EVALUATING AND PROMOTING OPEN DATA PRACTICES IN OPEN ACCESS JOURNALS

Eleni Castro
OpenBU and ETD Program Librarian, Boston University
<elenic@bu.edu>

Mercè Crosas
Director of Data Science, Institute for Quantitative Social Science (IQSS);
Harvard University
<mcrosas@iq.harvard.edu>

Alex Garnett
Research Data Management & Systems Librarian,
Simon Fraser University
<garnett@sfu.ca>

Kasey Sheridan
Research & Acquisitions Librarian at SunTrust
<kasey.sheridan@gmail.com>

Micah Altman, Head/Scientist, Program on Information Science;
Director of Research, MIT Libraries;
Massachusetts Institute of Technology
<escience@mit.edu>

[*For Submission to JSP -- Preprints Through PeerJ; SSRN*]

Abstract (150 words or less)

In the last decade there has been a dramatic increase in attention from the scholarly communications and research community to open access (OA) and open data practices. These are potentially related, because journal publication policies and practices both signal disciplinary norms, and provide direct incentives for data sharing and citation. However, there is little research evaluating the data policies of OA journals. In this study, we analyze the state of data policies in open access journals, by employing random sampling of the Directory of Open Access Journals (DOAJ) and Open Journal Systems (OJS) journal directories, and applying a coding framework that integrates both previous studies and emerging taxonomies of data sharing and citation. This study, for the first time, reveals both the low prevalence of data sharing policies and practices in OA journals, which differs from the previous studies of commercial journals' in specific disciplines.

Keywords (4-6 for indexing purposes)

Open Access; Open Data; Data Sharing; Data Citation

Introduction

With the Open Access (OA) movement celebrating its 15th anniversary in 2017, and Open Data (OD) moving more and more towards becoming an established research practice, we have sufficient information to observe how these two worlds intersect when it comes to the data sharing and citation practices of OA journals. In this study, we look at the prevalence of data sharing policies and analyze the overall data sharing and citation characteristics of OA journals. In our research we review previous studies on journal data policies from various disciplines, along with related efforts from the research data community, scholarly societies, funders, and flagship journals to help understand the prevalence of policies and best practices with regards to data sharing and citation. While a number of studies have analyzed data sharing policies and practices in scholarly journals (mostly commercial) within specific disciplines, little is known about the overall prevalence and characteristics of OA journals' data policies. In this study, we evaluate the state of data policies in open access journals, by employing random sampling of the Directory of Open Access Journals (DOAJ) and Open Journal Systems (OJS) journal directories; and applying a coding framework that integrates both previous studies and emerging taxonomies of data sharing and citation.

The Increasing Importance of Open Access and Open Data

Since the 2002 Budapest Open Access Initiative (BOAI), which provided a name to the free online sharing of research, there are now (as of March 2017) over 9,419 open access, peer-reviewed journals listed in the DOAJ. According to Peter Suber, director of the Harvard

Open Access Project, and one of the de facto leaders of the OA movement, “OA makes knowledge a public good in practice” and allows researchers to freely share knowledge and accelerate research without the economic and sharing restrictions put in place by the commercial publishing model (2012). Several recent studies have also shown that there are clear incentives for authors to choose OA publishing over the traditional commercial model. For example, Wagner’s (2010) annotated bibliography lists 39 studies where researchers had found a significant citation advantage for open access articles. Since then SPARC Europe (2015) has listed on their website 70 studies, of which 49 observed a citation advantage for articles published in OA journals from various disciplines. Furthermore, given the continuous rise in subscription costs for scholarly journals (Chavez, 2010), it is becoming increasingly important for OA journals to help make research more widely accessible in order for scholarly communication to thrive in the 21st century.

In parallel to the OA movement, the relevance and need for Open Data (OD)[1] as a tool for improving scholarly communication and research has also increased in importance over the past decade. This increase in importance can especially be seen with recent state, funder, and scholarly society mandates for researchers to make openly available their publicly funded data for other researchers to access and reuse (Jones, 2015). Even OA scholarly societies such as the Scholarly Publishing and Academic Resources Coalition (SPARC) include OD as part of their current policy priorities. SPARC’s 2017[2] plan discusses the need to "promote partnerships that leverage resources to sustain crucial infrastructure supporting Open Access, Open Data, and OER". More broadly, Open Data is claimed by advocates to promote transparency, innovation, and efficiency within the public and private sector. However, despite widespread support for data sharing, recent research has found that most academic researchers are not making their research data available to others, and that more direct incentives are needed to encourage data sharing (Fecher, Friesike, & Hebing, 2015).

Given the overall lack of strong data sharing policies in scholarly journals, which require authors to submit data with their article, OA journals can play a critical role in helping researchers openly publish research data associated with their articles (Gherghina & Katsanidou, 2013; McCullough, 2009). Over 15 years ago, OA journals had started a paradigm shift in publishing, and since they are already the best advocate for the public availability of research articles they can do the same with pushing for OD. This is made evident with, one of the most influential OA publishers, PLOS’ introduction of a data policy in 2014, which directly connected the OA mission to sharing data by stating that: “Access to research results, immediately and without restriction, has always been at the heart of PLOS’ mission and the wider Open Access movement.” Given the potential for OA to support OD and the alignment of interests, it is important to evaluate the current state of data sharing in OA journals. In the next section we

[1] According to the Open Knowledge Foundation’s Open Data Handbook (2015), “Open Data” is defined as that is open, without any restrictions on re-use.
[2] SPARC Open Data <https://sparcopen.org/open-data/>

review key elements of data policies in journals, and then apply this categorization to a sample of OA journals.

Key Characteristics of Formal Data Citation and Sharing Policies

Within the past decade, several studies have analyzed and characterized data citation and data sharing policies of mostly commercial journals within specific domains and sub-disciplines. Throughout these studies -- which are described in more detail below -- a number of key elements are shared and can be used to put together a discipline agnostic rubric to review OA journal policies from various disciplines.

Life & Environmental Sciences

In the biomedical domain, there have been multiple studies looking at journal policies related to data sharing and citation. Piwowar & Chapman (2008) looked at high-level characteristics of data sharing policies (i.e., policy was absent, weak, or strong) in journals from 2006, which primarily published articles on "gene expression profiling" and their policies on sharing microarray data. They found that at the time only 6% used an OA publishing model and so most of the journals analyzed were commercial. Their findings showed that data sharing prevalence was quite low, even for journals with very strict sharing requirements (persistent identifier or accession number *prior to publication)*. A few years later, Stodden, Guo & Ma (2012) found that some journals in bioinformatics and life sciences were: 1) making their requirements stricter by requiring data as a condition for publication (barring exceptions); 2) including a policy for sharing code, which would help with verifiability and allow others to more easily reuse this data or replicate the results. With regards to OA journals, the authors opined that with such small changes to open access policy from 2011 to 2012 it did not appear to be driving changes in data and code sharing policies. In a more recent study by Vasilevsky et al (2016), which used an adapted rubric from Stodden, Guo & Ma, they confirmed the results of earlier research that only a small number of biomedical journals require data sharing. They also found that OA journals “were not more likely to require data sharing than subscription journals”, and that most data sharing policies lacked any specific guidance on how to make data available and reusable. With respect to incentives, OD policies can be further strengthened by the results found in the Piwowar & Vision (2013) study, which concluded that for at least gene expression microarray data there is a robust citation benefit from OD and that it has been steadily increasing since 2003.

Within the Environmental Sciences, a 2010 study by Weber, Piwowar & Vision looked for the presence of data sharing and citation policies in journals, and discovered that some journals were also explicitly indicating where researchers should deposit and archive their data, as well as offering peer review guidelines. Furthermore, their analysis found that an overwhelming majority of journals (7 out of every 8) “fail to provide explicit directions for sharing and citing data.” The authors concluded that funding agencies and journals could encourage researchers to share more if they required data as a condition for publication, provided them with some guidelines or best practices, and most importantly made them more aware of the benefits of sharing, such as increased citation rates.

Social Sciences

Similarly, studies on data sharing practices in sub-disciplines within Social Sciences looked at the prevalence of data sharing and citation policies in scholarly journals. However, unlike other disciplines these studies also focused on the importance of replication policies, where in addition to transparency verifiability is also important in reviewing the quality of the research (King, 1995; McCullough, 2009; Ishiyama, 2014).

Despite the abovementioned studies stressing the importance of replication policies, a study by Gherghina & Katsanidou (2013) found that only 18/120 political science and international relations journals had such a policy. They also found that while many journals had mandatory data sharing policies not many of them provided specific guidelines for when and where to deposit data for long-term preservation and access. However, most of the journals they looked at did provide authors with guidelines on what they should make available (raw data, documentation, code, etc). In addition, a 2016 study by Key found that the strongest predictor of availability is whether a journal has a policy mandating that data and/or code be made publicly available at the time of publication (p.270).

Within the field of Economics, building on the work of the US economist B.D. McCullough, Vlaeminck (2013) found that out of 141 journals, 20.6% of them (only one of these was OA) had a data availability policy and even less (7.8%) had a replication policy. In addition to studying the extent of such policies, Vlaeminck also looked at the quality of the available policies. The author found that the majority of journals which had policies followed a similar policy to journals published by the American Economic Association (AEA) by making data submission mandatory (whenever possible) and specified exactly what data and files should be submitted to the journal prior to publication. Furthermore, although there were some journals which had a replication policy, none of them had any dedicated replication sections within their journals which would provide an additional incentive for authors to put in the effort to provide replication data for their articles.

Data & Methods

Since we wanted to look at the prevalence of data policies in OA scholarly journals in general, we selected the Directory of Open Access Journals (DOAJ) as a sampling frame, which is an actively maintained and well-established OA journal index with clear inclusion criteria. Once we defined our population of OA journals, we conducted a simple random sample of all scholarly journals in the DOAJ removing any that were predatory[3], theoretical, non-empirical, or non-English. As a comparison to see if we would find a similar prevalence in policies from a different data source, we also did a parallel random sample of all active[4] journals who are using OJS as their journal management system, since it is used by approximately 10,000 OA journals worldwide (Alperin et al., 2016). We conducted this sampling between January and May of

[3] Some of the journals we may have found as predatory in 2014 may no longer be listed in DOAJ today given their recent house cleaning efforts to remove any journals which did not meet their stricter criteria <http://www.nature.com/news/open-access-website-gets-tough-1.15674>

[4] We define “active” as an OJS journal having published at least 10 articles within the last year.

2015, with a targeted followup in March of 2017. For our coding, we included the data source, journal name, journal homepage URI, field of study, if the journal was: questionable/predatory, in the English language, and if they published empirical research. We manually reviewed each journal's website looking for relevant guidance by checking for submissions guidelines, journal policies, author guidelines, and similar terms -- if no guidance or reference to guidance was discoverable on the journal website we coded the journal as having no data policy. To identify relevant sections on the website we searched for "data", "citation/cite", "share", "sharing", "replication", "reproducible", "repository", "supplemental materials", or "supplemental data".

To code the strength of data sharing policies we adapt the Stodden, Guo & Ma (2012) five point scale, and in addition record whether a non-required but explicit policy actively encourages data sharing. We apply this same scale to measure the strength of data citation policies. For comparison with other data sharing studies, we measure additional characteristics of data-sharing policies, including: whether the place of deposit is specified (for comparability with Weber, Piwowar & Vision 2010); when data sharing is required (for comparability with Ghergina & Katsanidou 2013); exemptions to data policy (for comparability with Vlaemink 2013). For comparison with the broadly accepted Joint Declaration of Data Citation Principles (see Altman & Crosas 2013) we measure additional characteristics of citation policies, including: recommended/required location of data citation; recommended/required elements of data citation; and presence of example data citations. The full list of measures is documented in the replication dataset for this study.

To support replication and analysis, these coded data are permanently archived, and available through the Harvard Dataverse[5]. The dataset is accompanied by a codebook, which specifies the coding rules for each field. (Altman, et. al 2017)

Quantitative Results

As background, we calculated the distribution of characteristics journals in the initial sample of 50. As shown in Table 1, a relatively small percentage were potentially predatory, and/or or non-empirical. The initial OJS sample yielded a substantially larger percentage of non-English journals -- reflecting the popularity of the OJS system internationally, and especially in the global south.

	DOAJ	OJS
Questionable-Predatory	16%	0%
Empirical	80%	92%

[5] Harvard Dataverse <http://dataverse.harvard.edu>

English-Language	76%	40%

Table 1. Distribution of selection criteria characteristics in initial journal sample.

We eliminated journals that were non-empirical,non-English, or predatory and drew additional random samples, to obtain a sample of 50 randomly selected journals, stratified by database, that met all inclusion criteria. All further analysis below was performed using this set. As shown in Figure 1, below, these journals were distributed across all fields of science, although the sample showed a slight concentration in health sciences.

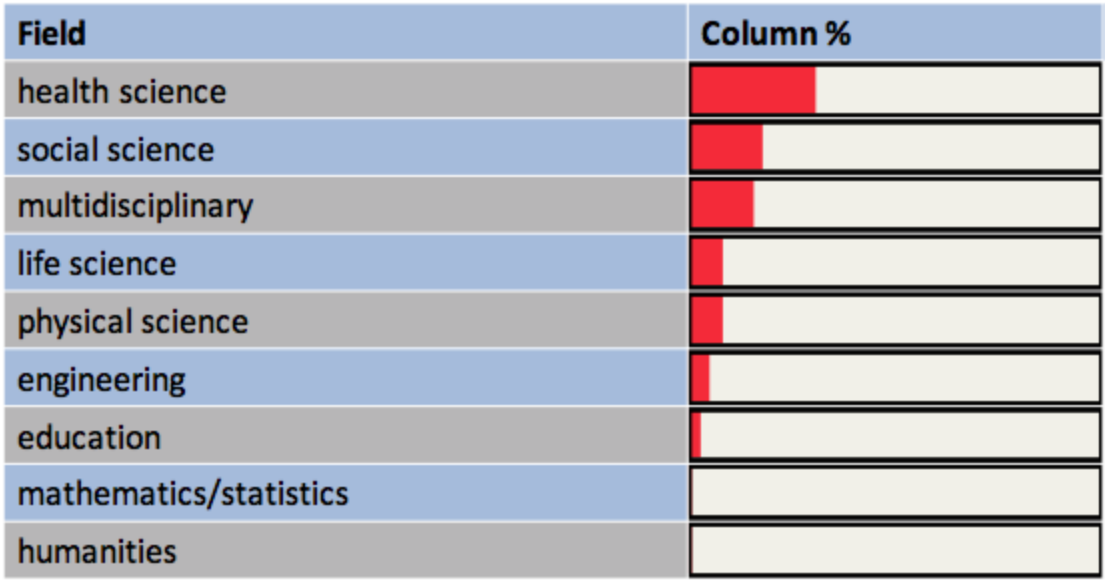
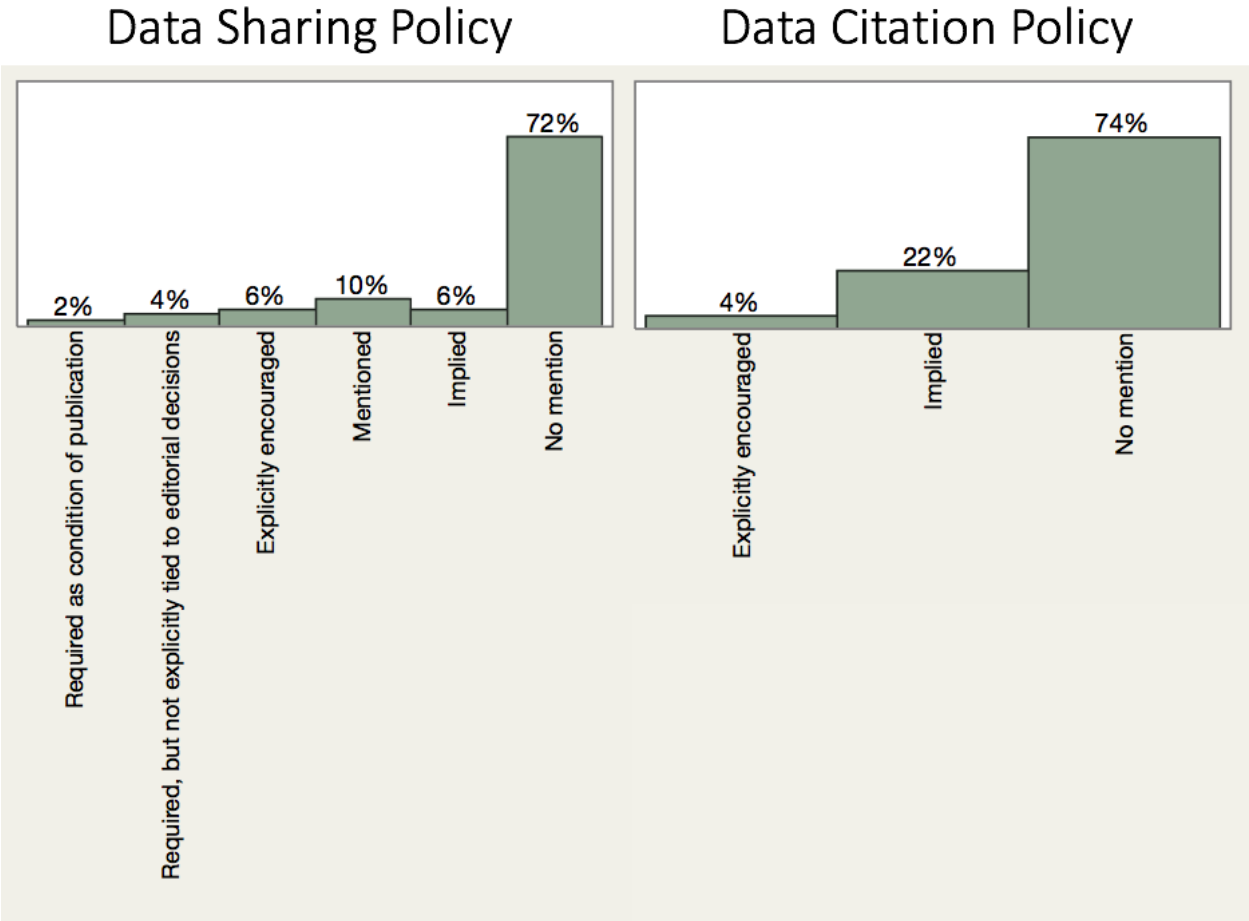


Figure 1. Distribution of journals across fields.

In Figures 2a and 2b we summarize the frequency of data sharing and citation policies across the random sample of OA journals. A number of patterns emerge from this distribution: The vast majority of OA journals sampled (74%) do not have any data policy -- even an implied one. Furthermore, only 6% of these journals require data sharing. Moreover, the journals' policies on data citation are even weaker: data citation is discussed in only 4% of cases sampled, and never explicitly required.



'Figure 2a and 2b. Distributions of citation and data sharing policies in OA journals.

Table 2 provides more detail on the specific elements of journal data policies. The vast majority of policies found in journals did not include requirements more specific than a general assurance of data availability. The most common specific policy details included an example data citation (7%), and specification of the place of deposit (4%).

Citation: Example citations provided	14%
Sharing: Place of deposit specified	8%
Sharing: Is deposit required	6%
Citation: Persistent ID required	6%
Sharing: Replication data required	4%

Table 2. Most frequent specific elements in journal data policy

To detect differences between population frames we compared the proportion of journals with any data policy between the DOAJ and OJS frames. This is displayed in figure 3. Data policies were approximately 25% more frequent in DOAJ journals than in OJS journals. This difference was only marginally statistically significant ($p < .10$), and should be considered suggestive, only. We conjecture that the difference may be due to the greater proportion of international journals in the OJS database.

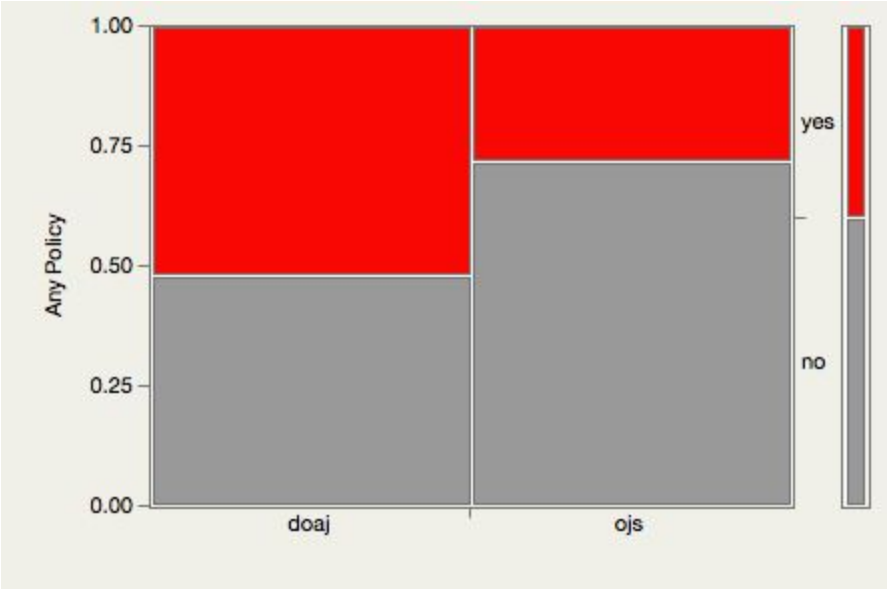


Figure 3. Frequency of data Policies in DOAJ vs. OAJ Journals

For comparison with previous work, we aggregated data sharing policies into three categories “strong” (a stated policy with any requirements), “weak” (any explicit or implied recommendations or referrals to the area, lacking specific requirements), and “none”, and compared the proportion of OA journals in our random sample with previous samples from four prior and contemporary studies.The results are displayed in Figure 4, below. Finally, we conducted a targeted followup in March 2017 to each of the journals in our samples, and evaluated the 2017 policy using only the three-level coding above. The results are displayed in Figure 5, below.

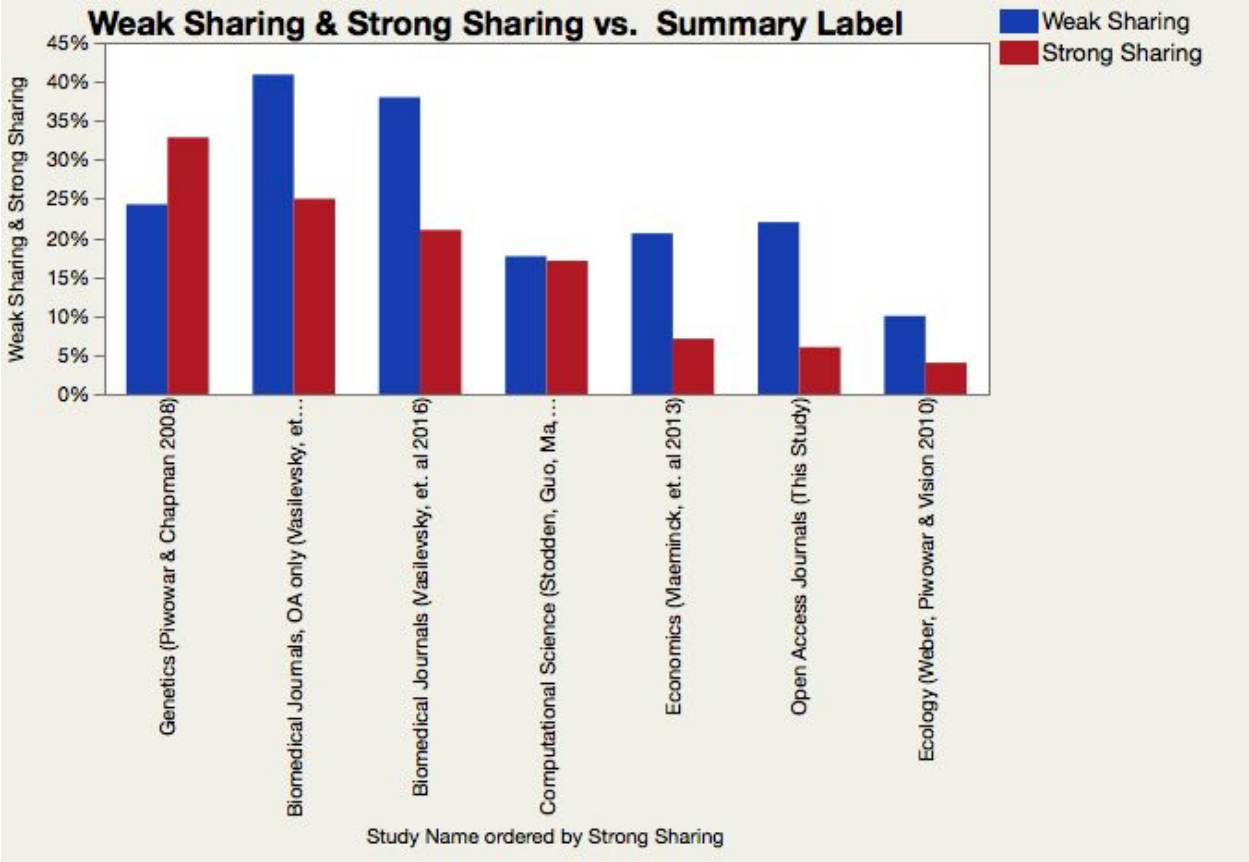


Figure 4. A Comparison of Policy Strength in Samples of different Journals Weak sharing is indicated in red, strong sharing in blue.

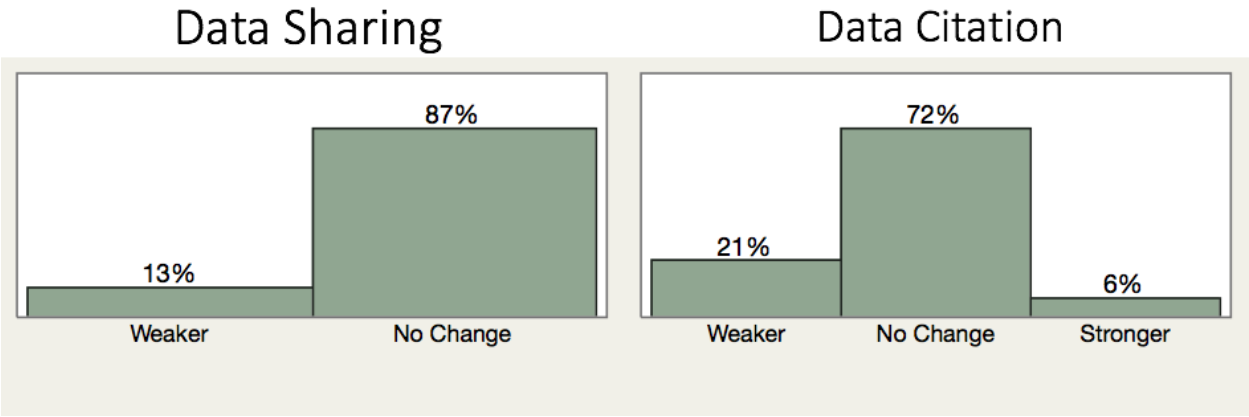


Figure 5. Change in Policy Strength 2015-2017

Several surprises are in evidence from these comparisons. First the overall level of strong data sharing in this sample is smaller than in studies of other samples of commercial and disciplinary journals. Second, OA journals in this sample are less likely to have strong policies than are the

commercial and disciplinary journals previously studied. Third, policies do not show strong signs of increasing in strength over time within these journals.

Comparison with Flagship Policies

As a point of comparison to the OA journals we sampled, we also identified some[6] exemplary data policies from major flagship journals and publishers (OA and commercial), disciplinary associations, scholarly societies, and funders of research grants.

Several notable disciplinary associations and scholarly societies have put together helpful guidelines for journals to adopt strong data policies. For example, in 2014 the American Political Science Association (APSA) worked with publishers to jointly publish: Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors (Lupia & Elman, 2014) in order to help further improve the quality of the data sharing, citation, replication policies and guidelines for authors submitting data to political science journals. Additionally, the APSA Section for Qualitative and Multi-Method Research has created its own website to address research transparency with qualitative research[7]. One discipline agnostic example is the TOP (Transparency and Openness Promotion) Guidelines[8], partly inspired by APSA's DA-RT[9], which were developed in 2015 by a group of journal editors and the Center for Open Science (COS), to help ensure that research published in scholarly journals is not only openly available but also reproducible. Since March 2017, 757 journals and 63 organizations have already become TOP signatories, many of them from OA publishers like Biomed Central and Ubiquity Press. They offer three different levels of transparency from milder/entry-level to stronger requirements for authors. For data citation policies in particular, both the FORCE11 Joint Declaration of Data Citation Principles (JDDCP)[10] (Altman et al, 2015), and Data Cite[11] provide guidance on community driven data citation practices that are both human understandable and machine-actionable, which include the requirement of a persistent identifier to the dataset, and a minimum amount of metadata to allow for attribution and reuse. In addition, several of the authors behind the JDDCP are currently working on a publisher-agnostic roadmap (now in preprint) with detailed instructions to help with implementing JDDCP-compliant data citation (Cousijin et al, 2017).

From the OA publisher perspective, several OA journals -- all of which appear to focus on the sciences -- have strong data sharing policies compared to journals in our sample. PLOS, as

---

[6] This is not an exhaustive or comprehensive list of what exemplary data policies are currently out there. The Journal Research Data Policy Bank (JoRD) is working towards developing such a database. See: Naughton, & Kernohan 2016 <http://doi.org/10.1629/uksg.284>
[7] APSA Qualitative and Multi-Method Research <https://www.qualtd.net/>
[8] TOP Guidelines <https://cos.io/top/>
[9] DA-RT on TOP Guidelines <http://www.dartstatement.org/2015-cos-top-guidelines>
[10] Force11 Joint Declaration of Data Citation Principles <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
[11] Data Cite - Cite Your Data <https://www.datacite.org/cite-your-data.html>

previously mentioned, has had a data policy since 2014 in which authors must provide a Data Availability Statement that is published with the accepted article. While this does not strictly speaking require the deposit of data in a publicly available repository, PLOS recommends repositories, and stresses that refusal to share data will be grounds for rejection. Since 2013 *BioMed Central*, has had an Open Data Policy that requires authors to apply a Creative Commons CC0 waiver to all published data in their articles (Hrynaszkiewicz, Bush & Cockerill, 2013), which ensures that data can more easily be reused by other researchers. GigaScience[12] -- an open access, open data, and open peer-review journal -- data policy requires authors to deposit the data in a publicly-accessible data repository, such as GigaDB, and that any data that is cited in their article follow the guidelines of JDDCP. Another exemplary publication worth mentioning is F1000Research -- an open publishing platform that provides transparent refereeing of articles -- who are unique amongst all the journals we reviewed since they provide a specific list of requirements for data repositories hosting data linked to one of their articles. Similar to the other exemplary journals, F1000Research require data to be made available, include detailed guidelines[13] for dataset submission.

From the flagship commercial journals we reviewed, we found a few exemplary data policies. In the sciences domain, the publisher Nature (now Springer Nature[14]) has had a mandatory data sharing policy since 2013 requiring that authors "make materials, data, code, and associated protocols promptly available to readers." In 2016, following the JDDCP, Nature introduced an updated policy adding data citation as mandatory, which encouraged including a persistent identifier to the dataset[15] (digital object identifier (DOI)), and the minimum information recommended by Data Cite. Differently, Elsevier's research data policy -- mindful of the challenges of sharing and making data accessible -- encourages open data rather than making it mandatory for publication[16]. In the social sciences several flagship journals have had long-standing data sharing policies. In the field of economics, the American Economic Review (AER), and by extension any of the journals from the American Economic Association (AEA), have a Data Availability Policy, in which the author is required to make the data available to reviewers[17]. Several political science journals take it one step further by making data sharing not just mandatory for reviewers but require that it be made open for the journal's readers. For example, the journal Political Analysis (PA) requires that authors make replication materials (data, code, and documentation) publicly available in the data repository Harvard Dataverse prior to publication, and appropriately cite all "original and archival" data (with citation examples

[12] GigaScience Policy <https://academic.oup.com/gigascience/pages/instructions_to_authors#Preparing Supporting Information>
[13] F1000Research Data Guidelines <https://f1000research.com/for-authors/data-guidelines>
[14] Springer Nature data policy <http://www.springernature.com/gp/group/data-policy>
[15] Nature's updated data policy <http://www.nature.com/news/announcement-where-are-the-data-1.20541>
[16] Elsevier's Research Data Policy <https://www.elsevier.com/about/company-information/policies/research-data>
[17] American Economics Association Data Availability Policy <https://www.aeaweb.org/journals/policies/data-availability-policy>

given)[18]. The American Journal of Political Science (AJPS), takes it one step further than most journals by having a Replication and Verification policy, which as part of the publication workflow requires that all articles submitted must be replicable and will be verified by a third-party to ensure this requirement is met prior to publication[19]. The AJPS also provide very detailed guidelines for which files and documentation authors should include to ensure that a study can be properly replicated.

In parallel, many funders have put out strong data policies for research they fund, or placed an emphasis on awarding grants to projects which look at making science more open. One notable example is the Laura and John Arnold Foundation, which make available Research Integrity[20] grants to help support transparency, reproducibility, and rigorous research standards. Such grants have helped organizations such as the Center for Open Science to help push for more transparent and open research practices. More recently, the Bill and Melinda Gates Foundation has updated their OA policy[21] to also include making any underlying data openly available immediately with no embargo as of Jan 1 , 2017, and much of their data is already shared through the public data repository Harvard Dataverse. From the federal funder perspective, a decade before the OSTP memo, the NIH has had a data sharing policy since 2003, and since then has continued to strengthen it, while also providing more guidance[22] for researchers on what data they should share depending on the kind of research they produce (e.g., Genomic Data Sharing).

Discussion

Given the revolutionary nature of the open access movement, which strives to make all research outputs open (Suber 2005), the data sharing policies in the OA journals we sampled are surprisingly weak. In comparison to studies of data sharing policies in commercial journals (discussed above) -- OA journals are less or no more likely to have a policy, and much less likely to have a strong one.

There seems to be a stark contrast between the desire for openness of published results with openness of process and evidence. Approximately ¾ of the OA journals we looked at have no data sharing policies at all -- even an implied one. Only 6% have a formal requirement, and data citation policies are even weaker -- rarely mentioned explicitly. We observed that the policies in place lack specificity and do not provide guidance for the researcher as to how they can share their data, which includes details on where to deposit, how to cite, and, where applicable, how to

[18] Political Analysis instructions to authors <https://academic.oup.com/pan/pages/Instructions_To_Authors>
[19] AJPS Replication Policy <https://ajps.org/ajps-replication-policy/>
[20] Research Integrity Grants <http://www.arnoldfoundation.org/initiative/research-integrity/>
[21] Gates OA policy <http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy>
[22] NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources <https://grants.nih.gov/policy/sharing.htm>

ensure the data can be replicated. According to McCullough (2009) this is paradoxical; considering the OA movement's "emphasis on making articles readily available, one would think that open access journals also would want to make data and code readily available."

What could be some of the possible reasons for such a low prevalence of policy? Excluding PLOS and some other outliers, generally OA journals lack the resources or backing of older established journals and publishers, which would be helpful to effectively push for strong data requirements from their authors. In addition, some OA journals also have an Author Processing Charge (APC), which would put an additional ask on authors if data sharing were to be required. A few studies of commercial journals also found that the lower the Impact Factor of a journal the less likely a journal is to have a data sharing requirement (Piwowar & Chapman 2010, Vasilevsky, et al 2016). In a 2015 study of data policies in commercial and OA economics journals, Vlaeminck and Hermann lucidly noted that in most cases, journals with strong data policies are among the top journals in their discipline and that they could afford to implement such guidelines, "while a medium or low-ranked journal planning to implement a DAP [data availability policy] could see a reduction in the amount of submissions it receives." (p. 154). Given the relatively young age of most OA journals, it may take more time for them to be fully established. Finally, some disciplines have embraced the culture of OD more than others, so it may take some time for the culture of data sharing to become more ingrained in certain disciplines, which would ultimately trickle down to individual OA journals.

Resources to Help OA Journals Adopt Data Policies

Conjectures aside, there are some notable existing resources for OA journals who wish to implement data sharing policies. These resources, which are detailed below, provide a variety of levels of assistance: from infrastructure and data curation services, to detailed guidance on policies, as well as standards and best practices for properly sharing data.

Although not specifically aimed at OA journals, there are several current projects and initiatives from the scholarly community at large, which are actively working on the development of best practices and guidelines on data sharing, data citation, and replication policies. For example, the previously mentioned largely-endorsed TOP Guidelines provide discipline agnostic instructions for journals to adopt various levels of data sharing and replication policies. OA journals can also use the exemplary data sharing policies of the previously mentioned flagship journals and publishers such as PLOS, AJPS, and Springer Nature. For guidance on data citation policy, journals should look at GigaScience, Springer Nature, and Political Analysis who provide helpful exemplary policy text on their respective websites. For further examples of citation policies, the FORCE11 JDDCP website has a list of many publisher signatories[23]. Furthermore, although this has yet to materialize, there are a large number of non-English OA journals (as evidenced from our initial OJS sample), which would benefit from having access to translated versions of exemplary journal data policies and guidelines from TOP.

[23] JDDCP Signatories List <https://www.force11.org/datacitation/endorsements>

OA journals also require resources to be able to implement and enforce such policies. For journals using OJS (v.2+) as their journal management system they can automatically deposit their data into the Harvard Dataverse repository, which is open to any scholar, regardless of institutional affiliation, to deposit their research data. Through OJS these journals can use the Dataverse plugin, which adds a data deposit step into the article submission workflow (Castro & Garnett 2014; Altman, et al 2015). This plugin automatically submits, via SWORD API, research data associated with a journal article into a Dataverse (King, 2007; Crosas, 2011)[24] repository and links it back to the journal article itself. Boilerplate data policies have also been included in OJS to help journals get started[25]. Journals can also directly partner with data archives and curated repositories, which provide services for data management, curation, and/or verification for replication. Given the large, growing amount of existing data repositories, journals can use Re3data.org, a large registry of data repositories, to help find (by content type, or subject) a suitable archive that can help with managing research data. Some publishers, such as PLOS and Springer Nature have also compiled lists of recommended repositories, which are recognized and trusted within their respective communities,divided into domain-specific and generalist data repositories[26] [27]. In addition to data management support, some data archives also provide curation services such as ICPSR for social science data (including the codebooks and documentation)[28], and Dryad[29] for data files associated with any published article in the sciences or medicine, as well as software scripts and other files. For data verification services, one notable example is the American Journal of Political Science's (AJPS) commitment that submitted replication materials will be verified to guarantee that they do reproduce the analysis results. AJPS' Replication policy has led to arrangements with: the University of North Carolina's Odum Institute for Research in Social Science to carry out the verifications for quantitative data; and the staff at the Qualitative Data Repository (QDR), at Syracuse University for the verification of qualitative analyses. Alternatively, if journals lack the resources to allow for this kind of verification, Key (2016) provides a different option which relies more on the scholarly community to voluntarily provide their own verification of the datasets that interest them, which also provides an opportunity for students to learn through replication.

---

[24] The Dataverse Project is an open source research data repository framework, which is developed at the Institute for Quantitative Social Science (IQSS) at Harvard University. For more details <http://dataverse.org>

[25] Boilerplate data policies in OJS <http://projects.iq.harvard.edu/files/ojs-dvn/files/journaldatapoliciesguidelinestemplateojsdataverseplugin.pdf>

[26] PLOS recommended repositories list <http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories>

[27] Springer Nature's list of recommended repositories <http://www.nature.com/sdata/policies/repositories>

[28] Overview of ICPSR's Data Management and Curation <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html>

[29] Dryad Digital Repository has archived datasets from over 560 journals. Their curation service and streamlined integration with journals is outlined on this webpage <http://datadryad.org/pages/faq>

Extensions to Research

The work in this study is preliminary in nature and so more research is necessary to help with developing more actionable conclusions. Below we describe a few potential more in-depth studies which can be done in order to help us further understand the unique needs of OA journals and data sharing. This research could help better inform international scholarly societies, and the OA community in general who already work with and provide support for OA publishing to help come up with creative ways to incentivize OA journals to start implementing strong data policies.

The design of our study enabled subgroup comparisons only between the DOAJ and OJS journals. Conducting studies with larger sample sizes would allow subgroup analysis of characteristics such as discipline, age of the journal, and peer review type in order to determine if certain characteristics of OA journals are associated with stronger data sharing.

Since our sample was from 2015 with a follow-up in 2017, a longitudinal study could be done in a few years time to see if there is any positive change with how OA journals (in DOAJ and OJS) are doing with regards to implementing strong data policies. Additionally, an analysis can be done of the journals which already had strong policies to see if they are enforcing them by looking for the presence of or link to data in the articles.

Summary

To summarize, our preliminary research on this topic has shown that there is a surprisingly weak adoption of data policies in OA journals (although many notable exceptions exist among publishers and journals, including PLOS, Biomed Central, and GigaScience). There are however many freely available tools and resources for OA journals to be able to easily start implementing data sharing, citation, and replication policies. In addition, there are plenty of opportunities to expand on the research we have done on the characteristics of OA journals and their ability or willingness to setup data policies.

Author Statement and Acknowledgments

We describe contributions to the paper using a standard taxonomy (Allen et. al 2014). Micah Altman and Eleni Castro were the lead authors, taking responsibility for content and revisions. Micah Altman authored the first draft of the manuscript, was responsible for the initial conceptualization, and for the data analysis. Kasey Sheridan provided data collection and management. All lead authors contributed to the conception of the report (including core ideas and statement of research questions), to the methodology, and to the writing through critical review and commentary. All authors contributed to review and commentary.

The writing of this report was supported by awards from the *Sloan Foundation.*

References

Allen L, Brand A, Scott J et al (2014) Credit where credit is due. Nature 508:312-313. 10.1038/508312a

Altman M, Borgman C, Crosas M, Martone M. An Introduction to the Joint Principles for Data Citation. Bulletin of the Association for Information Science and Technology [Internet]. 2015;41(3):43-44.

Altman, M., Crosas, M. (2013) The Evolution of Data Citation: From Principles to Implementation. IASSIST Quarterly 37:62

Altman, M., Castro, E., Crosas, M., Durbin, P., Garnett, A., & Whitney, J. (2015). Open Journal Systems and Dataverse Integration–Helping Journals to Upgrade Data Publication for Reusable Research. *Code4Lib Journal*, (30).

Altman, Micah; Castro, Eleni; Crosas, Merce; Garnett, Alex; Sherridan, Kasey, (2017), "Replication Data for: EVALUATING AND PROMOTING OPEN DATA PRACTICES IN OPEN ACCESS JOURNALS"", doi:10.7910/DVN/JPUJJC, MIT Program on Information Science Dataverse.

Alperin J, Stranack K, Garnett A (2016) On the Peripheries of Scholarly Infrastructure: A Look at the Journals Using Open Journal Systems. Proceedings of the 21st International Conference on Science and Technology Indicators. http://summit.sfu.ca/item/16763

American Political Science Association's Data Access and Research Transparency Group (2015) The (DA-RT) Data Access and Research Transparency Joint Statement. Available via <https://www.dartstatement.org/>

Castro E, Garnett A (2014) Building a Bridge Between Journal Articles and Research Data: The PKP-Dataverse Integration Project. International Journal of Digital Curation 9(1):1. 10.2218/ijdc.v9i1.0

Cousijin H. et al. (preprint: 2017) A Data Citation Roadmap for Scientific Publishers. bioRxiv. doi: https://doi.org/10.1101/100784

Crosas M. (2011). The dataverse network: An open-source application for sharing, discovering and preserving data. *D-lib Magazine*, *17*(1/2).

Chavez TA (2010) Numeracy: Open-Access Publishing to Reduce the Cost of Scholarly Journals. Numeracy 3(1):Article 8. 10.5038/1936-4660.3.1.8

Fecher B, Friesike S, Hebing M (2015) What Drives Academic Data Sharing?. PLoS ONE 10(2):e0118053

Gherghina S, Katsanidou A (2013) Data Availability in Political Science Journals. European Political Science

Goodhill GJ (2014) Open access: Practical costs of data sharing. Nature 509(7498):33. doi:10.1038/509033b

Hrynaszkiewicz I, Busch S, Cockerill MJ (2013) Licensing the future: report on BioMed Central's public consultation on open data in peer-reviewed journals. BMC Research Notes 6:318. doi:10.1186/1756-0500-6-318

Hrynaszkiewicz I, Cockerill MJ (2012) 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. 10.1186/1756-0500-5-494

Ishiyama J (2014) Replication, Research Transparency, and Journal Publications: Individualism, Community Models, and the Future of Replication Studies. PS: Political Science & Politics 47(01):78-83. 10.1017/S1049096513001765

Jones P (2015) The Scholarly Kitchen: Are We at a Tipping Point for Open Data?. Available via https://scholarlykitchen.sspnet.org/2015/03/18/are-we-at-a-tipping-point-for-open-data/

Key, E. M. (2016). How Are We Doing? Data Access and Replication in Political Science. PS: Political Science & Politics, 49 (2), 268–272. Cambridge University Press.

King, G. (1995) Replication, Replication. PS: Political Science & Politics 28(3):443-499

King, G. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." Sociological Methods and Research, 36: 173–199. Copy at http://j.mp/iHJcAa

Lupia, A., & Elman, C. (2014). Openness in Political Science: Data Access and Research Transparency: Introduction. PS: Political Science &; Politics, 47(1), 19-42. doi:10.1017/S1049096513001716

MacGregor J, Stranack K, Willinsky J (2014) The Public Knowledge Project: Open source tools for open access to scholarly communication. In: Anonymous . Springer International Publishing, pp 165-175

McCullough BD (2009) Open Access Economics Journals and the Market for Reproducible Economic Research. EAP : Economic Analysis and Policy : Journal of the Economic Society of Australia 39(1):117

Mooney H, Newton MP (2012) The Anatomy of a Data Citation: Discovery, Reuse, and Credit. Journal of Librarianship and Scholarly Communication 1(1):eP1035. 10.7710/2162-3309.1035

Open Knowledge Foundation (2015) Why Open Data?. In: Anonymous Open Data Handbook [online] http://opendatahandbook.org/

Piwowar HA, Chapman WW (2008) Identifying data sharing in biomedical literature. AMIA Annu Symp Proc:596-600

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. *Journal of informetrics*, *4*(2), 148-156.

Piwowar H, Vision TJ (2013) Data reuse and the open data citation advantage. PeerJ PrePrints 1(e1v1). http://dx.doi.org/10.7287/peerj.preprints.1v1

Piwowar HA, Chapman WW (2008) A review of journal policies for sharing research data. In: Anonymous Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing (ELPUB) June 25-27 2008, Toronto, Canada

PLOS, PLOS' New Data Policy: Public Access to Data, February 24, 2014 http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

SPARC Europe (2015) The Open Access Citation Advantage: Summary of results of studies [bibliography]. Available via . http://sparceurope.org/oaca_list/. Accessed 05/23 2015

Starr J, Castro E, Crosas M et al (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1. 10.7717/peerj-cs.1

Stodden V, Guo P, Ma Z (2012) How Journals Are Adopting Open Data and Code Policies. In: Anonymous Governing Pooled Knowledge Resources: Building Institutions for Sustainable Scientific, Cultural, and Genetic Resources Commons, 1st Thematic IASC Conference on the Knowledge Commons, Louvain-la-Neuve, Belgium. Governing Pooled Knowledge Resources: Building Institutions for Sustainable Scientific, Cultural, and Genetic Resources Commons, 1st Thematic IASC Conference on the Knowledge Commons

Suber P (2005) Open Access Overview http://legacy.earlham.edu/~peters/fos/overview.htm

Suber P (2012) Open Societies Foundations - Voices: Opening Access to Research

Suber P (2012) Open Access. MIT Press, Cambridge, MA

Tenopir C, Allard S, Douglass K et al (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6):e21101

Vlaeminck S (2013) Data Management in Scholarly Journals and Possible Roles for Libraries - Some Insights from EDaWaX. LIBER Quarterly 23(1)

Vlaeminck, S., & Herrmann, L. K. (2015). Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?. *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust*, 145-155.

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2016). *Reproducible and reusable research: Are journal data sharing policies meeting the mark?* (No. e2588v1). PeerJ Preprints.

Wagner AB (2010) Open Access Citation Advantage: An Annotated Bibliography. Issues in Science and Technology Librarianship Winter. 10.5062/F4Q81B0W