Journal of eScience Librarianship

Volume 9 | Issue 1 Article 3

2020-09-09

Open Data and Open Access Articles: Exploring Connections in the Life Sciences

Sarah C. Williams
University of Illinois at Urbana-Champaign

ISSN 2161-3974

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Open Data and Open Access Articles: Exploring Connections in the Life Sciences

Sarah C. Williams

University of Illinois at Urbana-Champaign, Champaign, IL, USA

Abstract

Objective: This small-scale study explores the current state of connections between open data and open access (OA) articles in the life sciences.

Methods: This study involved 44 openly available life sciences datasets from the Illinois Data Bank that had 45 related research articles. For each article, I gathered the OA status of the journal and the article on the publisher website and checked whether the article was openly available via Unpaywall and ResearchGate. I also examined how and where the open data was included in the HTML and PDF versions of the related articles.

Results: Of the 45 articles studied, less than half were published in Gold/Full OA journals, and while the remaining articles were published in Gold/Hybrid journals, none of them were OA. This study found that OA articles pointed to the Illinois Data Bank datasets similarly to all of the related articles, most commonly with a data availability statement containing a DOI.

Conclusions: The findings indicate that Gold OA in hybrid journals does not appear to be a popular option, even for articles connected to open data, and this study emphasizes the importance of data repositories providing DOIs, since the related articles frequently used DOIs to point to the Illinois Data Bank datasets. This study also revealed concerns about free (not licensed OA) access to articles on publisher websites, which will be a significant topic for future research.

Correspondence: Sarah C. Williams: scwillms@illinois.edu

Received: February 20, 2020 **Accepted**: May 28, 2020 **Published**: September 9, 2020

Disclosures: The author reports no conflict of interest.

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

Introduction

Scientific researchers must consider a myriad of options and expectations when preparing to disseminate their research results—publications and data. Their funding agency, such as those included in the Office of Science and Technology Policy's public access memo (Holdren 2013), may expect the resulting research data and publications to be made freely available to the public. Journal publishers in their field may encourage or require the research data underlying their article to be shared openly (Wiley 2018). Even their institution may have an open access (OA) policy, as many institutions have implemented or are developing OA policies (SPARC 2019). Researchers may need to follow standard practices in their discipline, and they may have their own preferences or convictions for how to disseminate their research. Plus, there are numerous options for sharing research data (e.g., supplementary materials, disciplinary and institutional data repositories) and a variety of publishing options (e.g., full OA journals, hybrid journals, subscription-only journals).

The publishing landscape continues to evolve. Some concepts and practices are well established, while others are emerging. The 2002 Budapest Open Access Initiative (BOAI) provided a foundational OA definition—making content freely available to readers without usage restrictions, and some OA categories are widely accepted (e.g., Gold and Green). With the Gold OA category, articles are openly available on the publisher website, either because the journal is completely open access (Gold/Full OA) or the author(s) paid an article processing charge (APC) for an article to be openly available in a subscription journal (Gold/Hybrid). With Green OA, subscription articles are self-archived in OA repositories, like institutional repositories or PubMed Central (the National Library of Medicine's free full-text archive of biomedical and life sciences journal literature). Bronze OA is a new category name coined by Piwowar et al. (2018). Bronze encompasses articles that are free (not licensed OA) on publisher websites, and while these article are free to read, they are not always free to reuse, so they do not necessarily meet the BOAI definition of OA.

In an earlier study exploring agricultural researchers' attitudes toward open access and data sharing (Williams et al. 2019), I was struck by the limited literature discussing these two topics in relation to each other. While some librarians might assume that open data and OA articles go hand-in-hand, scientists and librarians do not always share the same perspectives and practices (Imker 2017). To better understand this situation, I conducted a small-scale study to begin exploring the actual state of connections between open data and OA articles in the life sciences. Starting with the life sciences datasets in University of Illinois' institutional data repository, Illinois Data Bank, I identified dataset records with related articles. With those linked datasets and articles, I investigated the OA status of the articles and how the articles pointed to the open Illinois Data Bank data.

Literature Review

Of the limited literature linking open data and OA, a study by Teplitzky (2017) is particularly relevant. Focused on the earth sciences, Teplitzky (2017) investigated to what extent researchers who shared data via the Pangaea repository also made the related articles openly available. The percentage of articles linked to Pangaea datasets that were published in Gold/Full OA journals increased from 9.7% in 2010

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

to 30.9% in 2015. For select years, Teplitzky (2017) also explored OA more broadly to include articles openly available via hybrid journals, institutional repositories, and ResearchGate and found a majority of the articles were available without a subscription through multiple resources.

Castro et al. (2017) approached this linkage from a different direction by studying the data policies of OA journals. From the Directory of Open Access Journals and the Open Journal Systems, they sampled 50 OA journals from a variety of disciplines. They found a majority of the OA journals (74%) did not have a data policy in 2015 and the policies did not seem to strengthen when a targeted follow-up was completed in 2017.

Both articles (Castro et al. 2017; Teplitzky 2017) also provide valuable background information and references for the issues surrounding open data and OA, and there is a wealth of literature focused on these issues separately. Common themes are often found across open data and data sharing studies, regardless of whether the research was conducted at an institutional, disciplinary, national, or international level (Bishoff and Johnston 2015; Herold 2015; Kim and Zhang 2015; Tenopir et al. 2011); these common findings include variability in data sharing approaches and similarities in the barriers to, and motivations for, sharing data. Across the literature focused on scholars' practices and attitudes toward OA (Gaines 2015; Rowley et al. 2017; Tenopir et al. 2017; Xia 2010), a general sense of negativity often emerges, including uncertainty, fear, and ambivalence toward OA publishing.

Methods

This small-scale study started with the life sciences datasets in the Illinois Data Bank (https://databank.illinois.edu), which is the University of Illinois at Urbana-Champaign public access, research data repository that went live in 2016. From the 70 life sciences datasets as of April 30, 2019, I identified 44 dataset records that included at least one related publication, excluding pre-prints, theses, reports, books, and conference papers or proceedings (unless published in a journal). Information was gathered from the Illinois Data Bank from April 25-May 3, 2019.

In summer 2019, I gathered information about the related articles and the journals. For the journal OA status, I searched the Directory of Open Access Journals (DOAJ) and checked the journal website (i.e., about the journal and/or information for authors webpages). With the article DOI, I tried to access the article off-campus without a VPN connection to determine whether it was freely available on the publisher website. To see if an article was OA from a different source, I used University of Illinois' homegrown search system, Easy Search, which incorporates Unpaywall (a database of millions of scholarly articles that are legally OA). I searched the article DOI and if necessary, the article title and clicked the Easy Search Unpaywall link to see where it led. Similar to the study by Teplitzky (2017), I also looked for the articles in ResearchGate, an academic social media platform that facilitates research article sharing, whether the sharing is legal or infringes on copyright (McKenzie 2018). I searched the article title in the ResearchGate search box, and I only counted access if I could immediately download the article.

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

To check for references to the data in the related articles, I searched each article for "Illinois Data," "IDB," and a portion of the IDB DOI. I reviewed both the HTML and PDF versions of the articles, in case they were different. If I found a reference to the data, I used standard language to record how it was mentioned (e.g., "in data section with DOI link" or "in text with DOI link") and where it was mentioned (e.g., methods, discussion, acknowledgements).

Results

The Illinois Data Bank had 70 life sciences datasets published by April 30, 2019, and of those, 44 dataset records included at least one related article. The total number of related articles was 45. Some dataset records linked to more than one article, and some articles connected to more than one Illinois Data Bank record. Table 1 provides the publication years of the 45 articles. One of the articles was published in 1981, 30 years before the next oldest article, but regardless, the 1981 article from *Veterinary Pathology* was related to longitudinal data that was deposited into the Illinois Data Bank in 2017.

Table 1: Publication years of articles in this study.

Publication Year	Number of Articles (n = 45)
2019	11
2018	13
2017	13
2016	3
2015	2
2014	1
2011	1
1981	1

The articles were published in 36 different journals from 18 different publishers. Of the 18 publishers, 12 had just one journal, and six had multiple journals: Wiley (10), Elsevier (4), BMC (3), Springer (3), Nature (2), and Oxford (2). Thirty journals had one article each, while these six journals contained multiple articles: *PLoS ONE* (4), *BMC Genomics* (3), *Annals of Botany* (2), *Functional Ecology* (2), *GCB Bioenergy* (2), and *Scientific Reports* (2). Four of these journals were Gold/Full OA journals, and *Annals of Botany* and *Functional Ecology* were Gold/Hybrid journals. Among all 36 journals, 24 (66.6%) were Gold/Hybrid journals, and 12 (33.3%) were Gold/Full OA journals. The Appendix includes a complete list of the 36 journals, their publishers, OA status, and number of articles in each journal.

OA Status of Articles

A major goal of this study was investigating the OA status of the 45 articles related to the life sciences datasets openly available in the Illinois Data Bank. Nineteen of the articles were published in Gold/Full OA journals. Of the 26 remaining articles, 19 required a subscription to the Gold/Hybrid journals, and seven had Bronze (free, not licensed OA) access in Gold/Hybrid journals. One of

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

the free access articles was clearly in a sample issue of the journal, but in the other six cases, the reason for and the permanence of the free access was unclear. None of the 26 articles published in the Gold/Hybrid journals were available through licensed OA on the journal websites.

Many of the articles were openly available via other sources as well. I used University of Illinois' Easy Search system to search Unpaywall. Unpaywall linked to 24 articles on publisher websites—all 19 Gold/Full OA articles and 5 of Bronze articles, and it linked to two subscription articles that were openly available via PubMed Central (Green OA). I also investigated which articles were openly available in ResearchGate and found 22 articles were available for immediate download—14 Gold/Full OA articles, 3 Bronze articles, and 5 subscription articles.

Connections between Articles and Data

This study also examined how and where the open Illinois Data Bank data was included in the related articles. Of the 45 articles, 13 articles (29%) did not mention the data in the Illinois Data Bank. Some of these cases had a very logical explanation. For example, five of the related articles were published before the Illinois Data Bank existed, and in another case, the related article provides additional context for the data but was not directly related to the data.

For the 32 articles that did mention the Illinois Data Bank data, I tracked how and where the data was included in the article. Some of the articles referred to the data multiple times. The three most common ways to point to the Illinois Data Bank data were to: include a data availability statement with a DOI (18), mention the dataset and provide a DOI within the text of the article (13), and include the dataset with a DOI in the list of References (8). A few additional articles had a data availability statement or mentioned the data in the text of the article but did not include an Illinois Data Bank DOI. Outside of the data availability statement and references, the Illinois Data Bank datasets were mentioned throughout the text of the article: methods (8), acknowledgments (4), discussion (3), results (2) and conclusion (2), and one mentioned the data via a footnote near the article title.

Focusing specifically on the 19 Gold/Full OA articles, the findings are similar. Six of the articles (32%) did not mention the data in the Illinois Data Bank. Of the 13 articles that did point to the Illinois Data Bank data, nine had a data availability statement with a DOI and one more had a data availability statement without a DOI.

I also investigated whether the data was included similarly in the HTML and PDF versions of the articles. In almost every case, the HTML and PDF versions included the data in the same way. The only exception was an article in *Ecological Informatics*, where the HTML version had a data availability statement with a DOI but the PDF version had no data availability statement. In both versions of the article, the dataset was mentioned multiple times in the text—in the methods, results, discussion, and conclusion.

Discussion

The main limitation of this study is its small sample size—44 datasets and 45

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

related articles. The results cannot be used to make broad generalizations on the state of connections between open data and OA articles in the life sciences, but this study explores this important topic and points to several key findings that could be investigated in the future with larger sample sizes or with different methodologies.

OA Status of Articles

Starting with openly available datasets from the Illinois Data Bank, it is notable that only 19 of their 45 related articles were published in Gold/Full OA journals and none of the 26 articles published in Gold/Hybrid journals were available with an OA license on the journal websites. Other studies have found low OA adoption rates (1-4%) in hybrid journals, with exceptions for some disciplines or journals (Björk 2012; Kocher and Kelly 2016). In Teplitzky's study (2017) of OA articles connected to open data in Pangaea, Gold/Hybrid was the least common OA option in the two case study years—2010 and 2015. Gold OA in hybrid journals does not appear to be a popular option, even for articles connected to open data.

Seven of the articles in the Gold/Hybrid journals had Bronze (free, not licensed OA) access. While one of the free access articles was clearly in a sample issue of the journal, it was unclear why the other articles were freely available and how long they would remain freely available. In fact, when I checked the free access articles again three months later, two were no longer freely available—one in *Agricultural and Forest Meteorology* and one in *Annals of Botany*, while another free access article in *Annals of Botany* was still freely available.

Piwowar et al. (2018) coined a term for articles that are freely available on the publisher website without an explicit OA license: Bronze OA. Bronze is a combination of situations, including free/gratis access, delayed OA, and free-to-read journals that are not in the DOAJ and did not have clear license information. In two large-scale studies, Bronze was the most common method for publishers to make articles freely available (Martín-Martín et al. 2018; Piwowar et al. 2018). Martín-Martín et al. (2018) emphasized the precarious nature of Bronze OA and free/gratis access, and this small-scale study illustrates that point with the loss of free access to two of seven articles in only three-months time. Echoing the suggestions of Martín-Martín et al. (2018) and Piwowar et al. (2018), this will be an important topic for further research and discussion.

Connections between Articles and Data

I also explored how the articles pointed to the open Illinois Data Bank data, and there were a few notable findings. The subset of 19 Gold/Full OA articles handled data in ways similar to all of the articles. The Gold/Full OA articles had a similar percentage (around 30%) that did not mention the Illinois Data Bank data, and for those articles that did point to the data, the most common methods was to include it in a data availability statement with a DOI. In other words, for this small-scale study, Gold/Full OA articles do not have stronger connections to open data than other articles. This finding relates to the study by Castro et al. (2017), who analyzed OA journal data policies and compared their results to earlier studies of commercial and area-specific journal data policies. In their sample, the OA journals' data policies were not as strong as the data policies of commercial and area-specific journals (Castro et al. 2017).

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

Of the 32 articles that mentioned the Illinois Data Bank data, 30 of them included the data DOI. Such frequent use of DOIs, whether due to journal requirements or author preferences, reveals the importance of data repositories providing DOIs. Twenty-one of the 32 articles (66%) had a data availability statement (with or without a DOI). To have data sharing so commonly and clearly indicated is in sharp contrast to an earlier bibliographic study of articles published between 2001-2011, in which shared data (most commonly supplementary materials) was not always easy to find, because it was usually only mentioned in the text of the article (Williams 2012).

Future Research

In this small-scale study, many of the open datasets had related articles that were not OA, so future research could focus on the considerations and motivations of researchers' deciding whether or not to make their data and publications openly available. It would be interesting to explore why researchers make their data openly available but not the related articles. Admittedly, this is a complex situation, but it could be studied from several different angles, such as by examining funder and journal data requirements and institutional OA policies connected to research projects. Interviewing researchers about their open data and OA decisions would also be an interesting approach to this research question. The information gathered would help bridge differences in perspectives between librarians and scientists. Additionally, the methodology for this small-scale study could serve as a model for researchers to explore open data and OA article connections using other data repositories, and several of these smaller studies could be combined in a meta-analysis to more broadly examine the results in this study. Another future research project could conduct a larger-scale study of openly available life sciences datasets, similar to Teplitzky's (2017) study of earth sciences datasets in the Pangaea repository.

Conclusion

This study begins to explore the connections between open data and OA articles in the life sciences. Of the 45 articles investigated, less than half were published in Gold/Full OA journals, and while the remaining articles were published in Gold/ Hybrid journals, none of them were OA. The Gold/Full OA articles did not have stronger connections to open data than the non-OA articles. Similar to two larger-scale studies (Martín-Martín et al. 2018; Piwowar et al. 2018), this study revealed concerns about Bronze (free, not licensed OA) access to articles on publisher websites, which will be a significant topic for future research. In examining how and where the open Illinois Data Bank data was referred to in the articles, this study found that most articles used a data availability statement and frequently included the data DOI. This result emphasizes the importance of data repositories providing DOIs. Overall, exploring researchers' open data and OA decisions will be important for studying these connections in the future, especially since many of the open datasets in this study were related to articles that were not OA. A larger-scale study or a meta-analysis of comparable, smaller studies will be necessary to make broad generalizations on the state of connections between open data and OA articles in the life sciences.

JeSLIB 2020; 9(1): e1184
https://doi.org/10.7191/jeslib.2020.1184

Acknowledgments

I would like to thank Erin Kerby for providing valuable feedback on a draft of this article.

Supplemental Content

Appendix
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2020.1184 under "Additional Files".

References

Bishoff, Carolyn, and Lisa Johnston. 2015. "Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University." *Journal of Librarianship and Scholarly Communication* 3(2): eP1231. http://doi.org/10.7710/2162-3309.1231

Björk, Bo-Christer. 2012. "The Hybrid Model for Open Access Publication of Scholarly Articles: A Failed Experiment?" *Journal of the American Society for Information Science and Technology* 63(8): 1496-1504. https://doi.org/10.1002/asi.22709

Castro, Eleni, Mercè Crosas, Alex Garnett, Kasey Sheridan, and Micah Altman. 2017. "Evaluating and Promoting Open Data Practices in Open Access Journals." *Journal of Scholarly Publishing* 49(1): 66-88. https://doi.org/10.3138/jsp.49.1.66

Gaines, Annie M. 2015. "From Concerned to Cautiously Optimistic: Assessing Faculty Perceptions and Knowledge of Open Access in a Campus-Wide Study." *Journal of Librarianship and Scholarly Communication* 3(1): eP1212. https://doi.org/10.7710/2162-3309.1212

Herold, Philip. 2015. "Data Sharing Among Ecology, Evolution, and Natural Resources Scientists: An Analysis of Selected Publications." *Journal of Librarianship and Scholarly Communication* 3(2): eP1244. http://doi.org/10.7710/2162-3309.1244

Holdren, John P. 2013. "Increasing Access to the Results of Federally Funded Scientific Research." Accessed December 19, 2019. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Imker, Heidi J. 2017. "Overlooked and Overrated Data Sharing: Why Some Scientists are Confused and/or Dismissive." In *Curating Research Data: Practical Strategies for Your Digital Repository*, edited by Lisa R. Johnston, 127-150. Chicago: Association of College and Research Libraries.

Kim, Youngseek, and Ping Zhang. 2015. "Understanding Data Sharing Behaviors of STEM Researchers: The Roles of Attitudes, Norms, and Data Repositories." *Library & Information Science Research* 37(3): 189-200. https://doi.org/10.1016/j.lisr.2015.04.006

Kocher, Megan, and Julie Kelly. 2016. " Use of the Paid Open Access Option in Hybrid Open Access Journals in Agriculture: A Mixed-Methods Study." *Issues in Science and Technology Librarianship* 85. http://doi.org/10.5062/F47P8WDB

Martín-Martín, Alberto, Rodrigo Costas, Thed van Leeuwen, and Emilio Delgado López-Cózar. 2018. "Evidence of Open Access of Scientific Publications in Google Scholar: A Large-Scale Analysis." *Journal of Informatics* 12(3): 819-841. https://doi.org/10.1016/j.joi.2018.06.012

McKenzie, Lindsay. 2018. "Publishers Escalate Legal Battle Against ResearchGate." *Inside Higher Ed*, October 4, 2018. https://www.insidehighered.com/news/2018/10/04/publishers-accuse-researchgate-mass-copyright-infringement

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. 2018. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles." *PeerJ* 6: e4375. https://doi.org/10.7717/peerj.4375

Rowley, Jennifer, Frances Johnson, Laura Sbaffi, Will Frass, and Elaine Devine. 2017. "Academics' Behaviors and Attitudes Towards Open Access Publishing in Scholarly Journals." *Journal of the Association for Information Science and Technology* 68(5): 1201-1211. https://doi.org/10.1002/asi.23710

SPARC. 2019. "Coalition of Open Access Policy Institutions (COAPI) Members." Accessed December 19, 2019. https://sparcopen.org/coapi/members

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLoS ONE* 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, Carol, Elizabeth D. Dalton, Lisa Christian, Misty K. Jones, Mark McCabe, MacKenzie Smith, and Allison Fish. 2017. "Imagining a Gold Open Access Future: Attitudes, Behaviors, and Funding Scenarios among Authors of Academic Scholarship." *College & Research Libraries* 78(6): 824-843. https://doi.org/10.5860/crl.78.6.824

Teplitzky, Samantha. 2017. "Open Data, [Open] Access: Linking Data Sharing and Article Sharing in the Earth Sciences." *Journal of Librarianship and Scholarly Communication* 5(1): eP2150. http://doi.org/10.7710/2162-3309.2150

Wiley, Christie. 2018. "Data Sharing and Engineering Faculty: An Analysis of Selected Publications." *Science & Technology Libraries* 37(4): 409-419. https://doi.org/10.1080/0194262X.2018.1516596

Williams, Sarah C. 2012. "Data Practices in the Crop Sciences: A Review of Selected Faculty Publications." *Journal of Agricultural and Food Information* 13(4): 308-325. https://doi.org/10.1080/10496505.2012.717846

Williams, Sarah C., Shannon L. Farrell, Erin E. Kerby, and Megan Kocher. 2019. "Agricultural Researchers' Attitudes Toward Open Access and Data Sharing." *Issues in Science and Technology Librarianship* 91. https://doi.org/10.29173/istl4

Xia, Jingfeng. 2010. "A Longitudinal Study of Scholars Attitudes and Behaviors toward Open-Access Journal Publishing." *Journal of the American Society for Information Science and Technology* 61(3): 615-624. https://doi.org/10.1002/asi.21283