Research Data Publication and Citation Bibliography

Charles W. Bailey, Jr.

Houston: Digital Scholarship, 2022

The *Research Data Publication and Citation Bibliography* includes over 225 selected English-language articles and books that are useful in understanding the publication and citation of research data. It also provides limited coverage of closely related topics, such as research data identifiers (e.g., DOI) and scholarly metrics. It is available as a website and a website PDF with live links.

For an overview of data publication, see:

> Austin, Claire C., Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha K. Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte. "Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing." *International Journal on Digital Libraries* 18, no. 2 (2017): 77-92. https://doi.org/10.1007/s00799-016-0178-2

For an overview of data citation, see:

> Parsons, Mark A., Ruth E. Duerr, and Matthew B. Jones. "The History and Future of Data Citation in Practice." *Data Science Journal* 18, no. 1 (2019): p.52. https://doi.org/10.5334/DSJ-2019-052

This bibliography does not cover conference proceedings, digital media works (such as MP3 files), editorials, e-mail messages, interviews, letters to the editor, presentation slides or transcripts, technical reports. unpublished e-prints, and/or weblog postings.

Most sources have been published from January 2009 through December 2021; however, a limited number of earlier key sources are also included. The bibliography has links to included works. Where possible, it uses Digital Object Identifier System (DOI) URLs. All links are subject to change. Should a link be dead, try entering it in the Internet Archive Wayback Machine.

Some publishers may use nontraditional citation elements and patterns, and they may omit standard bibliographic elements and substitute new ones.

Abstracts are included in this bibliography if a work is under a Creative Commons Attribution License (BY and national/international variations), a Creative Commons public domain dedication (CC0), or a Creative Commons Public Domain Mark and this is clearly indicated in the work (see the "Note on the Inclusion of Abstracts" below for more details).

For over 200 works on the closely related topic of research data sharing and reuse, see:

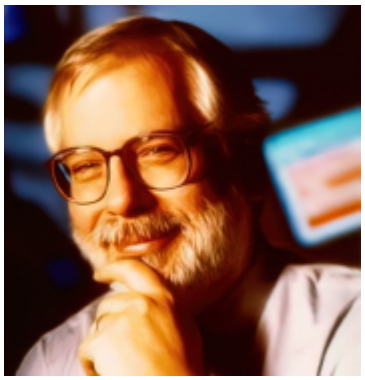
> Bailey, Charles W., Jr. *Research Data Sharing and Reuse Bibliography*. Houston: Digital Scholarship, 2021. http://digital-scholarship.org/rdsr/sharing.htm.

For an in-depth treatment of the curation of digital research data with over 800 references, see:

Bailey, Charles W., Jr. *Research Data Curation and Management Bibliography*. Houston: Digital Scholarship, 2021. http://digital-scholarship.org/rdcmb/rdcmb.htm.

Dedication

In memory of Paul Evan Peters (1947-1996), founding Executive Director of the Coalition for Networked Information, whose visionary leadership at the dawn of the Internet era fostered the development of scholarly electronic publishing.



Bibliography

Aalbersberg, IJsbrand Jan, Sophia Atzeni, Hylke Koers, Beate Specker, and Elena Zudilova-Seinstra. “Bringing Digital Science Deep inside the Scientific Article: The Elsevier Article of the Future Project.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 23, no. 4 (2014): 275-299. http://doi.org/10.18352/lq.8446

> In 2009, Elsevier introduced the “Article of the Future” project to define an optimal way for the dissemination of science in the digital age, and in this paper we discuss three of its key dimensions. First we discuss interlinking scientific articles and research data stored with domain-specific data repositories—such interlinking is essential to interpret both article and data efficiently and correctly. We then present easy-to-use 3D visualization tools embedded in online articles: a key example of how the digital article format adds value to scientific communication and helps readers to better understand research results. The last topic covered in this paper is automatic enrichment of journal articles through text-mining or other methods. Here we share insights from a recent survey on the question: how can we find a balance between creating valuable contextual links, without sacrificing the high-quality, peer-reviewed status of published articles?

Aalbersberg, IJsbrand, Judson Dunham, and Hylke Koers. “Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing.” *Data Science Journal* 12 (2013): pp.WDS235-WDS242. https://datascience.codata.org/articles/abstract/168/

> Researchers across disciplines are increasingly utilizing electronic tools to collect, analyze, and organize data. However, when it comes to publishing their work, there are no common, well-established standards on how to make that data available to other researchers. Consequently, data are often not stored in a consistent manner, making it hard or impossible to find data sets associated with an article—even though such data might be essential to reproduce results or to perform further analysis. Data repositories can play an important role in improving this situation, offering increased visibility, domain-specific coordination, and expert knowledge on data management. As a leading STM publisher, Elsevier is actively pursuing opportunities to establish links between the online scholarly article and data repositories. This helps to increase usage and visibility for both articles and data sets and also adds valuable context to the data. These data-linking efforts tie in with other initiatives at Elsevier to enhance the online article in order to connect with current researchers’ workflows and to provide an optimal platform for the communication of science in the digital era.

Abella, Alberto, Marta Ortiz-de-Urbina-Criado, and Carmen De-Pablos-Heredero. “The Process of Open Data Publication and Reuse.” *Journal of the Association for Information Science and Technology* 70, no. 3 (2019): 296-300. https://doi.org/https://doi.org/10.1002/asi.24116

Aleixandre-Benavent, Rafael, Antonia Ferrer Sapena, Silvia Coronado Ferrer, Fernanda Peset, and Alicia García García. “Policies Regarding Public Availability of Published Research Data in Pediatrics Journals.” *Scientometrics* 118, no. 2 (2019): 439-451. https://doi.org/10.1007/s11192-018-2978-1

Alsheikh-Ali, Alawi A., Waqas Qureshi, Mouaz H. Al-Mallah, and John P. A. Ioannidis. “Public Availability of Published Research Data in High-Impact Journals.” *PLOS ONE* 6, no. 9 (2011): e24357. https://doi.org/10.1371/journal.pone.0024357

Background

There is increasing interest to make primary data from published research publicly available. We aimed to assess the current status of making research data available in highly-cited journals across the scientific literature.

Methods and Results

We reviewed the first 10 original research papers of 2009 published in the 50 original research journals with the highest impact factor. For each journal we documented the policies related to public availability and sharing of data. Of the 50 journals, 44 (88%) had a statement in their instructions to authors related to public availability and sharing of data. However, there was wide variation in journal requirements, ranging from requiring the sharing of all primary data related to the research to just including a statement in the published manuscript that data can be available on request. Of the 500 assessed papers, 149 (30%) were not subject to any data availability policy. Of the remaining 351 papers that were covered by some data availability policy, 208 papers (59%) did not fully adhere to the data availability instructions of the journals they were published in, most commonly (73%) by not publicly depositing microarray data. The other 143 papers that adhered to the data availability instructions did so by publicly depositing only the specific data type as required, making a statement of willingness to share, or actually sharing all the primary data. Overall, only 47 papers (9%) deposited full primary raw data online. None of the 149 papers not subject to data availability policies made their full primary data publicly available.

Conclusion

A substantial proportion of original research papers published in high-impact journals are either not subject to any data availability policies, or do not adhere to the data availability instructions in their respective journals. This empiric evaluation highlights opportunities for improvement.

Altman, Micah, Christine Borgman, Mercè Crosas, and Maryann Matone. “An Introduction to the Joint Principles for Data Citation.” *Bulletin of the Association for Information Science and Technology* 41, no. 3 (2015): 43-45. https://doi.org/10.1002/bult.2015.1720410313

Altman, Micah, Eleni Castro, Mercè Crosas, Philip Durbin, Alex Garnett, and Jen Whitney. “Open Journal Systems and Dataverse Integration—Helping Journals to Upgrade Data Publication for Reusable Research.” *Code4Lib Journal*, no. 30 (2015). http://journal.code4lib.org/articles/10989

This article describes the novel open source tools for open data publication in open access journal workflows. This comprises a plugin for Open Journal Systems that supports a data submission, citation, review, and publication workflow; and an extension to the Dataverse system that provides a standard deposit API. We describe

the function and design of these tools, provide examples of their use, and summarize their initial reception. We conclude by discussing future plans and potential impact.

Altman, Micah, and Gary King. “A Proposed Standard for the Scholarly Citation of Quantitative Data.” *D-Lib Magazine* 13, no. 3/4 (2007). http://www.dlib.org/dlib/march07/altman/03altman.html

Aquino, Janine, John Allison, Robert Rilling, Don Stott, Kathryn Young, and Michael Daniels. “Motivation and Strategies for Implementing Digital Object Identifiers (DOIs) at NCAR’s Earth Observing Laboratory—Past Progress and Future Collaborations.” *Data Science Journal* 16, no. 7 (2017): p.7. http://doi.org/10.5334/dsj-2017-007

In an effort to lead our community in following modern data citation practices by formally citing data used in published research and implementing standards to facilitate reproducible research results and data, while also producing meaningful metrics that help assess the impact of our services, the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) has implemented the use of Digital Object Identifiers (DOIs) (DataCite 2017) for both physical objects (e.g., research platforms and instruments) and datasets. We discuss why this work is important and timely, and review the development of guidelines for the use of DOIs at EOL by focusing on how decisions were made. We discuss progress in assigning DOIs to physical objects and datasets, summarize plans to cite software, describe a current collaboration to develop community tools to display citations on websites, and touch on future plans to cite workflows that document dataset processing and quality control. Finally, we will review the status of efforts to engage our scientific community in the process of using DOIs in their research publications.

Arend, Daniel, Matthias Lange, Jinbo Chen, Christian Colmsee, Steffen Flemming, Denny Hecht, and Uwe Scholz. “E!DAL—A Framework to Store, Share and Publish Research Data.” *BMC Bioinformatics* 15, no. 214 (2014). https://doi.org/10.1186/1471-2105-15-214.

Background The life-science community faces a major challenge in handling “big data”, highlighting the need for high quality infrastructures capable of sharing and publishing research data. Data preservation, analysis, and publication are the three pillars in the “big data life cycle”. The infrastructures currently available for managing and publishing data are often designed to meet domain-specific or project-specific requirements, resulting in the repeated development of proprietary solutions and lower quality data publication and preservation overall.

Results *e!DAL* is a lightweight software framework for publishing and sharing research data. Its main features are version tracking, metadata management, information retrieval, registration of persistent identifiers (DOI), an embedded HTTP(S) server for public data access, access as a network file system, and a scalable storage backend. *e!DAL* is available as an API for local non-shared storage and as a remote API featuring distributed applications. It can be deployed“out-of-the-box”as an on-site repository.

Conclusions *e!DAL* was developed based on experiences coming from decades of research data management at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK). Initially developed as a data publication and documentation infrastructure for the IPK's role as a data center in the DataCite consortium, *e!DAL* has grown towards being a general data archiving and publication infrastructure. The *e!DAL* software has been deployed into the Maven Central Repository. Documentation and Software are also available at: http://edal.ipk-gatersleben.de.

Assante, Massimiliano, Leonardo Candela, Donatella Castelli, and Alice Tani. “Are Scientific Data Repositories Coping with Research Data Publishing?” *Data Science Journal* 15, no. 6 (2016): p.6. http://doi.org/10.5334/dsj-2016-006

> Research data publishing is intended as the release of research data to make it possible for practitioners to (re)use them according to “open science” dynamics. There are three main actors called to deal with research data publishing practices: researchers, publishers, and data repositories. This study analyses the solutions offered by generalist scientific data repositories, i.e., repositories supporting the deposition of any type of research data. These repositories cannot make any assumption on the application domain. They are actually called to face with the almost open ended typologies of data used in science. The current practices promoted by such repositories are analysed with respect to eight key aspects of data publishing, i.e., dataset formatting, documentation, licensing, publication costs, validation, availability, discovery and access, and citation. From this analysis it emerges that these repositories implement well consolidated practices and pragmatic solutions for literature repositories. These practices and solutions can not totally meet the needs of management and use of datasets resources, especially in a context where rapid technological changes continuously open new exploitation prospects.

Austin, Claire C., Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha K. Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte. “Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing.” *International Journal on Digital Libraries* 18, no. 2 (2017): 77-92. https://doi.org/10.1007/s00799-016-0178-2

Beckles, Zosia, Stephen Gray, Debra Hiom, Kirsty Merrett, Kellie Snow, and Damian Steer. “Disciplinary Data Publication Guides.” *International Journal of Digital Curation* 13, no. 1 (2018): 150-160. https://doi.org/10.2218/ijdc.v13i1.603

> Many academic disciplines have very comprehensive standard for data publication and clear guidance from funding bodies and academic publishers. In other cases, whilst much good-quality general guidance exists, there is a lack of information available to researchers to help them decide which specific data elements should be shared. This is a particular issue for disciplines with very varied data types, such as engineering, and presents an unnecessary barrier to researchers wishing to meet funder expectations on data sharing. This article outlines a project to provide simple, visual, discipline-specific guidance on data publication, undertaken at the University of Bristol at the request of the Faculty of Engineering.

Belter, Christopher W. “Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets.” *PLOS ONE* 9, no. 3 (2014): e92590. http://dx.doi.org/10.1371/journal.pone.0092590

> Evaluation of scientific research is becoming increasingly reliant on publication-based bibliometric indicators, which may result in the devaluation of other scientific activities —such as data curation—that do not necessarily result in the production of scientific publications. This issue may undermine the movement to openly share and cite data sets in scientific publications because researchers are unlikely to devote the effort necessary to curate their research data if they are unlikely to receive credit for doing so. This analysis attempts to demonstrate the bibliometric impact of properly curated and openly accessible data sets by attempting to generate citation counts for three data sets archived at the National Oceanographic Data Center. My findings suggest that all three data sets are highly cited, with estimated citation counts in most cases higher than 99% of all the journal articles published in *Oceanography* during the same years. I also find that methods of citing and referring to these data sets in scientific publications are highly inconsistent, despite the fact that a formal citation

format is suggested for each data set. These findings have important implications for developing a data citation format, encouraging researchers to properly curate their research data, and evaluating the bibliometric impact of individuals and institutions.

Bierer, Barbara E., Mercè Crosas, and Heather H. Pierce. “Data Authorship as an Incentive to Data Sharing.” *The New England Journal of Medicine* 376, no. 17 (2017): 1684-87. https://doi.org/10.1056/NEJMSB1616595

Buneman, Peter, Greig Christie, Jamie A. Davies, Roza Dimitrellou, Simon D. Harding, Adam J. Pawson, Joanna L. Sharman, and Yinjun Wu. “Why Data Citation Isn't Working, and What to Do about It.” *Database* 2020 (2020): baaa022. https://doi.org/10.1093/DATABA/BAAA022

We describe a system that automatically generates from a curated database a collection of short conventional publications—citation summaries—that describe the contents of various components of the database. The purpose of these summaries is to ensure that the contributors to the database receive appropriate credit through the currently used measures such as h-indexes. Moreover, these summaries also serve to give credit to publications and people that are cited by the database. In doing this, we need to deal with granularity—how many summaries should be generated to represent effectively the contributions to a database? We also need to deal with evolution—for how long can a given summary serve as an appropriate reference when the database is evolving? We describe a journal specifically tailored to contain these citation summaries. We also briefly discuss the limitations that the current mechanisms for recording citations place on both the process and value of data citation.

Callaghan, Sarah. “Preserving the Integrity of the Scientific Record: Data Citation and Linking.” *Learned Publishing* 27, no. 5 (2014): 15-24. https://doi.org/10.1087/20140504

———. “Research Data Publication: Moving Beyond the Metaphor.” *Data Science Journal* 18, no. 1 (2019): p.39. https://doi.org/10.5334/dsj-2019-039

Metaphors are a quick and easy way of grasping (often complicated) concepts and ideas, but like any useful tools, they should be used carefully. There are as many arguments about how datasets are like cakes as there are about how datasets aren't like cakes.

It can be easy to categorise a dataset as being a special class of academic paper. Positively, this means that the tools and services for scholarly publication can be utilised to transmit and verify datasets, improving visibility, reproducibility, and attribution for the dataset creators. Negatively, if a dataset doesn't fit within the criteria to meet the “academic publication” mould (e.g. because it is being continually versioned and updated, or it is still being collected and will be for decades) it might be considered to be of less value to the community.

It is often said that “all models are wrong, but some are useful” (Box, 1979). Hence we need to determine the usefulness and limits of models and metaphors, especially when trying to develop new processes and systems.

This paper further develops the metaphors outlined in Parsons and Fox (2013), and gives real world examples of the metaphors from scientific data stored in the Centre for Environmental Data Analysis (CEDA)—a discipline-specific environmental data repository, and the processes that created the datasets.

Callaghan, Sarah, Steve Donegan, Sam Pepler, Mark Thorley, Nathan Cunningham, Peter Kirsch, Linda Ault, Patrick Bell, Rod Bowie, Adam Leadbetter, Roy Lowry, Gwen Moncoiffé, Kate Harrison, Ben Smith-Haddon, Anita Weatherby, and Dan Wright. “Making Data a First Class Scientific Output: Data Citation and Publication by NERC’s Environmental Data Centres.” *International Journal of Digital Curation* 7, no. 1 (2012): 107-113. https://doi.org/10.2218/ijdc.v7i1.218

The NERC Science Information Strategy Data Citation and Publication project aims to develop and formalise a method for formally citing and publishing the datasets stored in its environmental data centres. It is believed that this will act as an incentive for scientists, who often invest a great deal of effort in creating datasets, to submit their data to a suitable data repository where it can properly be archived and curated. Data citation and publication will also provide a mechanism for data producers to receive credit for their work, thereby encouraging them to share their data more freely.

Callaghan, Sarah, Jonathan Tedds, John Kunze, Varsha Khodiyar, Rebecca Lawrence, Matthew S. Mayernik, Fiona Murphy, Timothy Roberts, and Angus Whyte. “Guidelines on Recommending Data Repositories as Partners in Publishing Research Data.” *International Journal of Digital Curation* 9, no. 1 (2014): 152-163. https://doi.org/10.2218/ijdc.v9i1.309

This document summarises guidelines produced by the UK Jisc-funded PREPARDE data publication project on the key issues of repository accreditation. It aims to lay out the principles and the requirements for data repositories intent on providing a dataset as part of the research record and as part of a research publication. The data publication requirements that repository accreditation may support are rapidly changing, hence this paper is intended as a provocation for further discussion and development in the future.

Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Sarah Callaghan. “On Research Data Publishing.” *International Journal on Digital Libraries* 18, no. 2 (2017): 73-75. https://doi.org/10.1007/s00799-017-0213-y

Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. “Data Journals: A Survey.” *Journal of the Association for Information Science and Technology* 66, no. 9 (2015): 1747-1762. http://dx.doi.org/10.1002/asi.23358

Colavizza, Giovanni, Iain Hrynaszkiewicz, Isla Staden, Kirstie J. Whitaker, and Barbara McGillivray. “The Citation Advantage of Linking Publications to Research Data.” *PLOS ONE* 15, no. 4 (2020): 1-18. https://doi.org/10.1371/JOURNAL.PONE.0230416

Efforts to make research results open and reproducible are increasingly reflected by journal policies encouraging or mandating authors to provide data availability statements. As a consequence of this, there has been a strong uptake of data availability statements in recent literature. Nevertheless, it is still unclear what proportion of these statements actually contain well-formed links to data, for example via a URL or permanent identifier, and if there is an added value in providing such links. We consider 531,889 journal articles published by PLOS and BMC, develop an automatic system for labelling their data availability statements according to four categories based on their content and the type of data availability they display, and finally analyze the citation advantage of different statement categories via regression. We find that, following mandated publisher policies, data availability statements become very common. In 2018 93.7% of 21,793 PLOS articles and 88.2% of 31,956 BMC articles had data availability statements. Data availability statements containing a link to data in a repository—rather than being available on request or included as supporting information files—are a fraction of the total. In 2017 and 2018, 20.8% of PLOS publications and 12.2% of BMC publications provided DAS containing a link to data in a repository. We also find an association between articles that include statements that link to data in a repository and up to 25.36% (± 1.07%) higher citation impact on average, using a citation prediction model. We discuss the potential implications of these results for authors (researchers) and journal publishers who

make the effort of sharing their data in repositories. All our data and code are made available in order to reproduce and extend our results.

Cook, Robert B., Suresh K. S. Vannan, Benjamin F. McMurry, Daine M. Wright, Y. Wei, Alison G. Boyer, and J. H. Kidder. “Implementation of Data Citations and Persistent Identifiers at the ORNL DAAC.” *Ecological Informatics* 33 (2016): 10-16. https://doi.org/https://doi.org/10.1016/j.ecoinf.2016.03.003

Costello, Mark J., and John Wieczorek. “Best Practice for Biodiversity Data Management and Publication.” *Biological Conservation* 173, no. 1 (2014): 68-73. http://dx.doi.org/10.1016/j.biocon.2013.10.018

Cousijn, Helena, Patricia Feeney, Daniella Lowenberg, Eleonora Presani, and Natasha Simons. “Bringing Citations and Usage Metrics Together to Make Data Count.” *Data Science Journal*, 18, no. 1 (2019): p.9. http://doi.org/10.5334/dsj-2019-009.

Over the last years, many organizations have been working on infrastructure to facilitate sharing and reuse of research data. This means that researchers now have ways of making their data available, but not necessarily incentives to do so. Several Research Data Alliance (RDA) working groups have been working on ways to start measuring activities around research data to provide input for new Data Level Metrics (DLMs). These DLMs are a critical step towards providing researchers with credit for their work. In this paper, we describe the outcomes of the work of the Scholarly Link Exchange (Scholix) working group and the Data Usage Metrics working group. The Scholix working group developed a framework that allows organizations to expose and discover links between articles and datasets, thereby providing an indication of data citations. The Data Usage Metrics group works on a standard for the measurement and display of Data Usage Metrics. Here we explain how publishers and data repositories can contribute to and benefit from these initiatives. Together, these contributions feed into several hubs that enable data repositories to start displaying DLMs. Once these DLMs are available, researchers are in a better position to make their data count and be rewarded for their work.

Cousijn, Helena, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, and Tim Clark. “A Data Citation Roadmap for Scientific Publishers.” *Scientific Data* 5, no. 180259 (2018). https://doi.org/10.1038/SDATA.2018.259.

This article presents a practical roadmap for scholarly publishers to implement data citation in accordance with the Joint Declaration of Data Citation Principles (JDDCP), a synopsis and harmonization of the recommendations of major science policy bodies. It was developed by the Publishers Early Adopters Expert Group as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH BioCADDIE program. The structure of the roadmap presented here follows the“life of a paper”workflow and includes the categories Pre-submission, Submission, Production, and Publication. The roadmap is intended to be publisher-agnostic so that all publishers can use this as a starting point when implementing JDDCP-compliant data citation. Authors reading this roadmap will also better know what to expect from publishers and how to enable their own data citations to gain maximum impact, as well as complying with what will become increasingly common funder mandates on data transparency.

Couture, Jessica L., Rachael E. Blake, Gavin McDonald, and Colette L. Ward. “A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing.” *PLOS ONE* 13, no.7 (2018): e0199789. https://doi.org/10.1371/journal.pone.0199789

Growth of the open science movement has drawn significant attention to data sharing and availability across the scientific community. In this study, we tested the ability to recover data collected under a particular funder-imposed requirement of public availability. We assessed overall data recovery success, tested whether

characteristics of the data or data creator were indicators of recovery success, and identified hurdles to data recovery. Overall the majority of data were not recovered (26% recovery of 315 data projects), a similar result to journal-driven efforts to recover data. Field of research was the most important indicator of recovery success, but neither home agency sector nor age of data were determinants of recovery. While we did not find a relationship between recovery of data and age of data, age did predict whether we could find contact information for the grantee. The main hurdles to data recovery included those associated with communication with the researcher; loss of contact with the data creator accounted for half (50%) of unrecoverable datasets, and unavailability of contact information accounted for 35% of unrecoverable datasets. Overall, our results suggest that funding agencies and journals face similar challenges to enforcement of data requirements. We advocate that funding agencies could improve the availability of the data they fund by dedicating more resources to enforcing compliance with data requirements, providing data-sharing tools and technical support to awardees, and administering stricter consequences for those who ignore data sharing preconditions.

Crosas, Mercè. “The Evolution of Data Citation: From Principles to Implementation.” *Journal of eScience Librarianship* 37, no. 1-4 (2013): 62-70. https://doi.org/10.29173/iq504

Dallmeier-Tiessen, Suenje, Varsha Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, and Angus Whyte. “Connecting Data Publication to the Research Workflow: A Preliminary Analysis.” *International Journal of Digital Curation* 12, no. 1 (2017): 88-105. https://doi.org/10.2218/ijdc.v12i1.533

The data curation community has long encouraged researchers to document collected research data during active stages of the research workflow, to provide robust metadata earlier, and support research data publication and preservation. Data documentation with robust metadata is one of a number of steps in effective data publication. Data publication is the process of making digital research objects ’FAIR’, i.e. findable, accessible, interoperable, and reusable; attributes increasingly expected by research communities, funders and society. Research data publishing workflows are the means to that end. Currently, however, much published research data remains inconsistently and inadequately documented by researchers. Documentation of data closer in time to data collection would help mitigate the high cost that repositories associate with the ingest process. More effective data publication and sharing should in principle result from early interactions between researchers and their selected data repository. This paper describes a short study undertaken by members of the Research Data Alliance (RDA) and World Data System (WDS) working group on Publishing Data Workflows. We present a collection of recent examples of data publication workflows that connect data repositories and publishing platforms with research activity ‘upstream’ of the ingest process. We re-articulate previous recommendations of the working group, to account for the varied upstream service components and platforms that support the flow of contextual and provenance information downstream. These workflows should be open and loosely coupled to support interoperability, including with preservation and publication environments. Our recommendations aim to stimulate further work on researchers’ views of data publishing and the extent to which available services and infrastructure facilitate the publication of FAIR data. We also aim to stimulate further dialogue about, and definition of, the roles and responsibilities of research data services and platform providers for the ’FAIRness’ of research data publication workflows themselves.

Dearborn, Carly C., Amy J. Barto, and Neal A. Harmeyer. “The Purdue University Research Repository: HUBzero Customization for Dataset Publication and Digital Preservation.” *OCLC Systems & Services: International Digital Llibrary Perspectives* 30, no. 1 (2014): 15-27. https://doi.org/10.1108/oclc-07-2013-0022

Dearborn, Dylanne, Steve Marks, and Leanne Trimble. “The Changing Influence of Journal Data Sharing Policies on Local RDM Practices.” *International Journal of Digital Curation* 12, no. 2 (2017): 376-389. https://doi.org/10.2218/ijdc.v12i2.583

The purpose of this study was to examine changes in research data deposit policies of highly ranked journals in the physical and applied sciences between 2014 and 2016, as well as to develop an approach to examining the institutional impact of deposit requirements. Policies from the top ten journals (ranked by impact factor from the Journal Citation Reports) were examined in 2014 and again in 2016 in order to determine if data deposits were required or recommended, and which methods of deposit were listed as options. For all 2016 journals with a required data deposit policy, publication information (2009-2015) for the University of Toronto was pulled from Scopus and departmental affiliation was determined for each article. The results showed that the number of high-impact journals in the physical and applied sciences requiring data deposit is growing. In 2014, 71.2% of journals had no policy, 14.7% had a recommended policy, and 13.9% had a required policy (n=836). In contrast, in 2016, there were 58.5% with no policy, 19.4% with a recommended policy, and 22.0% with a required policy (n=880). It was also evident that U of T chemistry researchers are by far the most heavily affected by these journal data deposit requirements, having published 543 publications, representing 32.7% of all publications in the titles requiring data deposit in 2016. The Python scripts used to retrieve institutional publications based on a list of ISSNs have been released on GitHub so that other institutions can conduct similar research.

Delikoura, Eirini, and Dimitrios Kouis. “Open Research Data and Open Peer Review: Perceptions of a Medical and Health Sciences Community in Greece.” *Publications* 9, no. 2 (2021): 14. https://doi.org/10.3390/PUBLICATIONS9020014.

Recently significant initiatives have been launched for the dissemination of Open Access as part of the Open Science movement. Nevertheless, two other major pillars of Open Science such as Open Research Data (ORD) and Open Peer Review (OPR) are still in an early stage of development among the communities of researchers and stakeholders. The present study sought to unveil the perceptions of a medical and health sciences community about these issues. Through the investigation of researchers’ attitudes, valuable conclusions can be drawn, especially in the field of medicine and health sciences, where an explosive growth of scientific publishing exists. A quantitative survey was conducted based on a structured questionnaire, with 179 valid responses. The participants in the survey agreed with the Open Peer Review principles. However, they ignored basic terms like FAIR (Findable, Accessible, Interoperable, and Reusable) and appeared incentivized to permit the exploitation of their data. Regarding Open Peer Review (OPR), participants expressed their agreement, implying their support for a trustworthy evaluation system. Conclusively, researchers need to receive proper training for both Open Research Data principles and Open Peer Review processes which combined with a reformed evaluation system will enable them to take full advantage of the opportunities that arise from the new scholarly publishing and communication landscape.

Dosch, Brianne, and Tyler Martindale. “Reading the Fine Print: A Review and Analysis of Business Journals’ Data Sharing Policies.” *Journal of Business & Finance Librarianship* 25, no. 3-4 (2020): 261-280. https://doi.org/10.1080/08963568.2020.1847549

Drachen, Thea Marie, Ole Ellegaard, Asger Væring Larsen, and Søren Bertil Fabricius Dorch. “Sharing Data Increases Citations.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 26, no. 2 (2016): 67-82. https://doi.org/10.18352/lq.10149

This paper presents some indications to the existence of a citation advantage related to sharing data using astrophysics as a case. Through bibliometric analyses we find a citation advantage for astrophysical papers in core journals. The advantage arises as indexed papers are associated with data by bibliographical links, and consists of papers receiving on average significantly more citations per paper per year, than do papers not associated with links to data.

Fenner, Martin, Merce Crosas, Jeffrey S. Grethe, David N. Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, and Tim Clark. “A Data Citation Roadmap for Scholarly Data Repositories.” *Scientific Data* 6, no. 1 (2019): 28-28. https://doi.org/10.1038/s41597-019-0031-8

This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles, a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Expert Group, as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH-funded BioCADDIE (https://biocaddie.org) project. The roadmap makes 11 specific recommendations, grouped into three phases of implementation: a) required steps needed to support the Joint Declaration of Data Citation Principles, b) recommended steps that facilitate article/data publication workflows, and c) optional steps that further improve data citation support provided by data repositories. We describe the early adoption of these recommendations 18 months after they have first been published, looking specifically at implementations of machine-readable metadata on dataset landing pages.

Fenner, Martin, Laurel L. Haak, Gudmundur A. Thorisson, Sergio Ruiz, and Jan Brase. “ODIN: The ORCID and DataCite Interoperability Network.” *International Journal of Knowledge and Learning* 9, no. 4 (2015): 305-325. https://doi.org/10.1504/ijkl.2014.069537

Force, Megan M., and Daniel M. Auld. “Data Citation Index: Promoting Attribution, Use and Discovery of Research Data.” *Information Services & Use* 34 (2014): 97-98. https://doi.org/10.3233/ISU-140737.

Goldstein, Justin C., Matthew S. Mayernik, and Hampapuram K. Ramapriyan. “Identifiers for Earth Science Data Sets: Where We Have Been and Where We Need to Go.” *Data Science Journal* 16 (2017): p.23. http://doi.org/10.5334/dsj-2017-023

Considerable attention has been devoted to the use of persistent identifiers for assets of interest to scientific and other communities alike over the last two decades. Among persistent identifiers, Digital Object Identifiers (DOIs) stand out quite prominently, with approximately 133 million DOIs assigned to various objects as of February 2017. While the assignment of DOIs to objects such as scientific publications has been in place for many years, their assignment to Earth science data sets is more recent. Applying persistent identifiers to data sets enables improved tracking of their use and reuse, facilitates the crediting of data producers, and aids reproducibility through associating research with the exact data set(s) used. Maintaining provenance —i.e., tracing back lineage of significant scientific conclusions to the entities (data sets, algorithms, instruments, satellites, etc.) that lead to the conclusions, would be prohibitive without persistent identifiers. This paper provides a brief background on the use of persistent identifiers in general within the US, and DOIs more specifically. We examine their recent use for Earth science data sets, and outline successes and some remaining challenges. Among the challenges, for example, is the ability to conveniently and consistently obtain data citation statistics using the DOIs assigned by organizations that manage data sets.

Gorman, Dennis M. “Availability of Research Data in High-Impact Addiction Journals with Data Sharing Policies.” *Science and Engineering Ethics* 26, no. 3 (2020): 1625-1632. https://doi.org/10.1007/s11948-020-00203-7

Grant, Rebecca, and Iain Hrynaszkiewicz. “The Impact on Authors and Editors of Introducing Data Availability Statements at Nature Journals.” *Journal of Digital Curation* 13, no. 1 (2018): 195-203. https://doi.org/10.2218/ijdc.v13i1.614

This article describes the adoption of a standard policy for the inclusion of data availability statements in all research articles published at the Nature family of journals, and the subsequent research which assessed the impacts that these

policies had on authors, editors, and the availability of datasets. The key findings of this research project include the determination of average and median times required to add a data availability statement to an article; and a correlation between the way researchers make their data available, and the time required to add a data availability statement.

Grant, Rebecca, Graham Smith, and Iain Hrynaszkiewicz. “Assessing Metadata and Curation Quality: A Case Study from the Development of a Third-Party Curation Service at Springer Nature.” *International Journal of Digital Curation* 14, no. 1 (2020): 238-249. https://doi.org/10.2218/ijdc.v14i1.599

Since 2017, the publisher Springer Nature has provided an optional Research Data Support service to help researchers deposit and curate data that support their peer-reviewed publications. This service builds on a Research Data Helpdesk, which since 2016 has provided support to authors and editors who need advice on the options available for sharing their research data. In this paper, we describe a short project which aimed to facilitate an objective assessment of metadata quality, undertaken during the development of a third-party curation service for researchers (Research Data Support). We provide details on the single-blind user-testing that was undertaken, and the results gathered during this experiment. We also briefly describe the curation services which have been developed and introduced following an initial period of testing and piloting.

Griffiths, Aaron. “The Publication of Research Data: Researcher Attitudes and Behaviour.” *International Journal of Digital Curation* 4, no. 1 (2009): 46-56. https://doi.org/10.2218/ijdc.v4i1.77

Grootveld, Marjan, and Jeff van Egmond. “Peer-Reviewed Open Research Data: Results of a Pilot.” *International Journal of Digital Curation* 7, no. 2 (2012): 81-91. https://doi.org/10.2218/ijdc.v7i2.231

Peer review of publications is at the core of science and primarily seen as instrument for ensuring research quality. However, it is less common to independently value the quality of the underlying data as well. In the light of the ’data deluge’ it makes sense to extend peer review to the data itself and this way evaluate the degree to which the data are fit for re-use. This paper describes a pilot study at EASY—the electronic archive for (open) research data at our institution. In EASY, researchers can archive their data and add metadata themselves. Devoted to open access and data sharing, at the archive we are interested in further enriching these metadata with peer reviews.

As a pilot, we established a workflow where researchers who have downloaded data sets from the archive were asked to review the downloaded data set. This paper describes the details of the pilot including the findings, both quantitative and qualitative. Finally, we discuss issues that need to be solved when such a pilot is turned into a structural peer review functionality for the archiving system.

Groth, Paul, Helena Cousijn, Tim Clark, and Carole A. Goble. “FAIR Data Reuse—The Path through Data Citation.” *Data Intelligence* 2, no. 1-2 (2020): 78-86. https://doi.org/10.1162/dint_a_00030.

One of the key goals of the FAIR guiding principles is defined by its final principle—to optimize data sets for reuse by both humans and machines. To do so, data providers need to implement and support consistent machine readable metadata to describe their data sets. This can seem like a daunting task for data providers, whether it is determining what level of detail should be provided in the provenance metadata or figuring out what common shared vocabularies should be used. Additionally, for existing data sets it is often unclear what steps should be taken to enable maximal, appropriate reuse. Data citation already plays an important role in making data findable and accessible, providing persistent and unique identifiers plus metadata on

over 16 million data sets. In this paper, we discuss how data citation and its underlying infrastructures, in particular associated metadata, provide an important pathway for enabling FAIR data reuse.

He, Lin, and Vinita Nahar. “Reuse of Scientific Data in Academic Publications.” *Aslib Journal of Information Management* 68, no. 4 (2016): 478-494. https://doi.org/10.1108/ajim-01-2016-0008

Helbig, Kerstin, Brigitte Hausstein, and Ralf Toepfer. “Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences.” *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015): eP1220. http://doi.org/10.7710/2162-3309.1220

INTRODUCTION As more and more research data becomes better and more easily available, data citation gains in importance. The management of research data has been high on the agenda in academia for more than five years. Nevertheless, not all data policies include data citation, and problems like versioning and granularity remain. SERVICE DESCRIPTION da|ra operates as an allocation agency for DataCite and offers the registration service for social and economic research data in Germany. The service is jointly run by GESIS and ZBW, thereby merging experiences on the fields of Social Sciences and Economics. The authors answer questions pertaining to the most frequent aspects of research data registration like versioning and granularity as well as recommend the use of persistent identifiers linked with enriched metadata at the landing page. NEXT STEPS The promotion of data sharing and the development of a citation culture among the scientific community are future challenges. Interoperability becomes increasingly important for publishers and infrastructure providers. The already existent heterogeneity of services demands solutions for better user guidance. Building information competence is an asset of libraries, which can and should be expanded to research data.

Herterich, Patricia, and Sünje Dallmeier-Tiessen. “Data Citation Services in the High-Energy Physics Community.” *D-Lib Magazine* 22, no. 1/2 (2016). http://www.dlib.org/dlib/january16/herterich/01herterich.html

Holt, Jade, Andrew Walker, and Phill Jones. “Introducing a Data Availability Policy for Journals at IOP Publishing: Measuring the Impact on Authors and Editorial Teams.” *Learned Publishing* 34, no. 4 (2021): 478-486. https://doi.org/https://doi.org/10.1002/leap.1386

Horsburgh, Jeffery S., Richard P. Hooper, Jerad Bales, Margaret Hedstrom, Heidi J. Imker, Kerstin A. Lehnert, Lea A. Shanley, and Shelley Stall. “Assessing the State of Research Data Publication in Hydrology: A Perspective from the Consortium of Universities for the Advancement of Hydrologic Science, Incorporated.” *WIREs Water* 7, no. 3 (2020): e1422. https://doi.org/https://doi.org/10.1002/wat2.1422

Hrynaszkiewicz, Iain, Aliaksandr Birukou, Mathias Astell, Sowmya Swaminathan, Amye Kenall, and Varsha Khodiyar. “Standardising and Harmonising Research Data Policy in Scholarly Publishing.” *International Journal of Digital Curation* 12, no. 1 (2017): 65-71. https://doi.org/10.2218/ijdc.v12i1.531

To address the complexities researchers face during publication, and the potential community-wide benefits of wider adoption of clear data policies, the publisher Springer Nature has developed a standardised, common framework for the research data policies of all its journals. An expert working group was convened to audit and identify common features of research data policies of the journals published by Springer Nature, where policies were present. The group then consulted with approximately 30 editors, covering all research disciplines within the organisation. The group also consulted with academic editors, librarians and funders, which informed development of the framework and the creation of supporting resources. Four types of data policy were defined in recognition that some journals and research

communities are more ready than others to adopt strong data policies. As of January 2017 more than 700 journals have adopted a standard policy and this number is growing weekly. To potentially enable standardisation and harmonisation of data policy across funders, institutions, repositories, societies and other publishers, the policy framework was made available under a Creative Commons license. However, the framework requires wider debate with these stakeholders and an Interest Group within the Research Data Alliance (RDA) has been formed to initiate this process.

Hrynaszkiewicz, Iain, Natasha Simons, Azhar Hussain, Rebecca Grant, and Simon Goudie. “Developing a Research Data Policy Framework for All Journals and Publishers.” *Data Science Journal* 19, no. 1 (2020): p.5. http://doi.org/10.5334/dsj-2020-005

More journals and publishers—and funding agencies and institutions—are introducing research data policies. But as the prevalence of policies increases, there is potential to confuse researchers and support staff with numerous or conflicting policy requirements. We define and describe 14 features of journal research data policies and arrange these into a set of six standard policy types or tiers, which can be adopted by journals and publishers to promote data sharing in a way that encourages good practice and is appropriate for their audience's perceived needs. Policy features include coverage of topics such as data citation, data repositories, data availability statements, data standards and formats, and peer review of research data. These policy features and types have been created by reviewing the policies of multiple scholarly publishers, which collectively publish more than 10,000 journals, and through discussions and consensus building with multiple stakeholders in research data policy via the Data Policy Standardisation and Implementation Interest Group of the Research Data Alliance. Implementation guidelines for the standard research data policies for journals and publishers are also provided, along with template policy texts which can be implemented by journals in their Information for Authors and publishing workflows. We conclude with a call for collaboration across the scholarly publishing and wider research community to drive further implementation and adoption of consistent research data policies.

Hunter, Jane. “Scientific Publication Packages—A Selective Approach to the Communication and Archival of Scientific Output.” *International Journal of Digital Curation* 1, no. 1 (2006): 33-52. https://doi.org/10.2218/ijdc.v1i1.4

The use of digital technologies within research has led to a proliferation of data, many new forms of research output and new modes of presentation and analysis. Many scientific communities are struggling with the challenge of how to manage the terabytes of data and new forms of output, they are producing. They are also under increasing pressure from funding organizations to publish their raw data, in addition to their traditional publications, in open archives. In this paper I describe an approach that involves the selective encapsulation of raw data, derived products, algorithms, software and textual publications within “scientific publication packages.” Such packages provide an ideal method for: encapsulating expert knowledge; for publishing and sharing scientific process and results; for teaching complex scientific concepts; and for the selective archival, curation and preservation of scientific data and output. They also provide a bridge between technological advances in the Digital Libraries and eScience domains. In particular, I describe the RDF-based architecture that we are adopting to enable scientists to construct, publish and manage “scientific publication packages”—compound digital objects that encapsulate and relate the raw data to its derived products, publications and the associated contextual, provenance and administrative metadata.

Jackson, Brian. “Open Data Policies among Library and Information Science Journals.” *Publications* 9, no. 2 (2021): 25. https://www.mdpi.com/2304-6775/9/2/25

Journal publishers play an important role in the open research data ecosystem. Through open data policies that include public data archiving mandates and data availability statements, journal publishers help promote transparency in research and

wider access to a growing scholarly record. The library and information science (LIS) discipline has a unique relationship with both open data initiatives and academic publishing and may be well-positioned to adopt rigorous open data policies. This study examines the information provided on public-facing websites of LIS journals in order to describe the extent, and nature, of open data guidance provided to prospective authors. Open access journals in the discipline have disproportionately adopted detailed, strict open data policies. Commercial publishers, which account for the largest share of publishing in the discipline, have largely adopted weaker policies. Rigorous policies, adopted by a minority of journals, describe the rationale, application, and expectations for open research data, while most journals that provide guidance on the matter use hesitant and vague language. Recommendations are provided for strengthening journal open data policies.

Jeong, Geum Hee. “Status of the Data Sharing Policies of Scholarly Journals Published in Brazil, France, and Korea and Listed in Both the 2018 Scimago Journal and Country Ranking and the Web of Science.” *Science Editing* 7, no. 2 (2020): 136-141. https://doi.org/10.6087/kcse.208

Purpose

The present study analyzed the current status of the data sharing policies of journals published in Brazil, France, and Korea that were listed in the 2018 Scimago Journal and Country Ranking and Web of Science Core Collection.

Methods

Web of Science journals were selected from the 2018 Scimago Journal and Country Ranking. The homepages of all target journals were searched for the presence of statements on data sharing policies, including clinical trial data sharing policies, the level of the policies, and actual statements of data availability in articles.

Results

Out of 565 journals from these three countries, 118 (20.9%) had an optional data sharing policy, and one had a mandatory data sharing policy. Harvard Dataverse was the repository of one journal. The number of journals that had adopted a data sharing policy was 11 (6.7%) for Brazil, 64 (27.6%) for France, and 44 (25.9%) for Korea. One journal from Brazil and 20 journals from Korea had adopted clinical trial data sharing policies in accordance with the International Committee of Medical Journal Editors. Statements of data sharing were found in articles from two journals.

Conclusion

Journals from France and Korea adopted data sharing policies more actively than those from Brazil. However, the actual implementation of these policies through descriptions of data availability in articles remains rare. In many journals that appear to have data sharing policies, those policies may just reflect a standard description by the publisher, especially in France. Actual data sharing was not found to be frequent.

Johnson, Jeremiah N., Keith A. Hanson, Caleb A. Jones, Ramesh Grandhi, Jaime Guerrero, and Jesse S. Rodriguez. “Data Sharing in Neurosurgery and Neurology Journals.” *Cureus* 10, no. 5 (2018): e2680. https://dx.doi.org/10.7759/cureus.2680

Jones, Catherine Mary, Brian Matthews, Ian Gent, Tom Griffin, and Jonathan Tedds. “Persistent Identification and Citation of Software.” *International Journal of Digital Curation* 11, no. 2 (2017): 104-114. https://doi.org/10.2218/IJDC.V11I2.422.

Software underpins the academic research process across disciplines. To be able to understand, use/reuse and preserve data, the software code that generated,

analysed or presented the data will need to be retained and executed. An important part of this process is being able to persistently identify the software concerned. This paper discusses the reasons for doing so and introduces a model of software entities to enable better identification of what is being identified.

The DataCite metadata schema provides a persistent identification scheme and we consider how this scheme can be applied to software. We then explore examples of persistent identification and reuse. The examples show the differences and similarities of software used in academic research, which has been written and reused at different scales. The key concepts of being able to identify what precisely is being used and provide a mechanism for appropriate credit are important to both of them.

Kansa, Eric C., Sarah Whitcher Kansa, and Benjamin Arbuckle. “Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology.” *International Journal of Digital Curation* 9, no. 1 (2014): 57-70. https://doi.org/10.2218/ijdc.v9i1.301

We present a case study of data integration and reuse involving 12 researchers who published datasets in Open Context, an online data publishing platform, as part of collaborative archaeological research on early domesticated animals in Anatolia. Our discussion reports on how different editorial and collaborative review processes improved data documentation and quality, and created ontology annotations needed for comparative analyses by domain specialists. To prepare data for shared analysis, this project adapted editor-supervised review and revision processes familiar to conventional publishing, as well as more novel models of revision adapted from open source software development of public version control. Preparing the datasets for publication and analysis required significant investment of effort and expertise, including archaeological domain knowledge and familiarity with key ontologies. To organize this work effectively, we emphasized these different models of collaboration at various stages of this data publication and analysis project. Collaboration first centered on data editors working with data contributors, then widened to include other researchers who provided additional peer-review feedback, and finally the widest research community, whose collaboration is facilitated by GitHub’s version control system. We demonstrate that the “publish” and “push” models of data dissemination need not be mutually exclusive; on the contrary, they can play complementary roles in sharing high quality data in support of research. This work highlights the value of combining multiple models in different stages of data dissemination.

Khan, Nushrat, Mike Thelwall, and Kayvan Kousha. “Measuring the Impact of Biodiversity Datasets: Data Reuse, Citations and Altmetrics.” *Scientometrics* 126, no. 4 (2021): 3621-3639. https://doi.org/10.1007/s11192-021-03890-6

Kim, Jihyun. “An Analysis of Data Paper Templates and Guidelines: Types of Contextual Information Described by Data Journals.” *Science Editing* 7, no. 1 (2020): 16-23. https://doi.org/10.6087/kcse.185 https://doi.org/10.6087/kcse.185

Purpose

Data papers are a promising genre of scholarly communication, in which research data are described, shared, and published. Rich documentation of data, including adequate contextual information, enhances the potential of data reuse. This study investigated the extent to which the components of data papers specified by journals represented the types of contextual information necessary for data reuse.

Methods

A content analysis of 15 data paper templates/guidelines from 24 data journals indexed by the Web of Science was performed. A coding scheme was developed

based on previous studies, consisting of four categories: general data set properties, data production information, repository information, and reuse information.

Results

Only a few types of contextual information were commonly requested by the journals. Except data format information and file names, general data set properties were specified less often than other categories of contextual information. Researchers were frequently asked to provide data production information, such as information on the data collection, data producer, and related project. Repository information focused on data identifiers, while information about repository reputation and curation practices was rarely requested. Reuse information mostly involved advice on the reuse of data and terms of use.

Conclusion

These findings imply that data journals should provide a more standardized set of data paper components to inform reusers of relevant contextual information in a consistent manner. Information about repository reputation and curation could also be provided by data journals to complement the repository information provided by the authors of data papers and to help researchers evaluate the reusability of data.

Kim, Jihyun, Soon Kim, Hye-Min Cho, Jae Hwa Chang, and Soo Young Kim. “Data Sharing Policies of Journals in Life, Health, and Physical Sciences Indexed in Journal Citation Reports.” *PeerJ* 8 (2020): e9924 https://doi.org/10.7717/peerj.9924

Background

Many scholarly journals have established their own data-related policies, which specify their enforcement of data sharing, the types of data to be submitted, and their procedures for making data available. However, except for the journal impact factor and the subject area, the factors associated with the overall strength of the data sharing policies of scholarly journals remain unknown. This study examines how factors, including impact factor, subject area, type of journal publisher, and geographical location of the publisher are related to the strength of the data sharing policy.

Methods

From each of the 178 categories of the Web of Science’s 2017 edition of Journal Citation Reports, the top journals in each quartile (Q1, Q2, Q3, and Q4) were selected in December 2018. Of the resulting 709 journals (5%), 700 in the fields of life, health, and physical sciences were selected for analysis. Four of the authors independently reviewed the results of the journal website searches, categorized the journals’ data sharing policies, and extracted the characteristics of individual journals. Univariable multinomial logistic regression analyses were initially conducted to determine whether there was a relationship between each factor and the strength of the data sharing policy. Based on the univariable analyses, a multivariable model was performed to further investigate the factors related to the presence and/or strength of the policy.

Results

Of the 700 journals, 308 (44.0%) had no data sharing policy, 125 (17.9%) had a weak policy, and 267 (38.1%) had a strong policy (expecting or mandating data sharing). The impact factor quartile was positively associated with the strength of the data sharing policies. Physical science journals were less likely to have a strong policy relative to a weak policy than Life science journals (relative risk ratio [RRR], 0.36; 95% CI [0.17-0.78]). Life science journals had a greater probability of having a weak

policy relative to no policy than health science journals (RRR, 2.73; 95% CI [1.05-7.14]). Commercial publishers were more likely to have a weak policy relative to no policy than non-commercial publishers (RRR, 7.87; 95% CI, [3.98-15.57]). Journals by publishers in Europe, including the majority of those located in the United Kingdom and the Netherlands, were more likely to have a strong data sharing policy than a weak policy (RRR, 2.99; 95% CI [1.85-4.81]).

Conclusions

These findings may account for the increase in commercial publishers' engagement in data sharing and indicate that European national initiatives that encourage and mandate data sharing may influence the presence of a strong policy in the associated journals. Future research needs to explore the factors associated with varied degrees in the strength of a data sharing policy as well as more diverse characteristics of journals related to the policy strength.

Klump, Jens, Roland Bertelmann, Jan Brase, Michael Diepenbroek, Hannes Grobe, Heinke Höck, Michael Lautenschlager, Uwe Schindler, Irina Sens, and Joachim Wächter. "Data Publication in the Open Access Initiative." *Data Science Journal*. p.5 (2006): 79-83. https://datascience.codata.org/articles/abstract/463/

The 'Berlin Declaration' was published in 2003 as a guideline to policy makers to promote the Internet as a functional instrument for a global scientific knowledge base. Because knowledge is derived from data, the principles of the 'Berlin Declaration' should apply to data as well. Today, access to scientific data is hampered by structural deficits in the publication process. Data publication needs to offer authors an incentive to publish data through long-term repositories. Data publication also requires an adequate licence model that protects the intellectual property rights of the author while allowing further use of the data by the scientific community.

Konkiel, Stacy. "Tracking Citations and Altmetrics for Research Data: Challenges and Opportunities." *Bulletin of the American Society for Information Science and Technology* 39, no. 6 (2013): 27-32. https://doi.org/10.1002/bult.2013.1720390610

Kratz, John, and Carly Strasser. "Data Publication Consensus and Controversies." *F1000Research* 3 (2014): 94. https://doi.org/10.12688/f1000research.3979.3

The movement to bring datasets into the scholarly record as first class research products (validated, preserved, cited, and credited) has been inching forward for some time, but now the pace is quickening. As data publication venues proliferate, significant debate continues over formats, processes, and terminology. Here, we present an overview of data publication initiatives underway and the current conversation, highlighting points of consensus and issues still in contention. Data publication implementations differ in a variety of factors, including the kind of documentation, the location of the documentation relative to the data, and how the data is validated. Publishers may present data as supplemental material to a journal article, with a descriptive "data paper," or independently. Complicating the situation, different initiatives and communities use the same terms to refer to distinct but overlapping concepts. For instance, the term published means that the data is publicly available and citable to virtually everyone, but it may or may not imply that the data has been peer-reviewed. In turn, what is meant by data peer review is far from defined; standards and processes encompass the full range employed in reviewing the literature, plus some novel variations. Basic data citation is a point of consensus, but the general agreement on the core elements of a dataset citation frays if the data is dynamic or part of a larger set. Even as data publication is being defined, some are looking past publication to other metaphors, notably"data as software,"for solutions to the more stubborn problems.

Kratz, John Ernest, and Carly Strasser. "Researcher Perspectives on Publication and Peer Review of Data." *PLOS ONE* 10, no. 2 (2015): e0123377.

https://doi.org/10.1371/JOURNAL.PONE.0117619

Data“publication”seeks to appropriate the prestige of authorship in the peer-reviewed literature to reward researchers who create useful and well-documented datasets. The scholarly communication community has embraced data publication as an incentive to document and share data. But, numerous new and ongoing experiments in implementation have not yet resolved what a data publication should be, when data should be peer-reviewed, or how data peer review should work. While researchers have been surveyed extensively regarding data management and sharing, their perceptions and expectations of data publication are largely unknown. To bring this important yet neglected perspective into the conversation, we surveyed ~ 250 researchers across the sciences and social sciences—asking what expectations“data publication”raises and what features would be useful to evaluate the trustworthiness, evaluate the impact, and enhance the prestige of a data publication. We found that researcher expectations of data publication center on availability, generally through an open database or repository. Few respondents expected published data to be peer-reviewed, but peer-reviewed data enjoyed much greater trust and prestige. The importance of adequate metadata was acknowledged, in that almost all respondents expected data peer review to include evaluation of the data's documentation. Formal citation in the reference list was affirmed by most respondents as the proper way to credit dataset creators. Citation count was viewed as the most useful measure of impact, but download count was seen as nearly as valuable. These results offer practical guidance for data publishers seeking to meet researcher expectations and enhance the value of published data.

Kwon, Hyun Jung, Yoon Joo Seo, Mi Yeon Kim, and Sue Yeon Chung. “Recommended Practices for Supplemental Data.” *Science Editing* 7, no. 1 (2020): 94-103. https://doi.org/10.6087/kcse.200

Since various forms of supplemental data (SD) have been introduced in academic publications, it has become necessary to establish guidelines to systematically process, indicate, and distribute such data. This material aims to help the science journals establish rational SD policies and guidelines and to ensure compliance with such policies and to manage them consistently. Generally, SD can be approached in a literal way by categorizing ‘appendices’ as ‘additional or separately added complementary materials’ and ‘supplements’ as ‘materials supplemental to the research in a comprehensive sense,’ rather than by viewing SD as an independent component of an article. The recommended practices of the National Information Standards Organization of USA advise the classification of SD into either ‘integral content’ or ‘additional content’ according to the content's functional relationship to the associated article. If a public depository is used for SD, the author can ensure the perpetuity of data accessibility by assigning a digital object identifier. Science journals should adopt appropriate SD policies and describe them in detail in the instructions for authors to ensure consistent compliance with those policies. Additionally, they should be able to inspect and maintain links, repositories, and metadata associated with the SD for specific articles on an ongoing basis.

Lammey, Rachael. “How Publishers Can Work with Crossref on Data Citation.” *Science Editing* 6, no. 2 (2019): 166-70. https://doi.org/10.6087/KCSE.165

It aims to explain why data citation is important, how publishers and data repositories can do this and what use will be made of the information they provide. There are large benefits to be accrued from sharing research data such as guarantee of reproducibility and transparency. Consistent citation practice around data is essential to helping these benefits to be realized. Data citation metadata is being disseminated and used through its application programming interfaces and the Event Data application programming interface. Event Data extracts this information into a separate service, so data citations are pre-filtered from the Crossref metadata. There are two methods by which publishers can register data citation information with Crossref. The first method is to deposit data citations in the citation section of the

metadata, i.e., the part containing the reference list of the article. The second method publishers can use to register data citations with Crossref is to use the relationships section of the metadata. There are a number of services already using Event Data to show information on data citation. To achieve the benefits of data citation, publishers or editors should have a data sharing and citation policy so that they share with their authors and readers.

Lawrence, Bryan, Catherine Jones, Brian Matthews, Sam Pepler, and Sarah Callaghan. “Citation and Peer Review of Data: Moving towards Formal Data Publication.” *International Journal of Digital Curation* 6, no. 2 (2011): 4-37. https://doi.org/10.2218/ijdc.v6i2.205

This paper discusses many of the issues associated with formally publishing data in academia, focusing primarily on the structures that need to be put in place for peer review and formal citation of datasets. Data publication is becoming increasingly important to the scientific community, as it will provide a mechanism for those who create data to receive academic credit for their work and will allow the conclusions arising from an analysis to be more readily verifiable, thus promoting transparency in the scientific process. Peer review of data will also provide a mechanism for ensuring the quality of datasets, and we provide suggestions on the types of activities one expects to see in the peer review of data. A simple taxonomy of data publication methodologies is presented and evaluated, and the paper concludes with a discussion of dataset granularity, transience and semantics, along with a recommended human-readable citation syntax.

Leadbetter, A., L. Raymond, C. Chandler, L. Pikula, P. Pissierssens, and E. Urban. *Ocean Data Publication Cookbook*. Oostende, Belgium: UNESCO, 2013. http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10574

Lee, Jungyeoun, and Jihyun Kim. “Korean Researchers’ Motivations for Publishing in Data Journals and the Usefulness of Their Data: A Qualitative Study.” *Science Editing* 8, no. 2 (2021): 145-152. https://doi.org/10.6087/kcse.246

Purpose

This study investigated the usefulness and limitations of data journals by analyzing motivations for submission, review and publication processes according to researchers with experience publishing in data journals.

Methods

Among 79 data journals indexed in Web of Science, we selected four data journals where data papers accounted for more than 20% of the publication volume and whose corresponding authors belonged to South Korean research institutes. A qualitative analysis was conducted of the subjective experiences of seven corresponding authors who agreed to participate in interviews. To analyze interview transcriptions, clusters were created by restructuring the theme nodes using Nvivo 12.

Results

The most important element of data journals to researchers was their usefulness for obtaining credit for research performance. Since the data in repositories linked to data papers are screened using journals’ review processes, the validity, accuracy, reusability, and reliability of data are ensured. In addition, data journals provide a basis for data sharing using repositories and data-centered follow-up research using citations and offer detailed descriptions of data.

Conclusion

Data journals play a leading role in data-centered research. Data papers are recognized as research achievements through citations in the same way as research papers published in conventional journals, but there was also a perception that it is difficult to attain a similar level of academic recognition with data papers as with research papers. However, researchers highly valued the usefulness of data journals, and data journals should thus be developed into new academic communication channels that enhance data sharing and reuse.

Li, Jiao, Si Zheng, Hongyu Kang, Zhen Hou, and Qing Qian. “Identifying Scientific Project-Generated Data Citation from Full-Text Articles: An Investigation of TCGA Data Citation.” *Journal of Data and Information Science* 1, no. 2 (2017): 32-44. https://doi.org/10.20309/JDIS.201612

Mathiak, Brigitte, and Katarina Boland. “Challenges in Matching Dataset Citation Strings to Datasets in Social Science.” *D-Lib Magazine* 21, no. 1/2 (2015). https://doi.org/10.1045/january2015-mathiak

Mayernik, Matthew S. “Data Citation Initiatives and Issues.” *Bulletin of the American Society for Information Science and Technology* 38, no. 5 (2012): 23-28. https://doi.org/10.1002/bult.2012.1720380508

Mayernik, Matthew S., Jennifer Phillips, and Eric Nienhouse. “Linking Publications and Data: Challenges, Trends, and Opportunities.” *D-Lib Magazine* 22, no. 5/6 (2016). https://doi.org/10.1045/may2016-mayernik

Mayo, Christine, Todd J. Vision, and Elizabeth A. Hull. “The Location of the Citation: Changing Practices in How Publications Cite Original Data in the Dryad Digital Repository.” *International Journal of Digital Curation* 11, no. 1 (2016): 150-155. https://doi.org/10.2218/ijdc.v11i1.400

While stakeholders in scholarly communication generally agree on the importance of data citation, there is not consensus on where those citations should be placed within the publication—particularly when the publication is citing original data. Recently, CrossRef and the Digital Curation Center (DCC) have recommended as a best practice that original data citations appear in the works cited sections of the article. In some fields, such as the life sciences, this contrasts with the common practice of only listing data identifier(s) within the article body (intratextually). We inquired whether data citation practice has been changing in light of the guidance from CrossRef and the DCC. We examined data citation practices from 2011 to 2014 in a corpus of 1,125 articles associated with original data in the Dryad Digital Repository. The percentage of articles that include no reference to the original data has declined each year, from 31% in 2011 to 15% in 2014. The percentage of articles that include data identifiers intratextually has grown from 69% to 83%, while the percentage that cite data in the works cited section has grown from 5% to 8%. If the proportions continue to grow at the current rate of 19-20% annually, the proportion of articles with data citations in the works cited section will not exceed 90% until 2030.

Missier, Paolo. “Data Trajectories: Tracking Reuse of Published Data for Transitive Credit Attribution.” *International Journal of Digital Curation* 11, no. 1 (2016): 1-16. https://doi.org/10.2218/ijdc.v11i1.425

The ability to measure the use and impact of published data sets is key to the success of the open data/open science paradigm. A direct measure of impact would require tracking data (re)use in the wild, which is difficult to achieve. This is therefore commonly replaced by simpler metrics based on data download and citation counts. In this paper we describe a scenario where it is possible to track the trajectory of a dataset after its publication, and show how this enables the design of accurate models for ascribing credit to data originators. A Data Trajectory (DT) is a graph that encodes knowledge of how, by whom, and in which context data has been re-used, possibly after several generations. We provide a theoretical model of DTs that is

grounded in the W3C PROV data model for provenance, and we show how DTs can be used to automatically propagate a fraction of the credit associated with transitively derived datasets, back to original data contributors. We also show this model of transitive credit in action by means of a Data Reuse Simulator. In the longer term, our ultimate hope is that credit models based on direct measures of data reuse will provide further incentives to data publication. We conclude by outlining a research agenda to address the hard questions of creating, collecting, and using DTs systematically across a large number of data reuse instances in the wild.

Mooney, Hailey, and Mark P. Newton. “The Anatomy of a Data Citation: Discovery, Reuse, and Credit.” *Journal of Librarianship and Scholarly Communication* 1, no. 1 (2012): p.eP1035. http://dx.doi.org/10.7710/2162-3309.1035

INTRODUCTION Data citation should be a necessary corollary of data publication and reuse. Many researchers are reluctant to share their data, yet they are increasingly encouraged to do just that. Reward structures must be in place to encourage data publication, and citation is the appropriate tool for scholarly acknowledgment. Data citation also allows for the identification, retrieval, replication, and verification of data underlying published studies. METHODS This study examines author behavior and sources of instruction in disciplinary and cultural norms for writing style and citation via a content analysis of journal articles, author instructions, style manuals, and data publishers. Instances of data citation are benchmarked against a Data Citation Adequacy Index. RESULTS Roughly half of journals point toward a style manual that addresses data citation, but the majority of journal articles failed to include an adequate citation to data used in secondary analysis studies. DISCUSSION Full citation of data is not currently a normative behavior in scholarly writing. Multiplicity of data types and lack of awareness regarding existing standards contribute to the problem. CONCLUSION Citations for data must be promoted as an essential component of data publication, sharing, and reuse. Despite confounding factors, librarians and information professionals are well-positioned and should persist in advancing data citation as a normative practice across domains. Doing so promotes a value proposition for data sharing and secondary research broadly, thereby accelerating the pace of scientific research.

Naudet, Florian, Charlotte Sakarovitch, Perrine Janiaud, Ioana Cristea, Daniele Fanelli, David Moher, and John P. A, Ioannidis. “Data Sharing and Reanalysis of Randomized Controlled Trials in Leading Biomedical Journals with a Full Data Sharing Policy: Survey of Studies Published in *the BMJ* and *PLOS Medicine*.” *BMJ* 360 (2018): k400. https://doi.org/10.1136/bmj.k400

Naughton, Linda, and David Kernohan. “Making Sense of Journal Research Data Policies.” *Insights* 29, no. 1 (2016): 84-89. http://doi.org/10.1629/uksg.284

This article gives an overview of the findings from the first phase of the Jisc Journal Research Data Policy Registry pilot (JRDPR), which is currently under way. The project continues from the initial study, ‘Journal of Research Data’ policy bank (JoRD), carried out by Nottingham University’s Centre for Research Communication from 2012 to 2014. The project undertook an analysis of 250 journal research data policies to assess the feasibility of developing a policy registry to assist researchers and support staff to comply with research data publication requirements. The evidence shows that the current research data policy ecosystem is in critical need of standardization and harmonization if such services are to be built and implemented. To this end, the article proposes the next steps for the project with the objective of ultimately moving towards a modern research infrastructure based on machine-readable policies that support a more open scholarly communications environment.

Novacescu, Jenny, Joshua E. G. Peek, Sarah Weissman, Scott W. Fleming, Karen Levay, and Elizabeth Fraser. “A Model for Data Citation in Astronomical Research Using Digital Object Identifiers (DOIs).” *Astrophysical Journal Supplement Series* 236, no. 1 (2018). https://doi.org/10.3847/1538-4365/AAB76A

Park, Hyoungjoo, and Dietmar Wolfram. “An Examination of Research Data Sharing and Re-Use: Implications for Data Citation Practice.” *Scientometrics* 111, no. 1 (2017): 443-461. https://doi.org/10.1007/s11192-017-2240-2

Park, Hyoungjoo, Sukjin You, and Dietmar Wolfram. “Informal Data Citation for Data Sharing and Reuse Is More Common than Formal Data Citation in Biomedical Fields.” *Journal of the Association for Information Science and Technology* 69, no. 11 (2018): 1346-1354. https://doi.org/10.1002/ASI.24049.

Parsons, M., and P. Fox. “Is Data Publication the Right Metaphor?” *Data Science Journal* 12 (2013): pp.WDS32-WDS46. https://datascience.codata.org/articles/abstract/63/

> International attention to scientific data continues to grow. Opportunities emerge to re-visit long-standing approaches to managing data and to critically examine new capabilities. We describe the cognitive importance of metaphor. We describe several metaphors for managing, sharing, and stewarding data and examine their strengths and weaknesses. We particularly question the applicability of a“publication”approach to making data broadly available. Our preliminary conclusions are that no one metaphor satisfies enough key data system attributes and that multiple metaphors need to co-exist in support of a healthy data ecosystem. We close with proposed research questions and a call for continued discussion.

Parsons, Mark A., Ruth E. Duerr, and Matthew B. Jones. “The History and Future of Data Citation in Practice.” *Data Science Journal* 18, no. 1 (2019): p.52. https://doi.org/10.5334/DSJ-2019-052

> In this review, we adopt the definition that 'Data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data' (TGDCSP 2013: CIDCR6). Furthermore, access should be enabled for both humans and machines (DCSG 2014). We use this to discuss how data citation has evolved over the last couple of decades and to highlight issues that need more research and attention.
>
> Data citation is not a new concept, but it has changed and evolved considerably since the beginning of the digital age. Basic practice is now established and slowly but increasingly being implemented. Nonetheless, critical issues remain. These issues are primarily because we try to address multiple human and computational concerns with a system originally designed in a non-digital world for more limited use cases. The community is beginning to challenge past assumptions, separate the multiple concerns (credit, access, reference, provenance, impact, etc.), and apply different approaches for different use cases.

Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. “Data Citation and Peer Review.” *Eos, Transactions American Geophysical Union* 91, no. 34 (2010): 297-298. https://doi.org/https://doi.org/10.1029/2010EO340001

Peer, Limor, and Stephanie Wykstra. “New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation.” *Journal of eScience Librarianship* 39, no. 4 (2015): 6-13. https://doi.org/10.29173/iq902

Pejša, Stanislav, Shirley J. Dyke, and Thomas J. Hacker. “Building Infrastructure for Preservation and Publication of Earthquake Engineering Research Data.” *International Journal of Digital Curation* 9, no. 2 (2014): 83-97. https://doi.org/10.2218/ijdc.v9i2.335

> The objective of this paper is to showcase the progress of the earthquake engineering community during a decade-long effort supported by the National Science Foundation in the George E. Brown Jr., Network for Earthquake Engineering Simulation (NEES). During the four years that NEES network operations have been headquartered at Purdue University, the NEEScomm management team has facilitated an unprecedented cultural change in the ways research is performed in earthquake engineering. NEES has not only played a major role in advancing the

cyberinfrastructure required for transformative engineering research, but NEES research outcomes are making an impact by contributing to safer structures throughout the USA and abroad. This paper reflects on some of the developments and initiatives that helped instil change in the ways that the earthquake engineering and tsunami community share and reuse data and collaborate in general.

Pepe, Alberto, Alyssa Goodman, August Muench, Merce Crosas, and Christopher Erdmann. “How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers.” *PLOS ONE* 9, no. 8 (2014): e104798. http://dx.doi.org/10.1371/journal.pone.0104798

We analyze data sharing practices of astronomers over the past fifteen years. An analysis of URL links embedded in papers published by the American Astronomical Society reveals that the total number of links included in the literature rose dramatically from 1997 until 2005, when it leveled off at around 1500 per year. The analysis also shows that the availability of linked material decays with time: in 2011, 44% of links published a decade earlier, in 2001, were broken. A rough analysis of link types reveals that links to data hosted on astronomers’ personal websites become unreachable much faster than links to datasets on curated institutional sites. To gauge astronomers’ current data sharing practices and preferences further, we performed in-depth interviews with 12 scientists and online surveys with 173 scientists, all at a large astrophysical research institute in the United States: the Harvard-Smithsonian Center for Astrophysics, in Cambridge, MA. Both the in-depth interviews and the online survey indicate that, in principle, there is no philosophical objection to data-sharing among astronomers at this institution. Key reasons that more data are not presently shared more efficiently in astronomy include: the difficulty of sharing large data sets; over reliance on non-robust, non-reproducible mechanisms for sharing data (e.g. emailing it); unfamiliarity with options that make data-sharing easier (faster) and/or more robust; and, lastly, a sense that other researchers would not want the data to be shared. We conclude with a short discussion of a new effort to implement an easy-to-use, robust, system for data sharing in astronomy, at theastrodata.org, and we analyze the uptake of that system to-date.

Peters, Isabella, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Gorraiz. “Research Data Explored: An Extended Analysis of Citations and Altmetrics.” *Scientometrics* 107, no. 2 (2016): 723-744. https://doi.org/10.1007/s11192-016-1887-4

Piwowar, Heather A. “Data Reuse and the Open Data Citation Advantage.” *PeerJ* 1, no. 1 (2013): e175. https://doi.org/10.7717/PEERJ.175.

Background. Attribution to the original contributor upon reuse of published data is important both as a reward for data creators and to document the provenance of research findings. Previous studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data. However, few previous analyses have had the statistical power to control for the many variables known to predict citation rate, which has led to uncertain estimates of the “citation benefit”, Furthermore, little is known about patterns in data reuse over time and across datasets.

Method and Results. Here, we look at citation rates while controlling for many known citation predictors and investigate the variability of data reuse. In a multivariate regression on 10,555 studies that created gene expression microarray data, we found that studies that made data available in a public repository received 9% (95% confidence interval: 5% to 13%) more citations than similar studies for which the data was not made available. Date of publication, journal impact factor, open access status, number of authors, first and last author publication history, corresponding author country, institution citation history, and study topic were included as covariates. The citation benefit varied with date of dataset deposition: a citation benefit was most clear for papers published in 2004 and 2005, at about 30%. Authors published most

papers using their own datasets within two years of their first publication on the dataset, whereas data reuse papers published by third-party investigators continued to accumulate for at least six years. To study patterns of data reuse directly, we compiled 9,724 instances of third party data reuse via mention of GEO or ArrayExpress accession numbers in the full text of papers. The level of third-party data use was high: for 100 datasets deposited in year 0, we estimated that 40 papers in PubMed reused a dataset by year 2, 100 by year 4, and more than 150 data reuse papers had been published by year 5. Data reuse was distributed across a broad base of datasets: a very conservative estimate found that 20% of the datasets deposited between 2003 and 2007 had been reused at least once by third parties.

Conclusion. After accounting for other factors affecting citation rate, we find a robust citation benefit from open data, although a smaller one than previously reported. We conclude there is a direct effect of third-party data reuse that persists for years beyond the time when researchers have published most of the papers reusing their own data. Other factors that may also contribute to the citation benefit are considered. We further conclude that, at least for gene expression microarray data, a substantial fraction of archived datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. “Sharing Detailed Research Data Is Associated with Increased Citation Rate.” *PLOS ONE* 2, no, (2007): e308. http://dx.doi.org/10.1371/journal.pone.0000308

Background

Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.

Principal Findings

We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly (p=0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

Significance

This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Pryor, Jennifer, Guy D. Eslick, Nicholas J. Talley, Kerith Duncanson, Simon Keely, and Emily C. Hoedt. “Clinical Medicine Journals Lag Behind Science Journals with Regards to ‘Microbiota Sequence’ Data Availability.” *Clinical and Translational Medicine* 11, no. 12 (2021): e656. https://doi.org/https://doi.org/10.1002/ctm2.656

Riedel, Nico, Miriam Kip, and Evgeny Bobrov. “Oddpub —A Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications.” *Data Science Journal* 19, no. 1 (2020): p.42. https://doi.org/10.5334/dsj-2020-042

Open research data are increasingly recognized as a quality indicator and an important resource to increase transparency, robustness and collaboration in science. However, no standardized way of reporting Open Data in publications exists, making it difficult to find shared datasets and assess the prevalence of Open Data in an automated fashion.

We developed ODDPub (Open Data Detection in Publications), a text-mining algorithm that screens biomedical publications and detects cases of Open Data. Using English-language original research publications from a single biomedical research institution (n = 8689) and randomly selected from PubMed (n = 1500) we iteratively developed a set of derived keyword categories. ODDPub can detect data sharing through field-specific repositories, general-purpose repositories or the supplement. Additionally, it can detect shared analysis code (Open Code).

To validate ODDPub, we manually screened 792 publications randomly selected from PubMed. On this validation dataset, our algorithm detected Open Data publications with a sensitivity of 0.73 and specificity of 0.97. Open Data was detected for 11.5% (n = 91) of publications. Open Code was detected for 1.4% (n = 11) of publications with a sensitivity of 0.73 and specificity of 1.00. We compared our results to the linked datasets found in the databases PubMed and Web of Science.

Our algorithm can automatically screen large numbers of publications for Open Data. It can thus be used to assess Open Data sharing rates on the level of subject areas, journals, or institutions. It can also identify individual Open Data publications in a larger publication corpus. ODDPub is published as an R package on GitHub.

Rousi, Antti M., and Mikael Laakso. “Journal Research Data Sharing Policies: A Study of Highly-Cited Journals in Neuroscience, Physics, and Operations Research.” *Scientometrics* 124, no. 1 (2020): 131-152. https://doi.org/10.1007/s11192-020-03467-9

The practices for if and how scholarly journals instruct research data for published research to be shared is an area where a lot of changes have been happening as science policy moves towards facilitating open science, and subject-specific repositories and practices are established. This study provides an analysis of the research data sharing policies of highly-cited journals in the fields of neuroscience, physics, and operations research as of May 2019. For these 120 journals, 40 journals per subject category, a unified policy coding framework was developed to capture the most central elements of each policy, i.e. what, when, and where research data is instructed to be shared. The results affirm that considerable differences between research fields remain when it comes to policy existence, strength, and specificity. The findings revealed that one of the most important factors influencing the dimensions of what, where and when of research data policies was whether the journal's scope included specific data types related to life sciences which have established methods of sharing through community-endorsed public repositories. The findings surface the future research potential of approaching policy analysis on the publisher-level as well as on the journal-level. The collected data and coding framework is provided as open data to facilitate future research and journal policy monitoring.

Rueda, Laura, Martin Fenner, and Patricia Cruse. “DataCite: Lessons Learned on Persistent Identifiers for Research Data.” *International Journal of Digital Curation* 11, no. 2 (2017): 39-47. https://doi.org/10.2218/ijdc.v11i2.421

Data are the infrastructure of science and they serve as the groundwork for scientific pursuits. Data publication has emerged as a game-changing breakthrough in scholarly communication. Data form the outputs of research but also are a gateway to new hypotheses, enabling new scientific insights and driving innovation. And yet stakeholders across the scholarly ecosystem, including practitioners, institutions, and funders of scientific research are increasingly concerned about the lack of sharing and reuse of research data. Across disciplines and countries, researchers, funders, and publishers are pushing for a more effective research environment, minimizing the duplication of work and maximizing the interaction between researchers. Availability, discoverability, and reproducibility of research outputs are key factors to support data reuse and make possible this new environment of highly collaborative research.

An interoperable e-infrastructure is imperative in order to develop new platforms and services for to data publication and reuse. DataCite has been working to establish and promote methods to locate, identify and share information about research data. Along with service development, DataCite supports and advocates for the standards behind persistent identifiers (in particular DOIs, Digital Object Identifiers) for data and other research outputs. Persistent identifiers allow different platforms to exchange information consistently and unambiguously and provide a reliable way to track citations and reuse. Because of this, data publication can become a reality from a technical standpoint, but the adoption of data publication and data citation as a practice by researchers is still in its early stages.

Since 2009, DataCite has been developing a series of tools and services to foster the adoption of data publication and citation among the research community. Through the years, DataCite has worked in a close collaboration with interdisciplinary partners on these issues and we have gained insight into the development of data publication workflows. This paper describes the types of different actions and the lessons learned by DataCite.

Schubert, Chris, Georg Seyerl, and Katharina Sack. “Dynamic Data Citation Service-Subset Tool for Operational Data Management.” *Data* 4, no. 3 (2019): 115. https://doi.org/10.3390/data4030115

In earth observation and climatological sciences, data and their data services grow on a daily basis in a large spatial extent due to the high coverage rate of satellite sensors, model calculations, but also by continuous meteorological in situ observations. In order to reuse such data, especially data fragments as well as their data services in a collaborative and reproducible manner by citing the origin source, data analysts, e.g., researchers or impact modelers, need a possibility to identify the exact version, precise time information, parameter, and names of the dataset used. A manual process would make the citation of data fragments as a subset of an entire dataset rather complex and imprecise to obtain. Data in climate research are in most cases multidimensional, structured grid data that can change partially over time. The citation of such evolving content requires the approach of “dynamic data citation”. The applied approach is based on associating queries with persistent identifiers. These queries contain the subsetting parameters, e.g., the spatial coordinates of the desired study area or the time frame with a start and end date, which are automatically included in the metadata of the newly generated subset and thus represent the information about the data history, the data provenance, which has to be established in data repository ecosystems. The Research Data Alliance Data Citation Working Group (RDA Data Citation WG) summarized the scientific status quo as well as the state of the art from existing citation and data management concepts and developed the scalable dynamic data citation methodology of evolving data. The Data Centre at the Climate Change Centre Austria (CCCA) has implemented the given recommendations and offers since 2017 an operational service on dynamic data citation on climate scenario data. With the consciousness that the objective of this topic brings a lot of dependencies on bibliographic citation research which is still under discussion, the CCCA service on Dynamic Data Citation focused on the climate domain specific issues, like characteristics of data, formats, software environment, and usage behavior. The current effort beyond spreading made experiences will be the scalability of the implementation, e.g., towards the potential of an Open Data Cube solution.

Seo, Sunkyung, and Jihyun Kim. “Data Journals: Types of Peer Review, Review Criteria, and Editorial Committee Members’ Positions.” *Science Editing* 7, no. 2 (2020): 130-135. https://doi.org/10.6087/kcse.207

Purpose

This study analyzed the peer review systems, criteria, and editorial committee structures of data journals, aiming to determine the current state of data peer review

and to offer suggestions.

Methods

We analyzed peer review systems and criteria for peer review in nine data journals indexed by Web of Science, as well as the positions of the editorial committee members of the journals. Each data journal's website was initially surveyed, and the editors-in-chief were queried via email about any information not found on the websites. The peer review criteria of the journals were analyzed in terms of data quality, metadata quality, and general quality.

Results

Seven of the nine data journals adopted single-blind and open review peer review methods. The remaining two implemented modified models, such as interactive and community review. In the peer review criteria, there was a shared emphasis on the appropriateness of data production methodology and detailed descriptions. The editorial committees of the journals tended to have subject editors or subject advisory boards, while a few journals included positions with the responsibility of evaluating the technical quality of data.

Conclusion

Creating a community of subject experts and securing various editorial positions for peer review are necessary for data journals to achieve data quality assurance and to promote reuse. New practices will emerge in terms of data peer review models, criteria, and editorial positions, and further research needs to be conducted.

Shaon, Arif, Sarah Callaghan, Bryan Lawrence, Brian Matthews, Timothy Osborn, Colin Harpham, and Andrew Woolf. “Opening Up Climate Research: A Linked Data Approach to Publishing Data Provenance.” *International Journal of Digital Curation* 7, no. 1 (2012): 163-173. https://doi.org/10.2218/ijdc.v7i1.223

Traditionally, the formal scientific output in most fields of natural science has been limited to peer-reviewed academic journal publications, with less attention paid to the chain of intermediate data results and their associated metadata, including provenance. In effect, this has constrained the representation and verification of the data provenance to the confines of the related publications. Detailed knowledge of a dataset’s provenance is essential to establish the pedigree of the data for its effective re-use, and to avoid redundant re-enactment of the experiment or computation involved. It is increasingly important for open-access data to determine their authenticity and quality, especially considering the growing volumes of datasets appearing in the public domain. To address these issues, we present an approach that combines the Digital Object Identifier (DOI)—a widely adopted citation technique—with existing, widely adopted climate science data standards to formally publish detailed provenance of a climate research dataset as an associated scientific workflow. This is integrated with linked-data compliant data re-use standards (e.g. OAI-ORE) to enable a seamless link between a publication and the complete trail of lineage of the corresponding dataset, including the dataset itself.

Shin, Nagai, Hideaki Shibata, Takeshi Osawa, Takehisa Yamakita, Masahiro Nakamura, and Tanaka Kenta. “Toward More Data Publication of Long-Term Ecological Observations.” *Ecological Research* 35, no. 5 (2020): 700-707. https://doi.org/https://doi.org/10.1111/1440-1703.12115

Data papers, such as those published by *Ecological Research*, encourage the retrieval and archiving of valuable unpublished, undigitized ecological observational data. However, scientists remain hesitant to submit their data to such forums. In this perspective paper, we describe lessons learned from the Long-Term Ecological

Research, the Global Biodiversity Information Facility and marine biological databases and discuss how data sharing and publication are both powerful and important for ecological research. Our aim is to encourage readers to submit their unpublished, undigitized ecological observational data then the data may be archived, published and used by other researchers to advance knowledge in the field of ecology. Coupling data sharing and syntheses with the development of innovative informatics would allow ecology to enter the realm of big science and provide seeds for a new and robust agenda of future ecological studies.

Sholler, Dan, Karthik Ram, Carl Boettiger, and Daniel S. Katz. “Enforcing Public Data Archiving Policies in Academic Publishing: A Study of Ecology Journals.” *Big Data & Society* 6, no. 1 (2019): 2053951719836258. https://doi.org/10.1177/2053951719836258

Siebert, Maximilian, Jeanne Fabiola Gaba, Laura Caquelin, Henri Gouraud, Alain Dupuy, David Moher, and Florian Naudet. “Data-Sharing Recommendations in Biomedical Journals and Randomised Controlled Trials: An Audit of Journals Following the ICMJE Recommendations.” *BMJ Open* 10, no. 5 (2020): e038887. https://doi.org/10.1136/bmjopen-2020-038887

Silvello, Gianmaria. “Theory and Practice of Data Citation.” *Journal of the Association for Information Science and Technology* 69, no. 1 (2018): 6-20. https://doi.org/10.1002/asi.23917

Simons, Natasha, Karen Visser, and Sam Searle. “Growing Institutional Support for Data Citation: Results of a Partnership Between Griffith University and the Australian National Data Service.” *D-Lib Magazine* 19, no. 11/12 (2013). https://doi.org/10.1045/november2013-simons

Smit, Eefke. “Eloise and Abelard: Why Data and Publications Belong Together.” *D-Lib Magazine* 17, no. 1/2 (2011). https://doi.org/10.1045/january2011-smit

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy Nurnberger, Stefan Proel, Andreas Rauber, Simone Sacchi, Arthur Smith, Mike Taylor, and Tim Clark. “Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications.” *PeerJ Computer Science* 1 (2015): e1. http://dx.doi.org/10.7717/peerj-cs.1

Reproducibility and reusability of research results is an important concern in scientific communication and science policy. A foundational element of reproducibility and reusability is the open and persistently available presentation of research data. However, many common approaches for primary data publication in use today do not achieve sufficient long-term robustness, openness, accessibility or uniformity. Nor do they permit comprehensive exploitation by modern Web technologies. This has led to several authoritative studies recommending uniform direct citation of data archived in persistent repositories. Data are to be considered as first-class scholarly objects, and treated similarly in many ways to cited and archived scientific and scholarly literature. Here we briefly review the most current and widely agreed set of principle-based recommendations for scholarly data citation, the Joint Declaration of Data Citation Principles (JDDCP). We then present a framework for operationalizing the JDDCP; and a set of initial recommendations on identifier schemes, identifier resolution behavior, required metadata elements, and best practices for realizing programmatic machine actionability of cited data. The main target audience for the common implementation guidelines in this article consists of publishers, scholarly organizations, and persistent data repositories, including technical staff members in these organizations. But ordinary researchers can also benefit from these recommendations. The guidance provided here is intended to help achieve widespread, uniform human and machine accessibility of deposited data, in support of significantly improved verification, validation, reproducibility and re-use of scholarly/scientific data.

Starr, Joan, and Angela Gastl. “isCitedBy: A Metadata Scheme for DataCite.” *D-Lib Magazine* 17, no. 1/2 (2011). http://www.dlib.org/dlib/january11/starr/01starr.html

Štebe, Janez, Maja Dolinar, Sonja Bezjak, and Ana Inkret. “Implementing the RDA Research Data Policy Framework in Slovenian Scientific Journals.” *Data Science Journal* 19, no 1 (2020): p.49. http://doi.org/10.5334/dsj-2020-049

The paper aims to present the implementation of the RDA research data policy framework in Slovenian scientific journals within the project RDA Node Slovenia. The activity aimed to implement the practice of data sharing and data citation in Slovenian scientific journals and was based on internationally renowned practices and policies, particularly the Research Data Policy Framework of the RDA Data Policy Standardization and Implementation Interest Group. Following this, the RDA Node Slovenia coordination prepared a guidance document that allowed the four pilot participating journals (from fields of archaeology, history, linguistics and social sciences) to adjust their journal policies regarding data sharing, data citation, adapted the definitions of research data and suggested appropriate data repositories that suit their disciplinary specifics. The comparison of results underlines how discipline-specific the aspects of data-sharing are. The pilot proved that a grass-root approach in advancing open science can be successful and well-received in the research community, however, it also pointed out several issues in scientific publishing that would benefit from a planned action on a national level. The context of an underdeveloped data sharing culture, slow implementation of open data strategy by the national research funder and sparse national data service infrastructure creates a unique environment for this study, the result of which can be used in similar contexts worldwide.

Steinhart, Gail. “DataStaR: A Data Sharing and Publication Infrastructure to Support Research.” *Agricultural Information Worldwide: An International Journal for the Information Specialists in Agriculture, Natural Resources, and the Environment* 4, no. 1 (2011). https://hdl.handle.net/1813/15035

Stockhause, Martina, and Michael Lautenschlager. “CMIP6 Data Citation of Evolving Data.” *Data Science Journal* 16, no. 30 (2017): p.30. http://doi.org/10.5334/dsj-2017-030

Data citations have become widely accepted. Technical infrastructures as well as principles and recommendations for data citation are in place but best practices or guidelines for their implementation are not yet available. On the other hand, the scientific climate community requests early citations on evolving data for credit, e.g. for CMIP6 (Coupled Model Intercomparison Project Phase 6). The data citation concept for CMIP6 is presented. The main challenges lie in limited resources, a strict project timeline and the dependency on changes of the data dissemination infrastructure ESGF (Earth System Grid Federation) to meet the data citation requirements. Therefore a pragmatic, flexible and extendible approach for the CMIP6 data citation service was developed, consisting of a citation for the full evolving data superset and a data cart approach for citing the concrete used data subset. This two citation approach can be implemented according to the RDA recommendations for evolving data. Because of resource constraints and missing project policies, the implementation of the second part of the citation concept is postponed to CMIP7.

Tazegul, Gokhan, and Emre Emre. “Scientometric Data and Open Access Publication Policies of Clinical Allergy and Immunology Journals.” *Cureus* 13, no. 2 (2021): e13564. https://doi.org/10.7759/cureus.13564

Tazegul, Gokhan, Emre Emre, Tahir Saygin, Tahir Saygin Öğüt, and Veli Yazisiz. “An Analysis of Scientometric Data and Publication Policies of Rheumatology Journals.”

Clinical Rheumatology 40, no. 11 (2021): 4693-4700. https://doi.org/10.1007/s10067-021-05824-2

Tedersoo, Leho, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann, and Tuul Sepp. "Data Sharing Practices and Data Availability upon Request Differ across Scientific Disciplines." *Scientific Data* 8, no. 1 (2021): 192. https://doi.org/10.1038/s41597-021-00981-0

> Data sharing is one of the cornerstones of modern science that enables large-scale analyses and reproducibility. We evaluated data availability in research articles across nine disciplines in Nature and Science magazines and recorded corresponding authors' concerns, requests and reasons for declining data sharing. Although data sharing has improved in the last decade and particularly in recent years, data availability and willingness to share data still differ greatly among disciplines. We observed that statements of data availability upon (reasonable) request are inefficient and should not be allowed by journals. To improve data sharing at the time of manuscript acceptance, researchers should be better motivated to release their data with real benefits such as recognition, or bonus points in grant and job applications. We recommend that data management costs should be covered by funding agencies; publicly available research data ought to be included in the evaluation of applications; and surveillance of data sharing should be enforced by both academic publishers and funders. These cross-discipline survey data are available from the plutoF repository.

Tegbaru, Dawit, Lisa Braverman, Anthony L. Zietman, Sue S. Yom, W. Robert Lee, Robert C. Miller, Isabel L. Jackson, Todd McNutt, and Andre Dekker. "ASTRO Journals' Data Sharing Policy and Recommended Best Practices." *Advances in Radiation Oncology* 4, no. 4 (2019): 551-558. https://doi.org/https://doi.org/10.1016/j.adro.2019.08.002

Teplitzky, Samantha. "Open Data, [Open] Access: Linking Data Sharing and Article Sharing in the Earth Sciences." *Journal of Librarianship and Scholarly Communication* 5, no. 1 (2017): eP2150. http://doi.org/10.7710/2162-3309.2150

> INTRODUCTION The norms of a research community influence practice, and norms of openness and sharing can be shaped to encourage researchers who share in one aspect of their research cycle to share in another. Different sets of mandates have evolved to require that research data be made public, but not necessarily articles resulting from that collected data. In this paper, I ask to what extent publications in the Earth Sciences are more likely to be open access (in all of its definitions) when researchers open their data through the Pangaea repository. METHODS Citations from Pangaea data sets were studied to determine the level of open access for each article. RESULTS This study finds that the proportion of gold open access articles linked to the repository increased 25% from 2010 to 2015 and 75% of articles were available from multiple open sources. DISCUSSION The context for increased preference for gold open access is considered and future work linking researchers' decisions to open their work to the adoption of open access mandates is proposed.

Tomaszewski, Robert. "Citations to Chemical Databases in Scholarly Articles: To Cite or Not to Cite?" *Journal of Documentation* 75 no. 6, (2019): 1317-1332. https://doi.org/10.1108/JD-12-2018-0214

Ulbricht, Damian, Kirsten Elger, Roland Bertelmann, and Jens Klump. "panMetaDocs, eSciDoc, and DOIDB—An Infrastructure for the Curation and Publication of File-Based Datasets for GFZ Data Services." *ISPRS International Journal of Geo-Information* 5, no. 3 (2016): 25. http://dx.doi.org/10.3390/ijgi5030025

Urban, Ed, Adam Leadbetter, Gwenaelle Moncoiffe, Peter Pissierssens, Lisa Raymond, and Linda Pikula. "Pilot Projects for Publishing and Citing Ocean Data." *Eos, Transactions*

American Geophysical Union 93, no. 43 (2012): 425-426. https://doi.org/https://doi.org/10.1029/2012EO430001

Van den Eynden, Veerle, and Louise Corti. “Advancing Research Data Publishing Practices for the Social Sciences: From Archive Activity to Empowering Researchers.” *International Journal on Digital Libraries* 18, no. 2 (2017): 113-121. https://doi.org/10.1007/s00799-016-0177-3

Sharing and publishing social science research data have a long history in the UK, through long-standing agreements with government agencies for sharing survey data and the data policy, infrastructure, and data services supported by the Economic and Social Research Council. The UK Data Service and its predecessors developed data management, documentation, and publishing procedures and protocols that stand today as robust templates for data publishing. As the ESRC research data policy requires grant holders to submit their research data to the UK Data Service after a grant ends, setting standards and promoting them has been essential in raising the quality of the resulting research data being published. In the past, received data were all processed, documented, and published for reuse in-house. Recent investments have focused on guiding and training researchers in good data management practices and skills for creating shareable data, as well as a self-publishing repository system, ReShare. ReShare also receives data sets described in published data papers and achieves scientific quality assurance through peer review of submitted data sets before publication. Social science data are reused for research, to inform policy, in teaching and for methods learning. Over a 10 years period, responsive developments in system workflows, access control options, persistent identifiers, templates, and checks, together with targeted guidance for researchers, have helped raise the standard of self-publishing social science data. Lessons learned and developments in shifting publishing social science data from an archivist responsibility to a researcher process are showcased, as inspiration for institutions setting up a data repository.

Vasilevsky, Nicole A., Jessica Minnier, Melissa A. Haendel, and Robin E. Champieux. “Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark?” *PeerJ* 5 (April 25, 2017): e3208. https://doi.org/10.7717/peerj.3208

Background

There is wide agreement in the biomedical research community that research data sharing is a primary ingredient for ensuring that science is more transparent and reproducible. Publishers could play an important role in facilitating and enforcing data sharing; however, many journals have not yet implemented data sharing policies and the requirements vary widely across journals. This study set out to analyze the pervasiveness and quality of data sharing policies in the biomedical literature.

Methods

The online author’s instructions and editorial policies for 318 biomedical journals were manually reviewed to analyze the journal’s data sharing requirements and characteristics. The data sharing policies were ranked using a rubric to determine if data sharing was required, recommended, required only for omics data, or not addressed at all. The data sharing method and licensing recommendations were examined, as well any mention of reproducibility or similar concepts. The data was analyzed for patterns relating to publishing volume, Journal Impact Factor, and the publishing model (open access or subscription) of each journal.

Results

A total of 11.9% of journals analyzed explicitly stated that data sharing was required as a condition of publication. A total of 9.1% of journals required data sharing, but did

not state that it would affect publication decisions. 23.3% of journals had a statement encouraging authors to share their data but did not require it. A total of 9.1% of journals mentioned data sharing indirectly, and only 14.8% addressed protein, proteomic, and/or genomic data sharing. There was no mention of data sharing in 31.8% of journals. Impact factors were significantly higher for journals with the strongest data sharing policies compared to all other data sharing criteria. Open access journals were not more likely to require data sharing than subscription journals.

Discussion

Our study confirmed earlier investigations which observed that only a minority of biomedical journals require data sharing, and a significant association between higher Impact Factors and journals with a data sharing requirement. Moreover, while 65.7% of the journals in our study that required data sharing addressed the concept of reproducibility, as with earlier investigations, we found that most data sharing policies did not provide specific guidance on the practices that ensure data is maximally available and reusable.

Vidal-Infer, Antonio, Rafael Aleixandre-Benavent, Rut Lucas-Domínguez, and Andrea Sixto-Costoya. “The Availability of Raw Data in Substance Abuse Scientific Journals.” *Journal of Substance Use* 24, no. 1 (2019): 36-40. https://doi.org/10.1080/14659891.2018.1489905

Vidal-Infer, Antonio, Beatriz Tarazona, Adolfo Alonso-Arroyo, and Rafael Aleixandre-Benavent. “Public Availability of Research Data in Dentistry Journals Indexed in Journal Citation Reports.” *Clinical Oral Investigations* (2017): 275-280. https://doi.org/10.1007/s00784-017-2108-0

Vlaeminck, Sven. “Dawning of a New Age? Economics Journals’ Data Policies on the Test Bench.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31, no. 1 (2021), 1-29. https://doi.org/10.53377/lq.10940

In the field of social sciences and particularly in economics, studies have frequently reported a lack of reproducibility of published research. Most often, this is due to the unavailability of data reproducing the findings of a study. However, over the past years, debates on open science practices and reproducible research have become stronger and louder among research funders, learned societies, and research organisations. Many of these have started to implement data policies to overcome these shortcomings. Against this background, the article asks if there have been changes in the way economics journals handle data and other materials that are crucial to reproduce the findings of empirical articles. For this purpose, all journals listed in the Clarivate Analytics Journal Citation Reports edition for economics have been evaluated for policies on the disclosure of research data. The article describes the characteristics of these data policies and explicates their requirements. Moreover, it compares the current findings with the situation some years ago. The results show significant changes in the way journals handle data in the publication process. Research libraries can use the findings of this study for their advisory activities to best support researchers in submitting and providing data as required by journals.

Vlaeminck, Sven, and Gert G. Wagner. “On the Role of Research Data Centres in the Management of Publication-Related Research Data.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 23, no. 4 (2014): 336-357. http://doi.org/10.18352/lq.9356

This paper summarizes the findings of an analysis of scientific infrastructure service providers (mainly from Germany but also from other European countries). These service providers are evaluated with regard to their potential services for the management of publication-related research data in the field of social sciences, especially economics. For this purpose we conducted both desk research and an

online survey of 46 research data centres (RDCs), library networks and public archives; almost 48% responded to our survey. We find that almost three-quarters of all respondents generally store externally generated research data—which also applies to publication-related data. Almost 75% of all respondents also store and host the code of computation or the syntax of statistical analyses. If self-compiled software components are used to generate research outputs, only 40% of all respondents accept these software components for storing and hosting. Eight out of ten institutions also take specific action to ensure long-term data preservation. With regard to the documentation of stored and hosted research data, almost 70% of respondents claim to use the metadata schema of the Data Documentation Initiative (DDI); Dublin Core is used by 30 percent (multiple answers were permitted). Almost two-thirds also use persistent identifiers to facilitate citation of these datasets. Three in four also support researchers in creating metadata for their data. Application programming interfaces (APIs) for uploading or searching datasets currently are not yet implemented by any of the respondents. Least common is the use of semantic technologies like RDF.

Concluding, the paper discusses the outcome of our survey in relation to Research Data Centres (RDCs) and the roles and responsibilities of publication-related data archives for journals in the fields of social sciences.

Walters, William H. “Data Journals: Incentivizing Data Access and Documentation within the Scholarly Communication System.” *Insights The UKSG Journal* 33, no. 1 (2020): 18. https://doi.org/10.1629/uksg.510

Data journals provide strong incentives for data creators to verify, document and disseminate their data. They also bring data access and documentation into the mainstream of scholarly communication, rewarding data creators through existing mechanisms of peer-reviewed publication and citation tracking. These same advantages are not generally associated with data repositories, or with conventional journals’ data-sharing mandates. This article describes the unique advantages of data journals. It also examines the data journal landscape, presenting the characteristics of 13 data journals in the fields of biology, environmental science, chemistry, medicine and health sciences. These journals vary considerably in size, scope, publisher characteristics, length of data reports, data hosting policies, time from submission to first decision, article processing charges, bibliographic index coverage and citation impact. They are similar, however, in their peer review criteria, their open access license terms and the characteristics of their editorial boards.

Williams, Sarah C. “Data Practices in the Crop Sciences: A Review of Selected Faculty Publications.” *Journal of Agricultural & Food Information* 13, no. 4 (2012): 308-325. https://doi.org/10.1080/10496505.2012.717846

Wiley, Chris. “Data Sharing: An Analysis of Medical Faculty Journals and Articles.” *Science & Technology Libraries* 40, no. 1 (2021): 104-115. https://doi.org/10.1080/0194262X.2020.1781740

Yoon, Jung Won, Eun Kyung Chung, Janet Schalk, and Jihyun Kim. “Examination of Data Citation Guidelines in Style Manuals and Data Repositories.” *Learned Publishing* 34, no. 2 (2021): 198-215. https://doi.org/10.1002/LEAP.1349

Zhao, Mengnan, Erjia Yan, and Kai Li. “Data Set Mentions and Citations: A Content Analysis of Full-text Publications.” *Journal of the Association for Information Science and Technology* 69, no. 1 (2018): 32-46. https://doi.org/10.1002/ASI.23919.

Zilinski, Lisa D., David Scherer, Darcy Bullock, Deborah Horton, Courtney Matthews. “Evolution of Data Creation, Management, Publication, and Curation in the Research Process.” *Transportation Research Record: Journal of the Transportation Research Board* 2414 (2014): 9-19. https://doi.org/10.3141/2414-02

Zwölf, Carlo Maria, Nicolas Moreau, Yaye-Awa Ba, and Marie-Lise Dubernet. "Implementing in the VAMDC the New Paradigms for Data Citation from the Research Data Alliance." *Data Science Journal* 18, no. 1 (2019): p.5. https://doi.org/10.5334/DSJ-2019-004.

> VAMDC [Virtual Atomic And Molecular Data Centre] bridged the gap between atomic and molecular (A&M) producers and users by providing an interoperable e-infrastructure connecting A&M databases, as well as tools to extract and manipulate those data. The current paper highlights how the new paradigms for data citation produced by the Research Data Alliance in order to address the citation issues in the data-driven science landscape, have successfully been implemented on the VAMDC e-infrastructure.

Note on the Inclusion of Abstracts

Abstracts are included in this bibliography if a work is under a Creative Commons Attribution License (BY and national/international variations), a Creative Commons public domain dedication (CC0), or a Creative Commons Public Domain Mark and this is clearly indicated in the publisher's current webpage for the article. Note that a publisher may have changed the licenses for all articles on a journal's website but not have made corresponding license changes in journal's PDF files. The license on the current webpage is deemed to be the correct one. Since publishers can change licenses in the future, the license indicated for a work in this bibliography may not be the one you find upon retrieval of the work.

Abstracts for works under the following types of Creative Commons Licenses (and their national/international variations) are not included:

- Attribution-NoDerivs
- Attribution-NonCommercial
- Attribution-NonCommercial-NoDerivs
- Attribution-NonCommercial-ShareAlike
- Attribution-ShareAlike

See the Creative Commons' *Frequently Asked Questions* for a discussion of how documents under different Creative Commons licenses can be combined.

About the Author

Charles W. Bailey, Jr. is the publisher of Digital Scholarship and a noncommercial digital artist (ORCID ID: https://orcid.org/0000-0001-8453-4402).

Bailey has over 44 years of information technology, digital publishing, and instructional technology experience, including 24 years of managerial experience in academic libraries. From 2004 to 2007, he was the Assistant Dean for Digital Library Planning and Development at the University of Houston Libraries. From 1987 to 2003, he served as Assistant Dean/Director for Systems at the University of Houston Libraries.

Previously, he served as Head, Systems and Research Services at the Health Sciences Library, The University of North Carolina at Chapel Hill; Systems Librarian at the Milton S. Eisenhower Library, The Johns Hopkins University; User Documentation Specialist at the OCLC Online Computer Library Center; and Media Library Manager at the Learning Resources Center, SUNY College at Oswego.

Bailey has discussed his career in an interview in *Preservation, Digital Technology & Culture*. See Bailey's vita for more details.

Bailey has been an open access publisher for over 32 years. In 1989, Bailey established PACS-L, a discussion list about public-access computers in libraries, and *The Public-Access Computer Systems Review*, the first open access journal in the field of library and information science. He served as PACS-L Moderator until November 1991 and as Editor-in-Chief of *The Public-Access Computer Systems Review* until the end of 1996.

In 1990, Bailey and Dana Rooks established *Public-Access Computer Systems News*, an electronic newsletter, and Bailey co-edited this publication until 1992.

In 1992, he founded the PACS-P mailing list for announcing the publication of selected e-serials, and he moderated this list until 2007.

In 1996, he established the *Scholarly Electronic Publishing Bibliography* (*SEPB*), an open access book that was updated 80 times by 2011.

In 2001, he added the *Scholarly Electronic Publishing Weblog*, which announced relevant new publications, to *SEPB*.

In 2001, he was selected as a team member of *Current Cites*, and he has was a frequent contributor of reviews to this monthly e-serial until 2020.

In 2005, he published the *Open Access Bibliography: Liberating Scholarly Literature with E-prints and Open Access Journals* with the Association of Research Libraries (also a website).

In 2005, Bailey established Digital Scholarship (http://digital-scholarship.org/), which provides information and commentary about digital copyright, digital curation, digital repository, open access, research data management, scholarly communication, and other digital information issues. Digital Scholarship's digital publications are open access. Its publications are under Creative Commons licenses.

At that time, he also established *DigitalKoans*, a weblog that covers the same topics as Digital Scholarship.

From April 2005 through May 2022, Bailey published the following books and book supplements: the *Scholarly Electronic Publishing Bibliography: 2008 Annual Edition* (2009), Digital Scholarship 2009 (2010), *Transforming Scholarly Publishing through Open Access: A Bibliography* (2010), the *Scholarly Electronic Publishing Bibliography 2010* (2011), the *Digital Curation and Preservation Bibliography 2010* (2011), the *Institutional Repository and ETD Bibliography 2011* (2011), the *Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works* (2012), the *Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works, 2012 Supplement* (2013), and the *Research Data Curation and Management Bibliography* (2021).

He also published and updated the following bibliographies and webliographies as websites with links to freely available works: the *Scholarly Electronic Publishing Bibliography* (1996-2011), the *Electronic Theses and Dissertations Bibliography* (2005-2021), the *Google Books Bibliography* (2005-2011), the *Institutional Repository Bibliography* (2009-2011), the *Open Access Journals Bibliography* (2010), the *Digital Curation and Preservation Bibliography* (2010-2011), the *E-science and Academic Libraries Bibliography* (2011), the *Digital Curation Resource Guide* (2012), the *Research Data Curation Bibliography* (2012-2019), the *Altmetrics Bibliography* (2013), the *Transforming Peer Review Bibliography* (2014), the *Academic Library as Scholarly Publisher Bibliography* (2018-2021), the *Research Data Sharing and Reuse Bibliography* (2021), and the *Research Data Publication and Citation Bibliography* (2022).

In 2011, he established the LinkedIn Digital Curation Group.

For more details, see the "A Look Back at 33 Years as an Open Access Publisher."

In 2010, Bailey was given a Best Content by an Individual Award by *The Charleston Advisor*. In 2003, he was named as one of *Library Journal*'s "Movers & Shakers." In 1993, he was awarded the first LITA/Library Hi Tech Award For Outstanding Communication for Continuing Education in Library and Information Science. In 1992, Bailey received a Network Citizen Award from the Apple Library.

In 1973, Bailey won a Wallace Stevens Poetry Award. He is the author of *The Cave of Hypnos: Early Poems*, which includes several poems that won that award.

Bailey has written over 30 papers about artificial intelligence, digital copyright, institutional repositories, open access, scholarly communication, and other topics.

He has served on the editorial boards of *Information Technology and Libraries*, *Library Software Review*, and *Reference Services Review*. He was the founding Vice-Chairperson of the LITA Imagineering Interest Group.

Bailey is a digital artist, and he has made over 600 digital artworks freely available on social media sites, such as Flickr, under Creative Commons Attribution-NonCommercial licenses. A list of his artwoks that includes links to high resolution JPEG images on Flickr is available.

He holds master's degrees in information and library science and instructional media and technology.

You can contact him at: publisher at digital-scholarship.org.

You can follow Bailey at these URLs:

Digital Artist weblog: https://charleswbaileyjr.name and RSS feed (https://charleswbaileyjr.name/feed)

DigitalKoans weblog: http://digital-scholarship.org/digitalkoans/

Flickr: https://www.flickr.com/photos/charleswbaileyjr/

Twitter (*DigitalKoans*): https://twitter.com/DigitalKoans

Citation

Charles W. Bailey, Jr., *Research Data Publication and Citation Bibliography* (Houston: Digital Scholarship, 2022), http://digital-scholarship.org/citation/citation.htm.

Bailey, Charles W., Jr. *Research Data Publication and Citation Bibliography*. Houston: Digital Scholarship, 2022. http://digital-scholarship.org/citation/citation.htm.