Do Journal Data Sharing Mandates Work? Life Sciences Evidence from Dryad[1]

Mike Thelwall, Kayvan Kousha: University of Wolverhampton

Purpose: Data sharing is widely thought to help research quality and efficiency. Since data sharing mandates are increasingly adopted by journals this paper assesses whether they work.

Design/methodology: This study examines two evolutionary biology journals, Evolution and Heredity, that have data sharing mandates and make extensive use of Dryad. It uses a quantitative analysis of presence in Dryad, downloads and citations.

Findings: Within both journals, data sharing seems to be complete showing that the mandates work on a technical level. Low correlations (0.15-0.18) between data downloads and article citation counts for articles published in 2012 within these journals indicate a weak relationship between data sharing and research impact. An average of 40-55 data downloads per article after a few years suggests that some use is found for shared life sciences data.

Research limitations: The value of shared data uses is unclear.

Practical implications: Data sharing mandates should be encouraged as an effective strategy.

Originality/value: This is the first analysis of the effectiveness of data sharing mandates.

Introduction

Scientific data sharing is increasingly encouraged or mandated by funders and journals (Mennes, Biswal, Castellanos, & Milham, 2013). Some organisations even fund projects to rescue data that had initially not been shared (Hsu, Lehnert, Goodwillie, Delano, Gill, et al., 2015). The advantages of shared data to science are that the results of any associated paper can be at least partially verified, reducing mistakes and fraud (Sandve, Nekrutenko, Taylor, & Hovig, 2013) and other researchers or the public can re-use the data for additional studies or purposes (Borgman, 2012; Caetano & Aisenberg, 2014). Thus, science can be both more replicable and efficient.

Most researchers seem to be willing to share data in principle (Tenopir, Allard, Douglass, Aydinoglu, Wu, et al., 2011; Tenopir, Dalton, Allard, Frame, Pjesivac, et al., 2015) and data sharing is common in forensic and evolutionary genetics (Anagnostou, Capocasa, Milia, & Bisol, 2013). Nevertheless, over 60% of scientists responding to one survey were unwilling to share their primary data before the final publication of their articles (Huang et. al., 2012). There are multiple possible reasons for individuals being reluctant to share: researchers need extra time to format data for sharing by adding provenance, describing variables in detail (Kroon-Batenburg & Helliwell, 2014), and anonymising if necessary (e.g., Hrynaszkiewicz, Norton, Vickers, & Altman, 2010); their work may be discredited if it is found to be mistaken; and others may publish research on their data that they were intending to do at a later stage. This last point is particularly important in research areas that produce data sets of lasting value that are difficult to collect. It is also disproportionally affects researchers in developing nations that may not have the resources to publish quickly

(Kenall, Harold, & Foote, 2014). Conversely, the advantages for those sharing the data include a higher perceived quality for their work if its associated data can be checked, higher visibility from the international community through data reuses, and perhaps also citations to their datasets or the associated paper from follow-up research (an important incentive: Tenopir, Allard, Douglass, Aydinoglu, Wu, et al., 2011). On balance, however, it seems that fields will not shift towards a culture of open data sharing without clear incentives and the most effective method seems to be data sharing mandates from journals and funders (Mennes, Biswal, Castellanos, & Milham, 2013). Despite these becoming increasingly widespread, it is not clear whether they are effective.

Academic journal author guidelines are the natural places in which data sharing policies should have the most effect and in some fields they are used for this. For example, by 2008, a third of journals publishing articles using gene expression microarray data had strong statements about the need for data sharing (Piwowar & Chapman, 2008). As an example of a field-wide attempt to develop a data sharing culture, several journals within the field of evolution have formulated and signed up to a Joint Data Archiving Policy (JDAP, see: datadryad.org/pages/jdap) in 2011, which requires all published articles to make all associated data publically available via a repository, possibly after an embargo period. Journals signing up to this policy include it within their author guidelines. The Dryad (datadryad.org) online digital repository has been central to this initiative. It started in 2007 and was originally designed for evolutionary biology research data (Greenberg, 2009; Krause, Clary, Ogletree, & Greenberg, 2015). Journals can integrate with Dryad by automatically notifying authors about how to submit data and then exchanging information with Dryad to ensure that information about an article matches its record in Dryad and giving reviewers access to the data (Kenall, Harold, & Foote, 2014; Miller, 2016).

Given that data has been shared, a second concern is whether it has been used. Since researchers' time is valuable, if very few shared datasets are ever used then the time and money devoted to archiving data and setting up repositories could be wasted (Borgman, 2012). Some research data is too specific to have new uses, data re-use can be hampered by the complexity of the data and incomplete understanding of its meaning (Tenopir et al., 2011; Borgman, 2012), and studies may be perceived as less important if they use recycled data (Curty & Qin, 2014). Datasets in Dryad are mainly cited by the papers from which they originate (He & Nahar, 2016). When re-used by another paper, the users are typically members of the originating research group, suggesting that data sharing is not widespread (He & Nahar, 2016) and this seems to be a common problem for data sharing (Borgman, 2012). Thus it seems possible that data sharing rarely leads to new uses. Similarly, a large scale study using the Data Citation Index has found that almost 90% of shared datasets had no identified citations (Robinson-García, Jiménez-Contreras, & Torres-Salinas, in press; see also: Ingwersen & Chavan, 2011).

A simple way to assess whether open data has been reused is to detect whether articles that share data tend to be more cited, even though it would be difficult to prove a direct cause-and-effect relationship. Moreover, reuse does not necessary lead to new citations. A citation advantage for publishing open data has previously been found in several different contexts. Considering only papers that created gene expression microarray data, those that openly published it attracted 9% more citations than those that did not, with a much larger difference of up to 30% for earlier papers (Piwowar & Vision, 2013; for the same data type see also: Ioannidis, Allison, Ball, Coulibaly, Cui, et al., 2009; Piwowar, Day, & Fridsma, 2007). This finding used multiple regression and factored out the publication date,

the journal impact factor, open access status, the number of authors, the publication history of the first and last authors, corresponding author country of origin, the citation history of the affiliated institutions, and the study topic. This is a strong finding but concerns a single specialist area in which data sharing is routine and it is not clear that the same would be true in other fields with different dynamics, including those for which data sharing is rare (e.g., Wallis, Rolando, & Borgman, 2013). In the social sciences, formal data sharing is uncommon but substantially boosts research productivity, even after accounting for researcher capabilities (Pienta, Alter, & Lyle, 2010). In international studies, data sharing is associated with increased citation rates after accounting for some researcher differences (Gleditsch, Metelits, & Strand, 2003). In astronomy, astrophysics, and paleoceanography, articles with associated public data tend to be 30-50% more cited than those without (Henneken & Accomazzi, 2011; Dorch, 2012; Sears, 2011) but these three studies did not take into account researcher capabilities so no cause-and-effect claim can be made.

Despite the studies discussed above, there do not seem to have been any previous investigations into data sharing mandates, which are a key strategy. The goal of this paper is therefore to add to knowledge about data sharing by investigating whether data sharing mandates associated with Dryad have been successful in terms of ensuring that authors share their data and assessing whether the shared data is used in any way.

Data and Methods

The overall research design was to download the records of all Dryad data files, identify the ten academic journals with the most records in Dryad and from this top ten examine in more detail the two journals with the lowest and highest apparent data sharing citation advantage. The issue of citation advantage was chosen in order to focus on journals that were most likely to share useful data, and journals with many articles in Dryad were selected to find evidence of successful data sharing mandates, if any.

The Dryad website sitemap was consulted for a complete list of URLs in the site and then this list of URLs was crawled with the free research crawler SocSciBot (socscibot.wlv.ac.uk) during 24-25 August 2016, obeying the robots.txt convention for polite crawling. A function was written to extract the key information from the crawled webpages and added to the free software Webometric Analyst (lexiurl.wlv.ac.uk; Services menu, DataDrayad menu item). From this information the ten academic journals mentioned most often in Dryad records were identified.

For each of the ten journals with the most Dryad records a complete list of documents of type journal article was identified using Scopus ISSN searches for the years 2010-2016 during September 15, 2016. The years 2010-2016 were selected because there were few hits before 2010. This excludes documents classified as reviews, editorials and other non-article types. The Scopus records were matched with the Dryad records by article DOI. This gave a complete list of journal articles for each of the ten journals 2010-2016 (September), together with their Dryad records, if any.

A simple heuristic was chosen to identify the two journals from this list with apparently the highest and lowest citation advantage for data sharing: Those with the most years in which data sharing articles had more/less citations than those without data sharing, respectively. Citation counts are skewed and so the geometric mean rather than the arithmetic mean was used for this calculation (Zitt, 2012). Since multiple journals had all years with a citation advantage of data sharing, the journal with the most articles was selected. This gave two journals for further examination: *Evolution* and *Heredity*.

Scopus citations and Dryad downloads were log transformed with $ln(1 + x)$ to reduce skewing because both were highly skewed. When geometric means are reported, these are the arithmetic means of the log transformed data, then transformed back with $exp(x) - 1$ and 95% confidence intervals were calculated using standard formulae for the normal distribution before the $exp(x) - 1$ transformation (Fairclough & Thelwall, 2015).

Analyses

The journals using Dryad most are (currently) in the life sciences (Table 1). A fifth or more of the articles since 2010 in at least eight journals have associated Dryad records. The average number of citations per article is not always higher for articles with Dryad data than for those without (Table 1). Regression analyses were not conducted since the number of articles per journal is too low to allow a sufficient number of relevant covariates to be included (cf. Piwowar & Vision, 2013). Overall, the results suggest that articles with shared data may tend to attract a few more citations but that the difference is not large and no cause-and-effect relationship can be claimed from this.

Table 1. The ten journals with the most articles with registered Dryad records. For each year, + indicates that articles with Dryad records are more cited than those without (comparing geometric means) and - indicates the opposite. Blanks indicate no records in Dryad for that year.

Journal	Dryad art.	2010	2011	2012	2013	2014	2015	2016
Molecular Ecology	1475 (59%)	-	+	-	-	-	-	+
Evolution	889 (46%)	+	+	+	+	+	+	+
Journal of Evolutionary Biology	637 (41%)	+	-	+	+	+	-	-
Heredity	352 (45%)		-	-	-	+	-	-
Molecular Ecology Resources	328 (36%)		+	+	+	+	+	+
Ecology and Evolution	316 (16%)		-	-	-	-	+	+
Biology Letters	302 (19%)			+	-	-	+	+
Systematic Biology	258 (63%)		+	+	+	+	+	+
Journal of Animal Ecology	231 (24%)				-	+	-	+
Functional Ecology	230 (23%)				-	+	-	-

Both Heredity and Evolution have signed up to JDAP[2]. JDAP has had an immediate and substantial effect, as can be seen in the rapid increase in the proportion of articles with Dryad records for both Evolution and Heredity (Figure 1, for the remaining journals, see figshare.com/articles/Ten_journals_sharing_data_in_Dryad/3859653) from the year of the agreement, 2011. Presumably, almost all articles published from 2012 onwards had to conform to the agreement. Examinations of articles without Dryad records suggested that they all reported their data in an online supplement, within the main body of the article (e.g., as a figure), re-used old data, posted the data to an institutional repository, posted the data to a gene sequence repository (e.g., GenBank, "an annotated collection of all publicly available DNA sequences" www.ncbi.nlm.nih.gov/genbank) or did not create useful data as part of the study (e.g., simulations). Thus, data sharing seems to be complete for these two journals, at least for studies that generate new non-simulated data.

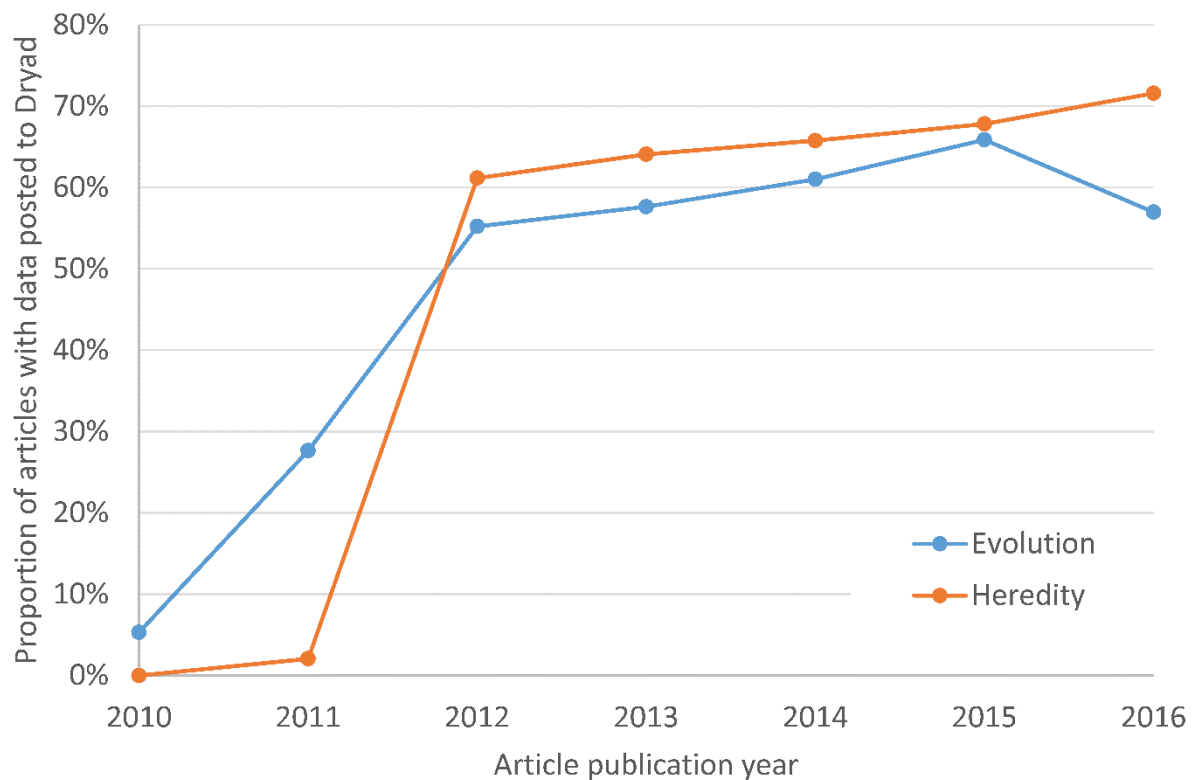[2] www.nature.com/hdy/for_authors.html, onlinelibrary.wiley.com/journal/10.1111/(ISSN)1558-5646

Figure 1. The proportion of articles in Evolution and Heredity with data posted to Dryad.

After a few years, Dryad datasets associated with articles in Heredity and Evolution can be expected to attract 40-55 downloads each (Figure 2 for the remaining journals, see figshare.com/articles/Ten_journals_sharing_data_in_Dryad/3859653). These downloads seem to accumulate slowly over time rather than mostly occurring shortly after the publication of an article, when the article is presumably most frequently read. This suggests that there is not a strong link between reading an article for interest in its content and accessing its associated data. This would be consistent with data accesses being primarily for uses other than checking the results of the published study. Dryad does not record reasons for accessing data (unlike some repositories: www.data-archive.ac.uk) and so the main usage evidence that can be systematically collected is from citations in other articles, which are rare (He & Nahar, 2016). Thus almost all data uses do not seem to directly help future studies.
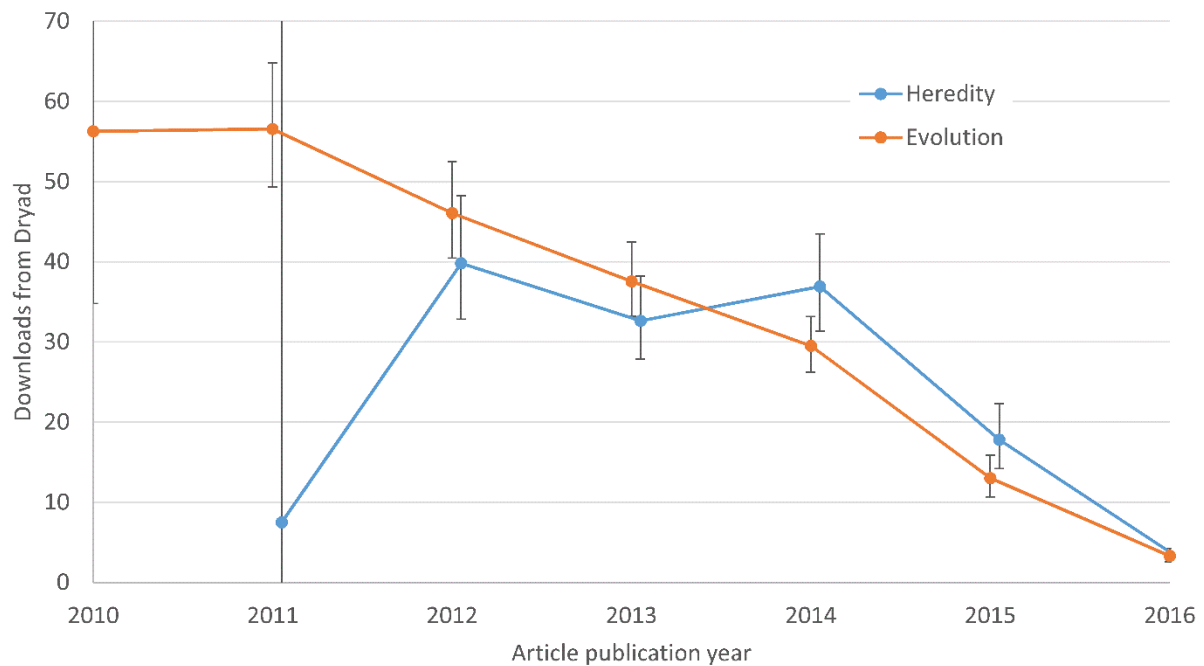
Figure 2. The geometric mean number of Dryad downloads per article for articles with data posted to Dryad. When multiple datasets are associated with an article, the maximum number of downloads for any individual dataset is used. Error bars represent 95% confidence intervals. The Heredity line is offset to avoid confidence interval overlaps.

Taking into account confidence limits for average citations per article, there is little difference in the citations attracted by articles with Dryad data and those without (Figures 3, 4). For the top ten journals, the difference between the two sets is statistically significant in only three separate cases, which is consistent with there being little or no underlying difference.
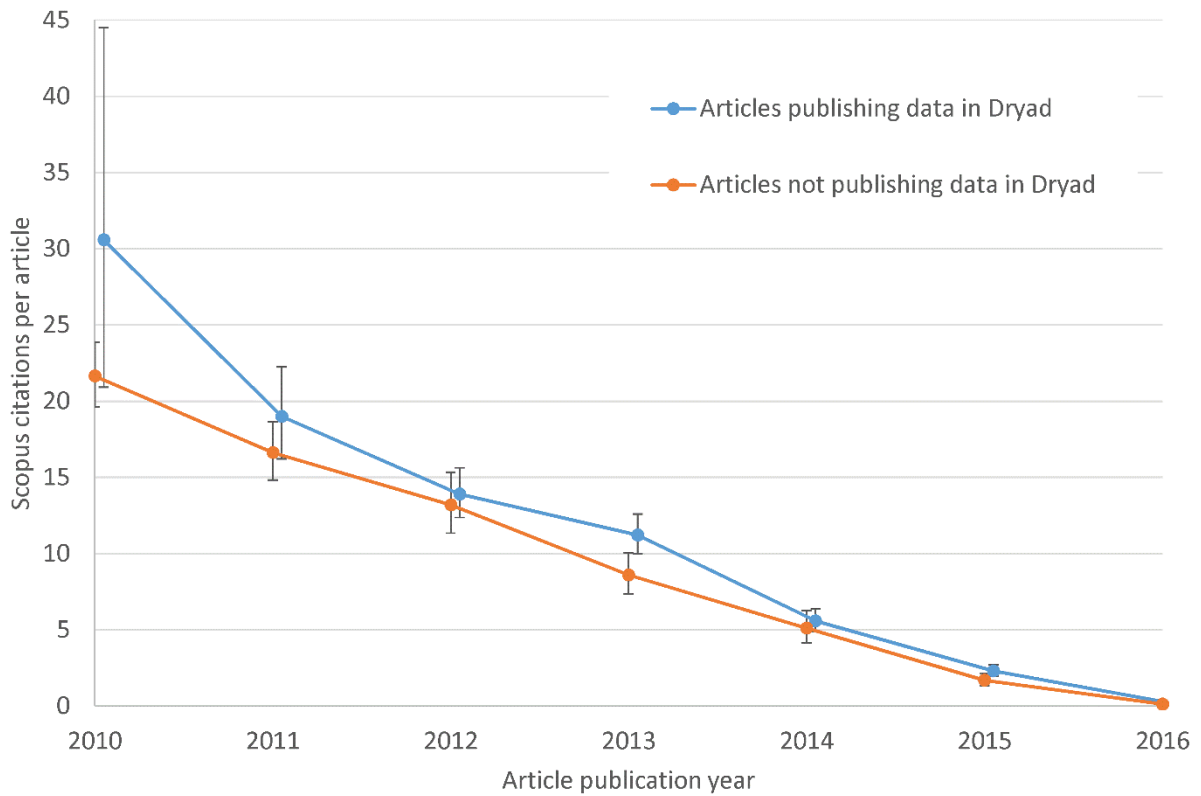
Figure 3. The geometric mean number of Scopus citations for articles in Evolution with and without an associated record in Dryad. Error bars represent 95% confidence intervals. The Heredity line is offset to avoid confidence interval overlaps.
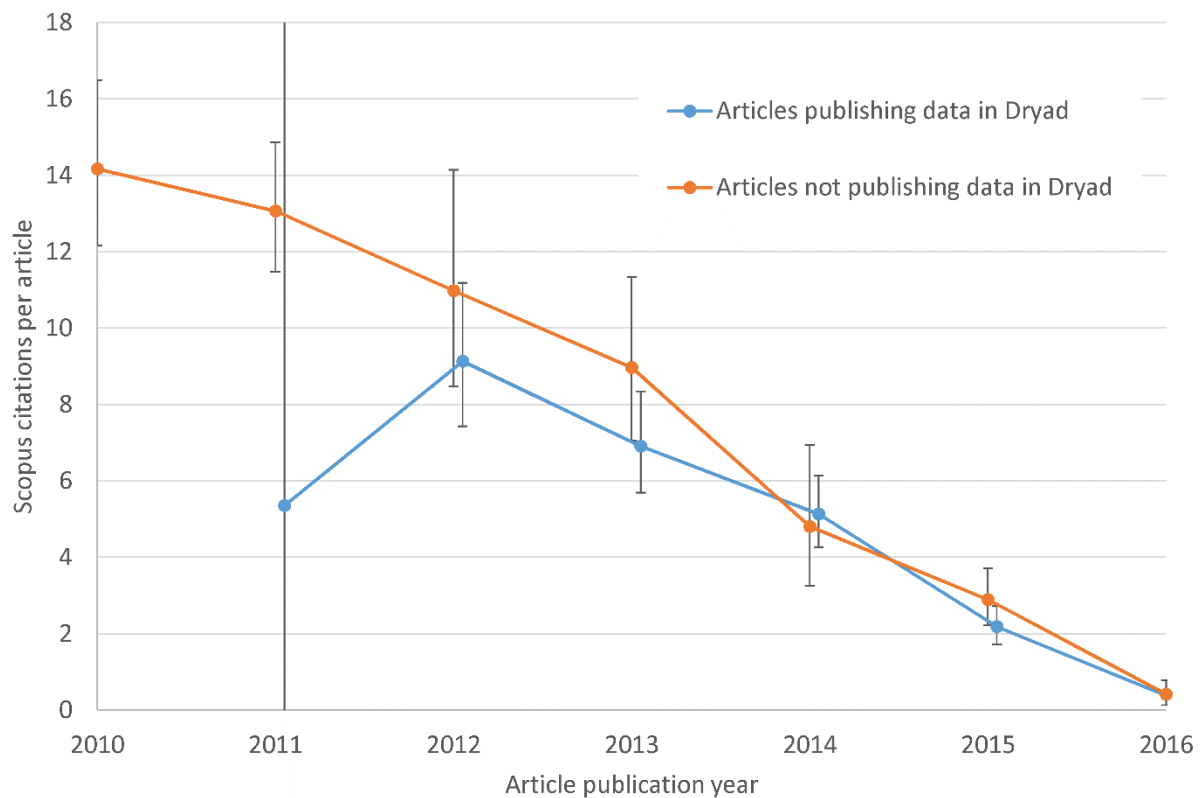

Figure 4. As Figure 3 for the journal Heredity.

Correlations between citation counts and downloads were calculated for the oldest year for which substantial data was available, 2012. For both Evolution and Heredity there is almost no relationship between citations and downloads (Figures 5, 6). Although Dryad data downloads rarely lead to citations (He & Nahar, 2016), if citations are a reasonably accurate reflection of the readership for an article then there should be a strong positive correlation between downloads and citations. Thus, downloads may depend mainly on the type of data deposited, such as whether it is in a simple format, and whether viewing it is necessary to understand the article.
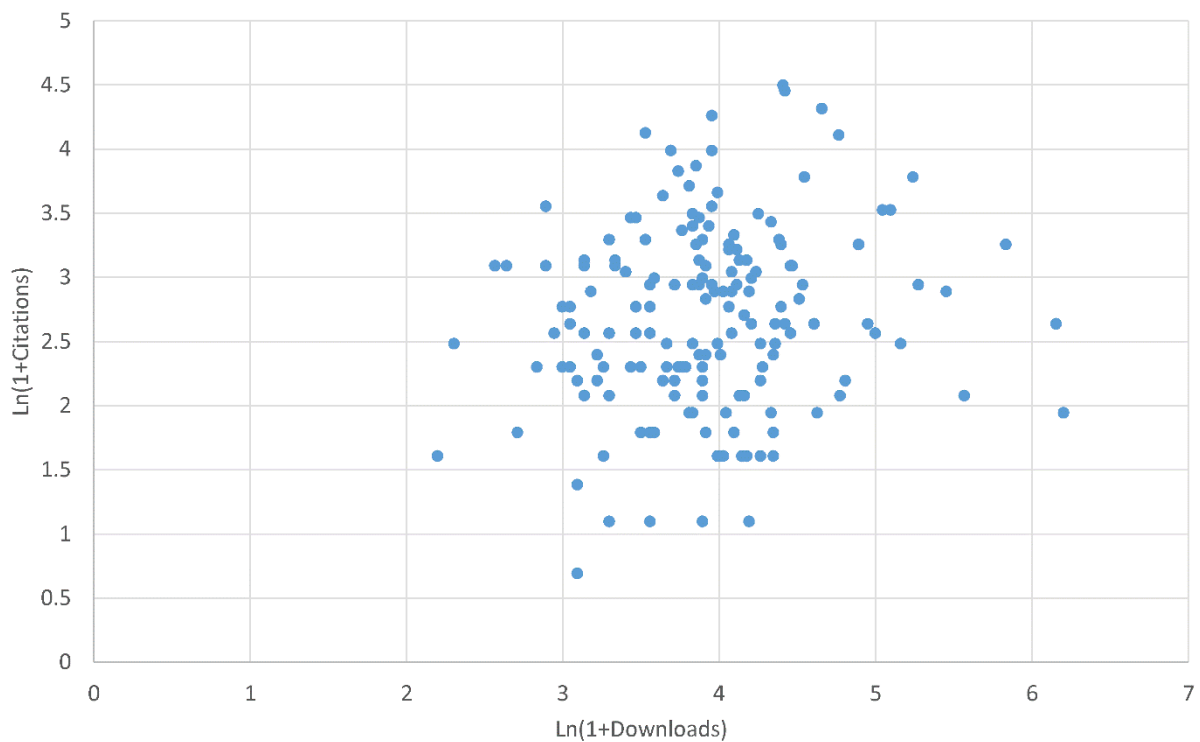

Figure 5. Scopus citations against Dryad downloads for articles from Evolution in 2012 with a record in Dryad (Pearson correlation on the log transformed data: 0.149, n=169).
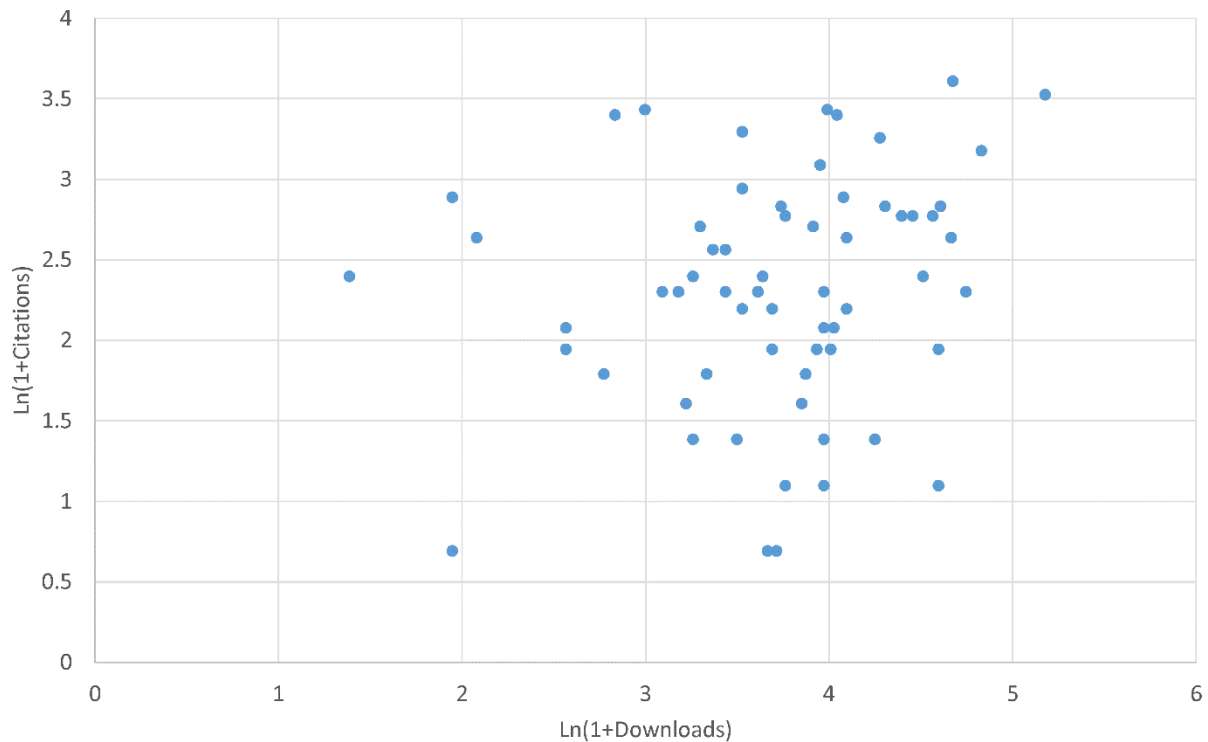
Figure 6. As Figure 5 for the journal Heredity (Pearson correlation on the log transformed data: 0.181, n=63).

Limitations and Conclusions

The results are limited to the coverage of journals that use Dryad extensively and are likely to be different in other fields and for journals that do not have a JDAP mandate. The main limitation is that no evidence is available about uses of Dryad data that do not lead to citations. In particular, it is possible that most downloads are for uses that do not reflect the wider impact of the work, such as web crawlers or authors and archivists checking data availability. In addition to these, there are many important ways of reusing data that do not lead to future citations, as discussed above. A limitation of the data is that article authors may share their data in repositories other than Dryad that were not tracked, weakening the statistical power of the tests above. If the reasons are systematic (e.g., a particular type of data is always stored in a data-specific repository) then this could also bias the findings.

The main finding from this paper is that data sharing mandates can be completely successful in evolutionary biology in terms of ensuring that all articles within a journal share their data. Of course, the fact that the mandate has been successful in some journals does not mean that it has not been successful in all and journals that have not signed up to JDAP presumably continue to publish evolutionary biology research without shared data. Nevertheless, the continued existence of major journals in the field that have operated the mandate since 2012 suggests that the field has transitioned to a state in which data sharing dominates or is at least an accepted mainstream activity.

In terms of usage for shared data, datasets associated with articles journals that extensively use Dryad are downloaded moderately often and probably on average at least 50 times in the long term. This is broadly in line with a geometric mean number of *views* of 76 for evolutionary biology datasets (from articles of all ages) in the general data and resource sharing repository Figshare (Thelwall & Kousha, 2016). The geometric mean

number of views for datasets in Figshare varied from 12 (education) to 101 (library and information studies). Adding this to the current paper, it seems that there is routinely a degree of interest in shared data even though this interest may rarely lead to citations. This supports a previous survey in which most researchers believed that shared data would be used in other ways than its intended use (Tenopir et al., 2011). Shared data may be used for new studies in the following ways at least.

- Meta analyses – researchers combine the results of multiple papers into a larger study. This can often be done with the summary results reported in a paper too.
- Comparison – researchers collect new data and compare their results with similar analyses on old data (e.g., new data on Asiatic lions compared to old data on African lions).
- Method improvement – researchers develop a new method to process existing data and publish the results on the existing data. This is common practice in computational linguistics. For example, many studies have reported different methods to automatically classify the polarity of movie reviews in the same dataset (Turney, 2002).

Reasons why shared data might be useful without leading to new citations include the following, which could be mainly characterised as "background" for research (Wallis, Rolando, & Borgman, 2013), as well as educational uses.

- Education – the data is used within a methods course as an example of real world data (e.g., "We will use ecological data from open access data repository Dryad [] to illustrate some of the statistical methods." From the *BOT 612: Ecological Statistics with R* syllabus[3]).
- Researcher training – students attempt to replicate a study in order to learn the methods.
- Results verification – readers replicate the data analysis to check that the findings are correct.
- Data understanding – readers examine the data to help them understand the paper or the methods.
- Method triangulation – readers repeat the data analysis with a different technique to check whether a different plausible approach would give the same results.

The above reasons may help to explain the weak correlations (0.15-0.18) between downloads and citations. The weak correlations might otherwise discourage researchers from data sharing by suggesting that it rarely has value.

In summary, given the apparent importance of data sharing in science, the "deluge" of data created by new methods (Borgman, 2012), the success of data sharing mandates, and the gap between the number of times that data is downloaded and citation counts, there is an increasingly urgent need to understand which uses are made of shared data and how much value they have (Borgman, 2015), especially the uses that do not lead to citations. This can help librarians and curators to support effective uses (MacMillan, 2014) as well as others evaluating the costs and benefits of this activity.

References

Anagnostou, P., Capocasa, M., Milia, N., & Bisol, G. D. (2013). Research data sharing: Lessons from forensic genetics. Forensic Science International: Genetics, 7(6), e117-e119.

---

[3] http://www2.hawaii.edu/~ogaoue/papers/BOT_612_StatR_2015.pdf

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Borgman, C. L. (2015). If data sharing is the answer, what is the question? ERCIM News, 100, 15-16.

Caetano, D. S., & Aisenberg, A. (2014). Forgotten treasures: the fate of data in animal behaviour studies. Animal Behaviour, 98(1), 1-5.

Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. Proceedings of the American Society for Information Science and Technology, 51, 1-4. http://onlinelibrary.wiley.com/doi/10.1002/meet.2014.14505101072/full

Dorch, S. B. F. (2012). On the citation advantage of linking to data: Astrophysics. https://hal-hprints.archives-ouvertes.fr/hprints-00714715/document/

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. Journal of Informetrics, 9(4), 895-906. doi: 10.1016/j.joi.2015.09.005

Gleditsch, N. P., Metelits, C., & Strand, H. (2003). Posting your data: Will you be scooped or will you be famous. International Studies Perspectives, 4(1), 89-97. http://dryadwiki.vm.duke.edu/images/a/ae/IntlStudiesSymp.pdf

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. Cataloging & Classification Quarterly, 47(3-4), 380-402.

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications: an investigation of Dryad Digital Repository. Aslib Journal of Information Management, 68(4), 478-494. Doi: 10.1108/AJIM-01-2016-0008

Henneken, E. A., & Accomazzi, A. (2011). Linking to data-effect on citation rates in astronomy. arXiv preprint arXiv:1111.3618.

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials, 11(9), http://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-11-9 doi:10.1186/1745-6215-11-9

Hsu, L., Lehnert, K. A., Goodwillie, A., Delano, J. W., Gill, J. B., Tivey, M. A., ... & Arko, R. A. (2015). Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. GeoResJ, 6, 108-114.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters, 5(5), 399-406.

Ingwersen, P., & Chavan, V. (2011). Indicators for the data usage index (DUI): An incentive for publishing primary biodiversity data through global information infrastructure. BMC Bioinformatics, 12(Suppl 15): S3.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... & Mangion, J. (2009). Repeatability of published microarray gene expression analyses. Nature genetics, 41(2), 149-155.

Kenall, A., Harold, S., & Foote, C. (2014). An open future for ecological and evolutionary data? BMC Ecology, 14(1), 10.

Krause, E. M., Clary, E., Ogletree, A., & Greenberg, J. (2015). Evolution of an application profile: advancing metadata best practices through the dryad data repository. In Proceedings of the 2015 International Conference on Dublin Core and Metadata Applications. New York, NY: ACM Press (pp. 63-75).

Kroon-Batenburg, L. M., & Helliwell, J. R. (2014). Experiences with making diffraction image data available: what metadata do we need to archive? Acta Crystallographica Section D: Biological Crystallography, 70(10), 2502-2509.

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. The Journal of Academic Librarianship, 40(5), 541-549.

Mennes, M., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2013). Making data sharing work: the FCP/INDI experience. Neuroimage, 82, 683-691.

Miller, G. W. (2016). Making data accessible: The dryad experience. Toxicological Sciences, 149(1), 2-3.

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The enduring value of social science research: the use and reuse of primary research data. In Organisation, Economics and Policy of Scientific Research workshop, Torino, Italy. https://deepblue.lib.umich.edu/handle/2027.42/78307

Piwowar, H.A., & Chapman, W. (2008). A review of journal policies for sharing research data. ELPUB2008. http://ocs.library.utoronto.ca/index.php/Elpub/2008/paper/view/684/0

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PLoS one, 2(3), e308.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175.

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (in press). Analyzing data citation practices using the Data Citation Index. Journal of the Association for Information Science and Technology. doi:10.1002/asi.23529

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLoS Comput Biol, 9(10), e1003285.

Sears, J. R. (2011). Data sharing effect on article citation rate in paleoceanography. In AGU fall meeting abstracts. http://adsabs.harvard.edu/abs/2011AGUFMIN53B1628S

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. PLoS ONEe, 6(6), e21101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE, 10(8), e0134826.

Thelwall, M. & Kousha, K. (2016). Figshare: A universal repository for academic resource sharing? Online Information Review, 40(3), 333-346. doi:10.1108/OIR-06-2015-0190

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics. (pp. 417-424).

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS one, 8(7), e67332.

Zitt, M. (2012). The journal impact factor: Angel, devil, or scapegoat? A comment on JK Vanclay's article 2011. Scientometrics, 92(2), 485-503.