2016

An Examination of Research Data Sharing and Re-Use: Implications for Data Citation Practice

Hyoungjoo Park
University of Wisconsin-Milwaukee, park32@uwm.edu

Dietmar Wolfram
University of Wisconsin - Milwaukee School of Information Studies, dwolfram@uwm.edu

Recommended Citation
Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. Scientometrics, 111(1), 443-461.

This is a preprint of an article published in *Scientometrics*

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics, 111*(1), 443-461.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-017-2240-2

An Examination of Research Data Sharing and Re-Use: Implications for Data Citation Practice

Hyoungjoo Park*, Dietmar Wolfram

School of Information Studies, University of Wisconsin – Milwaukee

P.O. Box 413, Milwaukee, WI USA 53201

* Corresponding author: Email: park32@uwm.edu

Tel: 01 414 229 4707

Fax: 01 414 229 6699

Keywords: citation analysis, data citation, data sharing, data re-use, citer-based analysis, research data

Abstract This study examines characteristics of data sharing and data re-use in Genetics and Heredity, where data citation is most common. This study applies an exploratory method because data citation is a relatively new area. The Data Citation Index (DCI) on the Web of Science was selected because DCI provides a single access point to over 500 data repositories worldwide and to over two million data studies and datasets across multiple disciplines and monitors quality research data through a peer review process. We explore data citations for Genetics and Heredity, as a case study by examining formal citations recorded in the DCI and informally by sampling a selection of papers for implicit data citations within publications. Citer-based analysis is conducted in order to remedy self- citation in the data citation phenomena. We explore 148 sampled citing articles in order to identify factors that influence data sharing and data re-use, including references, main text, supplementary data/information, acknowledgments, funding information, author information, and web/author resources. This study is unique in that it relies on a citer-based analysis approach and by analyzing peer-reviewed and published data, data repositories, and citing articles of highly productive authors where data sharing is most prevalent. This research is intended to provide a methodological and practical contribution to the study of data citation.

Introduction

In this era of big data and open science, data citation is increasingly important with regard to shared research data that are linked to published outputs in datasets, data repositories and articles. Today's researchers work in a computational, heavily data-intensive and collaborative environment in order to further scientific discovery across and within various fields. It is

becoming routine for researchers (i.e. authors and data publishers) to submit their research data, such as datasets and computer code, as supplementary information in order to comply with the data sharing requirements of major funding agencies, high profile journals and data journals (National Cancer Institute, 2006; National Institutes of Health, 2003).

Major funding agencies now require a data management plan for data sharing. In 2013, the National Science Foundation (NSF) announced that any application would be rejected or terminated if the requisite data management plan was not provided. High profile journals such as *Nature* and data publications such as the *PLOS* family of journals also require researchers to submit such supplementary information as datasets and/or computer code and thus to share their data. However, researchers have been hesitant to do so because of concerns about misuse, scooping and receiving sufficient credit for their work. From the perspective of data sharers, data scooping, planarization or loss of publication opportunities may be real concerns. Individuals' perceptions that current rewards systems do not generate credit, in the form of social recognition, promotions, tenure and successful grant applications, that is commensurate with their time and effort may also help to explain their reluctance to share their data.

From the perspective of data re-users, collecting data themselves may be more attractive than re-using shared data produced by other researchers because of the time and effort needed in order to understand and reanalyze other researchers' published data for secondary analysis. The absence of universally accepted standards for data citation may also be an issue. This situation creates challenges for how the citation of published, peer-reviewed research data and literature can be measured and enhanced appropriately in order to give proper credit to those who expend the time and effort to share their work and, thereby, create the potential for the future re-use of their data. Assessing shared data and their potential future re-use for secondary research is, therefore,

particularly important in the data-intensive and collaborative research environments. It is also critical where scientists in the current era of big data are urged to collaborate with colleagues from different disciplines (i.e. interdisciplinary research) in order to solve "complex" problems. These considerations drive the research questions addressed in this study:

RQ1. In an environment where published, peer-reviewed data sharing is most common, how prevalent is data re-use as measured by data citation?

RQ2. To what extent do authors formally and informally document data citation?

RQ3. What are the ongoing challenges to studying data citation and re-use?

Literature review

There are gaps in the research literature mainly in the realms of data sharing, data re-use and data citation. Previous research regarding data sharing has been limited by a reliance on survey and interview methods that approach data sharing behavior on the individual level and within few investigations of multiple disciplines. Relatively recently, the social sciences have been actively studied as a domain regarding data sharing and data re-use (Curty, 2015; Fear, 2013; Yoon, 2015). Kim (2013) has studied multiple disciplines across the STEM (Science, Technology, Engineering, and Math) fields because, as he noted, without consideration of disciplinary factors, scientific data sharing behavior in general cannot be studied. Gaps in previous studies regarding data re-use include a focus on users' trust judgment (Yoon, 2015) and persistent identification (Lee, 2015); in these cases, interviews represent the main method used. With regard to data citation from the perspective of data re-use, there are relatively few studies because research has instead focused on data sharing (Helbig, Hausstein, & Toepfer, 2015), for example in the context of GIS data citation (LaBonte, 2015). One study by Fear (Fear, 2013) analyzed data re-

use in the social sciences from the Interuniversity Consortium for Political and Social Research (ICPSR).

Despite their limitations, previous studies have yielded a number of findings, which can be summarized briefly as follows:

1. Research data sharing makes research data citable and re-usable for secondary research (Helbig, Hausstein, & Toepfer, 2015)
2. Articles with shared research data have increased citation rates, which lends them greater impact (Helbig, Hausstein, & Toepfer, 2015; Piwowar & Vision, 2013; Piwowar, Day, & Fridsma, 2007)
3. The same authors tend to use the same shared data repeatedly (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015), which could mean a high rate of self-citation
4. Each discipline has distinct data sharing practices (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014) that need to be studied separately
5. Within scientific communities, the actual rate of data sharing varies from discipline to discipline (Tenopir, et al., 2011)
6. Worldwide, data sharing and re-use practices and perceptions among scientists differ among age groups and geographic regions (Tenopir, et al., 2015; Helbig, Hausstein, & Toepfer, 2015; Peters I. , Kraker, Lex, Gumpenberger, & Gorraiz, 2015). Certain types of data, such as surveys and aggregated and sequence data are more often cited and receive higher altmetrics scores (Peters, Kraker, Lex, Gumpenberger and Gorraiz 2015).

These findings have helped to inform the research methods of the current study. The review of the relevant data citation literature is organized into several themes: general, standards and principles, data journals, practices, disciplinary focus, metadata, peer reviewed data and dataset granularity.

General

Previous literature on data citation has discussed the topics of data citation principles, standardization, peer review for data publication, practices, infrastructure, metadata elements that are associated with a dataset rather than embedded (such as provenance metadata rather than descriptive metadata), DOI (Digital Object Identifiers, for both unique and persistent identifiers that include a time-stamp and version history), technical infrastructure, quality control for reliable data re-use, flexibility for interoperability across communities, policies regarding repositories and data journals, data management practices best suited to research, the high incidence of self-citation, citation protocols, altmetrics and linked data (Lawrence, Jones, Mattews, Pepler, & Callaghan, 2011; Task Group on Data Citation Standards Practices, 2013). Regarding data sharing practices, previous studies have focused on recommendations for data citation provided by data repositories. The Thomsen Reuters Data Citation Index (DCI) of the Web of Science (WoS) has been mainly studied (Peters I. , Kraker, Lex, Gumpenberger, & Gorraiz, 2016) as a scholarly database. The limitation of studying the DCI is that data citation, which corresponds to the "isCitedBy" scheme of the DCI, is measured only with regard to data repositories and not articles (Starr & Gastl, 2011), which can be a major concern for data citation when it comes to counting bi-directional links among journal publishers, datasets and repositories. Some studies have investigated Google Scholar; however, because of Google's

agreement with publishers, it may not be easy for this search engine to count data citations (Data Citation Synthesis Working Group, 2014,). Neither citation tools, reference management software tools nor data citation, whether open source or proprietary, have been actively studied so far. On the other hand, association rule discovery, community discovery, hub/authority analysis and co-citation analysis have been examined from the perspectives of data mining technologies (Task Group on Data Citation Standards Practices, 2013). Concerns and challenges have also been discussed by a number of researchers regarding the importance of data citation (Green, 2009). Dynamic datasets can be a big challenge without open time series owing to their lack of ambiguity and of persistent identifiers (Green, 2009).

Standards and principles

Implementing data citation standards for scholarly works is, then, an important aspect of ensuring that relevant published data are cited and that the citation is beneficial to those who publish it. Data citation may come to represent a new form of credit that researchers who publish their data receive when they are required by major funding agencies and/or the policies of influential journals to share their data. As discussed, however, the lack of universally accepted standards for publishers, journal editors and funding agencies represents a barrier for researchers, though establishing such standards for peer-reviewed data publication is not a simple task. One reason might be the lack of standards for the peer review process during data publication for both data journals and regular journals. Owing to the lack of clear standards, principles or mechanisms for doing so, researchers are reluctant to make public the data that they have produced. Principles for data citations are importance, credit and attribution, unique

identification, access, persistence, specificity and verifiability, interoperability and flexibility (Data Citation Synthesis Working Group, 2014).

The principles of dynamic data citation are currently being discussed in the context of the Permanent IDentifier (PID) assigned query with the prerequisite of time-stamping, re-writing, hashing and data versioning in order to cite arbitrary subsets of data and data that is dynamic (DataCite Metadata Working Group, 2015). Several organizations work on data citation, including DataCite, World Data System, the Committee on Data for Science and Technology, Research Data Alliance, National Information Standards Organization, the Dataverse Network, International Council for Scientific and Technical Information, Creative Commons, STM-Association and the UK's digital curation center (Mayernik, 2012).

Data journals

Based on the review of the literature there has been little research on data citation (i.e., data citation in data sharing and data re-use) in data journals. Rather, data journals have been studied from the perspectives of journal policies and the practices of scholarly databases and external data repositories. Data journals have been launched in order to meet the increasingly recognized need to give credit and rewards to authors who share their research data as well as the need for re-users (e.g., data consumers) to know how data is produced and what quality control has been performed (Nature Publishing Group, 2013). *Scientific Data*, an open-access and online-only data journal established by the Nature Publishing Group, provides a "data descriptor" for datasets. A data descriptor links related journal articles to actual data files stored in external and recommended data repositories in various communities (Nature Publishing Group, 2013)

because *Nature* does not itself host data, though *Scientific Data* of course requires the release of datasets. The host for data journals can be the journals themselves, publisher data repositories and/or external data repositories. A data journal may require that authors submit an article (document) and one or more datasets at the same time. Current practices of data journals in scientific communities can be distinguished as follows: (1) data contained within journals (e.g. tables, graphs, plotting, etc.), (2) data in journal supplements (restricted or unrestricted), (3) journals that store data themselves, (4) journals that store dynamic/interactive data in public repositories (e.g., Elsevier's data viewer, which works within the article but uses data in public repositories) and (5) data-only publications (Reilly, et al., 2011). Examples of data journals include, in addition to *Scientific Data* from the Nature Publishing Group, the *Biomedical Data Journal*, *PLOS*, *F1000Research*, *Scientific Data*, *BMC Research Notes*, *Giga Science*, *Data Science Journal*, *Journal of Open Archaeology Data*, *Biodiversity Data Journal*, *Journal of Open Psychology* and *Open Health Data*.

Practices

Common practices in data citation have not yet been broadly implemented that give due credit by means of bibliographic references to published research data (Task Group on Data Citation Standards Practices, 2013). Deficiencies include the absence of links to data within an article, persistent identifiers for data in footnotes, metadata, peer review for data that is if submitted or of standardized "copyediting" routines for data, so that data sharing is left up to researchers. Published research data is regarded as supplementary material that resides in publishers' hosted repositories or in external data repositories (Task Group on Data Citation Standards Practices, 2013). It is argued, however, that data citation should accompany such published works as articles in a references or "literature cited" section (Altman, 2012; Callaghan, 2012) in order to

give due credit to data sharers. Access to data repositories (open access data repositories), whether unrestricted, limited or restricted, should be studied in the context of data sharing and the potential future re-use of data. Within previous literature, data identifiers are one area that has been actively studied; however, as mentioned above, dynamic datasets can be a big challenge in the absence of open time series owing to lack of ambiguity and of persistent identifiers (Green, 2009, p. 13).

Disciplinary focus

As alluded to above, the study of data citation needs to be conducted differently within each discipline (discipline-specific) rather than across disciplines (interdisciplinary); and because each discipline has its own practices regarding data citation (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014), each should be studied separately. To advance scientific discoveries, each discipline's dependency on access to specialized materials and equipment should be taken into account from the perspective of the economics of science (Stephan, 2010). Furthermore, the increasing need of scientists who have narrow expertise and specializations to realize significant scientific outcomes (Jones, 2009) might lead to innovative breakthroughs in the analysis of data in "big science" (large, long-lived, projects that depend on extensive instrumentation).

Discipline-specific studies of data citation have been conducted in the fields of astronomy (Kurtz, 2012), earth and physical sciences (Callaghan, 2012), humanities (Sperberg-McQueen, 2012), life sciences (Bourne, 2012) and social sciences (Fear, 2013; Vardigan, 2012). Geographic Information Systems (GIS), for example, make use of large datasets that are often

combined with other datasets, which indicates the importance of the citation rate of GIS data even in the absence of results relevant to determining the citation rates by analyzing peer reviewed articles (LaBonte, 2015). The multidisciplinary approach has been discussed in the context of efforts to identify options for effective data citation practices and standards across both the natural and social sciences (Uhlir, 2012).

Metadata

Previous literature noted that metadata in data citation need to be studied from the perspective of consistency, quality and sustainability (Helbig, Hausstein, & Toepfer, 2015; Starr & Gastl, 2011; Task Group on Data Citation Standards Practices, 2013). Metadata in data citation is currently inconsistent, and needs to take into account contexts such as *administrative* or *methodological* metadata rather than descriptive metadata (Starr & Gastl, 2011). Quality control has focused on the reliable re-use of data for reproducibility. Research has emphasized the importance of metadata openness, platform-independence and effective recognition (Task Group on Data Citation Standards Practices, 2013). Sustainability is another concern in with regard to maintaining metadata (Helbig, Hausstein, & Toepfer, 2015). Thus, the DataCite Metadata Schema v3.1 has been designed as "a list of core metadata properties chosen for the accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions" (DataCite Metadata Working Group, 2015). Methods metadata have also been studied in soil science by examining such common methods-related elements of journal articles as description, citation and sampling (Chao, 2015).

Peer reviewed data

Peer review is important for scientific consensus. Peer review and formal publication help to ensure the quality of data, since it must be checked by domain experts whose review of it takes into account the discipline and data type (Lawrence, Jones, Mattews, Pepler, & Callaghan, 2011). Consistent methods for proper data citation currently in use are simply not sufficient, and appropriate basic management issues such as standardized format and data validation need to be addressed in the data management community through collaborative research (Parsons, Duerr, & Minster, 2010).

Dataset granularity

Granularity refers to "the level of detail of datasets, version control tracks revisions to those datasets (regardless of their granularity level)" (Task Group on Data Citation Standards Practices, 2013, p. 35). Granularity in data citation has not been studied actively despite the fact that it is a significant feature with regard to such parameters as collection level, item level and/or data level. Buneman (2006) noted that more than a single level of granularity is needed for the citation system. Concerns in this respect include "issues of granularity, version control, microattribution (fine-grained and unambiguous credit), contributor identifiers, and facilitation of reuse" (Task Group on Data Citation Standards Practices, 2013).

In summary, the study of data sharing, re-use and citation has focused largely on surveys of researchers engaged in these activities, or issues arising from the study of these practices. The study of data from data citation databases or the publications themselves remains relatively unexplored. The present study makes an original contribution to this area.

Methods

This research applies an exploratory method to the study of data sharing, re-use and citation. The DCI was selected as a data source because this index: (1) provides a single access point to over 500 repositories world-wide and to over two million data studies and datasets across multiple disciplines (http://wokinfo.com/products_tools/multidisciplinary/dci/), and (2) it monitors the quality of research data through editorial review of the repositories that house the data across multiple disciplines around the world (Swoger, 2012). In order to explore data citation as data sharing and re-use, this study applied citer-based methods in multiple Subject Categories, and exploratory data analysis. Based on a sample of highly cited authors, we explored references, main text, acknowledgement, supplementary information (e.g., supplementary materials, supplementary data, and web resources) and author information manually.

All 156 WoS Subject Categories for the sciences were analyzed in order to identify the top subject category with the highest numbers of records in the DCI. Genetics and Heredity represented the top subject category with almost 2.3 million records. Summary data for the subject area, including citations, document type (dataset, data study, repository), prevalence of DOIs and the distribution of citations over time were tabulated. Records were then sorted by the "most highly cited" in order to identify 15 datasets for further exploration. We identified the first and the last author for each of the 15 datasets, totaling 30 authors, for further analysis. Selection of the first and last authors was based on the assumption that the first author is the one who made the most significant contribution and the last author is the senior researcher with the most prestigious reputation (Wren, J. D., Kozak, K. Z., Johnson, K. R., Deakyne, S. J., Schilling, L.

M., & Dellavalle, R. P., 2007). We applied citer-based analysis, similar to the method used by Lu, Ajiferuke and Wolfram (Lu, Ajiferuke, & Wolfram, 2014) using the DCI- collected data and citing articles to these authors from the All Collections of the WoS. Bibliographic references for the citing articles for each publication were collected by using the 'Create citation report' feature. Then, the 'Analyze results" function for citing articles was used in order to identify the citers for each publication. All of the retrieved results (i.e., all of the citing articles) of the 30 authors were saved in tabular form and subjected to systematic sampling of every 10^{th} citing article of the 30 authors. In cases where the citing articles could not be obtained, the next observation in the systematic sample was selected among the 2,368 records. The sample size totaled 148 (n=148). Some authors had 0 citing articles and others had 50 citing articles only. We manually examined the 148 citing articles for evidence of data sharing and re-use the references, main text, acknowledgements, supplementary information (e.g., supplementary materials, supplementary data, and web resources) and author information in order to identify formal (i.e., cited) and informal (i.e., mentioned in passing or implied) data sharing and re-use.

Results

Figure 1 summarizes the top 10 Subject Categories where data sharing is most prevalent. The top subject category is Genetics and Heredity. The distribution for data sharing is quite skewed. Data cited in repositories were mostly available as unrestricted datasets rather than restricted/limited or embargoed datasets. Data sharing was very diverse depending on the subject category. For instance, some subject categories had more than 1 million shared datasets, others had 0 shared datasets in the DCI. Shared data have a low percentage of Digital Object Identifiers (DOIs).

DOIs represent unique identifiers for objects such as electronic documents, which simplifies the process of tracking digital objects. The proportion of data with DOIs in the Genetics and Heredity subject category was 4.63% (n=4,628 datasets). This low percentage makes it difficult to track automatically data citations. Generics and Heredity research often requires large amounts of data and data collection over time, thereby encouraging a culture of data sharing and re-use. We could not identify with the current datasets whether other disciplines have similar cultures . The DCI reported citation only in journal articles rather than conference proceedings or books, which may limit the record of data citation and re-use. Considering that conference proceedings are regarded as primary dissemination venues in some sciences, the impact of not indexing conference proceedings in the DCI needs to be studied further.

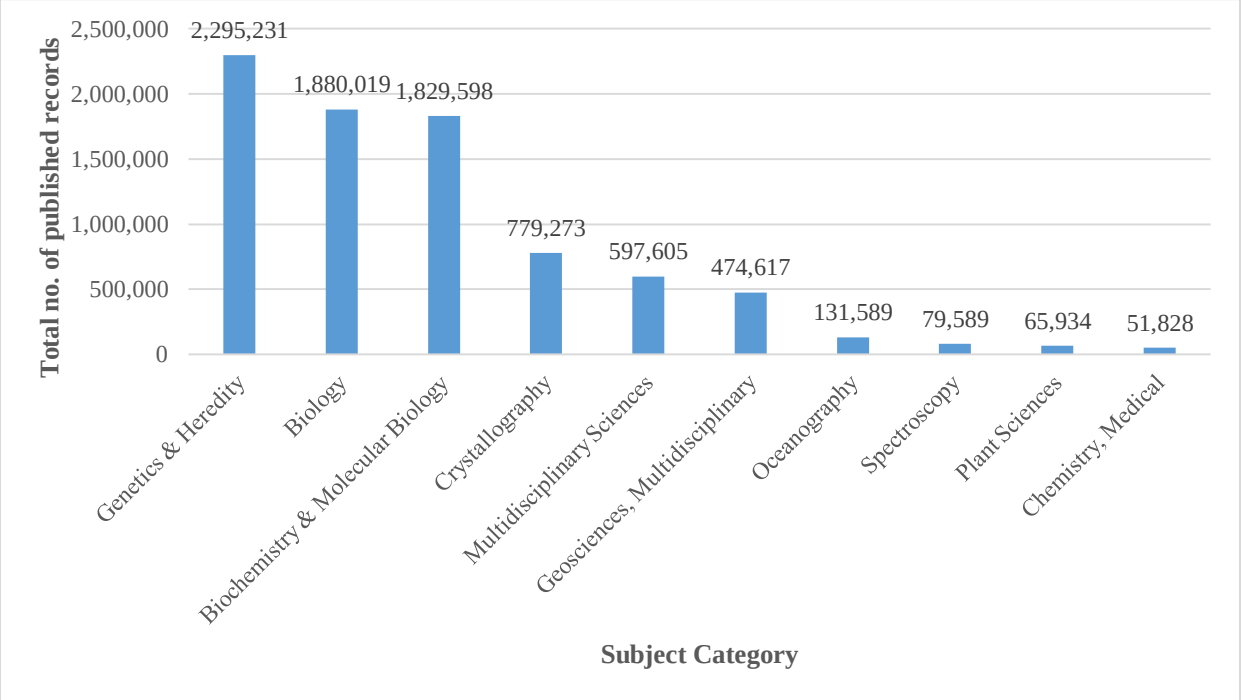

Fig. 1 Top 10 Subject Categories where data sharing is most prevalent

The data cited in the published articles analyzed were mainly housed in journal publishers' internal websites rather than in external/third-party data repositories. This may be due to the strict policies of journal publishers and/or related to the maintenance of data integrity. Neither the form of repositories nor the form of data studies were commonly found in the DCI.

Table 1 presents a summary of the distribution of citations for Genetics and Heredity based on the different types of citable units (dataset, data study, repository) for all the records studied and for those specifically represented with a DOI. No repositories have corresponding DOIs, although, on average, as citable units, the 39 repositories resulted in the highest average citations (mean = 41.9 citations, minimum =1 citation, max=701 citations). Based on the result, repository is the document type where most citations are received (# total items = 39, total citations = 1,633) among data set, data study and repository.

Table 1 Overview of citation distribution of Genetics and Heredity in the DCI (n=100,000 items)

	Document Type	# Total items	Total Citations	Mean Citations	Standard Deviation	Variance	Maximum Citations	Minimum Citations
All	Total	100,000	115,585	1.2	4.32	18.7	701	1
	Data set	56,599	65,828	1.2	0.94	0.9	121	1
	Data study	43,362	48,124	1.1	5.1	25.6	643	1
	Repository	39	1,633	41.9	129.8	16,846.8	701	1
With DOI	Total	4,528	4,531	1	0.03	0.001	3	1
	Data set	4,526	4,529	1	0.03	0.001	3	1

	Data study	2	2	1	0	0	1	1
	Repository	0	0	0	0	0	0	0

Table 2 summarizes the general results obtained from the DCI-based citation analysis for the last 35 years. The total amount of research data has dramatically increased since 1999. This analysis reveals that both uncitedness and items without DOIs are prevalent. The high level of uncitedness of research data corresponds to the findings of Torres-Salinas et al. (2014) and Peters et al. (2016). The number of items with a DOI is comparatively higher after the year 2000. Items with a DOI and at least 1 citation represent 0% of citations over the past 35 years. This may mean the DOI is not an important factor for the increase of research data citations, although further research is needed. The percentage of research data with a DOI is lower than we had expected. However, the increase of data published since 2000 with DOIs may confirm the interest in newer research data of the Genetics and Heredity in recent years. This result corresponds to the findings of Peters et al. (2016) for the interest in newer research data and increase of the scientific community in recent years.

Table 2 Results of the DCI-based citation information for 35 years (n=100,000 items)

	DCI	1980-1989	1990-1999	2000-2009	2010-2015
Total	Total # of items	2	6	36,841	63,151
Cited Reference Count	# items with >2 cited references (%)	0 (0%)	0 (0%)	2 (0%)	228 (0.4%)
	# items with at least 1 cited reference (%)	0 (0%)	0 (0%)	738 (2%)	2,255 (3.6%)
	Uncited (%)	2 (100%)	6 (100%)	36,103 (98%)	60,773 (96.2%)

Total Citations (All Sources)	# items with >2 total citations (%)	2 (100%)	6 (100%)	1,495 (4.1%)	4,278 (6.8%)
	# items with at least 1 citation (%)	2 (100%)	6 (100%)	35,347 (100%)	2,255 (100%)
	Uncited (%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
DOI	Items with DOI	0 (0%)	0 (0%)	2,150 (5.8%)	2,378 (3.8%)
	Items with DOI and at least 1 citation	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Items without DOI	2 (100%)	6 (100%)	34,691 (94.2%)	60,773 (96.2%)

Citer-based analysis

In order to explore co-author self-citation or re-citation, we applied a citer-based method. The Genetics and Heredity subject category is in the sciences, where collaboration (e.g., co-authorship) is more prevalent than is the case in the social sciences or humanities (Larivière, Gingras, Sugimoto & Tsou, 2015), and hyperauthorship may be more prevalent. Hyperauthorship (Cronin, 2001) refers to the practice of publishing papers with large numbers of co-authors, as many as 100 or even 500, which can inflate the number of people influenced by a given work in citing articles. Hyperauthorship was relatively common in the area of Genetics and Heredity. In terms of collaboration, citer-based analysis may represent a remedy for co-author self-citation. Ajiferuke, Lu, and Wolfram (2010) extended the definition of self-citation "to include citations originating from publications authored by one of the coauthors of the cited publication of interest, or coauthor self-citations" (p. 2089) because citations usually do not address the origin of the citation beyond self-citations.

A small number of highly cited authors may be unduly influence data citation counts. This study revealed that self-citation, including co-author self-citation, is prevalent in data citation. Table 3 illustrates that co-author data self-citation or recitation is more common than bibliographic self-citation. The average percentage of self-citation at the dataset level (8%) in Genetics and Heredity was much higher than the average self-citation at the publication (bibliographic citation) level (1.2%), meaning that the same data was cited (e.g., measured as a citation by the DCI) more than once owing to self-citation. This result corresponds to finding of (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015).

Table 3 Summary of self-citation for citer-based analysis

Subject Category	Totals					
	Publication level (i.e., data)			Article level (i.e., citing article)		
		DCI		WoS all databases		
	Total datasets in subject category	Total citations without self-citations	Total citations including self-citations	Total citing articles	Total citations without self-citations	Total citations including self-citations
Genetics and Heredity	11,514	384	418 (8%)	8,419	8,314	105 (1.2%)

Table 4 summarizes the results of the manual analysis of data sharing and re-use in the 148 sampled articles. An outlier that had 62 total citations in the main text was removed from the sample. It would have had a dramatic influence on the mean values obtained for the analysis. Note that the total citations of data sharing are higher than the total citations of data re-use in all citing articles whether it is main text, reference, supplementary information or acknowledgement.

Table 4 Overview of citation distribution of Genetics and Heredity in citing articles regarding data re-use and data sharing (n=148)

	Citing articles	Total citations	Mean citations	Standard deviation	Variance	Maximum citations	Minimum citations
Data re-use	Main text	29	0.2	0.62	0.39	4	0
	Reference	17	0.11	0.47	0.23	3	0
	Supplementary information	16	0.11	0.73	0.53	8	0
	Acknowledgement	4	0.03	0.2	0.04	2	0
Data sharing	Main text	173	1.17	3.45	11.91	24	0
	Reference	71	0.48	1.25	1.57	8	0
	Supplementary information	60	0.41	1.11	1.24	10	0
	Acknowledgement	12	0.08	0.53	0.28	6	0

Location of data citations

Examples of data sharing and re-use were most common in the main text of the articles, followed by the Reference and Supplementary information sections, respectively, with far fewer examples Acknowledgement section of the publications. Examples of data sharing and re-use appearing in different sections of publications follow.

Data citations appearing in the reference section of an article occur less frequently than in the main text, making it difficult to identify the reward and credit for data authors (i.e., data sharers). Consistent data citation formats could not be found. Current data citation practices do not (yet) benefit data sharers because only one sample has placed data citation with an accession number within the references (i.e., GenBank accession # AF336231) that might mean that data producers' publications are regularly cited rather than citing datasets directly (Fig. 2). References to data journals could be counted as possible re-use. Data re-use was mainly found when terms such as

"data," "survey" or ".gov" appear in hyperlink format (e.g., Available: http://www.cdc.gov/brfss/technical_infodata/surveydata/ 2007.htm) in references. In the references, generalized rules (e.g., DataCite) are not used. 'Suppl' was used in order to make it easier to find supplementary information. Also, data citation was sometimes not located in the references of an article in order to record scholarly records, but was instead located in the supplementary information, outside of the references. Data that had been re-used was often not acknowledged in the reference lists, but was rather hidden in the representation of data (e.g., tables, figures, images, graphs, and other elements), which may be a consequence of the fact that data citation practices are not yet common in scholarly communications. Computer code was not shared by data creators in any citing articles.

- Fattovich G, Stroffolini T, Zagni I,.Donato F. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. Gastro- enterology 2004;127:**Suppl 1**:S35-S50
- Lui Z, Lin J, ChenW, Jia Z, Pan D, Xu A. 2001. Sequence of complete exon 2 and partial intron 2 of HLA-DPB1*8001 allele. (**GenBank accession # AF336231**).
- **Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10, 5–6**
- Anon., 2009a. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey **Questionnaire** 2007 [Online]. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Available: http://www.cdc.**gov**/brfss/technical_infodata/**surveydata**/ 2007.htm [Accessed July 2009].
- Centers for Disease Control and Prevention. National Health and Nutrition Examination **Survey**: Surplus Sera Laboratory Component: Racial/Ethnic Variation in Sex Steroid

Hormone Concentrations Across Age in US Men (October 2006). Atlanta, GA: Centers for Disease Control and Prevention; 1997. (http://www.cdc.**gov**/nchs/nhanes/nh3**data**.htm). (Accessed June 1, 2009)

- National Oceanic and Atmospheric Administration. 2011. National Climatic Data Center. Protecting the Past, Revealing the Future. Available: http://www.ncdc.**noaa**.**gov**/oa/ncdc.html [accessed 19 December 2011].

Fig. 2 References

Journal publishers' internal websites, rather than an external/third-party data repositories or institutional repositories, were identified as indicators of data sharing in supplementary data/information (Fig. 3). Data citation was captured in the supplementary information/materials of citing articles that did not give due credit for researchers' use of shared data. Online articles provided supplementary data, although the same articles in offline formats (e.g., PDF) did not provide any information regarding supplementary data in common. This situation can be problematic because researchers store articles in common in an offline format in a local storage site (e.g., Mendeley desktop).

- Raw data from ASD family (**accession** phs000267.v1.p1) and SAGE control (**Accession**: phs000092.v1.p1) genotyping are at NCBI dbGAP.
- **439_2010_911_MOESM1_ESM.doc (131584) Supplementary** material 1 (DOC 128 kb)
- **Supplementary** Information is linked to the **online version of the paper** at www.nature.com/nature.
- **Supplemental** Data include three figures and two tables and **can be found with this**

article online at **http://www.cell.com/AJHG**/.

- The DNA resequencing data for SULT1E1 described in this manuscript have been **deposited** in the **NIH**-sponsored Pharmacogenetics Research Network **database**
- GBB_608_sm_figureS3.tif 855K **Supporting** info item

Fig. 3 Supplementary data/Information

The institutional homepage (e.g., http://pga.gs.washington.edu), third-party data repository (e.g., GenBank) and funding agencies' websites (e.g., National Institutes of Health or NIH) were identified as supplementary websites (Fig. 4). Websites were public and accessible to those not affiliated with the institution. Pages of individual researchers' websites, such as those of authors, were usually referred to with a URL only, i.e., without URIs or DOIs in the reference list, which confounds working with supplementary datasets or websites that are independent of journal publishers' websites, making automatic tracking or measurements of datasets difficult.

- http://pga.gs.**washington.edu**; Seattle SNPs Web site. http://genome.perlegen.com/browser/download.html; Perlegen Web site. http://genome.**ucsc.edu**/cgi-bin/hgGateway; UCSC Genome Browser.
- Accession numbers and URLs for data presented herein are as follows:

dbSNP Home Page, http://www.ncbi.nlm.**nih**.gov/SNP/index.html (for tagSNPs 790 [rs3093058], 1440 [rs3091244], 1919 [rs1417938], 2667 [rs1800947], 3006 [rs3093066], 3872 [rs1205], and 5237 [rs2808630]) **GenBank**, http://www.ncbi.nlm.**nih**.gov/Genbank/ (for the CRP gene [**accession number** AF449713]) SeattleSNPs Program for Genomic Applications, http://pga.gs.**washington.edu**/protocols/dnapanel_protocol.html TFSEARCH, http://www.cbrc.jp/research/db/TFSEARCH.html

Fig. 4 Supplementary website

Sections for acknowledgments or funding information are used for neither the indication of data sharing nor the indication of data re-use. The NIH was mainly used as a repository in acknowledgments. This may be due to the NIH's relatively early data sharing requirements, which date back to 2002. Funding agencies' websites (e.g., that of the NIH), institutional websites (e.g., wustl.edu), third-party data repositories (e.g., PharmGKB) and personal acquaintances (e.g., Dr. Donald Capra) were found in the acknowledgments as indicators of data sharing (Fig. 5).

- **Phenotypic and genotypic data** are **stored** in the NIDA Center for Genetic Studies (NCGS) at http://zork.**wustl.edu/** under NIDA Contract HHSN271200477451C (PIs J Tischfield and J Rice)
- Data will be **deposited** into the **Pharmacogenetics Knowledge Base (PharmGKB)**, supported by **NIH**/NIGMS Pharmacogenetics Research Network and **Database** Grant U01GM61374, http://pharmgkb.org
- We are **indebted to** Dr. Donald Capra for providing the amino acid analysis data
- Mortality data for the Netherlands were **obtained from** ''Statistics Netherlands

Fig. 5 Acknowledgments

The main text of publications, specifically the methodology sections, such as data collection or data analysis, provides the most direct evidence of data re-use (Fig. 6). Data re-use was mainly found as narrative content embedded in the main text or in the representation of the data. In the main text, specific terms and/or phrases were found as indicators of data re-use and data sharing. Indicator terms and/or phrases for data re-use include "samples," "sample sets," "donated from/by," "obtained from," "purchased from," "donated from," "used," "repository," "gift," "lab/laboratories," "commercial," "Corp.," "Inc.," and "Ltd." Data re-use was not only from data

originating from scholarly communications but also from proprietary datasets that may require re-users to purchase quality datasets. The "donated from/by" indicates that the direct contacts of acquainted authors were used for obtaining data for secondary analysis. When the main text was reviewed, authors described their data as having been purchased from "Corp.," "Inc." or "Ltd.". Indicator terms and/or phrases for data sharing were identified as "supplemental," and "supplemental material," and "repository." The reproducibility of data re-use was mostly hidden in the representation of data (e.g., figures, tables, images, graphs, and other elements) within the main text--places where due credit to contributors of shared/published data is more difficult to verify and/or identify. Indicating terms and/or phrases were mostly found in the methods portion, such as data collection or data analysis in the main text. Ways of collecting data for data re-use that may save considerable time and effort for researchers who re-use data for secondary analysis were identified mainly in the data collection process in the methods portion, namely (1) directly downloading data from (restricted/unrestricted) repositories, (2) data purchase from companies/labs, and (3) obtaining data through personal acquaintances (e.g. donation).

- Table 1. Baseline Characteristics of Participants in the Third National Health and Nutrition Examination **Survey** (1988-1991) and the Multi-Ethnic Study of Atherosclerosis (2002)
- DNA **samples** from 60 AA and 60 CA subjects (**sample sets** HD100AA and HD100CAU) were **obtained** from the Coriell Institute Cell **Repository** (Camden, NJ, USA).
- … with 100 individuals stemming from the Coriell Cell **Repositorie** (HD100CAU; Coriell Institute of Medical Research, Camden, NJ) and
- Restriction endonucleases were **purchased from** Bethesda Research Laboratories **Inc**.,

New England Bio**labs**, Boehringer Corporation **Ltd** and Miles Laboratories **Inc**. Phage T4-DNA ligase was either a **gift** from K. and N. E. Murray or **purchased from** Bethesda Research Laboratories **Inc**. or New England Biolabs. DNA polymerase (Klenow fragment) …

- Population samples DNA samples from human populations were **obtained** from the Coriell Cell **Repository**. Sequence variation was surveyed in DNA samples from three human populations: 24 European-Americans (**Repository numbers** NA17206–8, 17211–17, 21, 24, 26, 28, 30, 32, 34–36, 38, 40, 43–45), 24 African-Americans (NA17101–116, NA17133–40), and 24 Asians (10 Han Chinese: NA16654, 88, 89, 17014–20; 10 Japanese: NA17051–60; and four southeast Asians: NA17081–84). T4 DNA ligase was a generous **gift** of 0. Danos, all other enzymes were **purchased from** New England **Laboratories** or Boehringer Mannheim and **used** according to the manufacturers' instructions. [y-3 P]ATP (3000 Ci/mmole), [a - P] dXTP (3000 Ci/mmole), [ax- P] cordycepin triphosphate (3000 Ci/mmole) were from Amersham and [35S] methionine (1000 Ci/mmole) was **from** New England Nuclear. Chemicals used for DNA sequencing were of the highest grade commercialy available. Chemicals used for protein sequencing were **from** Beckman.
- A plasmid pMCR561 was kindly **donated by** T. Miki (Yamaguchi University, School of Medicine, Japan) (11). An expression plasmid pPL-X that carries the PL promoter and N gene on a 1215-base pair (bp') segment of the genome inserted between the EcoRI and BamHI site of pBR322 and its host strain N4830 (12) were **obtained from** Pharmacia/P-L Biochemicals.
- Enzymes and Reagents-Various DNA-modifying and restriction enzymes were **commercial** products. [LT-~'P]~ATP (>400 Ci/mmol, 1 Ci = 37 GBq) was **purchased**

from Amersham **Corp**. Dideoxy-NTPs and deoxy-NTPs were **obtained** from P-L Biochemicals and Sigma, respectively. Other reagents were commercial products of analytical grade.

- **Supplemental** Material can be found at: http://jn.nutrition.org/content/suppl/2008/11/20/138.12.2422.DC1.htm
- The details of the model building procedure are presented in the **Supplemental** Material, p. 4 (http:// dx.doi.org/10.1289/ehp.1104447)
- The genomic sequences 20 bases upstream and downstream of each LPA SNP of interest were **downloaded from** the UCSC Genome Browser (http://genome.**ucsc**.**edu/)**

Fig. 6 Main text

Although not common, author information is included when datasets are stored publicly. Thus, for example, both the project website (e.g., http://www.1000genomes.org) and major funding agencies' websites (e.g., NIH) were used as indicators of data sharing (Fig. 7).

Primary sequence reads, mapped reads, variant calls, inferred genotypes, estimated haplotypes and new independent validation data are **publicly available** through the **project website** (http://www.1000genomes.org); filtered sets of variants, allele frequencies and genotypes are also deposited in dbSNP (http:// www.ncbi.nlm.**nih**.gov/snp).

Fig. 7 Author information

Discussion

The availability of data citation may encourage data authors to make their peer reviewed data discoverable for re-use by others in order to increase data authors' recognition and rewards in

scholarly communications. In answer to RQ1, the frequency analysis of the WoS subject categories in which data citation is taking place reveals that the formally recorded citations are largely concentrated in a small number of disciplines in the biomedical sciences and selected physical sciences. We cannot conclude from this that data citation is only predominant in these fields, but rather that these fields may have greater data repository representation in the DCI.

Although the growth of formal data citation over the past 35 years has been impressive, these formal citations represent only a subset of data sharing and re-use practice. Data citation is assumed to be a prerequisite for data re-use, but does not necessarily reflect actual data re-use or the reach of the public data. Conversely, data re-use may not be captured through data citation because authors may not formally cite the data being re-used. Standard methods used in citation analysis allow us to explore community, collaboration, recitation and self-citations in scholarly communications. Measuring scholarly impact is important for the reproducibility of research (e.g., data re-use for secondary analysis) in scholarly communications (e.g., scientific community). The analysis of the sample of publications from the Genetics and Heredity area reveals that in addition to formal data citation, which may or may not be indexed in the DCI, there is substantial informal data citation and re-use taking place (RQ 2). The DCI does not capture these references because the citing authors themselves are not formally citing the data sources or their re-use, or the DCI does not index all of the repositories or datasets used. With respect to data sharing, shared data indexed in the DCI were housed mostly in repositories with unrestricted use. Shared data in the citing articles were housed mostly on servers of journal publishers that may be restricted due to subscription requirements. Funding agencies' websites (e.g., NIH) are actively used as a repository of data sharing for preservation or curation.

Ongoing challenges remain in identifying and documenting data citation. First, the practice of informal data citation presents a challenge for accurately documenting data citation practice (RQ 3). As the investigation of the 148 articles revealed, formal and informal data citation take place in different areas of articles. It would be reasonable to expect data citations to appear alongside standard bibliographic citations as acknowledgment of the author utilization of data citation. Based on the analysis of data citations and the citing literature, data citation, if included, may be found in supplementary materials or acknowledgements. Furthermore, citations may be informally included in the main text of a document. These forms of acknowledgement can be as simple as the re-use of figures that summarize data from earlier papers by the authors themselves or others.

Second, data recitation by one or more co-authors of earlier studies (i.e., self-citation) is common, which reduces the broader impact of data sharing by limiting much of the re-use to the original authors. This observation represents a key challenge to the identification of data re-use without analyzing the content of the citing document to determine if data re-use actually took place. Citer-based analysis merits consideration as an alternative to citation-based analysis for collaboration, recitation, and self-citations (Ajiferuke, Lu, & Wolfram, 2010). This study reveals that co-author self-citations among highly cited authors are common in data citation. This finding demonstrates that an increase in citations does not necessarily indicate new and unique citers. Co-author self-citation needs to be studied in further detail in data citation.

Third, data citation may not indicate inquiries into phenomena associated with a rapidly advancing area, such as in the hard sciences or computer engineering because works were heavily associated with journal articles. Around 90 percent of works were journal articles. In a rapidly advancing area, conference proceedings can have greater importance than journal articles

or books as research dissemination venues. Unlike conference proceedings, reviewing time for articles or books may take more than a year depending on the journal or publisher. This may be because high profile journals have policies of strict data sharing requirements, while conference proceedings or books do not currently have strict data sharing policies. Genetics and Heredity represents the field with the greatest volume of data sharing, as recorded by the DCI. However, data citation and re-use are still relatively infrequent and data recitation is common. We cannot conclude that this would be the case for all disciplines.

Fourth, the number of authors associated with shared datasets raises questions of the ownership of and responsibility for a collective work, although some journals require one author to be responsible for the data used in the study. Hyperauthorship is common in some areas, such as biomedical research, because, in big science, large research teams are commonplace. These situations raise questions regarding the identification of universal indicators for interdisciplinary research in big science and make clear the vital importance of discipline-specific research owing to diverse citation behaviors in different disciplines. There is a need to consider whether this is practical for data citation, however, owing to data reduction metrics in regular journals or in data journals. A data journal (e.g., the one supported by *Nature*) providing a specific "data citation" section within its articles starting from 2016 can play an important role in data authorship. Data authorship for sequencing and/or verifying authors should be given careful consideration because courtesy authorship (i.e., a contributing role as an author in the acknowledgments or in supplementary information/materials apart from references) may be more complex than sequencing and/or verifying an article (i.e., a single work), since it does not give credit to contributors. Version control is also important, since citing an article with associated datasets

(i.e., a single article or work having multiple associated datasets) may create additional challenges for data citation.

Manipulation and/or duplication of research resources, such as image files, as a form of data re-use can be identified by providing unique searchable identifiers of exact resources, namely Research Resource Identifiers (RRIDs). Examples include model organisms, antibodies, reagents or tools used for the experimental procedures or supplemental experimental procedures in articles that have RRIDs. RRIDs in the methods sections of the main text in articles may improve automatic machine tracking of data re-use for data citation in terms of both identifiability and reproducibility. As this study has revealed, the methods portion in the main text, such as data collection or data analysis, provides the most direct indicators of data re-use.

The current study represents an initial exploration. Limitations include the focus on first and last authors, with the assumption that the first author is the one whose contribution was greatest and the last is the senior and most prestigious researcher. Secondary impact of data re-use could not be identified in the current study. Funding agencies' data sharing requirements are major imperatives for data sharing rather than the requirements of journal publishers, although this is not necessarily indicated. Furthermore, the limitation of the current study to the field of Heredity and Genetics prevents us from generalizing the findings to other disciplines. However, if the practices observed are common in the subject area exhibiting the highest level of data sharing activity, this in itself attests to the need for greater standardization of data citation. Finally, for authors in disciplinary areas that are now beginning to use open data and data citation, they may wish to learn from the challenges outlined for Genetics and Heredity, where data sharing is already common.

Conclusion

This study explored data sharing, re-use and citation characteristics in the WoS category Genetics and Heredity, the WoS subject area with the highest level of data citation. The practice of citation indicates scholarly influence. It supports the idea that data is an important research output. More than 2,000,000 peer-reviewed data publications and their citing articles in the Genetics and Heredity subject category of the DCI were explored.

Challenges remain for effectively and more comprehensively recording data citation so that authors of datasets received appropriate attribution. Systematic recording of data citation is still lacking, which creates barriers for researchers interested in studying data citation and for author of open datasets who may not receive attribution for their data contributions. Sources, like the DCI, have begun to capture instances of data citation, but currently index only a fraction of the data citation activity. Similarly, authors do not systematically document data sharing and re-use through formal citation, although it may be captured informally within publications, but not in a standardized way. Consistent data citation format usage by authors could not be found. Higher levels of data citation activity are currently limited to a small number of disciplines, as recorded by the DCI. Data self-citation was found to be relatively common, where one or more co-authors of public datasets re-used data in subsequent publications. Data re-use is not always clear. Research data re-use cultures already exist, though they are not prevalent. Identifiers of exact resources that have been re-used, such as images, antibodies, organisms or tools (e.g., RRIDs), could not be identified in the current study. Re-use of quantitative datasets are more active than the re-use of qualitative datasets although that may be due to the proportionate difference in the number of quantitative datasets. Methods sections such as data collection or data analysis can be indicators of data re-use. Data re-use for secondary analysis was primarily found in the

representation of data (e.g., tables, figures, graphs, and images) of published articles rather than in the narrative content embedded in articles (e.g., main text). This highlights the importance of version control because data citation with re-used research data was primarily found to occur in cases of co-author self-citations. Furthermore, identifying unique authors may not be easy because authors with a researcher identifier, Open Researcher and Contributor IDentifier (ORCID) numbers, in the DCI were rare. Citing articles were in high profile journals rather than conference proceedings, books, or low profile journals that may demand subscription. The format of data cited from the citing-articles in the DCI was mostly in the form of datasets (e.g., accession numbers when stored in repositories). Future research will investigate additional subject categories in order to identify similarities and differences in data sharing, citation and re-use practice within (i.e., discipline-specific) and across disciplines (i.e., interdisciplinary).

References

Ajiferuke, I., Lu, K., & Wolfram, D. (2010). A comparison of citer and citation-based measure outcomes for multiple disciplines. *Journal of the American Society for Information Science and Technology, 61*(10), 2086-2096.

Altman, M. (2012). Data citation in the Dataverse Network. In *For attribution: Developign scientific data attribution and data citation practices and standards: Summary of an international workshop* (pp. 99-106). Washington, D.C.: National Academies Press.

Bourne, P. (2012). Towards data attribution and citation in the life sciences. In P. F. Uhlir (Ed.), *For attribution: Developing scientific data attribution and data citation practices and*

standards: Summary of an international workshop (p. 2012). Washington, D.C.: National Academies Press.

Buneman, P. (2006). How to cite curated databases and how to make them citable. *The 18th International Conference on Scientific and Statistical Database Management* (pp. 195-203). Los Alamitos: IEEE Computer Society. Retrieved January 26, 2016, from http://homepages.inf.ed.ac.uk/opb/homepagefiles/harmarnew.pdf

Callaghan, S. (2012). Data citation in the earth and physical sciences. In P. F. Uhlir (Ed.). Washington, D.C.: D.C.

Chao, T. (2015). Mapping methods metadata for research data. *International Journal of Digital Curation, 10*(1), 82-94.

Coleman. (2012). *Data citation.* Retrieved from http://www.bl.uk/aboutus/stratpolprog/digi/datasets/workshoparchive/LAC_Citation_May2012.pdf

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Jounal of the American Society for Information Science and Technology, 52*(7), 558-569.

Curty, R. G. (2015). *Beyond "data thrifting": An investigation of factors influencing research data.* Syracuse, NY: Syracuse University.

Data Citation Synthesis Working Group. (2014, February). *Joint declaration of data citation principles-Final*. Retrieved 2016, from www.force11.org/datacitation

DataCite Metadata Working Group. (2015). *DataCite metadata schema for the publication and citation of research data*. doi:10.5438/0010

Fear, K. (2013). *Measuring and anticipating the impact of data reuse.* Ann Arbor: University of Michigan.

Green, T. (2009). *We need publishing standards for datasets and data tables.* OECD Publishing. doi:10.1787/787355886123

Helbig, K., Hausstein, B., & Toepfer, R. (2015). Supporting data citation: Experiences and best practices of a DOI allocation agency for social sciences. *Journal of Librarianship and Scholarly Communication, 3*(2), eP1220. doi:10.7710/2162-3309.1220

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economics Studies, 76*(1), 283-317.

Kim, Y. (2013). *Institutional and individual influences on scientists' data sharing behaviors.* Syracuse, NY: Syracuse University.

Kurtz, M. J. (2012). Linking, finding and citing data in astronomy. In P. F. Uhlir (Ed.), *For attribution: Developing scientific data attribution and data citation practices and standards: Summary of an international workshop* (pp. 161-166). Washington, D.C., U.S.A.: National Academies Press.

LaBonte, K. B. (2015). Data citation rates: GIS data in the marine sciences and publisher citation requirements. *International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC).* doi:http://hdl.handle.net/1912/7402

Larivière, V. G. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66*(7), 1323-1332.

Lawrence, B., Jones, C., Mattews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation, 6*(2), 4-37.

Lee, D. J. (2015). *Research data curation practices in institutional repositories and data identifiers.* FL: Florida State University.

Lu, K., Ajiferuke, I., & Wolfram, D. (2014). Extending citer analysis to journal impact evaluation. *Scientometrics, 100*(1), 245-260.

Mayernik, M. S. (2012). Session summary: The RDAP (Research Data Alliance & Preservation) 12 data citation panel. *Bullentin of the American Society for Information Science and Technology, 38*(5), 31.

National Cancer Institute. (2006). *Data sharing policy.* Retrieved from http://ctep.cancer.gov/protocolDevelopment/docs/data_sharing_policy.pdf

National Institutes of Health. (2003). *NIH data sharing policy and implementation guideline*. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Nature Publishing Group. (2013). Announcement: Launch of an online data journal. *Nature, 502*(7074), 142. doi:10.1038/502142a

Parsons, M. A., Duerr, R., & Minster, J. B. (2010). Data citation and peer review. *Transactions American Geophysical Union, 91*(34), 297-298.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research data explored: Citations versus altmetrics. *arXiv preprint*. doi:arXiv:1501.03342

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics, 107*(2), 723-744.

Piwowar, H. A., & Vision, T. (2013). Data reuse and the open data citation advantage. *PeerJ, 1*, E175.

Piwowar, H. A., Day, R., & Fridsma, D. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, e308. doi:10.1371/journal.pone.0000308

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wikinson, M., & European Commission. (2011). *Report on integration of data and publications.*

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23529

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1-43. doi:10.1002/aris.2009.1440430113

Sperberg-McQueen, M. (2012). Data citation in the humanities: What's the problem? In P. F. Uhlir (Ed.), *For attribution: Developing scientific data attribution and data citation practices and standards: Summary of an international workshop* (pp. 59-64). Washington: National Academies Press.

Starr, J., & Gastl, A. (2011). IsCitedBy: A metadata scheme for DataCite. *D-Lib Magazine, 17*(1/2). doi:10.1045/january2011-starrto

Stephan, P. (2010). The economics of science. In B. Hall, & N. Rosenberg (Eds.). Amsterdam: Elsevier.

Swoger, B. (2012, December). Thomson Reuters Data Citation Index. *Library Journal*. Retrieved from http://wokinfo.com/media/pdf/dci-libjrnl-review.pdf

Task Group on Data Citation Standards Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal, 12*, CIDCR1-CIDCR75.

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

Tenopir, C., Dalton, E., 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*, e0134826. doi:doi:10.1371/journal.pone.0134826

Torres-Salinas, D., Jiménez-Contreras, E., & Robinson-García, N. (2014). How many citations are there in the Data Citation Index? *arXiv preprint*. doi:arXiv:1409.0753

Uhlir, P. F. (2012). *Developing data attribution and citation practices and standards: Summary of an international workshop.* The National Academic Press.

Vardigan, M. (2012). Data citation for the social sciences. In P. F. Uhlir (Ed.), *For attribution: Developing scientific data attribution and data citation practices and standards: Summary of an international workshop* (pp. 55-58). Washington, D.C.: National Academies Press.

Wren, J. D., Kozak, K. Z., Johnson, K. R., Deakyne, S. J., Schilling, L. M., & Dellavalle, R. P. (2007). The write position. *EMBO reports, 8*(11), 988-991.

Yoon, A. (2015). *Data reuse and users' trust judgments: Toward trusted data curation.* Chapel Hill, NC: University of North Carolina.