



## Library Hi Tech

Do usage counts of scientific data make sense? An investigation of the Dryad repository

Lin He, Zhengbiao Han,

### Article information:

To cite this document:

Lin He, Zhengbiao Han, (2017) "Do usage counts of scientific data make sense? An investigation of the Dryad repository", Library Hi Tech, Vol. 35 Issue: 2, pp.332-342, <https://doi.org/10.1108/LHT-12-2016-0158>

Permanent link to this document:

<https://doi.org/10.1108/LHT-12-2016-0158>

Downloaded on: 10 May 2018, At: 01:59 (PT)

References: this document contains references to 44 other documents.

To copy this document: [permissions@emeraldinsight.com](mailto:permissions@emeraldinsight.com)

The fulltext of this document has been downloaded 425 times since 2017\*

### Users who downloaded this article also downloaded:

(2017), "Research data management in Turkey: perceptions and practices", Library Hi Tech, Vol. 35 Iss 2 pp. 271-289 <a href="https://doi.org/10.1108/LHT-11-2016-0134">https://doi.org/10.1108/LHT-11-2016-0134</a>

(2017), "An analysis of the changing role of systems librarians", Library Hi Tech, Vol. 35 Iss 2 pp. 303-311 <a href="https://doi.org/10.1108/LHT-08-2016-0092">https://doi.org/10.1108/LHT-08-2016-0092</a>

Access to this document was granted through an Emerald subscription provided by emerald-srm:395687 []

### For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit [www.emeraldinsight.com/authors](http://www.emeraldinsight.com/authors) for more information.

### About Emerald [www.emeraldinsight.com](http://www.emeraldinsight.com)

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

\*Related content and download information correct at time of download.

# Do usage counts of scientific data make sense? An investigation of the Dryad repository

Lin He and Zhengbiao Han  
*Nanjing Agricultural University, Nanjing, China*

332

Received 28 December 2016  
Revised 18 April 2017  
Accepted 18 April 2017

## Abstract

**Purpose** – The purpose of this paper is to evaluate the impact of scientific data in order to assess the reliability of data to support data curation, to establish trust between researchers to support reuse of digital data and encourage researchers to share more data.

**Design/methodology/approach** – The authors compared the correlations between usage counts of associated data in Dryad and citation counts of articles in Web of Science in different subject areas in order to assess the possibility of using altmetric indicators to evaluate scientific data.

**Findings** – There are high positive correlations between usage counts of data and citation counts of associated articles. The citation counts of article's shared data are higher than the average citation counts in most of the subject areas examined by the authors.

**Practical implications** – The paper suggests that usage counts of data could be potentially used to evaluate scholarly impact of scientific data, especially for those subject areas without special data repositories.

**Originality/value** – The study examines the possibility to use usage counts to evaluate the impact of scientific data in a generic repository Dryad by different subject categories.

**Keywords** Bibliometrics, Data sharing, Citation counts, Dryad repository, Scientific data, Usage counts

**Paper type** Research paper

## Introduction

Researchers are required to share scientific data produced in their research to public, as the requirements of funding agencies and journal publishers (The National Science Foundation, 2011). Data repositories are developing in rapid speeds and lots of repositories have been in use in various domains, which play a significant role with regard to data preservation, data sharing and data reuse (Hey *et al.*, 2009; Pham-Kanter *et al.*, 2014; Borgman, 2012). However, the curators and researchers are facing some problems in data preservation, sharing and reuse. They always struggle with making judgments for which data sources are of enough value to be collected from tremendous amounts of digital data (Uhlir, 2010). More generally, most of researchers encounter problems when they reuse or re-analyze data due to lack of evidence of the credibility of scientific data (Pham-Kanter *et al.*, 2014). Despite data sharing increasingly widespread, it is not clear how to evaluate whether scientific data are effective and valuable. On the other hand, data sharing is often seen as an additional time-consuming effort, some researchers are not willing to share their primary data. Researchers need to provide academic impact benefits to encourage them to share more data in practice (Tenopir *et al.*, 2011). It is important to find a way to assess the perceived quality of the shared data and the visibility of those data from academic user communities through data repositories.

Over the past few decades, widely accepted impact evaluation indicators have been built for academic publications which fundamentally depend on citation counts (Garfield, 1979; Norris and Oppenheim, 2010). However, it would not be wise to evaluate the academic impact of scientific data based on their citation counts in the articles (Ingwersen and Chavan, 2011; Ingwersen, 2014) because they are rarely cited or cited instead of the associated article (Fear, 2013; Piwowar *et al.*, 2007). Moreover, general and consistent standardized data citation criteria have not been in use (CODATA-ICSTI Task Group, 2013; Spengler, 2012; Ball and Duke, 2015; California Digital Library, 2015; Moritz *et al.*, 2011) and



centralized citation index databases are still in progress (Costas *et al.*, 2013; Thomson Reuters, 2012).

Most data repositories, such as Dryad, publish the counts of views and downloaded of data sets on their websites and enable long-term, sustainable preservation of data for researchers (Greenberg, 2009). Alternative online metrics, such views, saves, and recommendations are more informative for digital resources (Konkiel, 2013) and they are widely used for impact evaluation of academic publications (Thelwall and Kousha, 2015a, b). So it is possible to use them to analyze the impact of scientific data. Although there are a few studies that have shown that there is an association between data sharing and citation counts (Piwowar *et al.*, 2007, Piwowar, 2011; Piwowar and Chapman, 2010; Belter, 2014), they all focused on individual disciplines and types of resource. The possibility of alternative online metrics in general multidisciplinary repositories is still needed in order to obtain advice for data curators and researchers about how useful the data are in their user communities.

This paper explores the possibility of using altmetric indicators to evaluate the impact of scientific data based on a general multidisciplinary repository. The Dryad repository was chosen as it is a widely accepted open archive for general-purpose unstructured scientific data and has been widely recommended as one of the best choices of non-specific repositories by many journals and funding agencies. The object of this paper is to assess the influence of data usage counts on article citation counts and the influence of data usage counts on the extent to which articles attract more citations. The following research questions drive the study:

- RQ1.* Do usage counts of data in Dryad associate with citation counts of their corresponding articles in Web of Science (WoS)?
- RQ2.* Do articles which deposited data in Dryad receive more citation counts than articles published in each WoS subject area on average?

These questions are addressed by investigating the relationship between usage statistics of research data sets in Dryad and citation counts of associated articles in WOS. WoS was used to access citation counts of publications because of its wider use in academic communications. If the answers to the questions were positive, the altmetric indicators in data repositories would be potential bibliometric factors to help evaluate academic the impact of data. Also, it would encourage researchers to share their data actively if more academic credit could be obtained by archiving scientific data in the repository.

### Literature review

Some qualitative research methods have been used to assess the trustworthiness of data repositories and the qualification of data reuse (Consultative Committee for Space Data Systems, 2012; Dobratz *et al.*, 2007). The indicators usually include data completeness, relevancy, accessibility and credibility as well as document quality and data producer reputation (Faniel *et al.*, 2015). However, bibliometric indicators are essential for obtaining quantitative measures for the academic impact of articles and research data. They are more objective and overcome the problems of subjective human intervention caused by the different educational backgrounds and qualifications of involved individuals.

Impact indicators derived from the web are widely used for academic articles and other academic outputs, which are not restricted by the citation index database and citation standards (Thelwall *et al.*, 2013; Costas *et al.*, 2013; Priem *et al.*, 2010). These web indicators are available for outputs which are recently published to have attracted many citations (Thelwall and Wilson, 2016). There are some research studies using bibliometric or altmetric indicators to assess the impact of scientific data (Ingwersen, 2014; Harris, 2014; Hicks *et al.*, 2015). For example, views and shares of resources were used to analyze the use of the

resources in the generic repository Figshare (Thelwall and Kousha, 2016). Ingwersen and Chavan (2011) designed Data Usage Index (DUI) to measure the impact of shared data, which are usage indicators including being accessed and used by user communities. But the DUI constitutes the only feasible indicators for biodiversity data sets, they are difficult to elicit explicitly from scientific publications in other research fields (Ingwersen, 2014). Fear (2013) proposed to measure the reuse counts of research data in social science repository named ICPSR and suggested that reuse counts may not be an especially useful metric for such data. The existing studies focused on individual disciplines and types of resource, each of them has their own features in scientific sharing and reuse. So, it is necessary to explore the possibility of impact indicators of scientific data in multidisciplinary repositories based on social utility-based research metrics.

Although half of the academic journals seem to have a data sharing policy for submitted articles (Sturges *et al.*, 2014), contrasting with huge amount of producing data in scientific research, data shared in public is still much lower (Ochsner *et al.*, 2008). One of the most important reasons is that some researchers do not believe that they could receive benefits if they spent much time on data sharing (Tenopir *et al.*, 2011). Therefore, researchers would be more likely to share data if others could benefit from their sharing data and their academic impacts could be improved (Hedstrom and Niu, 2008; European Commission, 2011). In some specific research areas, such as microarray and oceanography, studies have identified that data set sharing frequencies are associated with journal impact factors, and citation counts of articles with sharing data are higher than those without data shared (Piwowar *et al.*, 2007; Piwowar, 2011; Piwowar and Chapman, 2010; Belter, 2014). However, in more disciplines, it is necessary to find more evidence to show the association between data sharing and citation counts of publications.

### Data and methods


The research design is to download the metadata records of Dryad data sets from Dryad website ([www.datadryad.org/](http://www.datadryad.org/)), identify the download times of data sets and the publications which made references to the data sets in Dryad. The subsequent citation counts of the publications were obtained from WoS due to its comprehensive coverage of academic publications. We accessed all of the data sets in December of 2015.

#### *Dryad repository*



If a researcher's publication is ready for peer-review process, the supporting data in multiple formats should be submitted to Dryad before submitting the publication as the general requirement of journal publishers. After submitting data associated with a publication is identified by a Dryad curator, the data sets will receive persistent DOIs for use in citation. Usually, the publication of data is published much earlier than the articles to the readers due to delay of peer-review and publications period of journal publishers. The data have been downloaded many times before the associated article is cited by other publications. An example of data described in Dryad is shown in Figure 1.

A total of 10,820 metadata records of data sets in Dryad (until December of 2015) were downloaded by programming according to their URLs. First, the Dryad website sitemap was consulted for a complete list of URLs which contained the DOIs of metadata records of data sets. Then we use regular expressions written in Python to extract all the DOIs of metadata of data sets contained in web pages. According to the DOIs of data, we downloaded the metadata of data sets by using automatic crawling software written in Python. The downloaded metadata of data sets included original journal information of associated articles, publication date of articles, titles of the articles, DOIs of articles, DOIs of usage counts of associated data sets in Dryad.

**Data from: Parasitic plants have increased rates of molecular evolution across all three genomes**



**Files in this package**

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.  

Title	Sister Clade Comparisons
Downloaded	10772 times
Description	Tree files, alignments, PAML executables and associated command files for sister pair rates estimation of parasite and nonparasite clades. Sequence data compiled from GenBank accessions (see paper for details). Additional information included in README file
Download	<a href="#">README.txt (7.558 Kb)</a>
Download	<a href="#">Comparisons.zip (25.69 Mb)</a>
Details	<a href="#">View File Details</a>

---

When using this data, please cite the original publication:

Bromham L, Cowman PF, Lanfear R (2013) Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evolutionary Biology 13: 126. <http://dx.doi.org/10.1186/1471-2148-13-126>

Additionally, please cite the Dryad data package:

Bromham L, Cowman PF, Lanfear R (2013) Data from: Parasitic plants have increased rates of molecular evolution across all three genomes. Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.fc74k>

[Cite](#) | [Share](#)

**Figure 1.**  
Interface of dryad  
data description

### *Citation counts in WoS*

Citation counts of articles associated with shared data sets in Dryad were obtained from the website of WoS, according to the DOIs of the articles identified from metadata records of data sets from Dryad. In order to get a bird's-eye view of average citation counts, we also accessed the number of publications in each journal from 2010 to 2015 and their total citation counts of each journal per year.

Here, we chose the taxonomy of Journal Citation Reports (JCR) as our fundamental principles to identify the subject categories of publications and data sets. Not all of the articles with Dryad data sets records are indexed by WOS due to its coverage limitation.

Only 9,333 out of 10,820 of associated articles are indexed by WOS. Therefore, in this paper, statistical analyses related to subject areas are all based on these 9,333 indexed articles.

For average citation counts in different subject areas, we first got the journal lists of each subject area from JCR, and then summed up the number of publications and total citation counts of each journal in the same subject area by year. Eventually, average citation counts of each subject area by year could be calculated in order to make a comparison of citation counts between articles which shared data in Dryad and the overall articles in each subject.

## Results

### *General statistics of publications in Dryad*

In total, 10,820 of downloaded data were grouped by publication year (up to December 2015), and the number of publications in each year are shown in Figure 2. Before 2010, there are only few publications that shared data sets in Dryad; however, since 2010, there is an exponential growth trend in the number of depositing data sets in Dryad.

Table I shows the number of publications in Dryad in different subject areas as well as the total counts of publications in different subject areas in WoS. It is easy to find that there is not any data sets with zero download times, whereas most of the associated articles are uncited in WOS.

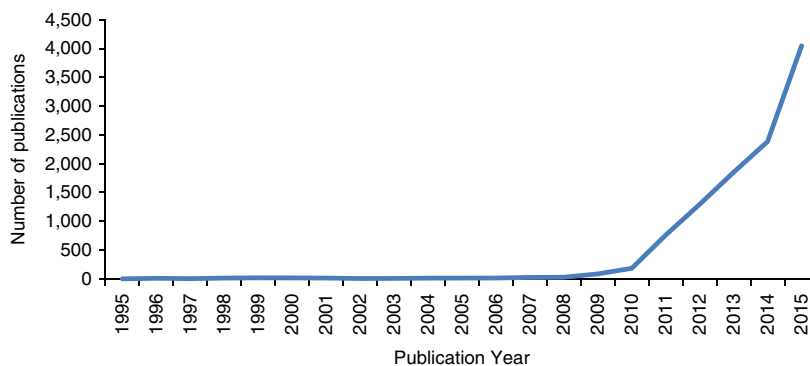
### *Correlation analysis between usage counts in Dryad and citation counts in WoS*

Pearson correlation was used to calculate the relationship between citation counts of articles and download times of the associated data sets in Dryad (Table II). From Table II, it shows that there are 11 subject areas which have significant correlations between citation counts and download times among 15 subject areas we examined. There are only four subject areas which have no significant correlations. That is to say, there are positive correlations between download times of data sets in Dryad and citation counts of associated articles.

Average citation counts are also reported in contrast to downloaded times in general. Geometric mean of citation counts and download times in each year from 2004 to 2015 are shown in Figure 3. The shape of the citation counts graph is substantially similar to the shape of download times graph by publication time.

### *Citation counts of publications with associated data sets in Dryad compared with total citations in different subject areas*

Average citation counts were calculated for each subject area at the level of all journals indexed in WoS from 2010 to 2015, for the purpose of making a comparison with the citation counts of publications which deposited data in Dryad. We calculated the citation counts of



**Figure 2.**  
Number of data  
published in Dryad  
in different years

Subject area	Data associated articles published in Dryad	Total articles in WoS	Percentage of articles in Dryad accounting for articles in WoS	Zero downloaded counts of data in Dryad	Percentage of zero citation counts of articles in WoS
Behavioral sciences	41	8,674	0.47	0	15
Biochemical research methods	23	2,622	0.88	0	57
Biochemistry and molecular biology	1,873	22,576	8.30	0	15
Biodiversity conservation	147	5,779	2.54	0	25
Biology	329	16,559	1.99	0	22
Cell biology	42	50,674	0.08	0	7
Ecology	3,686	38,673	9.53	0	25
Evolutionary biology	723	5,425	13.33	0	22
Genetics and heredity	320	15,822	2.02	0	30
Medicine, general and internal	123	36,368	0.34	0	44
Microbiology	32	14,282	0.22	0	31
Multidisciplinary sciences	1,291	185,912	0.69	0	45
Paleontology	88	5,351	1.64	0	33
Plant sciences	163	23,606	0.69	0	33
Zoology	74	14,114	0.52	0	20

**Table I.**  
The subject areas and the number of publications in Dryad compared with general number of publications in WoS

Subject categories	Pearson coefficient	<i>P</i>
Behavioral sciences	0.593**	0.000
Biochemical research methods	0.059	0.791
Biochemistry and molecular biology	0.0139**	0.000
Biodiversity conservation	0.364**	0.000
Biology	0.487**	0.000
Cell biology	0.205	0.193
Ecology	0.397**	0.000
Evolutionary biology	0.032	0.386
Genetics and heredity	0.172**	0.002
Medicine, general and internal	0.165	0.068
Microbiology	0.369**	0.038
Multidisciplinary sciences	0.066*	0.019
Paleontology	0.303**	0.004
Plant sciences	0.553**	0.000
Zoology	0.09	0.444

**Notes:** \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

**Table II.**  
Pearson correlations between downloaded times in Dryad and citation counts in WOS

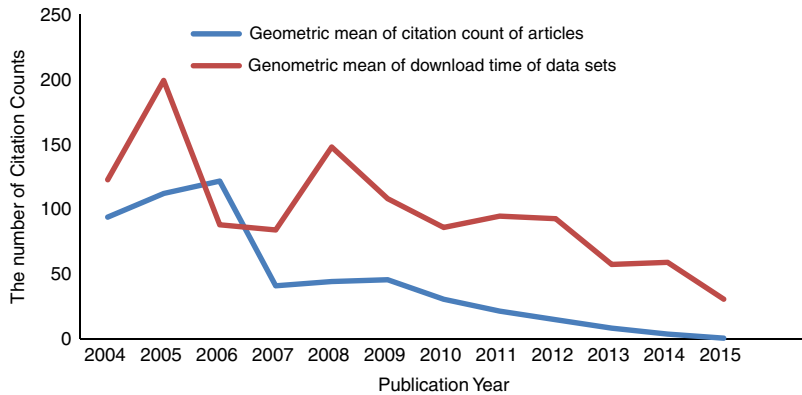
publications which published data in Dryad that were higher than the overall average citation counts in each subject area from 2010 to 2015. The figure in Table III shows the numbers of publications shared data in Dryad whose citation counts are higher than average citation counts in each subject area.

## Discussion

To answer the first research question, the Pearson coefficients were calculated. Among 15 subject areas we examined, there are 11 subject areas accounting for 73 percent of all subject areas, which have significant correlations between articles in WOS and their associated data sets in Dryad. In medicine, general and internal, biochemical research

methods, cell biology, evolutionary biology and zoology, there are no significant correlations between download times of data sets in Dryad and associated articles in WOS. One possible reason would be that shared data could be used in other ways than leading to new citations (Tenopir *et al.*, 2011; Wallis *et al.*, 2013). The shared data could be used for result verification, data understanding and method triangulation in different subject areas and other educational or background used for research (Thelwall and Kousha, 2017). Another reason to explain the different roles of data sets in research is the number of download times of data sets in Dryad compared with the number of citation counts of associated articles in WoS. All of the shared data sets are downloaded by researchers at least several times, however, lots of associated articles in WoS are uncited. So probably we can say that the data sets play different roles in researches rather than leading to new citations.

In particular, a more interesting phenomenon is that the citation counts trend of publications in WoS in the past decade is rather similar to the trend of downloaded times in Dryad shown in Figure 3. More specific, we can find that the shape of geometric download time is earlier than the shape of geometric citation counts in WoS. This could be used to



**Figure 3.** The shapes of geometric mean of citation count of papers in WOS and download times of Dryad data from 2004 to 2015

Subject areas	2010	2011	2012	2013	2014	2015
Behavioral sciences	N	-	-	+	+	+
Biochemical research methods	+	-	-	-	-	+
Biochemistry and molecular biology	-	+	+	-	+	+
Biodiversity conservation	N	-	-	-	+	-
Biology	+	+	+	+	+	-
Cell biology	N	+	+	+	+	+
Ecology	+	-	-	+	-	+
Evolutionary biology	+	+	+	+	+	+
Genetics and heredity	N	-	-	-	-	-
Medicine, general and internal	-	-	-	-	-	-
Microbiology	N	+	+	+	+	-
Multidisciplinary sciences	N	+	+	+	+	+
Paleontology	+	N	+	+	+	+
Plant sciences	+	+	-	+	+	+
Zoology	+	+	+	+	+	+

**Table III.** Statistic of citation counts of articles with Dryad data sets compared with geometric mean in each subject areas from 2010 to 2015

**Notes:** + indicate that the numbers of citations of articles with Dryad data sets are higher than geometric means of that subject areas; - indicates the opposite, and N indicates that there is no data sets in Dryad for that year



measure academic impact of research data in early stage if the significance of usage counts could be accepted in the future. Although citation counts should be obtained after a period of longer time due to publication release life cycle, downloaded times could be shown in short period of time. Normally, the researchers submit the data related to publications to Dryad before peer review of the publications, as the requirement of publications journals for better understanding of research arguments.

For the second research question, it is encouraging that in most of the subject areas in Dryad, the number of citation counts of articles with data sets is higher than average citation counts. It could benefit researchers in their additional time-consuming effort to share data. However, in terms of usage for shared data, not all the articles with associated data sets shared in Dryad could lead to more citation counts among all the subject areas we examined, for example in medicine, general and internal and genetics and heredity. However, generally in medicine, general and internal, genetics and heredity, scientific data types are nucleic acid sequence, protein sequence, molecular and supramolecular structure, neuroscience or omics. There are much field-specific repositories, for example, deposition of microarray data in ArrayExpress or GEO, deposition of gene sequences in GenBank, EMBL or DDBJ. Data sets in those subject areas deposited in Dryad may be more general for further explanation or augment for the articles.

Actually, Dryad[1] is a generalist repository that can handle a wide variety of data, and may also be appropriate for storage of associated analyses, or experimental-control data, supplementing the primary data record for some unstructured data. From the perspective of association between data sharing and citation counts, citation counts of publications in subject areas with specific data repositories are not in typical significance level in Dryad repository compared with that without specific data repositories consequently. Generally, Dryad is more popular for evolutionary biology, ecology and some general-purpose research areas. Thus, we can find that the subject areas in those fields with higher levels of Dryad usage have a significant performs on higher citation counts of articles compared with averages level in the same subject areas in fields with higher levels of Dryad usage have a significant performs on higher citation counts of articles compared with averages level in the same subject areas. This could help curators to think about the effective uses and evaluating the benefits and costs of shared data.

### Limitations

The first limitation of this paper is the coverage of Dryad. Although Dryad is a general-purpose scientific data repository for multi-disciplines, the most of deposited data are heavily concentrated on ecology, evolutionary biology; and there are only few data in social science and humanities science, which are reflected in the fourth column of Table I. The second limitation is that not all of the shared data in Dryad are indexed in WoS due to the coverage limitation. So the statistical results are not based on the whole data in Dryad, which would decrease the accuracy of the number to a certain degree. Another reason to lose some records in WoS is that we recognized publications in WoS by DOIs downloaded from Dryad. Some fake DOIs by mistake would mislead the searching of related publications in WoS.

### Conclusions

This paper examined 9,333 out of 18,620 publications in Dryad repository which have subject categories in WoS indexed by JCR. There are higher positive correlations between usage counts of depositing data in Dryad and citation counts of articles in WoS. More interestingly, the shape of downloaded times in Dryad is substantially much similar to the shape of citation counts in WoS in the last decade from 2004 to 2015, but the primary time trend in downloaded times is earlier than in citation counts, which gives impact evidence at an earlier stage than is possible with citation counts. Based on our statistical

results, usage counts of data in Dryad could be a potential positive indicator to measure the impact of their associated articles as well as the influence of scientific data. However, it should be very cautious to use this indicator, since it could be manipulated very easily by person. In addition, it is not meaningful to compare usage counts of data in Dryad with citation counts of articles in WoS. Because citation counts are closely related to ongoing research and usage counts of data merely demonstrate that readers are more interested in the research, both of them would reflect different types of impacts if they were used in applications.

In most of the subject areas we estimated in this paper, citation counts of publications shared data in Dryad are higher than average citation counts in the same subject area in WoS based on the publications published from 2010 to 2015. For those subject areas which are in higher levels of Dryad use or have no specific data repositories, the publications would get higher citations counts if researchers deposit data records associated with publications in Dryad. The conclusion also suggests that it has more advantage to promote academic impact of research articles if associated data are shared in appropriate repository. The conclusion also coincides with earlier studies about the association between research credits and data sharing in genomic research (Piwowar and Chapman, 2010).

#### Note

1. [www.nature.com/sdata/data-policies/repositories](http://www.nature.com/sdata/data-policies/repositories)

#### References

- Ball, A. and Duke, M. (2015), *How to Track the Impact of Research Data with Metrics*, DCC How-to Guides, Digital Curation Centre, Edinburgh, available at: [www.dcc.ac.uk/resources/how-guides](http://www.dcc.ac.uk/resources/how-guides) (accessed January 9, 2016).
- Belter, C.W. (2014), "Measuring the value of research data: a citation analysis of oceanographic data sets", *PLoS ONE*, Vol. 9 No. 3, p. e92590.
- Borgman, C.L. (2012), "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078.
- California Digital Library (2015), "Why Use EZID?", available at: <http://ezid.cdlib.org/home/why> (accessed December 11, 2015).
- CODATA-ICSTI Task Group (2013), "Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data", *Data Science Journal*, Vol. 12, pp. CIDCR1-CIDCR75, available at: <http://doi.org/10.2481/dsj.OSOM13-043>
- Consultative Committee for Space Data Systems (2012), "Space data and information transfer systems – audit and certification of trustworthy digital repositories", Standard No. ISO 16363:2012 (CCSDS 652-R-1), Consultative Committee for Space Data Systems, Washington, DC, available at: [www.iso.org/iso/catalogue\\_detail.htm?csnumber=56510](http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510) (accessed January 6, 2016).
- Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013), "The value of research data – metrics for datasets from a cultural and technical point of view", *A Knowledge Exchange Report*, pp. 1-48.
- Dobratz, S., Schoger, A. and Strathmann, S. (2007), "The Nestor catalogue of criteria for trusted digital repository evaluation and certification", *Journal of Digital Information*, Vol. 8 No. 2, pp. 75-86, available at: <http://journals.tdl.org/jodi/index.php/jodi/article/view/199/180> (accessed January 11, 2016).
- European Commission (2011), "Digital agenda: turning government data into gold", available at: [http://europa.eu/rapid/press-release\\_IP-11-1524\\_en.htm](http://europa.eu/rapid/press-release_IP-11-1524_en.htm) (accessed January 8, 2016).
- Faniel, I.M., Kriesberg, A. and Yake, E. (2015), "Social scientists' satisfaction with data reuse", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 6, pp. 1404-1416.

- Fear, K.M. (2013), "Measuring and anticipating the impact of data reuse", dissertation, University of Michigan, Ann Arbor, MI, available at: <http://deepblue.lib.umich.edu/handle/2027.42/102481> (accessed January 3, 2016).
- Garfield, E. (1979), *Citation Indexing – its Theory and Application in Science, Technology, and Humanities*, John Wiley & Sons, Ltd, New York, NY.
- 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*, Vol. 47 Nos 3/4, pp. 380-402.
- Harris, S. (2014), "Acquisition opens up altmetrics options. Research information", available at: [www.researchinformation.info/features/feature.php?feature\\_id=490](http://www.researchinformation.info/features/feature.php?feature_id=490) (accessed January 2, 2016).
- Hedstrom, M. and Niu, J. (2008), "Incentives for data producers to create 'archive-ready' data: implications for archives and records management", National Digital Information Infrastructure and Preservation Program (NDIPP) Partners Meeting, Berkeley, CA.
- Hey, T., Tansley, S. and Tolle, K. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA, available at: <http://research.microsoft.com/en-us/collaboration/fourthparadigm> (accessed January 2, 2016).
- Hicks, D., Wouters, P., Waltman, L., Rijcke, S. de and Rafols, I. (2015), "Bibliometrics: the Leiden Manifesto for research metrics", *Nature*, Vol. 520 No. 7548, pp. 429-431, doi: 10.1038/520429a.
- Ingwersen, P. (2014), "Scientific datasets: informetric characteristics and social utility metrics for biodiversity data sources", *Library and Information Sciences*, Springer, Berlin, Heidelberg, pp. 107-117.
- Ingwersen, P. and Chavan, V. (2011), "Indicators for the data usage index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure", *BMC Bioinformatics*, Vol. 12 No. 15, p. S3.
- Konkiel, S. (2013), "Tracking citations and altmetrics for research data: challenges and opportunities", *Bulletin of the American Society for Information Science and Technology*, Vol. 39 No. 6, pp. 27-32.
- Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, Y., Cockerill, M. and Chavan, V. (2011), "Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF data publishing framework task group", *BMC Bioinformatics*, Vol. 12 No. S15, pp. 1-10.
- Norris, M. and Oppenheim, C. (2010), "The h-index: a broad review of a new bibliometric indicator", *Journal of Documentation*, Vol. 66 No. 5, pp. 681-705, doi: 10.1108/00220411011066790.
- Ochsner, S.A., Steffen, D.L., Stoeckert, C.J. and McKenna, N.J. (2008), "Much room for improvement in deposition rates of expression microarray datasets", *Nature Methods*, Vol. 5 No. 12, p. 991.
- Pham-Kanter, G., Zinner, D.E. and Campbell, E.G. (2014), "Codifying collegiality: recent developments in data sharing policy in the life sciences", *PLoS ONE*, Vol. 9 No. 9, p. e108451.
- Piwovar, H.A. (2011), "Who shares? Who doesn't? Factors associated with openly archiving raw research data", *PLoS ONE*, Vol. 6 No. 7, p. e18657.
- Piwovar, H.A. and Chapman, W.W. (2010), "Public sharing of research datasets: a pilot study of associations", *Journal of Infometrics*, Vol. 4 No. 2, pp. 148-156.
- Piwovar, H.A., Day, R.S. and Fridsma, D.B. (2007), "Sharing detailed research data is associated with increased citation rate", *PLoS ONE*, Vol. 2 No. 3, p. e308.
- Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010), "Altmetrics: a manifesto", available at: <http://altmetrics.org/manifesto/> (accessed January 9, 2016).
- Spengler, S. (2012), *Data Citation and Attribution: A Funder's Perspective*, National Academies Press, Washington, DC, pp. 177-188.
- Sturges, P., Bamkin, M., Anders, J. and Hussain, A. (2014), "Access to research data: addressing the problem through journal data sharing policies", *Proceedings of the IATUL Conferences, Helsinki, June 2-5*, available at: <http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2012&context=iatul> (accessed January 9, 2016).

- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E. and Frame, M. (2011), "Data sharing by scientists: practices and perceptions", *PLoS ONE*, Vol. 6 No. 6, p. e21101.
- The National Science Foundation (2011), "Dissemination and sharing of research results: national science foundation office of budget, finance, and award management. government", available at: [www.nsf.gov/bfa/dias/policy/dmp.jsp](http://www.nsf.gov/bfa/dias/policy/dmp.jsp) (accessed January 15, 2016).
- Thelwall, M. and Kousha, K. (2015a), "Web indicators for research evaluation, part 1: citations and links to academic articles from the web", *El Profesional de la Información*, Vol. 24 No. 5, pp. 587-606, doi: 10.3145/epi.2015.sep.08.
- Thelwall, M. and Kousha, K. (2015b), "Web indicators for research evaluation, part 2: social media metrics", *El Profesional de la Información*, Vol. 24 No. 5, pp. 607-620, doi: 10.3145/epi.2015.sep.09.
- Thelwall, M. and Kousha, K. (2016), "Figshare: a universal repository for academic resource sharing?", *Online Information Review*, Vol. 40 No. 3, pp. 333-346, doi: 10.1108/OIR-06-2015-0190.
- Thelwall, M. and Kousha, K. (2017), "Do journal data sharing mandates work? Life sciences evidence from Dryad", *Aslib Journal of Information Management*, Vol. 69 No. 1, pp. 36-45, doi: 10.1108/AJIM-09-2016-0159.
- Thelwall, M. and Wilson, P. (2016), "Mendeley readership altmetrics for medical articles: an analysis of 45 fields", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 8, pp. 1962-1972.
- Thelwall, M., Haustein, S., Larivière, V. and Sugimoto, C.R. (2013), "Do altmetrics work? Twitter and ten other social web services", *PLoS ONE*, Vol. 8 No. 5, p. e64841.
- Thomson Reuters (2012), "Repository evaluation, selection, and coverage policies for the data citation index within Thomson Reuters web of knowledge", available at: [http://wokinfo.com/products\\_tools/multidisciplinary/dci/selection\\_essay/](http://wokinfo.com/products_tools/multidisciplinary/dci/selection_essay/) (accessed December 12, 2015).
- Uhlir, P.F. (2010), "Information Gulags, intellectual straightjackets, and memory holes: three principles to guide the preservation of scientific data", *Data Science Journal*, Vol. 9 No. 6, pp. ES1-ES5, available at: <http://doi.org/10.2481/dsj.Essay-001-Uhlir>
- Wallis, J.C., Rolando, E. and 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*, Vol. 8 No. 7, p. e67332.

### Further reading

- Larsen, P.O. and von Ins, M. (2009), "The steady growth of scientific publication and the declining coverage provided by science citation index", in Larsen, B. and Leta, J. (Eds), *Proceedings of ISSI 2009 – 12th International Conference of the International Society for Scientometrics and Informetrics*, Vol. 2, Int Soc Scientometrics & Informetrics-Issi, Leuven, pp. 597-606.
- Pendlebury, D.A. (2008), *Using Bibliometrics in Evaluating Research*, Research Department, Thomson Scientific, Philadelphia, PA.
- Zahedi, Z., Costas, R. and Wouters, P. (2014), "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications", *Scientometrics*, Vol. 101 No. 2, pp. 1491-1513.

### Corresponding author

Lin He can be contacted at: [helin@njau.edu.cn](mailto:helin@njau.edu.cn)

---

For instructions on how to order reprints of this article, please visit our website:

[www.emeraldgroupublishing.com/licensing/reprints.htm](http://www.emeraldgroupublishing.com/licensing/reprints.htm)

Or contact us for further details: [permissions@emeraldinsight.com](mailto:permissions@emeraldinsight.com)