Aslib Journal of Information Management

Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository

Lin He Vinita Nahar

Reuse of scientific data in academic publications

An investigation of Dryad Digital Repository

Lin He
Department of Information Science, Nanjing Agricultural University, Nanjing, China, and
Vinita Nahar
Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK

Abstract

Purpose – In recent years, a large number of data repositories have been built and used. However, the extent to which scientific data are re-used in academic publications is still unknown. The purpose of this paper is to explore the functions of re-used scientific data in scholarly publication in different fields.
Design/methodology/approach – To address these questions, the authors identified 827 publications citing resources in the Dryad Digital Repository indexed by Scopus from 2010 to 2015.
Findings – The results show that: the number of citations to scientific data increases sharply over the years, but mainly from data-intensive disciplines, such as agricultural, biology science, environment science and medicine; the majority of citations are from the originating articles; and researchers tend to reuse data produced by their own research groups.
Research limitations/implications – Dryad data may be re-used without being formally cited.
Originality/value – The conservatism in data sharing suggests that more should be done to encourage researchers to re-use other's data.

Keywords Citation analysis, Data repositories, Academic publications, Data reuse, Dryad Digital Repository, Research data

Paper type Research paper

Introduction

With the rapid growth of science and technology, there is a significant inclination toward data-driven research. Data-driven research depends heavily on large data sets, which cannot easily be produced independently. Typically, these research fields using data-driven approaches include life sciences, earth sciences and geographical sciences, etc. It is in the interest of all funding agencies, scientific institutions and research communities to deposit scientific data, which have been produced in the process of research, in open access data repositories. Depositing scientific data in public repositories has several advantages from advancing research innovation to retaining data integrity by well-managed and long-term data preservation. Researchers can reuse shared data to reproduce research, validate research results and propose new research in relation to existing relevant data. At the same time, citation counts of the publications would increase, if the relevant data were shared publicly (Borgman, 2012; Piwowar *et al.*, 2007; Piwowar, 2011).

By 2014, re3data.org[1] had indexed over 1,000 research data repositories from all over the world, which makes it the largest and the most comprehensive online catalogue of research data repositories on the web. These indexed data can be differentiated in institutional, disciplinary, multidisciplinary and project-specific

Aslib Journal of Information Management
Vol. 68 No. 4, 2016
pp. 478-494

2050-3806
DOI 10.1108/AJIM-01-2016-0008

scientific data repositories (Pampel *et al.*, 2013). Prominent examples of discipline-based scientific data repositories are GenBank[2] in genetic sequences, PANGAEA[3] in earth and environmental science and HEASARC[4] in astronomy science. Figshare[5], Dryad[6] and LabArchives[7] serve for the multidisciplinary research needs of scientific data deposition.

For domain-specific repositories, data are managed by disciplinary or national infrastructures that are responsible for collecting, storing, preserving and providing data to researchers. It has been investigated that data repositories have played crucial roles in some data-intensive areas (Pham-Kanter *et al.*, 2014). However, compared to the huge investment in discipline-based scientific data repositories, data repositories for multidisciplinary research needs have not got enough attention, and the sharing of research data remains a limited activity (Cragin *et al.*, 2010). Little is known about how and why researchers re-used data shared by others in different research fields, particularly from the perspective of bibliometric analysis. In order to get a bird's eye view of the wide range of research areas, a general-purpose widely accepted open archive of the scientific data should be selected as a data source. Hence, in this paper, Dryad Digital Repository (DDR)[8] is selected as the data source, which is a curated resource that makes the data underlying scientific publications freely discoverable, reusable and citable for a wide diversity of data types. It has been widely recommended as one of the best choices if a non-specific repository is selected by many journals or funding agencies (Nature, 2015).

This paper will address three research questions taking DDR, a typical multidisciplinary repository, as an example:

RQ1. Have scientific data in DDR been widely re-used in different fields when data are available publically?

RQ2. What are the main functions of the re-used data in DDR if researchers cite re-used data in their publications?

RQ3. What proportion of shared data are re-used among depositing data in DDR?

In order to answer these research questions, this paper aims to examine how scientific data are formally cited in different disciplines within academic publications. The paper also aims to explore the reasons behind citing data produced by others, and the functions of re-used data in the new research articles. In accordance with these research objectives and questions, the rest of the paper is organized as follows. Second section outlines the background and related research on the development of data sharing and data reuse, for the benefit of policy makers, journals and funding agencies. Third section briefly explains the research methodology, which is based on the citation references to the data repository from Scopus by using bibliometric analysis. Fourth section presents the statistical results of data citations, and fifth section provides discussions and key findings on the function of the re-used data in different fields. Sixth section discusses the limitation of this paper. Finally, seventh section provides concluding remarks.

Background and related work

In the last decade, digital scientific data preservation in a variety of research fields has increased in number and in scope by the effort of policy makers, funding bodies, publishing agencies and scientists (Hey *et al.*, 2009). A recent survey shows that third-party repositories and online supplements, as well as data sharing requirements of

funding agencies, particularly the National Institutes of Health and the National Human Genome Research Institute, were perceived by scientists to have a significant impact on facilitating data sharing (Pham-Kanter *et al.*, 2014). Policy makers, publishing bodies and funding agencies also strongly believe that shared data are important and useful for researchers, which benefit the development of science (Borgman, 2012).

Some data-intensive research fields (such as the life sciences and earth sciences) with a long history of data sharing, have some strong examples to demonstrate that data sharing and data reuse have benefitted their scientific research to a great extent due to their distinct features in data production (Kenall *et al.*, 2014; Kaye *et al.*, 2009; Ochsner *et al.*, 2008). Many successful cases that re-used shared data to produce new research have been reported in the past, such as species records in biodiversity research (Faith *et al.*, 2013; Moritz *et al.*, 2011; Barve, 2014) and human biological samples (Chen, 2013). Researchers discovered three new species of the bacterial endosymbiont *Wolbachiapipientis* in the three different species of fruit fly using the raw data deposited in Trace Archive (Salzberg *et al.*, 2005). The study focussed on the benefits to researchers of having publicly available raw data. Johan Rung and Alvis Brazma retrieved publications that had used public gene expression data from ArrayExpress Archive (Rung and Brazma, 2013). They found that 38 publications (42 percent) had directly or indirectly used the third-party open archived data for new research. Moreover, new collaboration can also be developed by sharing and reuse of the scientific data in open archives (Kenall *et al.*, 2014).

However, there are still many research fields in which scientific data sharing and reuse are less common, which stands in contrast to research fields such as genomics with positive examples of data reuse benefiting researchers. Borgman (2013) surveyed 1,700 researchers about their data sharing behaviors, and the result shows that only 22.6 percent of researchers usually use or browse published data, and 21.4 percent of them occasionally make use of that data, while 56 percent of them never use or browse publically shared data. It was reported that the reuse of mammography images is very difficult because the data are very hard to interpret if they were separated from the related context (Hartswood *et al.*, 2012). In seismology, researchers must verify whether the data are trustworthy (Faniel and Jacobsen, 2010), assuming that the more metadata the document included, the more reliable the data are (Faniel and Jacobsen, 2010). Because of the long tail theory in "small science," it is still difficult to find proper re-used data (Wallis *et al.*, 2013). The reproducibility of studies from data deposited in the archives is still limited, largely owing to the lack of sufficient annotations for scientific data (Rung and Brazma, 2013). Also, there are a few other inhibitors of data reuse by researchers, which include the quality of documents, reliability of data, interpretation of data and application context to specific problems.

Several studies have previously conducted bibliometric analyses on scientific data re-used in academic papers (Moed, 2010). Piwowar examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The research results show that sharing data are associated with an increase in citations of the paper describing the data (Piwowar *et al.*, 2007; Piwowar, 2011). Belter (2014) investigated citation counts of three oceanographic data sets curated by National Oceanographic Data Center . The finding reveals that the three data sets are highly cited, with estimated citation counts in most cases higher than 99 percent of all the journal articles published in *Oceanography* during the same years. Parsons *et al.* (2010) used Google Scholar to search for mentions of snow cover data sets archived at the National Snow and Ice Data Center. They found that mention times in research paper increased from 100 to 600 mentions between 2002 and 2009.

Existing studies have shown that although there are many obstacles in data reuse, data archived in open repositories have been re-used well in some specific research areas. This fact is encouraging not only for the data stakeholders but also for the data producers. However, because of long tail of data sharing and data reuse, how and why data are re-used at the overall level, has not been discussed in detail. In this paper, we take DDR as an example and intend to discover how and why researchers reuse scientific data from the open archives in a wide range of research areas. The findings will contribute to enabling policy makers or journals to provide improved guidelines to promote data sharing and data reusability. In this paper, the citations to the DDR in reference to the publications will be used as an evidence of the re-used scientific data. Recently, Kousha and Thelwall have successfully used the URL-citing method to explore the use of YouTube videos in publications (Kousha *et al.*, 2012).

Methods

To address the research questions, we extracted URL citations to DDR from academic publications indexed by Scopus from 2010 to 2015 (up to August of 2015). We downloaded some metadata of DDR for further analysis of the extracted URLs.

The data set for citation analysis

Scopus is used to access the academic publications that cite scientific data present in DDR. The search interface of the Scopus database provides the search field for Reference (REF). REF indexes all types of references including URL citations. Unlike Scopus, Thomson Reuters Web of Science (WoS) does not enable URL citations searching for references. We used the keywords "dryad.*" and "doi" to retrieve publications via REF. A total of 827 citation results were obtained. Some citations are not valid because they do not contain full details of referential URLs. We filtered out those invalid citations. After refining, 550 valid URL citations are saved in the database.

According to valid URL citations, we downloaded the corresponding metadata fields of URL citations from the Dryad website. The metadata fields include data title, data types, downloaded times, keywords, descriptions and original journal names where data were published. For the purpose of exploring the functions of data re-used in citing publications, full-texts of citing publications are also downloaded.

In this paper, the types of scientific data citations are defined as self-citation and non-self-citation in terms of their relations to the cited scientific data. Non-self-citation refers to the title of cited scientific data is the same as a citing publications. For example, in Figure 1, the citation from the article in Procedure A to the scientific data in Procedure B is non-self-citation. On the contrary, if the title of cited data exactly matches the title of its citing publication, then the type of the citation is self-citation. For example, the following scientific data (D):

> Hoy, S.R., Petty, S.J., Millon, A., Whitfield, D.P., Marquiss, M., Davison, M., Lambin, X. (2014), "Data from: age and sex-selective predation as moderators of the overall impact of predation", Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.h1289

The data were cited by the behind article (A):

> Hoy, S.R., Petty, S.J., Millon, A., Whitfield, D.P, Marquiss, M., Davison, M., Lambin, X. (2015), "Age and sex-selective predation as moderators of the overall impact of predation", *Journal of Animal Ecology*, Vol. 84, No. 3, pp. 692-701. http://dx.doi.org/10.1111/1365-2656.12310

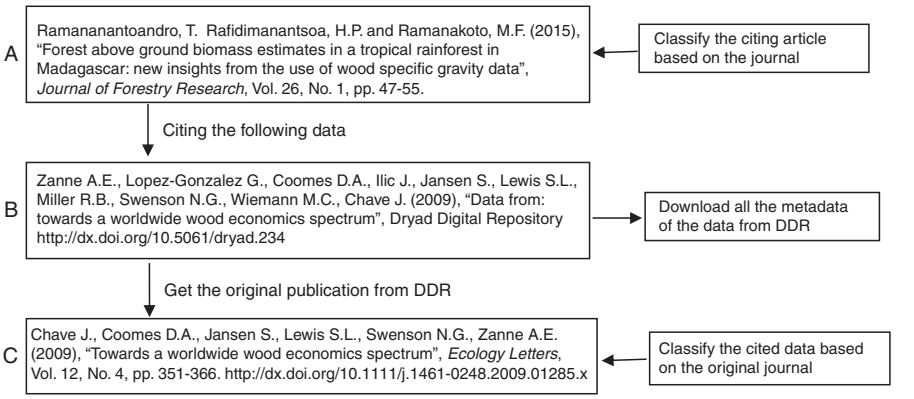


Figure 1.
Brief procedures of data processing

The citation (D is cited by A) is a self-citation because the title of scientific data and citing article is the same.

The categories of different subject areas

In order to find the distribution of scientific data used in different research fields, it is necessary to select a classification scheme of subject areas from many existing categories of discipline classification. In this study, we use the classification scheme of Scopus for journals[9] to classify the subject areas of citing and cited publications. The classification scheme is referred to as SCSJ in this paper. The category of an article depends on the category of its published journals. If an article is published by journal J, and the journal J belongs to a category C, then the article will also be assigned to category C.

For example, the citing publication in Figure 1 (Procedure A) would be classified according to its publishing journal. In the classification scheme SCSJ, *Journal of Forestry Research* was assigned to class forestry coded with 1107. Based on our classification rule, the article was classified to 1107 as well. There are 36 categories such as 1100 and 300 subcategories such as 1101, 1102 and so on in the SCSJ. However, the number of citations is only 550 for analysis, so the narrow subcategories such as 1101 and 1102 are merged into their parent class (broader upper category) in order to get more concentrated broader research fields. As a result, the final category of the example citation article is 1100, which is the parent class code of 1107. The cited data (Procedure C in Figure 1) were classified to class 1100 based on published journal of *Ecology Letters* using the same classification rule.

If an article was assigned to more than one category, we made the simplifying assumption that all categories had contributed equally. Hence, for an article with three categories C1, C2 and C3, the proportion of the article to each category (C1, C2, C3) is 1/3, respectively. The proportion P_c, a contribution of a category to article a with n categories is: $P_{c,a} = 1/n$. Let A denote the set of all articles in the collection. Then the number of citations belonging to category C, which is the sum of contributions to each article by proportion, is given by: $n_c = \sum_{a \in A} P_{c,a}$.

The function of re-used data cited by publications

The function of re-used data refers to the reason for authors to cite these data in their publications. The classical theory of citation motivation (Garfield, 1979) is used to differentiate the role of re-used data in the new publications.

We chose 30 percent of the total 550 citations to analyze the function of re-used data. In total, 165 articles were chosen by using a random algorithm to ensure the selection of data sample. In total, 130 full-texts among these 165 articles can be accessed from Google Scholar, Elsevier, EBSCO and Springer.

We invited two annotators who are researchers in biology science and library science to index the function of data in citing publications according to the citation motivation theory (Garfield, 1979). The annotators agreed on 119 of the 130 publications, which were tabulated for further analysis.

The contents of re-used scientific data

To most of the data curators, scientific data are generally classified into five categories: observational data, experimental data, simulation data, derived or complied data and reference or canonical data[10],[11] to present their data contents. The contents of scientific data are indexed in the metadata field description in DDR. The 119 publications for function analysis above were used as samples. We extracted the keywords of data types in the metadata of description. The details are shown in the results section.

Results

Citing and cited data of DDR in different research areas

Table I contains information pertaining to scientific data of DDR (citing papers) and DDR data citations cited by papers in Scopus (cited papers) in different subject areas between 2010 and 2015 (only to August). More than 95 percent of publications citing scientific data of DDR are research articles. column 4 shows the number of Dryad citations cited by publications of Scopus, and column 5 is the number of citations of DDR for each publication in different fields. We can see that the number of citations varies in different research fields and the amount of depositing data is skewed in different research fields as well. However, the quantity of cited data is far lower than the quantity of depositing data in DDR.

Data citations and data depositing over time

From Figure 2, we can see that there has been a consistent upward trend in citing scientific data in DDR by publications in Scopus since 2010. From 2010 to 2013, there is a steady upward trend with an increase of nearly 3 percent every year. Particularly from 2013, the citation trend sharply increased to 36 percent. The citation counts in 2015 are only up to August, but this does not contradict the pattern of an increasing trend.

Figure 3 shows the time span from data being published in DDR to being cited by publications in Scopus. Almost 50 percent of scientific data published in DDR were cited by publications indexed in Scopus in the same year, whereas, almost 20 percent of citations are cited in the following year. This means that more than 70 percent of scientific data were cited immediately within two years.

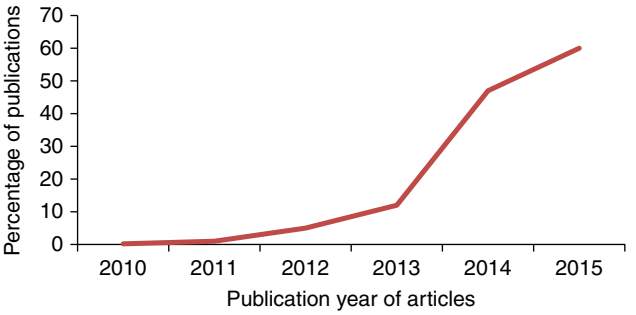
Re-used data citation type in the citing publications

From Table II, we found that 84 percent of scientific data citations are self-citing, and only 16 percent of citations are non-self-citing. This trend can be observed in the majority of research fields.

Research field	Statistic of citing papers: No. of publications with DDR citations	Statistic of citing papers: No. of articles with DDR citations[a]	Statistic of cited data: No. of Dryad resources cited	Statistic of cited data: Dryad citation per publication	Statistic of DDR: No. of Dryad resources totally	Statistic of DDR: Percentage of Dryad resources re-used
1000 general	1	1	1	1	292	0.3
1100 agricultural and biological science	264	257	247	1.04	6,657	4
1200 arts and humanities	1	1	1	1	3	33
1300 biochemistry, genetics and molecular biology	36	35	33	1.06	3,062	1.2
1600 chemistry	1	1	1	1	11	9
1700 computer science	1	1	0	1	9	11
1900 earth and planetary sciences	1	1	1	1	138	0.7
2000 economics, econometrics and finance	1	1	1	1	1	100
2100 energy	1	1	0	0	1	100
2300 environmental science	113	108	77	1.4	394	29
2400 immunology and microbiology	8	8	6	1.28	108	8
2700 medicine	115	108	104	1.04	805	14
2800 neuroscience	6	6	2	2.83	48	13
3200 psychology	1	1	1	1	6	17
Total	550	530	475	1.04	11,535	5

Notes: [a]Omitting reviews, conference papers, editorials, letters and notes

Table I.
General statistics for citing and cited sources to data of DDR



Note: Number of citations from 2015 is not complete because downloads were only made up until August of 2015

Figure 2.
Number of publications citing DDR data over time

File types of re-used data cited in publications

The file types of data are generally represented by the metadata field format in DDR. We downloaded them according to the URL citations from the references list of the citing publications. Figure 4 shows the counts of different file types across research fields.

It is apparent that the top three ranks of file types cited in SCOPUS are in the formats of .xls, .csv and .txt. All of them are text-based files as well as illustrative-types such as tables, figures and texts. The functions of these data are normally to give further illustrations to research arguments or to certify the credibility of research results.

Contents of re-used scientific data

The contents of re-used data were extracted from the metadata fields descriptions in DDR, and the statistical results are shown in Table III. The column type in Table III refers to the data generated for different purposes, which are described in the method section. The column details refers to the source of the data.

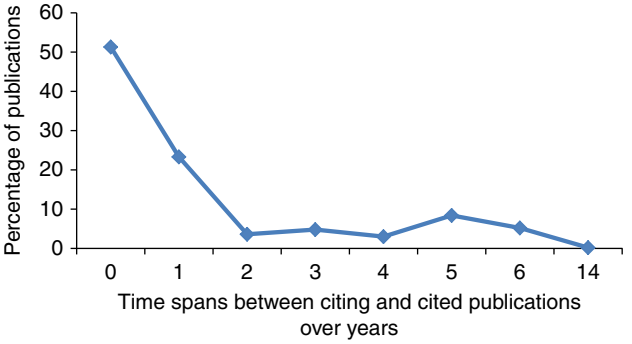


Figure 3.
Time span of publications between citing and cited publications over years

Subject area	Self-citation The number of self-citation	Self-citation rate (%)	Non-self-citation The number of non-self-citation	Non-self-citation rate (%)
1000 general	0	0	1	100
1100 agricultural and biological science	223	84	41	16
1200 arts and humanities	2	100	0	0
1300 biochemistry, genetics and molecular biology	30	84	6	16
1600 chemistry	1	100	0	0
1700 computer science	0	0	1	100
1900 earth and planetary sciences	1	100	0	0
2000 economics, econometrics and finance	1	100	0	0
2100 energy	0	0	1	100
2300 environmental science	86	76	27	24
2400 immunology and microbiology	8	100	0	0
2700 medicine	106	93	8	7
2800 neuroscience	5	83	1	17
3200 psychology	1	100	0	0
Total	464	84	86	16

Table II.
Citation types of data in DDR cited in publications

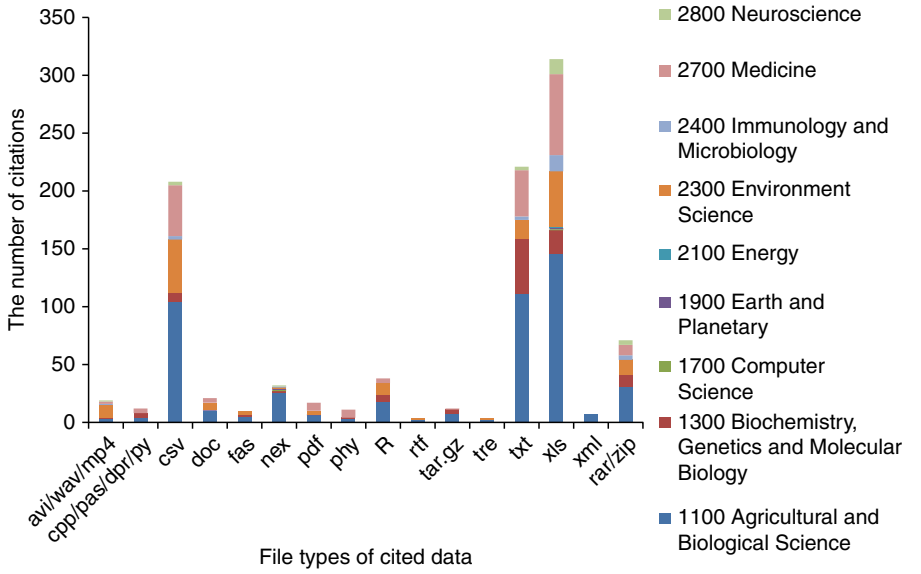


Figure 4.
File types of data cited by publications in different subject areas

Type	Counts	Details
Observational data (31%)		
Survey data	21 (18%)	15 (species), 6 (surroundings)
Sample data	15 (13%)	15 (plants)
Experimental data (45%)		
Gene sequence	32 (27%)	22 (species), 10 (plants)
Field data	22 (18%)	18 (plants), 4 (surroundings)
Simulation data (3%)	4 (3%)	2 (species), 2 (surroundings)
Derived or complied data (20%)		
Text and data mining	22 (18%)	18 (species), 4 (plants)
3D models	3 (2%)	3 (surroundings)

Table III.
Contents of cited scientific data by others

Experimental data are mostly re-used (45 percent) by researchers. This kind of data are generated in a controlled environment from the laboratory equipment, such as gene sequences, chromatograms and spectroscopy or toroid magnetic field data. The other kind of data is observational data (31 percent). They are mainly captured in real-time from the fields, farmlands, greenhouses or other natural environmental conditions reflecting the features of nature, such as sensor data, survey data, sample data or neurological images. Derived or compiled data account for the proportion of 20 percent, which are the analytical intelligence of further and refined analysis to specific research questions. For example, phylogenetic trees for genes or species are widely cited with a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities.

The main function of re-used data cited by publications

The full-texts of citing publications were downloaded from Scopus, Google Scholar and WoS. Two annotators classified the functions of scientific data in 130 citing publications as described in the section methods. The agreed 119 papers are analyzed in Table IV. The first column function describes the reasons for reusing scientific data of others in the new publications, and the last column positions in articles explains where citation appeared in new publications.

The citations of re-used data appearing in the section of methods and materials of publications, accounted for 75 percent of cases. This is the most important section for explaining research methodology or argument in general. Among these citations, 28 percent were directly used as raw data. Typically in bioinformatics, researchers combine many different data sets from other research to address a new research question without generating new data. 30 percent of the existing data were combined with new generated data to address a new research problem. And, 17 percent of the data were used as a comparison to assess the value of a new method. Another kind of reuse, accounting for 21 percent of cases, appears in the section discussion/evaluation/results of new publications. In some cases, the data are used as a baseline to evaluate the performance of new research results, whereas, sometimes they are used as meta-analysis of summary-level data, such as p values or effect sizes from compared conditions to support an argument. Such data reuses are the most popular way to evaluate the performance of new experimental results. The third kind of reuse (4 percent) is in a review to explain related research work, usually appearing in the sections of related research.

Discussion

Analysis of re-used data in different research areas

Although DDR is a general-purpose repository, citation analysis results (Table I) show that there is a significant difference in citing scientific data across different research areas. In total, 85 percent of citations are mainly distributed in three fields having data-intensive features. They are agricultural and biology science (55.9 percent), environment science (16 percent) and medicine (13.6 percent). Figure 5 shows the number of re-used data in different research areas in terms of self-citation or non-self-citation. The quantity of data re-used in these research areas are much larger than in other research areas. Originally, data sharing began from these three research fields, which are regarded as pioneers in the development of infrastructures, resources and policies to promote data sharing. In these three domains, many standards and criteria have been incrementally developed for collecting, storing, preserving, accessing and citing scientific data (Kaye *et al.*, 2009).

Function	Numbers	Position in the article
Giving credit for related work	4 (4%)	Related research
Evaluating analysis	16 (13%)	Discussion/evaluation/result
Meta-analysis of summary data	10 (8%)	
Evaluating analysis method	20 (17%)	Methods and materials
Supporting data for new studies	36 (30%)	
Raw data	33 (28%)	

Table IV. Function of DDR citation in the citing publications

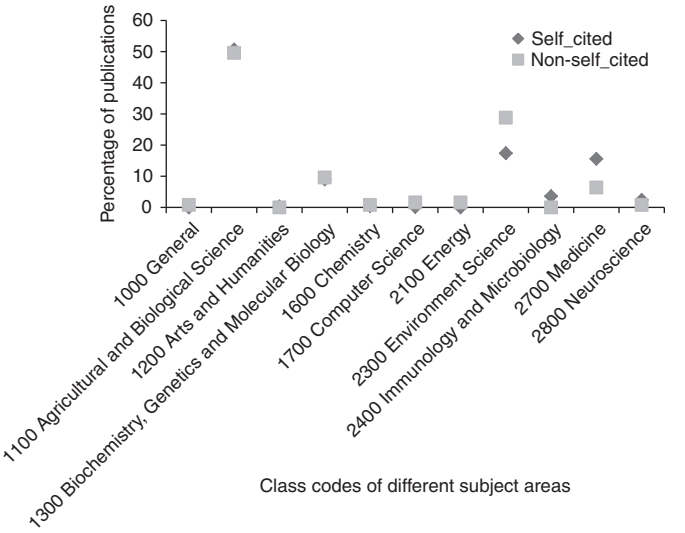


Figure 5.
Distribution of data re-used among different subject areas

We performed statistical analysis on the archiving policy of publishing journals in which scientific data are highly re-used (top 15) among these research areas. The strength of policy on data archiving and the association with Dryad are clearly shown in Table V. There is a strong correlation between promoting policy of journals or fund agencies with data sharing and reusing behavior. In Table V, the value mandatory of the field data archiving policy, indicates that journals require an explicit data accessibility statement about manuscripts including archiving policy and depositing locations. We find that almost all of the journals have mandatory requirements on data archiving, with a detailed explanation on depositing and citation of data sharing. Since 2011, a number of ecology and evolution journals known as the Joint Data Archiving Policy[12] declares specific detailed requirements on data

The name of journals	Data archiving policy	Recommendation
American Naturalist	Mandatory	Y
Journal of Ecology	Mandatory	Y
Journal of Animal Ecology	Mandatory	Y
Functional Ecology	Mandatory	Y
Molecular Ecology	Mandatory	Y
Biological Journal of the Linnean Society	Mandatory	N
PLoS ONE	Mandatory	Y
BMC Evolutionary Biology	Mandatory	Y
Methods in Ecology and Evolution	Mandatory	Y
PLoS Biology	Mandatory	Y
Proceedings of the Royal Society B: Biological Sciences	Mandatory	Y
eLife	Recommended	N
Biology Letters	Mandatory	Y
Evolution	Mandatory	Y

Note: The column of recommendations refers to whether journals recommended DDR as a premier choice for data depositing

Table V.
Description of data archiving policy of the top 15 journals with highest data re-used

archiving policy along with the journal submission. Similarly, the BMC journals also drafted a policy[13], and the Royal Society journals also announced data submission policy[14].

Furthermore, Figure 6 shows the number of re-used data in accordance with the country of authorship. The top ranked countries in quantity, including USA, UK, Australia and Canada, are all advocates of data sharing and data reuse in scientific research. Therefore, there is a strong association between policy leading tendency and actively data sharing and reusing behavior. It is the policy of funding agency and journals to enforce the development of data sharing and data reusing in these data-intensive research areas.

Analysis of the functions of scientific data re-used by researchers

Since it is the contribution of funding agencies or journals to promote data sharing through mandatory policies, two significant questions arise: are data sharing behaviors putting researchers under pressure?; and, what are the main functions of the re-used data in a different research filed in terms of current policy?

Most journals have supportive policies for encouraging contributors to submit as much data as possible related to the manuscripts for the benefits of the peer-reviewers and readers. However, a few other journals have policies on the limited amount of supplementary information that authors are allowed to submit since 2010 (Borowski, 2011; Maunsell, 2010). Thus, in practice, researchers are more likely to deposit research data generated during research into open data repositories. Depositing data into scientific repositories has much more advantages than in supplementary files. For example, it can give more priorities on the storage file size, format and preservation time.

Due to the requirements of journals and funding agencies, it is easy to understand the reason why the majority of citations to DDR are self-cited, shown in Table II. Citations to DDR are mainly self-cited as further illustrations to demonstrate and support their arguments or to increase reader's confidence in the reliability of the research. Therefore, most of them are in the formats of the text-based spreadsheets, tables or figures. As a consequence, most of the scientific data present in DDR have become supplements to the written records of research due to the increased pressure of scientific data open access as the requirements of journals and funding agencies.

However, some researchers tend to publish new research articles by reusing research data produced by others. As shown in Table II, 16 percent of the total data cited in references of publications are "real data reuse," which means that the shared scientific data are re-used by others either within the same research group or from the different researchers. As we can see from Table IV, more than 50 percent of non-self-citations

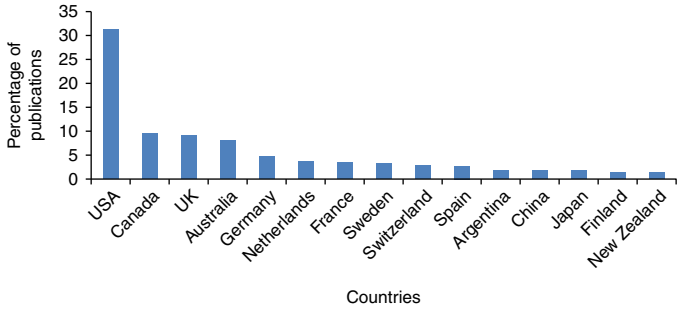


Figure 6.
The number of re-used data in accordance with the country of authorship

appeared in the method and material section in papers. In general, method and material is considered as the most important section of a research article. Non-self-citation data always are cited as supporting data for a new research or as a raw data for a new research directly. It is clearly demonstrated that some re-used data are making a significant role to promote new researches, nonetheless this kind of data reuse only account for a small proportion.

Analysis on the preferences of scientific data re-used by researchers

Some experimental or observational data received more citations compared to stimulated or derived data shown in Table III. Typically, observatories are important sources of data distributions of natural phenomena, and similarly, experimental data are the essential records to replicate the experiments. These are the general types of primary data, which are mainly acquired from the lab equipment and captured in specific environments. The citation analysis shows that, the "rawest" scientific data have received more non-self-citations. This indicates that primary data with less further analysis will have greater value than secondary data specific to certain questions. They will be more likely to be re-used in later time compared to those with much analysis for specific questions.

In most of the research areas including agricultural and biological science (1100), biochemistry, genetics and molecular biology (1300), medicine (2700), the majority of re-used data belongs to or are shared by the authors' own research groups. This shows that researchers prefer to reuse the data published by their own group to support new research. One of the possible reasons is that data are more interpretable and creditable within the same research group where data are produced. Interpretation and the trustworthiness of the data are the two main factors that impact the reusability of the scientific data (Faniel and Jacobsen, 2010). The data produced within the same research group are more creditable because there are more detailed contexts recorded to insure the quality of the data. For better assurance of interpretation and credibility, descriptive metadata should contain more information about and contexts in which data are generated and their usability. We investigated the metadata of Dryad which uses Dublin Core to describe scientific data, e.g., title, doi, published journal of related article, keywords, description and download times. However, we found that not all metadata fields of description contain fully detailed information indexed by different researchers with different research backgrounds. Therefore, less information about the quality and interpretation can be obtained from the metadata provided at present. Typically, text content of the publications is the only way for the readers to have better understanding of data. Until now, data are still acting as supplementary materials of research articles, not independent resource to articles. This reduces the probability of reusing data because fewer contexts are available for interpretation and quality control for data (Figure 7).

Limitations

This paper assesses show scientific data are cited in research publications and to understand the value of scientific data in scholarly communication. Unfortunately, only a few literature databases index non-bibliographic citations such as web URL citations, enable searching for them. Surprisingly, SCOPUS provides access to manipulate the searching of references in various formats. However, there is still not a common citation standard for the scientific data citation format, thus many authors are not sure how to

cite the scientific data in proper formats. As a result, most of the papers provide a footnote or an explanation at the end of the publications. In some cases, the citations of scientific data are unavailable for the entire indexing information including titles or web URLs. Therefore, only parts of the scientific data citations were collected because of incomplete bibliographic descriptions in the references.

Another limitation of this paper is the scope of DDR in terms of research areas. We selected a typical general-purpose scientific data repository widely used by researchers as the case study to examine the reuse of scientific data in scholarly communication. However, it is rather difficult to find a perfect repository for "small science." Although DDR is a general-purpose and wide-diverse scientific data repository, it seems that data in ecology and evolutionary science account for more proportion of all re-used data.

Conclusions

In recent years, the amount of depositing data in DDR has been increasing exponentially. By the end of 2014, the quantity of data sets had grown to 7,185 by comparison with the number of 181 in 2010. DDR is providing a free open platform for multi-discipline data sharing. Data present in DDR has been widely accepted as a reliable, public scientific data repository by researchers, journal publishers and funding agencies. However, we should also raise awareness to the fact that the number of data reuse is falling behind the fast increasing speed of depositing data in DDR.

From the citation analysis of research data from DDR cited in Scopus, we find that the majority of data reuse type is still self-cited, that is, to say, most researchers tend to reuse their own data. There are several reasons for the conservatism in reusing DDR's data. Firstly, the policy on data sharing and reuse is one of the most important driving forces to encourage researchers to deposit their data in DDR. More and more funding agencies and journal publishers require depositing entire data sets related to the submitted articles or research projects. In this circumstance, most of the shared data consists of further illustrations or demonstrations to support arguments, increasing readers' confidence in the reliability of the research. Therefore, data for more specific purposes have little value to other researchers. The possibility of reusing these data will be very low in the future. The other reason for the conservatism in data reusing is that data generated by researchers themselves are more interpretable and reliable. It is very difficult to understand the creation process of the data and to use the data if insufficient contexts or explanations are given.

In conclusion, although there is a steady upward growth in re-used DDR's data, the amount of data re-used is still very low compared to data deposition. Data curators

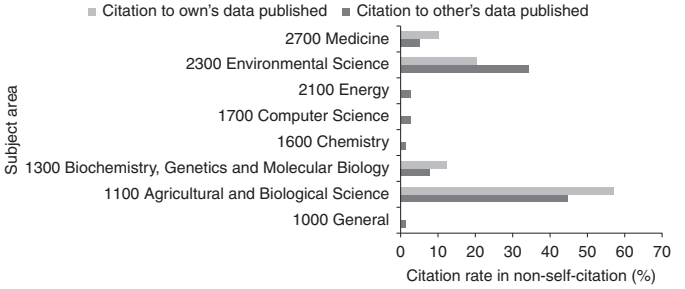


Figure 7.
Distribution of data reuse types about data provenances

should exploit more feasible approaches to encourage researchers to use the other's data. More solutions should be proposed to improve better understanding of the contexts of scientific data on their generation and use, as reliability of data will greatly improve the value of them in academic research.

Acknowledgments

The authors express their greatest gratitude to Mike Thelwall, Professor in School of Mathematics and Computer Science, University of Wolverhampton, UK for his constructive feedback and valuable suggestions. He provided ongoing encouragement and guidance throughout this study that substantially shaped the intellectual direction of this paper.

This paper is supported by the grant of Chinese National Social Science on organization and application of scientific data in the knowledge service environment (No. 13CTQ035) and evaluation research of scientific influence based on altermetrics method (No. 15BTQ061).

Notes

1. re3data.org: www.re3data.org/
2. www.ncbi.nlm.nih.gov/genbank
3. www.pangaea.de
4. http://heasarc.gsfc.nasa.gov/
5. http://figshare.com
6. http://datadryad.org/
7. www.labarchives.com
8. Dryad: http://datadryad.org/
9. http://files.sciverse.com/documents/xlsx/title_list.xlsx
10. www.bu.edu/datamanagement/background/whatisdata/
11. http://guides.library.oregonstate.edu/data-management-types-formats
12. http://datadryad.org/pages/jdap
13. www.biomedcentral.com/about/editorialpolicies#DataandMaterialRelease
14. https://royalsociety.org/journals/ethics-policies/#question6

References

Barve, V. (2014), "Discovering and developing primary biodiversity data from social networking sites: a novel approach", *Ecological Informatics*, Vol. 24 No. 1, pp. 194-199.

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. doi: 10.1371/journal.pone.0092590.

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.

Borgman, C.L. (2013), "Big data, little data, no data: the contested landscape of data sharing and reuse", Trends in Society and Information Technology Seminar Series. University of California, Irvine, CA.

Borowski, C.L. (2011), "Enough is enough", *The Journal of Experimental Medicine*, Vol. 208 No. 7, p. 1337, available at: http://jem.rupress.org/content/208/7/1337.full.pdf

Chen, H. (2013), "Governing international biobank collaboration: a case study of China Kadoorie Biobank", *Science Technology & Society*, Vol. 18 No. 3, pp. 321-338.

Cragin, M.H., Palmer, C.L., Carlson, J.R. and Witt, M. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society*, Vol. 368 No. 1926, pp. 4023-4038.

Faith, D., Collen, B., Ariño, A., Koleff, P., Guinotte, J., Kerr, J. and Chavan, V. (2013), "Bridging the biodiversity data gaps: recommendations to meet users' data needs", *Biodiversity Informatics*, Vol. 8 No. 2, pp. 41-58.

Faniel, I.M. and Jacobsen, T.E. (2010), "Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data", *Computer Supported Cooperative Work*, Vol. 19 Nos 3-4, pp. 355-375. doi: 10.1007/s10606-010-9117-8.

Garfield, E. (1979), *Citation Indexing: its Theory and Application in Science, Technology and Humanities*, John Wiley, New York, NY.

Hartswood, M., Procter, R., Taylor, P., Blot, L., Anderson, S., *et al.* (2012), "Problems of data mobility and reuse in the provision of computer-based training for screening mammography", *Proceedings of the 2012 Annual Conference on Human Factors in Computing Systems: ACM Conference on Human Factors in Computing Systems (CHI)*, ACM Press, New York, NY, pp. 190-190.

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 September 28, 2015).

Kaye, J., Heeney, C., Hawkins, N., de Vries, J. and Boddington, P. (2009), "Data sharing in genomics – re-shaping scientific practice", *Nature Reviews Genetics*, Vol. 10 No. 5, pp. 331-335. doi: 10.1038/nrg2573.

Kenall, A., Harold, S. and Foote, C. (2014), "An open future for ecological and evolutionary data?", *BMC Evolutionary Biology*, Vol. 14 No. 66, pp. 1-6, available at: http://doi.org/10.1186/1471-2148-14-66

Kousha, K., Thelwall, M. and Abdoli, M. (2012), "The role of online videos in research communication: a content analysis of YouTube videos cited in academic publications", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 9, pp. 1710-1727.

Maunsell, J. (2010), "Announcement regarding supplemental material", *The Journal of Neuroscience*, Vol. 30 No. 32, pp. 10599-10600.

Moed, H.F. (2010), "Measuring contextual citation impact of scientific journals", *Journal of Informetrics*, Vol. 4 No. 3, pp. 256-277.

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., 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. 5528-5533.

Nature (2015), "Availability of data, material and methods", available at: www.nature.com/authors/policies/availability.html (accessed September 20, 2015).

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.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., *et al.* (2013), "Making research data repositories visible: the re3data.org registry", *PLoS ONE*, Vol. 8 No. 11, p. e78080. doi: 10.1371/journal.pone.0078080.

Parsons, M.A., Duerr, R. and Minster, J.B. (2010), "Data citation and peer review", *Eos, Transactions American Geophysical Union*, Vol. 91 No. 91, pp. 297-298. doi: 10.1029/2010eo340001.

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. doi: 10.1371/journal.pone.0108451.

Piwowar, H.A. (2011), "Who shares? Who doesn't? Factors associated with openly archiving raw research data", *PLoS ONE*, Vol. 6 No. 7, p. e18657. doi: 10.1371/journal.pone.0018657.

Piwowar, 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. doi: 10.1371/journal.pone.0000308.

Rung, J. and Brazma, A. (2013), "Reuse of public genome-wide gene expression data", *Nature Review Genetics*, Vol. 14 No. 2, pp. 89-99.

Salzberg, S.L., Hotopp, J.C.D., Delcher, A.L., Pop, M., Smith, D.R., Eisen, M.B. and Nelson, W.C. (2005), "Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species", *Genome Biology*, Vol. 6 No. 3, p. R23, available at: http://doi.org/10.1186/gb-2005-6-3-r23

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. doi: 10.1371/journal.pone.0067332.

Further reading

Kell, D. and Oliver, S. (2004), "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era", *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology*, Vol. 26 No. 1, pp. 99-105.

About the authors

Lin He is a Professor in the College of Information Science, University of Nanjing Agricultural University. Her interesting includes information retrieval, knowledge organization and text mining. Lin He is the corresponding author and can be contacted at: helin@njau.edu.cn

Vinita Nahar is a Post Doctoral Research Fellow in Natural Language Processing at the Research Group in Computational Linguistics (RGCL), University of Wolverhampton. Her interesting includes data mining, text mining, machine learning, information retrieval, pattern recognition, natural language processing, sentiment analysis, information system, sensitive event detection, cyberbullying detection, and social networks analysis and prediction.