IF associated with the journals with a data sharing requirement.
263 This result was similar to other studies [26,33–35]. As has been noted, prestigious journals may
264 be better positioned and be more willing to impose new requirements and practices on authors
265 [17,26].

266

267 The importance and benefits of data sharing are often linked to and discussed within the larger
268 context of open access to research results, specifically the published literature. Public access to
269 both peer reviewed articles and data are regarded as necessary elements for addressing problems
270 within the scientific enterprise and realizing the full value of research investments [2]. While we
271 found that an open access citable item is much more likely to be published in a journal with a
272 data sharing requirement, we did not find that open access journals are any more likely to require
273 data sharing than subscription journals. This result is in contrast with a previous finding from
274 Piwowar and Chapman [26]. However, we analyzed a greater number of journals and a greater
275 number of open access journals. We hypothesize some open access journals may be less willing
276 to impose additional requirements, because they lack the prestige or prominence of more
277 established journals and publishers. Smaller and independent open access journals may also lack
278 the resources to facilitate and enforce data sharing.

279

280 How data is managed and shared affects its value. If a data set is difficult to retrieve or
281 understand, for example, replication studies can't be performed and researchers can't use the
282 data to investigate new questions. While 65.7% of the journals in our study that required data
283 sharing addressed the concept of reproducibility, as with earlier investigations [26,34] we found
284 that most data sharing policies did not provide specific guidance on the practices that ensure data
285 is maximally available and reusable [36][37]. For example, the majority of journals that
286 addressed data sharing (DSM 1-5) recommended depositing data in a public repository; however,
287 only a handful of journals provided guidelines or requirements related to licensing considerations
288 or retention timeframes. While a higher IF was associated with the presence of a data sharing
289 requirement, overall the policies did not provide guidelines or specificity to facilitate
290 reproducible and reusable research. This result is similar to a previous study in which we showed
291 that the majority of biomedical research resources are not uniquely identifiable in the biomedical
292 literature, regardless of journal Impact Factor [38].

293
294 Our study confirms earlier investigations which observed that only a minority of biomedical
295 journals require data sharing, and a significant association between higher Impact Factors and
296 journals with a data sharing requirement. Our approach, however, included several limitations.
297 Only journals in the top quartiles by volume or Impact Factor for the World of Science
298 categories we identified as belonging to the biomedical corpus were analyzed, which introduced
299 some inherent biases. Additionally, in hindsight, it would have been valuable to have
300 systematically analyzed more nuanced aspects of the policies' quality characteristics, such as
301 whether minimal information or metadata standards were addressed and if the shared data was
302 reviewed in the peer review process. Finally, it should be noted that many of the policies we
303 reviewed were difficult to interpret. While the study's authors are confident that the data sharing
304 scores we assigned reflect the most accurate interpretation of each journal's policy at the time of
305 our data collection, the policies in general included ambiguous and fragmented information. It is
306 possible, therefore, that there are gaps between the scores we assigned to some policies and their
307 editorial intent.

308
309 As a continuation of this work, several follow-up activities are being pursued. We plan to build a
310 community curated and regularly updated public database of journal data sharing policies. In
311 addition to providing a searchable resource of journal data sharing policies, the database's
312 curation schedule will facilitate an understanding of policy changes over time and inform them.
313 A follow-up study will look at the data availability for articles associated with the journals in this
314 study. Finally, building upon recommendations outlined by the Journal Research Data (JoRD)
315 Project [34] and Lin and Strasser [17], we intend to convene a community of stakeholders to
316 further work on recommendations and template language for strengthening and communicating
317 journal data sharing policies. Maximally available and reusable data will not be achieved via the
318 implementation of vague data sharing policies that lack specific direction on where data should

319 be shared, how it should be licensed, or the ways in which it should be described. On the
320 contrary, such specificity is essential.

321 **Conclusions**

322 We observed a two-pronged problem with journal data sharing policies. First, given the attention
323 the benefits of data sharing have received from the biomedical community, it is problematic only
324 a minority of journals have implemented a strong data sharing requirement. Second, among the
325 policies that do exist, guidelines vary and are relatively ambiguous. Overall, the biomedical
326 literature is lacking policies that would ensure that the data underlying it is maximally available
327 and reusable.

328

329 This is problematic in regards to affecting the kinds of outcomes and improvements open data is
330 supposed to facilitate. This study adds to a growing body of work aimed at analyzing and
331 improving journal data sharing policies.

332 **Acknowledgements**

333 Thanks to the following colleagues for their curation assistance or visualization advice: Steven
334 Bedrick, PhD; Heather Coates, MLIS; Jill Emery, MLIS; Erin Foster, MLIS; Danielle Robinson;
335 Chris Shaffer, MLIS; Kate Thornhill, MLIS; Jackie Wirz, PhD.

336 **References**

337

- 338 1. NIH (2016) Principles and Guidelines for Reporting Preclinical Research | National
339 Institutes of Health (NIH). Available: [https://www.nih.gov/research-training/rigor-](https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research)
340 [reproducibility/principles-guidelines-reporting-preclinical-research](https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research). Accessed 20 October
341 2016.
- 342
- 343 2. Holdren JP (2012) Increasing Access to the Results of Federally Funded Scientific
344 Research. Available:
345 [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)
346 [2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- 347
- 348 3. Research Councils UK (2011) RCUK Common Principles on Data Policy - Research
349 Councils UK. Available: <http://www.rcuk.ac.uk/research/datapolicy/>. Accessed 25 October
350 2016.
- 351
- 352 4. Drazen JM, Morrissey S, Malina D, Hamel MB, Champion EW (2016) The Importance -
353 and the Complexities - of Data Sharing. *N Engl J Med* 375: 1182–1183.
354 doi:10.1056/NEJMe1611027.
- 355

- 356 5. Medium.com (2016) Inspiring a New Generation to Defy the Bounds of Innovation: A
357 Moonshot to Cure Cancer – Cancer Moonshot. Available: [https://medium.com/cancer-](https://medium.com/cancer-moonshot/inspiring-a-new-generation-to-defy-the-bounds-of-innovation-a-moonshot-to-cure-cancer-fbdf71d01c2e#.gx72sbluo)
358 moonshot/inspiring-a-new-generation-to-defy-the-bounds-of-innovation-a-moonshot-to-
359 cure-cancer-fbdf71d01c2e#.gx72sbluo. Accessed 5 November 2016.
360
- 361 6. Borgman CL (2012) The conundrum of sharing research data. *Acta Anaesthesiol Scand* 63:
362 1059–1078. doi:10.1002/asi.22634.
363
- 364 7. Collins FS, Tabak LA (2014) Policy: NIH plans to enhance reproducibility. *Nature* 505:
365 612–613. doi:10.1038/505612a.
366
- 367 8. Fischer BA, Zigmond MJ (2010) The essential nature of sharing in science. *Sci Eng Ethics*
368 16: 783–799. doi:10.1007/s11948-010-9239-x.
369
- 370 9. NSF (n.d.) Dissemination and Sharing of Research Results | NSF - National Science
371 Foundation. Dissemination and Sharing of Research Results. Available:
372 <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>. Accessed 20 October 2016.
373
- 374 10. NIH (n.d.) National Institutes of Health Genomic Data Sharing Policy. National Institutes
375 of Health Genomic Data Sharing Policy. Available: <https://gds.nih.gov/03policy2.html>.
376 Accessed 20 October 2016.
377
- 378 11. European Commission (2016) FAIR Data Management in Horizon 2020. Available:
379 <http://aims.fao.org/activity/blog/guidelines-fair-data-management-horizon-2020>. Accessed
380 20 October 2016.
381
- 382 12. Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Matthews B, et al. (2014) Enabling
383 Sharing and Reuse of Scientific Data. *New Review of Information Networking* 19: 16–43.
384 doi:10.1080/13614576.2014.883936.
385
- 386 13. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data sharing by
387 scientists: practices and perceptions. *PLoS ONE* 6: e21101.
388 doi:10.1371/journal.pone.0021101.
389
- 390 14. Longo DL, Drazen JM (2016) Data Sharing. *N Engl J Med* 374: 276–277.
391 doi:10.1056/NEJMe1516564.
392
- 393 15. LeClere F (2010) Too many researchers are reluctant to share their data. *The Chronicle of*
394 Higher Education. Available:
395 <http://www.scienceofsciencepolicy.net/sites/default/files/attachments/Too%20Many%20Re>

- 396 searchers%20Are%20Re...pdf.
397
- 398 16. Savage CJ, Vickers AJ (2009) Empirical study of data sharing by authors publishing in
399 PLoS journals. PLoS ONE 4: e7078. doi:10.1371/journal.pone.0007078.
400
- 401 17. Lin J, Strasser C (2014) Recommendations for the role of publishers in access to data.
402 PLoS Biol 12: e1001975. doi:10.1371/journal.pbio.1001975.
403
- 404 18. Nature Publishing Group (2013) Reporting checklist for life sciences articles. Available:
405 <http://www.nature.com/authors/policies/checklist.pdf>.
406
- 407 19. AAAS S (n.d.) Science: editorial policies | Science | AAAS. Available:
408 <http://www.sciencemag.org/authors/science-editorial-policies>. Accessed 21 October 2016.
409
- 410 20. Nature (n.d.) Availability of data, material and methods. Available:
411 <http://www.nature.com/authors/policies/availability.html>. Accessed 21 October 2016.
412
- 413 21. PLoS (n.d.) PLoS Data Availability. Available: [http://journals.plos.org/plosone/s/data-](http://journals.plos.org/plosone/s/data-availability)
414 [availability](http://journals.plos.org/plosone/s/data-availability). Accessed 21 October 2016.
415
- 416 22. The Royal Society (n.d.) Royal Society Data Sharing and Mining. Available:
417 <https://royalsociety.org/journals/ethics-policies/data-sharing-mining/>. Accessed 21 October
418 2016.
419
- 420 23. Clarivate Analytics (n.d.) Journal Citation Reports. Available:
421 <http://ipscience.thomsonreuters.com/product/journal-citation-reports/>. Accessed 3
422 November 2016.
423
- 424 24. Stodden V, Guo P, Ma Z (2013) Toward Reproducible Computational Research: An
425 Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS ONE 8: e67111.
426 doi:10.1371/journal.pone.0067111.
427
- 428 25. R Foundation for Statistical Computing RCT (n.d.) R: A Language and Environment for
429 Statistical Computing. Available: <https://www.R-project.org>. Accessed 2 November 2016.
430
- 431 26. Piwowar, Chapman HA, Wendy W (2008) A review of journal policies for sharing
432 research data. Available:
433 <http://ocs.library.utoronto.ca/index.php/Elpub/2008/paper/view/684/0>.
434
- 435 27. Barbui C (2016) Sharing all types of clinical data and harmonizing journal standards. BMC

- 436 Med 14: 63. doi:10.1186/s12916-016-0612-8.
437
- 438 28. McCain KW (1995) Mandating Sharing: Journal Policies in the Natural Sciences. *Sci*
439 *Commun* 16: 403–431. doi:10.1177/1075547095016004003.
440
- 441 29. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum
442 information about a microarray experiment (MIAME)-toward standards for microarray
443 data. *Nat Genet* 29: 365–371. doi:10.1038/ng1201-365.
444
- 445 30. Hrynaszkiewicz I, Khodiyar V, Hufton AL, Sansone S-A (2016) Publishing descriptions of
446 non-public clinical datasets: proposed guidance for researchers, repositories, editors and
447 funding organisations. *Research Integrity and Peer Review* 1. doi:10.1186/s41073-016-
448 0015-6.
449
- 450 31. Piwowar HA, Chapman WW (2010) Public sharing of research datasets: a pilot study of
451 associations. *Journal of informetrics* 4: 148–156. doi:10.1016/j.joi.2009.11.010.
452
- 453 32. Lariviere V, Kiermer V, MacCallum CJ, McNutt M, Patterson M, et al. (2016) A simple
454 proposal for the publication of journal citation distributions. *BioRxiv*. doi:10.1101/062109.
455
- 456 33. Stodden V, Guo P, Ma Z (2012) How Journals are Adopting Open Data and Code Policies.
457 In: N A, editor. *Proceedings for the Governing Pooled Knowledge Resources*. Digital
458 Library of the Commons: University of Indiana. Available:
459 <http://hdl.handle.net/10535/9584>.
460
- 461 34. Sturges P, Bamkin M, Anders JHS, Hubbard B, Hussain A, et al. (2015) Research data
462 sharing: Developing a stakeholder-driven model for journal policies. *Journal of the*
463 *Association for Information Science and Technology* 66: 2445–2455.
464 doi:10.1002/asi.23336.
465
- 466 35. Magee AF, May MR, Moore BR (2014) The dawn of open access to phylogenetic data.
467 *PLoS ONE* 9: e110268. doi:10.1371/journal.pone.0110268.
468
- 469 36. DataONE (n.d.) All Best Practices | DataONE. Available: [https://www.dataone.org/all-](https://www.dataone.org/all-best-practices)
470 [best-practices](https://www.dataone.org/all-best-practices). Accessed 25 October 2016.
471
- 472 37. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, et al. (2015) Achieving human and
473 machine accessibility of cited data in scholarly publications. *PeerJ Computer science* 1.
474 doi:10.7717/peerj-cs.1.
475

476 38. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, et al. (2013) On the
477 reproducibility of science: unique identification of research resources in the biomedical
478 literature. PeerJ 1: e148. doi:10.7717/peerj.148.
479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507
508
509
510
511
512
513
514
515
516

Figure 1

Figure 1: Percentage of journals per each data sharing mark (DSM).

The top bar shows the percentage of all journals for each data sharing mark. The middle bar shows the percentage of citable items from each journal (including PLoS One) for each data sharing mark. The lower bar shows the percentage of citable items for each journal (excluding PLoS One) for each data sharing mark. Because of the journal PLoS One's high publishing activity, we analyzed the percentage of citable items for each data sharing mark including and excluding PLoS One.

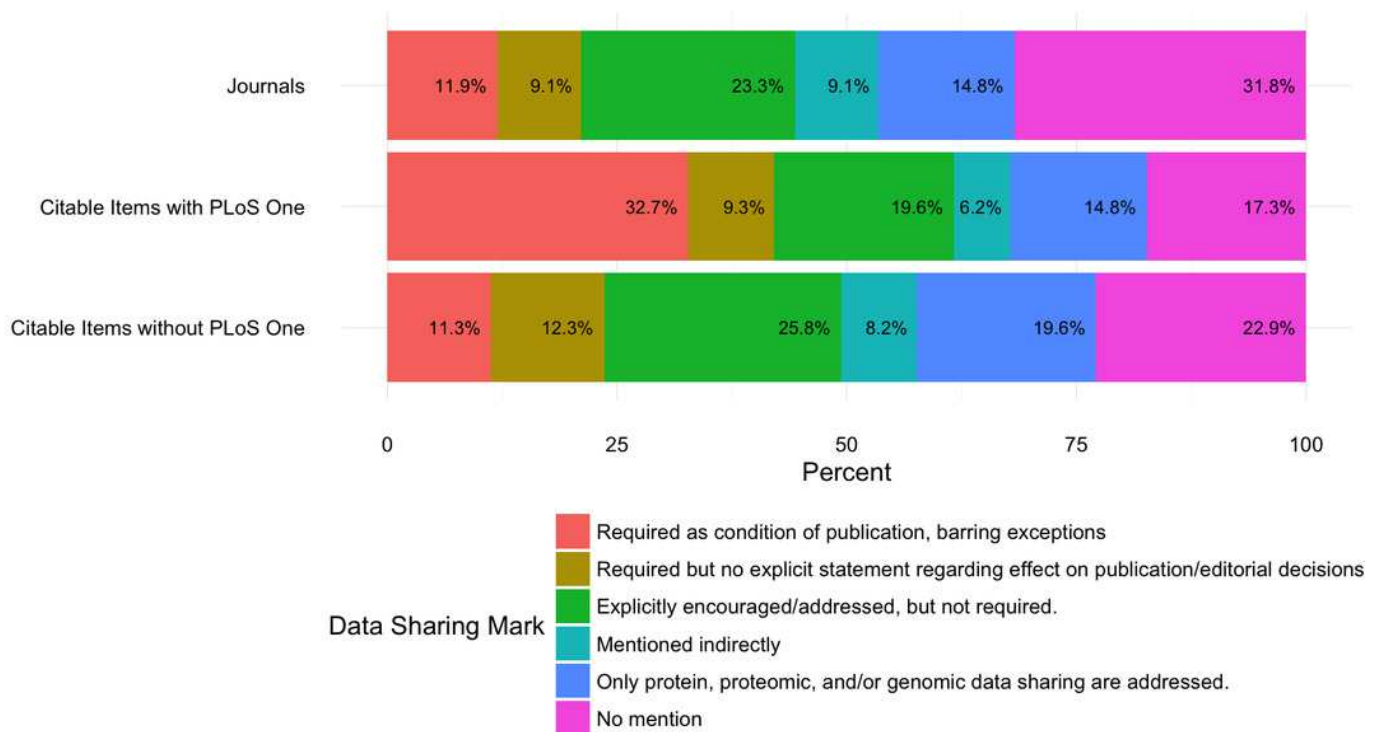


Figure 2

Figure 2: Impact factors were higher for journals with the strongest data sharing policies (DSM 1) compared to journals with no mention of data sharing (DSM 6).

The median Impact Factor was calculated for the journals with each data sharing mark for each report year (light color=2013, dark color=2014). The lower and upper hinges of the boxplots represent the first and third quartiles of journal Impact Factor, the horizontal line represents the median, the triangle represents the mean, and the upper and lower whiskers extend from the hinge to the highest (lowest) value that is within 1.5 times the interquartile range of the hinge, with journals outside this range represented as points.

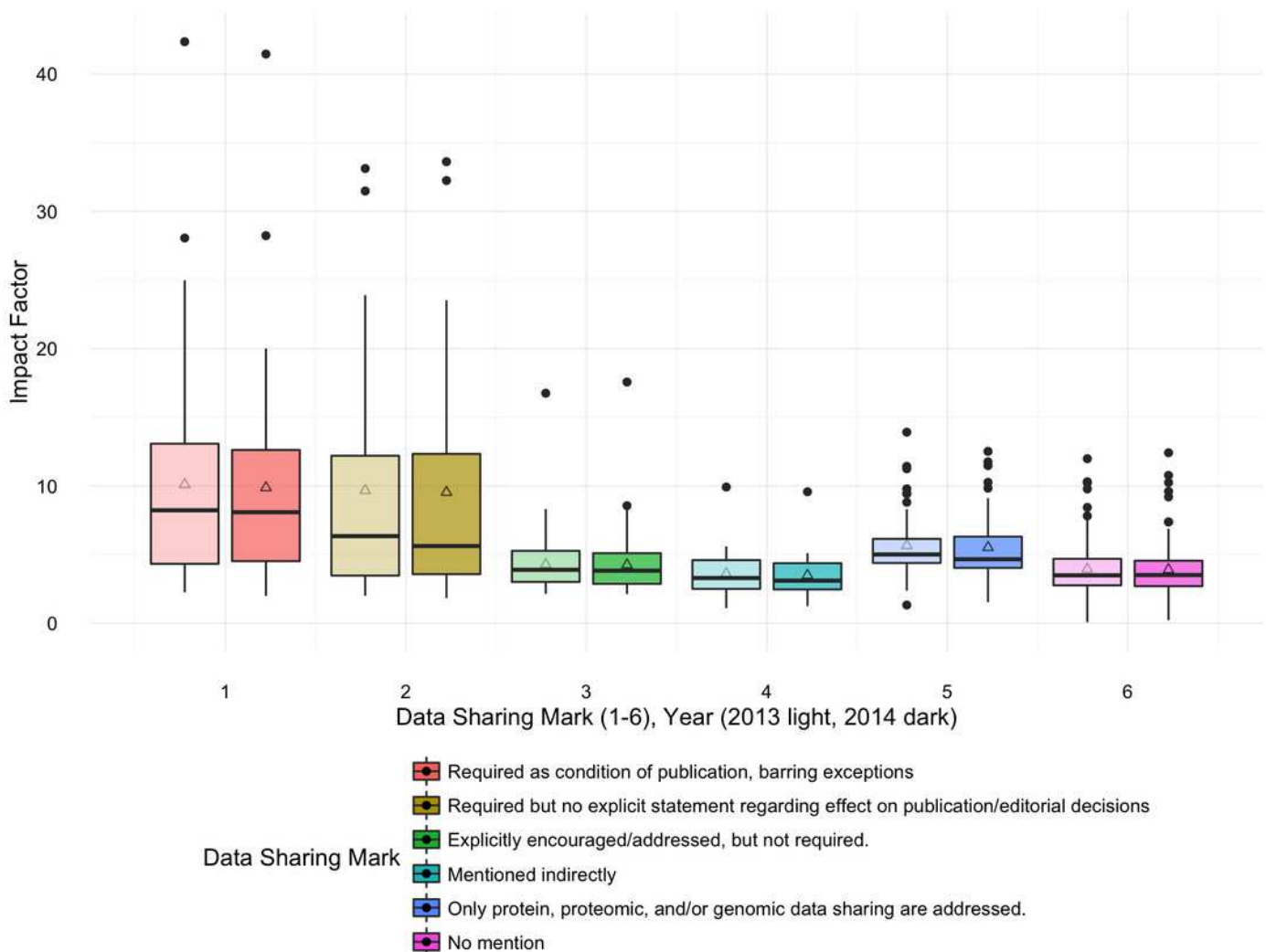


Figure 3

Figure 3: Recommended data sharing method by data sharing mark (DSM) 1-5.

The number (percent) of journals with each recommended data sharing method is represented by each tile, with brighter blue shades denoting higher percentages of journals with the given data sharing method.

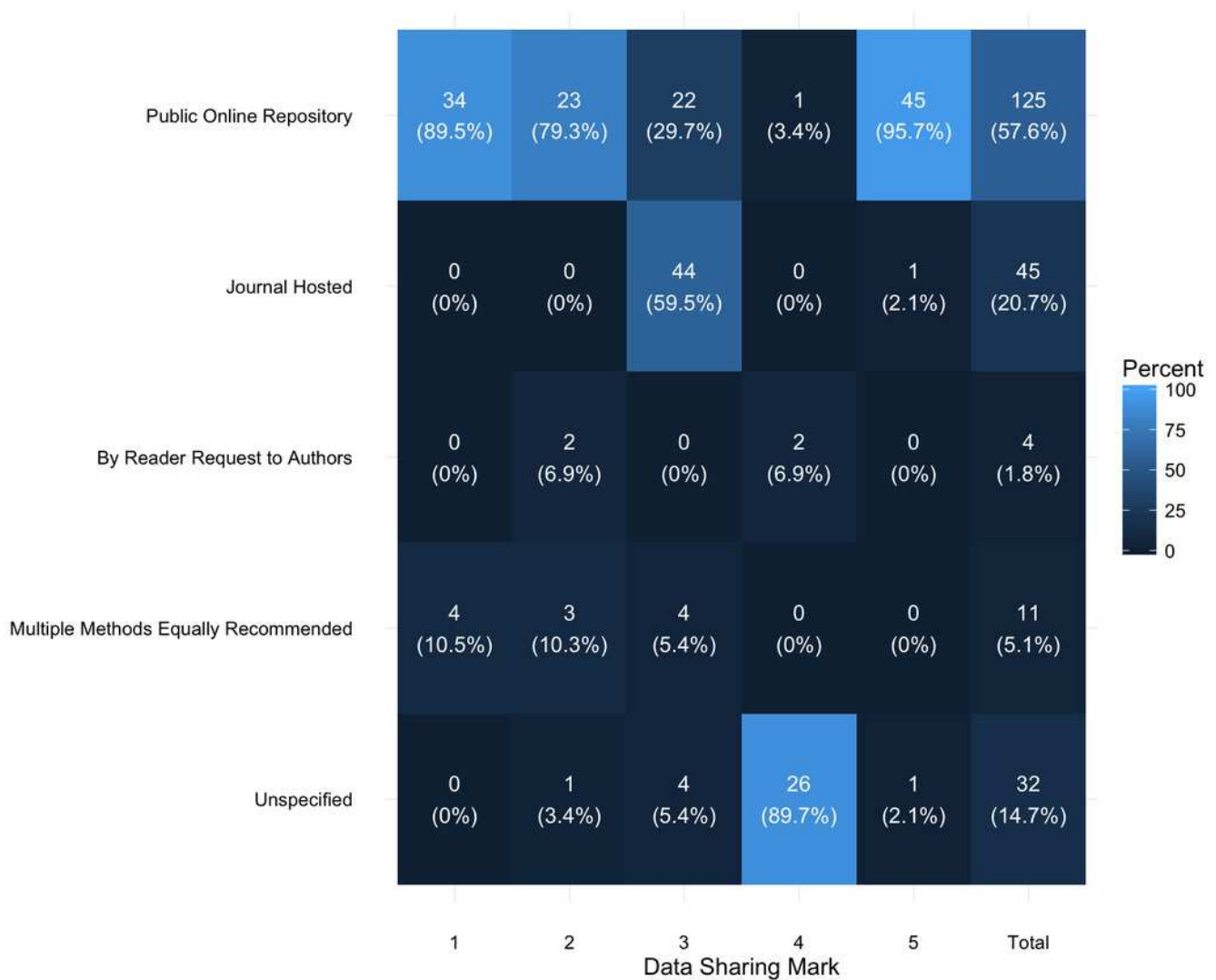


Table 1 (on next page)

Table 1: Journal Impact Factor Category

1

Table 1: Journal Impact Factor Category	<i>N (%)</i>
<2	19 (6%)
2-3.99	125 (39.3%)
4-5.99	102 (32.1%)
6-7.99	25 (7.9%)
8-9.99	15 (4.7%)
10-29.99	29 (9.1%)
30-43	3 (0.9%)

2

Table 2 (on next page)

Table 2: Number of citable items per journal

1

Table 2: Number of citable items per journal	<i>N (%)</i>
<100	42 (13.2%)
100-500	239 (75.2%)
500-1000	28 (8.8%)
1000-32000	9 (2.8%)

2

3

Table 3 (on next page)

Table 3: Journal scoring rubric used in this study, adapted from Stodden et al., 2013.

Data Sharing Mark	
1	Required as condition of publication, barring exceptions
2	Required but, no explicit statement regarding effect on publication/editorial decisions
3	Explicitly encouraged/addressed, but not required.
4	Mentioned indirectly
5	Only protein, proteomic, and/or genomic data sharing are addressed.
6	No mention
Journal Access Mark (Whole Journal Model, Does Not Consider Hybrid Publishing)	
1	Open access
0	Subscription
Protein, Proteomic, Genomic Data Sharing Required with Deposit to Specific Data Banks	
a	Yes
b	No
Recommended Sharing Method	
A	Public Online Repository
B	Journal Hosted
C	By Reader Request to Authors
D	Multiple methods equally recommended
E	Unspecified
If Journal Hosted	
a	Journal will host regardless of size
b	Journal has data hosting file/s size limit
c	Unspecified
Copyright/Licensing of Data	
a	explicitly stated or mentioned
b	no mention

Archival/Retention Policy (Statement about how long the data should be retained).	
a	explicitly stated
b	no mention
Reproducibility or Analogous Concepts Noted as Purpose of Data Policy	
a	explicitly stated
b	no mention

1

Table 4 (on next page)

Table 4: Number of journals per open access

1

Table 4: Number of journals per open access							
<i>Open Access</i>	<i># Journals (%)</i>	<i>Median # Citable Items per Journal 2013</i>	<i># Citable Items 2013 (%)</i>	<i># Citable Items 2013, Remove PLoS One (%)</i>	<i>Median # Citable Items per Journal 2014</i>	<i># Citable Items 2014 (%)</i>	<i># Citable Items 2014, Remove PLoS One (%)</i>
Open Access	44 (13.8%)	199.5	43789 (33.6%)	12293 (12.4%)	207	45831 (35.0%)	15791 (15.6%)
Subscription	274 (86.2%)	246.5	86541 (66.4%)	86541 (87.6%)	240	85276 (65.0%)	85276 (84.4%)

2

Table 5 (on next page)

Table 5: Publishing Volume by Data Sharing Mark

Table 5: Publishing Volume by Data Sharing Mark								
DSM	DSM Description	# Journals (%)	Median # Citable Items per Journal 2013	# Citable Items 2013 (%)	# Citable Items 2013, Remove PLoS One (%)	Median # Citable Items per Journal 2014	# Citable Items 2014 (%)	# Citable Items 2014, Remove PLoS One (%)
1	Required as condition of publication, barring exceptions	38 (11.9%)	230.5	42,669 (32.7%)	11,173 (11.3%)	220	42,794 (32.6%)	12,754 (12.6%)
2	Required but no explicit statement regarding effect on publication/editorial decisions	29 (9.1%)	209	12,138 (9.3%)	12,138 (12.3%)	227	12,436 (9.5%)	12,436 (12.3%)
3	Explicitly encouraged/addressed, but not required.	74 (23.3%)	259.5	25,519 (19.6%)	25,519 (25.8%)	282.5	26,026 (19.9%)	26,026 (25.8%)
4	Mentioned indirectly	29 (9.1%)	256	8,062 (6.2%)	80,62 (8.2%)	225	7,894 (6%)	7,894 (7.8%)
5	Only protein, proteomic, and/or genomic data sharing are addressed.	47 (14.8%)	277	19,339 (14.8%)	19,339 (19.6%)	316	19,080 (14.6%)	19,080 (18.9%)
6	No mention	101 (31.8%)	211	22,603 (17.3%)	22,603 (22.9%)	213	22,877 (17.4%)	22,877 (22.6%)
<i>Publishing Volume by Data Sharing Requirement</i>								
DSM 1&2	Required	67 (21.1%)	226	54,807 (42.1%)	23,311 (23.6%)	221	55,230 (42.1%)	25,190 (24.9%)
DSM 3-6	Not Required	251 (78.9%)	248	75,523 (57.9%)	75,523 (76.4%)	244	75,877 (57.9%)	75,877 (75.1%)

<i>Publishing Volume in All Journals</i>								
Total	All Journals	318 (100%)	243	130,330 (100%)	98,834 (100%)	237.5	131,107 (100%)	101,067 (100%)

2

Table 6 (on next page)

Table 6: Open Access Journals and Citable Items by Data Sharing Mark

1

Table 6: Open Access Journals and Citable Items by Data Sharing Mark	<i>Subscription</i>	<i>Open Access</i>	<i>% Open Access</i>
<i>Data Sharing Mark</i>	<i># Journals (# Citable Items)</i>	<i># Journals (# Citable Items)</i>	<i>% Journals (% Citable Items)</i>
1- Required as condition of publication, barring exceptions	29 (7,709)	9 (34,960; 3464*)	23.7% (81.9%; 31%*)
2- Required but no explicit statement regarding effect on publication/editorial decisions	27 (11,864)	2 (274)	6.9% (2.3%)
3- Explicitly encouraged/addressed, but not required.	63 (22,884)	11 (2,635)	14.9% (10.3%)
4- Mentioned indirectly	29 (8,062)	0 (0)	0% (0%)
5- Only protein, proteomic, and/or genomic data sharing are addressed.	40 (17,401)	7 (1,938)	14.9% (10.0%)
6- No mention	86 (18,621)	15 (3,982)	14.9% (17.6%)
<i>Data Sharing Requirement</i>			
DSM 1&2 - Required	56 (19,573)	11 (35,234; 3,738*)	16.42% (64.29%; 16.04%*)
DSM 3-6 - Not Required	218 (66,968)	33 (8,555)	13.15% (11.33%)

2 * After removing PLoS One

3