Are Researchers Citing Their Data? A Case Study from The U.S. Geological Survey

RESEARCH PAPER

GRACE C. DONOVAN
MADISON L. LANGSETH
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Data citation promotes accessibility and discoverability of data through measures carried out by researchers, publishers, repositories, and the scientific community. This paper examines how a data citation workflow has been implemented by the U.S. Geological Survey (USGS) by evaluating publication and data linkages. Two different methods were used to identify data citations: examining publication structural metadata and examining the full text of the publication. A growing number of USGS researchers are complying with publisher data sharing policies aimed to capture data citation information in a standardized way within associated publications. However, inconsistencies in how data citation information is documented in publications has limited the accessibility and discoverability of the data. This paper demonstrates how organizational evaluations of publication and data linkages can be used to identify obstacles in advancing data citation efforts and improve data citation workflows.

CORRESPONDING AUTHOR:
Grace C. Donovan
U.S. Geological Survey, Core Science Systems, Science, Analytics, and Synthesis, Denver, Colorado 80225, US
gdonovan@usgs.gov

KEYWORDS:
Data citation; structural metadata; data linkages; data sharing; open data

TO CITE THIS ARTICLE:
Donovan, G C and Langseth, M L 2024 Are Researchers Citing Their Data? A Case Study from The U.S. Geological Survey. *Data Science Journal*, 23: 24, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2024-024

INTRODUCTION

Data citations promote increased transparency and credit attribution for published data (ESIP Data Preservation and Stewardship Committee 2019; Parks et al. 2018; Zhao et al. 2017, Huang et al. 2015). These citations incorporate several components: author name, publication year, data release title, version number (if applicable), publisher name, and a digital object identifier (DOI) (USGS Data Management 2022). Similar to citations for published manuscripts, data citations ensure that contributors receive credit for their work (Mooney 2011) and allow contributors to track the impact of their data. Additionally, data citations enable the use and reuse of data by providing users with information to identify and access data (Lafia et al. 2023). Digital Object Identifiers (DOIs) assigned to data products are a primary means of tracking publication and data linkages (Zhao et al. 2017; Belter 2014). DOIs for data products also act as a 'standard mechanism for retrieval of metadata about the object' (Wilkinson et al. 2016).

Groups are working to promote data citation in research through community engagement. For example, Make Data Count is a global, community-led initiative, focused on incentivizing data sharing by developing 'open research data assessment metrics' (Make Data Count 2022). Two contributing organizations to Make Data Count are DataCite and Crossref. DataCite is a DOI and metadata registration organization focusing primarily on research data (DataCite 2022). Similarly, Crossref is a DOI and metadata registration organization focusing primarily on manuscripts and reports (Wilkinson 2022). Together, these organizations ensure the accessibility and discoverability of data and associated research artifacts through their partnership in linking publications registered with Crossref to data DOIs (Lin 2016).

Make Data Count (2022) outlines the ideal data citation workflow as follows:

1. Researchers include data citation in their publications according to journal data policies.
2. Publishers send data citation to Crossref as part of the publications' DOI metadata.
3. Repositories send publication references to DataCite as part of the datasets' DOI metadata.
4. Crossref and DataCite share DOI metadata with the research community through Application Programming Interfaces (APIs), such as Event Data (Rittman 2020).
5. Research community can access metrics related to links between datasets and publications using the Crossref and DataCite APIs.

DOI metadata is the foundation of the Make Data Count Initiative and data citation workflows. Crossref and DataCite document information about their DOIs in structural metadata. Structural metadata is machine-readable information that outlines the 'structure, type, and relationships of data' (Melton & Buxton 2006). While the infrastructure to support data citation is in place, variations in data citation practices have introduced complexities into data citation tracking (Gregory et al. 2023). Organizations like Crossref and DataCite, as well as some publishers, encourage researchers to include data citations within reference lists through data citation policies (Gregory et al. 2023; Farley 2022). However, several studies demonstrate that researchers continue to cite data in 'informal' ways (i.e., the data is mentioned within the full text of publications) that may not be included in publication structural metadata (Parks et al. 2018; Zhao et al. 2017; Belter 2014). Parks et al. (2018), Zhao et al. (2017), and Lafia et al. (2023) found that several inconsistencies in how researchers cite data were due to a lack of understanding regarding how to cite data and the importance and implications of citing data. However, researchers are not solely responsible for creating consistent data citations. Publishers also have a large role to play in data citation. For example, even though publishers are responsible for submitting reference lists to Crossref, some publishers may not have developed workflows necessary to include reference lists in the Crossref structural metadata. Deviations from the ideal data citation workflow ultimately impede our 'ability to consistently analyze, detect, and quantify data citations' (Irrera et al. 2023) through structural data analysis methods.

While it may be impossible to assess whether data citations are missing from a corpus of works using these methods alone, it may be possible to gauge uptake of data citations within a smaller research community using additional methods like text and data mining. Previous studies have demonstrated the efficacy of text and data mining techniques in identifying data citations within the full text of publications (Kafkas et al. 2013; Parks et al. 2018; Parsons et al.

2019). In this analysis, we leverage two text and data mining tools, Publink and xDD, to identify data citations that may not be present in structural metadata records. Publink is a Python package that allows users to find relationships between publications and data (Wieferich et al. 2020). In cases where references are not included in the publication's DOI structural metadata, Publink can be used to see if researchers are referencing their data by searching for mentions of data DOIs in the full text of publications included in the eXtract Dark Data (xDD) digital library. xDD, formerly known as GeoDeepDive, is a cyberinfrastructure that compiles data on published literature and provides users with the ability to perform full text searches of published literature using the xDD API (Peters et al. 2021a). As of 2021, xDD contained over 14 million commercial and open access publications of scientific works. While xDD initially compiled Earth science publications, it currently aims to be discipline agnostic.

In this analysis, publications authored by U.S. Geological Survey (USGS) researchers were evaluated to determine the presence of data citations. The USGS is a research agency that provides science about natural hazards, natural resources, ecosystems and environmental health, and the effects of climate and land-use change (USGS 2022). USGS research is disseminated through various types of publications, including USGS-authored journal articles through external publishers and series reports published by the USGS (USGS OSQI 2021c). An agreement between USGS and xDD has enabled xDD to index USGS series reports (Peters et al. 2021b). Publink and xDD are ideal tools for examining data mentioned within the full text of USGS series reports as well as USGS-authored publications indexed in xDD. Additionally, USGS researchers, through an instructional memorandum, were encouraged to publicly release data associated with their scholarly publications as of 2015 (USGS OSQI 2017). This instructional memorandum became policy and went into full effect in 2016 (USGS OSQI 2016). USGS policy requires that these data be assigned a DOI, be accompanied by a citation, and be referenced from the associated publication (USGS OSQI 2017). When USGS researchers acquire a DOI for their data through the USGS DOI Tool, they are asked to provide the DOI for the associated publication. The data DOI structural metadata offers access to a corpus of publications that should include data citations in some form (i.e., within the structural metadata or the full text). Considering these factors, the USGS presents a unique case study to evaluate the current state of data citations within a subset of the scientific research community. Our analysis shows how combined data citation tracking methods can be used to evaluate the extent to which researchers, publishers, and repositories have adhered to the ideal data citation workflow. This evaluation can help identify areas for improvement in data discoverability and accessibility.

METHODS

Metrics on data citations in publications produced by USGS authors were collected and analyzed using the USGS data DOI database (USGS DOI Tool), xDD, and the Crossref Application Programming Interface (API) in Jupyter Notebooks. These data were used to create a baseline analysis of how often researchers have cited the associated data in publications. Publications released from 2016 through 2022 were included in the collection. Publications released prior to 2016 were not included in the collection on account of the USGS instructional memorandum (USGS OSQI 2016) that became policy and went into full effect in 2016. Using the USGS DOI Tool API, we created an initial dataset by extracting data DOIs whose metadata included a related primary publication DOI. Additional related primary publication DOIs were identified through quality checks that captured incorrectly formatted DOIs (e.g., related primary publication DOIs not being stored in the DOI URL format) or placeholder DOIs (e.g., https://doi.org/10.xxxxxxx.xxxxxx) (Donovan & Langseth 2024). In total, there were 2,772 publications included in the analysis dataset. Links from a data DOI to a related primary publication are manually supplied by data authors in the USGS DOI Tool and are not required. Additionally, not all USGS publications use newly generated data to support their conclusions, which means that their authors are not minting USGS DOIs for data referenced in the publication. Therefore, the related primary publications included in the analysis dataset represent only a subset (around 16%) of all USGS publications (17,841) between 2016 and 2022.[1]

1 Total USGS publication count retrieved from the USGS Publications Warehouse, which catalogs all USGS series publications and articles published through external journals. https://pubs.er.usgs.gov/search?q=&startYear=2016&endYear=2022&subtypeName=Journal+Article&subtypeName=USGS+Numbered+Series&subtypeName=USGS+Unnumbered+Series.

First, we checked if a formal data citation was present in the publication's Crossref structural metadata. We obtained the article title, publication year, and publisher, using the habanero Python library (Chamberlain et al. 2022), based on the primary publication DOI. We also documented whether the Crossref structural metadata contained references. If references were included, the 'reference-count' value in the Crossref structural metadata was greater than zero and the publication was recorded as having references (Figure 1). For cases where the 'reference count' value was greater than zero, the publication was recorded as citing the data DOI if the associated data DOI was included in the 'doi' element of a reference in the Crossref structural metadata (Figure 2) (Donovan & Langseth 2024). Only publications with references in the Crossref structural metadata could be definitively recorded as citing the data DOI. For example, a publication could have a human-readable references section that included a data citation with a data DOI; however, for the purposes of this study, if the data DOI was not included in the 'doi' element of a reference in the Crossref structural metadata, then the data DOI would not be found using this method and would not count as a cited data DOI.

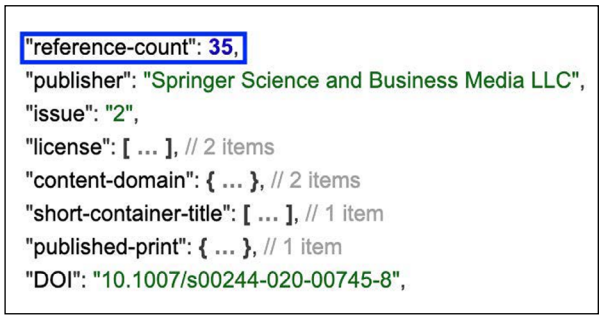

Figure 1 Crossref API call (https://api.Crossref.org/works/10.1007/s00244-020-00745-8) indicating Crossref structural metadata contains references.

Figure 2 Crossref API call (https://api.Crossref.org/works/10.1007/s00244-020-00745-8) indicating the data DOI is listed in the 'doi' element in the reference of the Crossref structural metadata.

Second, we checked if there was a data citation in the full text of the publication, rather than in the publication's structural metadata. For publications with full text available in xDD (49% of the full publication list), the presence of a data DOI mentioned anywhere in the full text was identified using the Publink python package, built on top of the xDD API (Wieferich et al. 2020; Donovan & Langseth 2024).

Information on Crossref references and data DOIs captured within the Crossref references was used to create three subsets to analyze the data between 2016 and 2022 (Figure 3):

- Publications with Crossref references that contained data DOIs
- Publications with Crossref references that did not contain data DOIs
- Publications without Crossref references

Binomial Generalized Linear Models (GLMs) were used to examine trends in the proportion of publications with data DOIs captured in the Crossref reference(s) of their associated publications between 2016 and 2022.

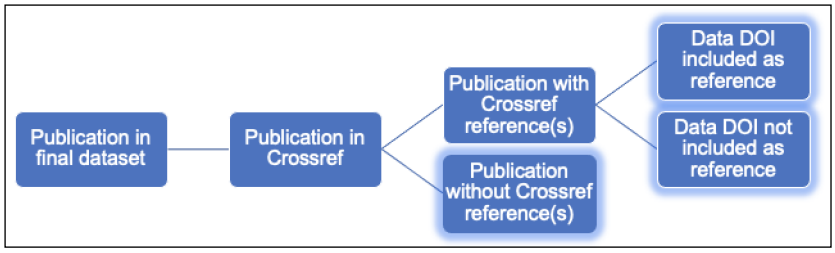

Figure 3 Overview of Crossref analysis method demonstrating how publications were subset and data DOIs were identified in Crossref structural metadata.

Similarly, information on publications in xDD and data DOIs mentioned within the full text of the publications was used to subset the data into three categories for analysis between 2016 and 2022 (Figure 4):

- Publications in xDD that mentioned the data DOI
- Publications in xDD that did not mention the data DOI
- Publications that were not in xDD

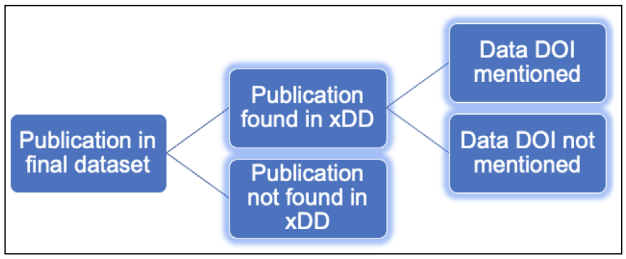

Figure 4 Overview of the xDD analysis method demonstrating how publications were subset and data DOIs mentions were identified in the full text of publications indexed in xDD.

Binomial GLMs were used to examine trends in the number of publications with data DOIs mentioned in publications found in xDD between 2016 and 2022.

We examined differences in data citations for different publishers to understand how different publisher data policies may have contributed to data access and data citation efforts. Web searches were also performed to assess publishers' publicly documented data policies.

RESULTS

CROSSREF REFERENCES

Fifty-three percent of the publications in the analysis dataset included references in their Crossref structural metadata, whereas 47% of the publications did not include references. The lack of references in the publication structural metadata does not necessarily imply that a given publication is devoid of references in its full text. However, missing references from structural metadata may point to an obstacle with the implementation of the ideal data citation workflow. The percentage of publications with indexed Crossref reference(s) fluctuated between 2016 and 2022 (Figure 5). However, this did not represent a statistically significant trend ($p = 0.41$).

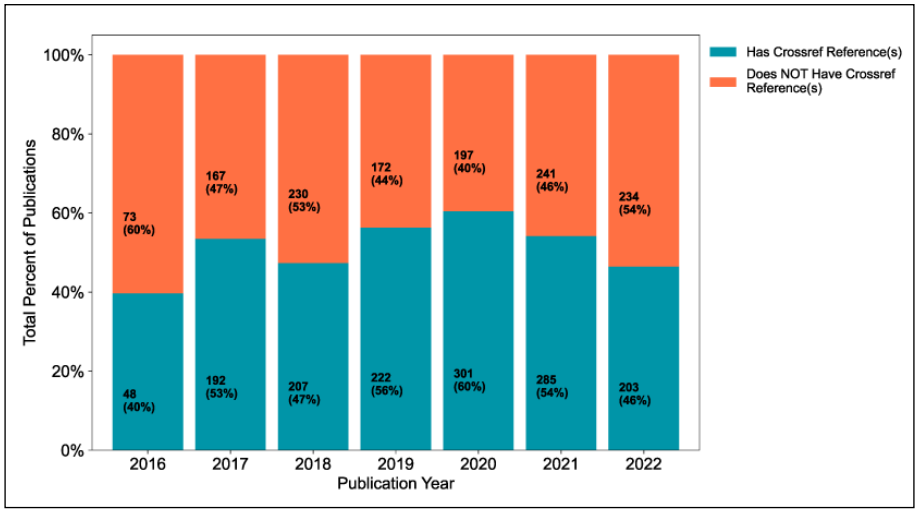

Figure 5 Percentage of publications with indexed Crossref reference(s) in their Crossref structural metadata by publication year.

Two hundred and thirty-nine publications included data DOIs within the Crossref references, which accounted for 9% of publications in the analysis dataset and 16% of publications with references included in the Crossref structural metadata (Figure 6). The percentage of publications with data DOIs included in the Crossref structural metadata's references grew between 2016 and 2022 from 4% to 30%, representing a statistically significant trend ($p < 0.001$) (Figure 6).

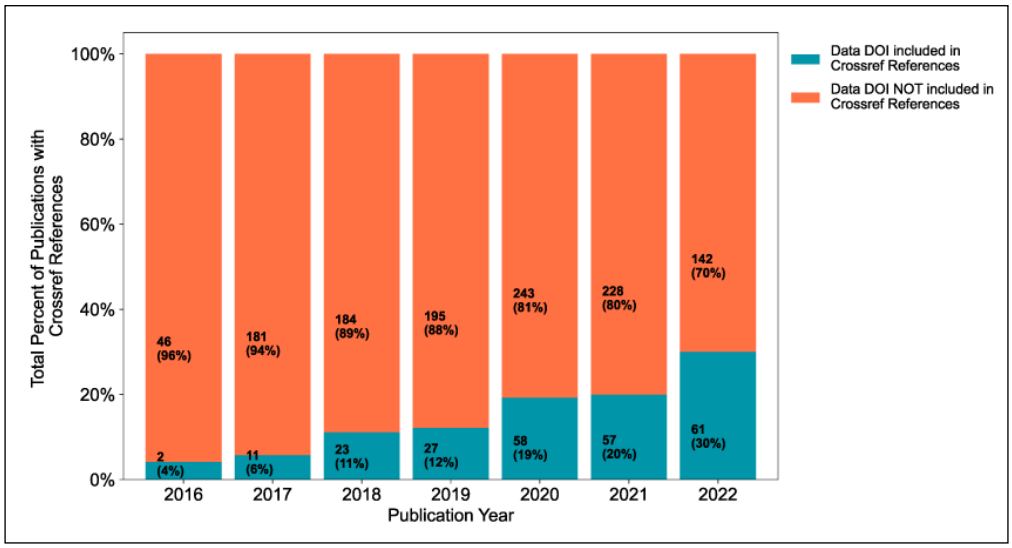

Figure 6 Percentage of publications with indexed Crossref references that cite or do not cite their associated data DOI in their Crossref structural metadata by publication year.

xDD MENTIONS

Forty-nine percent of the publications included in the analysis dataset had their full text indexed in xDD (Figure 7). Over three quarters of the publications with full text indexed in xDD (77%) mentioned their data DOI (Figure 7).

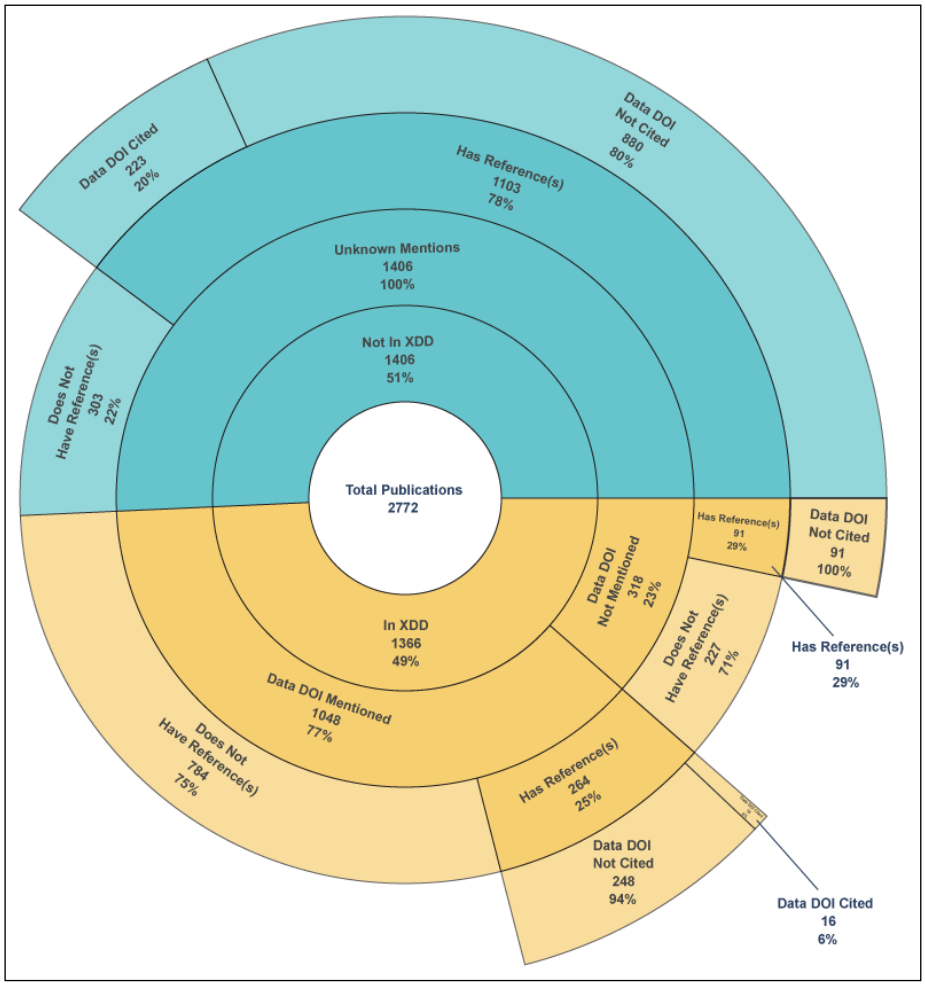

Figure 7 Publications subset by Crossref and xDD analysis method results, demonstrating the percentage of publications that mention a data DOI in their full text and/or cite a data DOI in their Crossref structural metadata references.

Between 2016 and 2022, there was an overall increase in the number of publications mentioning their data DOIs (from 63% to 82%); however, there was no statistically significant trend in the increase in number of publications per year within this period (p = 0.53) (Figure 8).

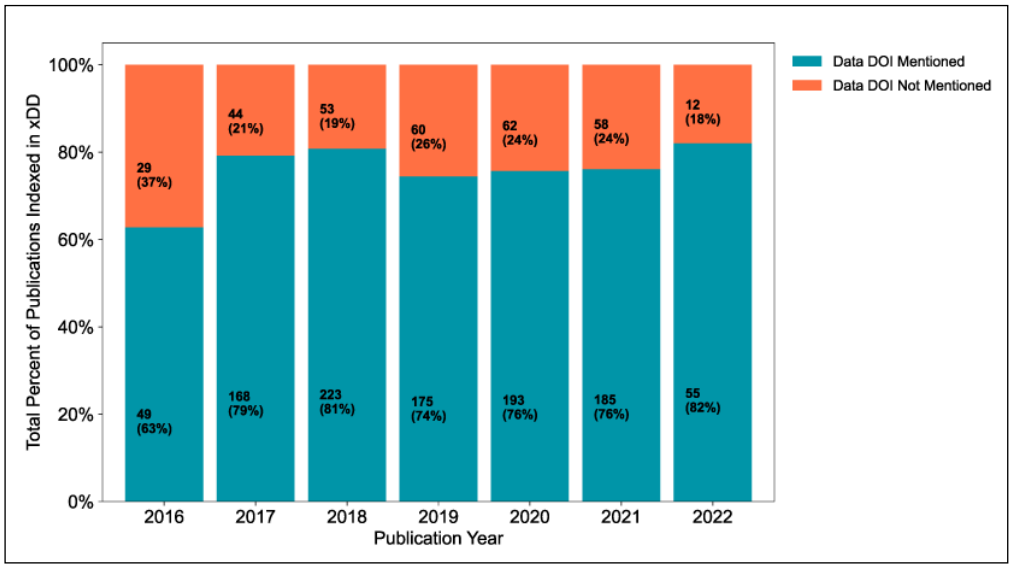

Figure 8 Percentage of publications with full text indexed in xDD with and without data DOI mentioned by publication year.

EFFECT OF PUBLISHER DATA POLICY

Fifty-eight different publishers released the 2,772 publications included in the analysis dataset. Eight out of the 58 publishers have the full text of their publications indexed in xDD. The proportion of publications found in xDD that mentioned a data DOI were analyzed by these publishers (Figure 9).

The top 10 publishers in this analysis published over 90% of the publications in the analysis dataset. The data availability policy for each of the top 10 publishers and all publishers with their full text indexed in xDD was analyzed (Table 1).

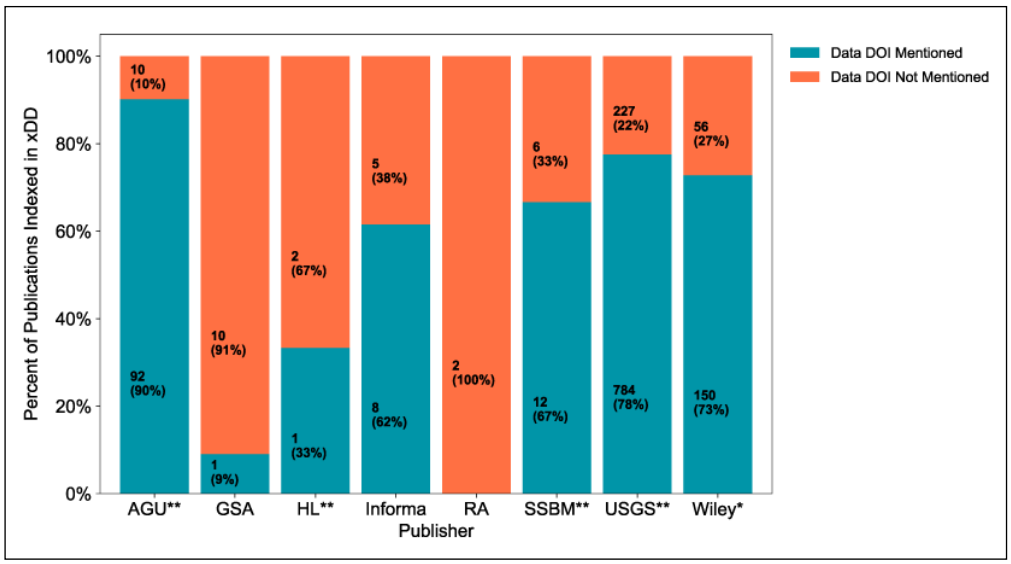

Figure 9 Percentages of publications with full text indexed in xDD that mention or do not mention their associated data DOI (see publisher abbreviations table above for publisher names). **Indicates publishers with data policies encouraging either a data availability statement or data citations in their reference lists.

The sample size by publishers varied greatly, with some having an extremely small number of publications in xDD. It may be possible to discern the significance of publisher data policies requiring or encouraging data availability statements or data citation and their impact on whether data DOIs are mentioned within the full text of publications for the publishers with smaller numbers of publications in xDD within the analysis dataset by contacting individual publishers directly. Yet, based on the criteria selected and the methodology used, it was not possible to link the data policies to the results in this analysis for the publishers with small sample sizes of publications in xDD. However, publishers with larger sample sizes (i.e., AGU, USGS, Wiley) in the analysis dataset, all had some version of data policy (Table 1), and more than 70% of their publications mentioned data DOIs.

PUBLISHER	NUMBER OF PUBLICATIONS IN ANALYSIS DATASET	DATA AVAILABILITY STATEMENTS	DATA CITATIONS IN REFERENCES LIST	LINK TO POLICY
Regional Euro-Asian Biological Invasions Centre Oy (REABIC)	29	Not Mentioned	Not Mentioned	None Found
Oxford University Press (OUP)	34	Not Mentioned*	Not Mentioned	https://academic.oup.com/pages/open-research/research-data
Frontiers Media SA	49	Required	Required	https://www.frontiersin.org/guidelines/policies-and-publication-ethics
American Chemical Society (ACS)	58	Encouraged*	Encouraged*	https://publish.acs.org/publish/data_policy
Public Library of Science (PLoS)	69	Required	Encouraged	https://journals.plos.org/plosone/s/data-availability
MDPI	132	Required	Not Mentioned	https://www.mdpi.com/ethics
American Geophysical Union (AGU)	135	Required	Required	https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors
Springer Science and Business Media LLC (SSBM)	234	Required	Not Mentioned	https://www.springer.com/gp/editorial-policies/data-availability-statement
Wiley	521	Encouraged*	Encouraged*	https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html
U.S. Geological Survey (USGS)	1237	Not Mentioned	Required	https://www.usgs.gov/office-of-science-quality-and-integrity/fundamental-science-practices-fsp-guide-data-releases-or

Table 1 Information on data policies for top 10 publishers of publications in analysis dataset and publishers with full text indexed in xDD. *For publishers with different data availability policy levels, the most lenient policy level is documented.

Eight of the top ten publishers included references in their Crossref structural metadata (Table 2). The analysis showed that the USGS and Regional Euro-Asian Biological Invasions Centre did not send references to Crossref between 2016 and 2022. Out of all the publishers, 18 (31%) have not sent any references to Crossref, seven (12%) have sent some references, and 33 (57%) have sent references for all of their publications.

PUBLISHER	PUBLICATIONS WITH INDEXED REFERENCES	PUBLICATIONS WITHOUT INDEXED REFERENCES
American Chemical Society (ACS)	58	0
American Geophysical Union (AGU)	135	0
Frontiers Media SA	49	0
MDPI	131	1
Oxford University Press (OUP)	34	0
Public Library of Science (PLoS)	69	0
Regional Euro-Asian Biological Invasions Centre Oy (REABIC)	0	29
Springer Science and Business Media LLC (SSBM)	234	0
U.S. Geological Survey (USGS)	0	1,236
Wiley	517	4

Table 2 The number of publications with and without indexed references for each of the top 10 publishers.

Numerous publications released by the top 10 publishers that contained references within the Crossref structural metadata did not include data DOIs within the 'doi' element (Figure 10). Publishers that require or encourage data citations in the reference section of their publications through data policies had a lower proportion of publications with data DOIs in their Crossref structural metadata (e.g., American Geophysical Union (AGU) and Wiley) compared to publishers that do not require or encourage data citations in the reference section of their publications (e.g., MDPI and Springer Science and Business Media LLC (SSBM)). The results also indicate that SSBM (45%) and MDPI AG (41%) released the largest percentage of publications with data DOIs

included as references within the Crossref structural metadata. Missing data DOIs from the 'doi' element in Crossref structural metadata did not necessarily mean that a reference to the data was not made in the references section of the paper or as unstructured text in the Crossref structural metadata. Publishers with publications within the analysis dataset included data references in the Crossref structural metadata in various ways:

- Data DOI listed along with all citation fields (e.g., title, authors) in 'unstructured' element in Crossref references
- Data reference included in Crossref references without the DOI

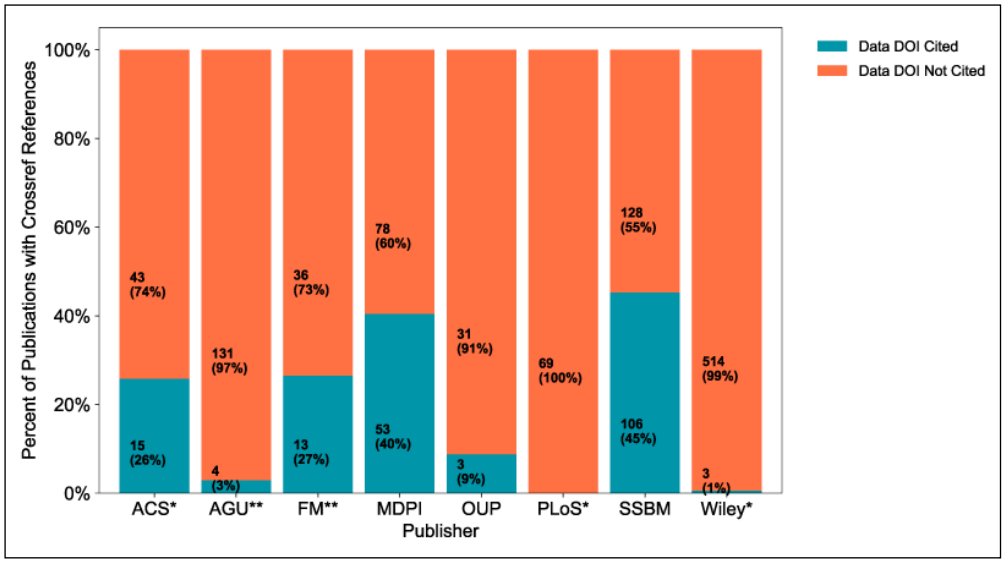

Figure 10 Percentages of publications with data DOIs cited and not cited in the publication's Crossref structural metadata for the eight out of the top ten publishers with Crossref references (see publisher abbreviations table above for publisher names). **Indicates publishers with data policies requiring data citations in their reference lists. *Indicates publishers with data policies encouraging data citations in their reference lists.

DISCUSSION

This assessment of data DOI mentions and citations within scholarly works and associated Crossref structural metadata provides insight into the implementation of the ideal data citation workflow for USGS authored publications. With over 2,000 publications analyzed, the analysis dataset provided a sample of USGS scholarly works between 2016 and 2022 expected to have data citations for known USGS data DOIs. This analysis revealed that not all USGS researchers have included a DOI for data within the references of their publications. However, a considerable portion of USGS researchers (77%) have included data DOIs in their publications, at least for the publications that were indexed in xDD (Figure 7). These data DOI mentions could be found anywhere within the publication, not only in the reference list. Given current methods using Crossref and DataCite structural metadata to track citations, it was difficult to assess how the data DOIs were being referenced within publications (within the reference list, a data availability statement, or within the body of the publication). Despite a high percentage (77% of publications in xDD) of data DOI mentions (Figure 8), there is still work, such as policy updates, outreach campaigns, and adoption of consistent reference sharing methods, that could be done to ensure that USGS researchers are meeting USGS policy requiring that publications reference their data (USGS OSQI 2017; USGS OSQI 2021a; USGS OSQI 2021b).

Many research institutions such as government agencies and universities have embraced the movement toward scientific reproducibility and transparency (Kretser et al. 2019), prompting publishers to 'adapt their workflows to enable data citation practices and provide tools and guidelines that improve the implementation process for authors and editors, and relieve stress points around compliance' (Cousijn et al. 2018). The addition of USGS Survey Manual Chapter 1100.2 (USGS OSQI 2021b; USGS OSQI 2021a) aims to support researchers through the implementation of procedures to verify data are cited in USGS series publications during the editorial review process. Hardwicke et al. (2018), suggest that this type of implementation of dedicated staff and resources geared towards assessing data citations, has the potential to improve policy compliance and ensure that data are cited properly. Given that the USGS Survey Manual Chapter 1100.2 was released in 2021, future analysis could determine if the Survey Manual is helping to increase the number USGS data citations. Regardless of this undertaking

by the USGS or similar efforts among research organizations, other publishers of scientific content may not incorporate this step in their editorial process. Without this level of assistance, researchers are solely responsible for ensuring that any associated data are cited properly. As Belter (2014) suggests, publishers that are not already working with researchers to ensure proper citation of data in their publications may consider becoming involved in this process to support data sharing.

Data citation outreach campaigns within organizations, such as the USGS, could be used to inform researchers about the importance and benefits of including data citations in their works, as well as how to include references to their data to maximize citation tracking efforts. Many publishers are making strides to promote the ideal data citation workflow by informing researchers about their responsibilities related to providing access to and citing their data (Table 1). Although our results do not definitively link publisher data citation policies to an increase in the occurrence of data citations in their publications, other studies (Colavizza et al. 2020) suggest this type of impact from such policies. Publishers also play a large role in ensuring that any data that researchers cite in their publications get included in the structural metadata sent to Crossref. As part of the ideal data citation workflow, publishers are strongly encouraged to send data citations to Crossref as part of their publications' structural metadata references. Publishers are responsible for maintaining structural metadata, which supply key information about publication and data relationships (Wilkinson 2022; Mooney 2011) and offer a means of programmatically tracking these relationships. Most publishers in this analysis (69%) are sending references to Crossref for all or some of their publications. Yet, there is a notable percentage of publications that did not include reference(s) in their Crossref structural metadata between 2016 and 2022 (Figure 5). These missing references suggest a breakdown in step two of the ideal data citation workflow, where publishers may not be including references in the publication DOI metadata that they send to Crossref. USGS, which is the publisher that makes up 45% of publications included in the analysis dataset, does not send any references to Crossref. The authors of this paper are working with the USGS Library and USGS SPN to develop a workflow for sending references to Crossref.

Despite these data policies and the fact that some of these publishers are sending references to Crossref, this does not necessarily translate to data DOIs appearing in the Crossref references in a consistent manner (within the 'doi' element). Crossref encourages publishers to use the 'doi' element whenever possible for more precise linking (Farley 2022). However, Crossref also states that data and software references can be included in the 'unstructured_citation' element. This approach is likely much easier for publishers to achieve, instead of parsing data and software citations in individual elements, which may be different than the process for parsing their citations for publications. However, using the 'unstructured_citation' element is less useful for data citation tracking efforts such as this analysis because the content within the element is not structured and may not always contain the data DOI. Cases where certain elements from the data citation were included (e.g., 'title') but the data DOI was excluded, were also identified. This approach is less useful for data citation tracking efforts because there is no way to find the data DOI using the Crossref metadata. AGU staff recently uncovered some issues in data citation workflows that may be partially responsible for many Crossref references not listing the data DOI in the 'doi' element (S. Stall, personal communication, July 19, 2023). They have published a preprint describing the steps publishers need to take to improve their workflows (Stall et al. 2022). Until publisher workflows are aligned with this new guidance, and for cases where the data DOI is either not captured or not easily parsed, data citation tracking efforts can be supplemented by using workflows involving literature databases such as xDD and associated tools like Publink.

xDD allows users to discover relationships between publications and data that may not be captured in the Crossref and DataCite structural metadata (Wieferich et al. 2020). Although only half of the publications in the total dataset were in xDD (Figure 7), more mentions of data DOIs were found through the xDD method than through the Crossref method. Using xDD, 38% of all publications in the dataset were identified as having mentioned the data DOI. Whereas, using the Crossref methods, only 9% of publications were identified with links to the data DOIs. By combining the Crossref and xDD methods, links to the data DOIs in 1,271 publications (46% of the analysis dataset) were identified. While the most ideal approach to finding connections between data and publications would be through DataCite and Crossref structural metadata,

it may take time for smaller publishers, such as USGS, to develop workflows to document and maintain this information. xDD can be used to discover data citation information in publications where these connections are missing in the DataCite and Crossref structural metadata. xDD also provides the means to retroactively add information about data and publication linkages to DataCite structural metadata through tools like Publink (Wieferich et al. 2020). Although xDD may not contain an all-inclusive library of all publications, it can be used in tandem with structural metadata infrastructures to inform users about relationships between publications and associated data. Advancements in these tools and infrastructures could promote more in-depth analysis of data citation practices and be used to identify gaps more clearly in resources or opportunities for data citation training.

Data accessibility is fundamental to the transparency and integrity of published research. Without clear linkages between publications and their associated data, data may be inaccessible, stifling data sharing and the reproducibility of scientific findings. Incorporation of data citations in publications allow users access to data while ensuring that researchers can track the impact of their data and receive credit for their work. The roles defined in the Make Data Count Initiative's ideal data citation workflow describe how researchers, publishers, repositories, and the scientific community can take steps to ensure data and publications are linked through data citations. Although the results of this analysis indicate that portions of the ideal citation workflow are being implemented within this subset of the scientific community, improvements can be made to fully satisfy the objective of the ideal data citation workflow. For instance, it would be beneficial to continue to encourage USGS researchers to follow publisher data-sharing policies and for publishers to consider adopting consistent reference-sharing methods with repositories. As the scientific community continues to improve data and publication linkages, coupled data citation tracking methods can offer information to further refine implementations of the ideal data citation workflow.

ABBREVIATIONS

TERM	ABBREVIATION
White House Office of Science and Technology Policy	OSTP
Office of Management and Budget	OMB
Office of Science Quality and Integrity	OSQI
U.S. Geological Survey	USGS
Trusted Digital Repository	TDR
Digital Object Identifier	DOI
USGS Fundamental Science Practices	FSP
Joint Declaration of Data Citation Principles	JDDCP
eXtract Dark Data	xDD
Application Programming Interface	API

PUBLISHER ABBREVIATIONS

PUBLISHER NAME	ABBREVIATION
American Chemical Society	ACS
American Geophysical Union	AGU
Frontiers Media SA	FM
Geological Society of America	GSA
Hindawi Limited	HL
Informa UK Limited	Informa

(Contd.)

PUBLISHER NAME	ABBREVIATION
MDPI AG	MDPI
Oxford University Press (OUP)	OUP
Public Library of Science	PLoS
Regional Euro-Asian Biological Invasions Centre Oy	REABIC
Springer Science and Business Media LLC	SSBM
U.S. Geological Survey	USGS
Wiley	Wiley

DATA ACCESSIBILITY STATEMENT

Data used to support conclusions in this study about data DOI mentions and citations within USGS authored publications are available at: Donovan, G.C., & Langseth, M.L., 2024, U.S. Geological Survey Data Citation Analysis, 2016–2022: U.S. Geological Survey data release, https://doi.org/10.5066/P9CPC9M2.

ACKNOWLEDGEMENTS

We would like to thank Max Joseph and Taylor Hunt from the University of Colorado Boulder's Earth Lab for their help developing the initial Python code used to gather data from the Crossref API. We would also like to thank Dalton Hance and Karen Ryberg for assistance in determining appropriate statistical analyses for our data and Shelley Stall for helping us understand challenges associated with getting data citations to Crossref from the publishers' perspective. Finally, we would like to thank our reviewers, Leslie Hsu, Katharine Dahm, and Daniel Wieferich, who provided invaluable feedback on this manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Grace C. Donovan orcid.org/0000-0002-6632-4564
U.S. Geological Survey, Core Science Systems, Science, Analytics, and Synthesis, Denver, Colorado 80225, US
Madison L. Langseth orcid.org/0000-0002-4472-9106
U.S. Geological Survey, Core Science Systems, Science, Analytics, and Synthesis, Denver, Colorado 80225, US

REFERENCES

Belter, C W 2014 Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE,* 9(3): e92590. DOI: https://doi.org/10.1371/journal.pone.0092590

Chamberlain, S, Maupetit, J, Peak, S, et al. 2022 *Habanero version 1.2.2*. Available at https://github.com/sckott/habanero [Last accessed 29 January 2022].

Colavizza, G, Hrynaszkiewicz, I, Staden, I, et al. 2020 The citation advantage of linking publications to research data. *PLoS ONE,* 15(4): e0230416. DOI: https://doi.org/10.1371/journal.pone.0230416

Cousijn, H, Kenall, A, Ganley, E, et al. 2018 A data citation roadmap for scientific publishers. *Scientific Data,* 5(1): 180259. DOI: https://doi.org/10.1038/sdata.2018.259.

DataCite 2022 *Welcome to DataCite.* Available at https://datacite.org/index.html [Last accessed 14 September 2022].

Donovan, G C and **Langseth, M L.** 2024 U.S. Geological Survey Data Citation Analysis, 2016–2022: U.S. Geological Survey data release. DOI: https://doi.org/10.5066/P9CPC9M2

ESIP Data Preservation and Stewardship Committee 2019 *Data Citation Guidelines for Earth Science Data, Version 2*. ESIP. DOI: https://doi.org/10.6084/m9.figshare.8441816.v1

Farley, I 2022 *References*. Available at https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/references/ [Last accessed 15 June 2023].

Gregory, K, Ninkov, A, Ripp, C, et al. 2023 Tracing data: A survey investigating disciplinary differences in data citation. *Quantitative Science Studies*, 4(3): 622–649. DOI: https://doi.org/10.1162/qss_a_00264

Huang, Y H, Rose, P W and **Hsu C N** 2015 Citing a data repository: A case study of the protein data bank. *PLoS ONE*, 10(8): e0136631. DOI: https://doi.org/10.1371/journal.pone.0136631

Irrera, O, Mannocci, A, Manghi, P, et al. 2023 tracing data footprints: Formal and informal data citations in the scientific literature. *Springer*. DOI: https://doi.org/10.1007/978-3-031-43849-3_7

Kafkas, Ş, Kim, J H and **McEntyre, J R** 2013 Database citation in full text biomedical articles. *PLoS One*, 8: e63184. DOI: https://doi.org/10.1371/journal.pone.0063184

Kretser, A, Murphy, D, Bertuzzi, S, et al. 2019 Scientific integrity principles and best practices: Recommendations from a scientific integrity consortium. *Science and Engineering Ethics*, 25(2): 327–355. DOI: https://doi.org/10.1007/s11948-019-00094-3

Lafia, S, Thomer, A, Moss, E, et al. 2023 How and why do researchers reference data? A study of rhetorical features and functions of data references in academic articles. *Data Science Journal*, 22(1): 10. DOI: https://doi.org/10.5334/dsj-2023-010

Lin, J 2016 Linking Publications to Data and Software. *Crossref Blog*. Available at https://www.Crossref.org/blog/linking-publications-to-data-and-software/#:~:text=Crossref%20and%20DataCite%20have%20partnered%20to%20provide%20automatic,research%20information%20network%20with%20full%20and%20accurate%20metadata [Last accessed 14 September 2022].

Make Data Count 2022 *Make Data Count*. Available at https://makedatacount.org/ [Last accessed 18 November 2022].

Melton, J and **Buxton, S** 2006 *Querying XML: XQuery, Xpath, and SQL/XML in context*. Elsevier Science. pp. 67–84. DOI: https://doi.org/10.1016/B978-155860711-8/50005-8

Mooney, H 2011 Citing data sources in the social sciences: Do authors do it? *Learned Publishing*, 24(2): 99–108. DOI: https://doi.org/10.1087/20110204

Parks, H, You, S and **Wolfram, D** 2018 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(11): 1346–1354. DOI: https://doi.org/10.1002/asi.24049

Parsons, M A, Duerr, R E, Jones, M B 2019 The history and future of data citation in practice. *Data Science Journal*, 18(1): 52. DOI: https://doi.org/10.5334/dsj-2019-052

Peters, S E, Ross, I A, Rekatsinas, T, et al. 2021a *xDD: A Digital Library and Cyberinfrastructure Facilitating the Discovery and Utilization of Data & Knowledge in Published Documents*. Available at https://geodeepdive.org [Last accessed on 28 December 2021].

Peters, S E, Ross, I A, Rekatsinas, T, et al. 2021b *xDD: About*. Available at https://geodeepdive.org/about.html [Last accessed on 28 December 2021].

Rittman, M 2020 *Event Data*. Available at https://www.crossref.org/services/event-data/ [Last accessed 15 June 2023].

Stall, S, Bilder, G, Cannon, M, et al. 2022 Journal production guidance for software and data citations. *ESS Open Archive*. DOI: https://doi.org/10.22541/essoar.167252601.17695321/v1

U.S. Geological Survey (USGS) 2022 *Who We Are*. Available at https://www.usgs.gov/about/about-us/who-we-are [Last accessed on 18 November 2022].

U.S. Geological Survey Data Management (USGS Data Management) 2022 *Data Citation | U.S. Geological Survey*. Available at https://www.usgs.gov/data-management/data-citation [Last accessed on 14 September 2022].

U.S. Geological Survey Office of Science Quality and Integrity (USGS OSQI) 2016 *Public Access to Results of Federally Funded Research at the U.S. Geological Survey: Scholarly Publications and Digital Data*. Available at http://sparcopen.org/wp-content/uploads/2016/04/USGS-PublicAccessPlan-APPROVED.pdf [Last accessed on 14 September 2022].

U.S. Geological Survey Office of Science Quality and Integrity (USGS OSQI) 2017 *502.8 – Fundamental Science Practices: Review and Approval of Scientific Data for Release*. Available at https://www.usgs.gov/survey-manual/5028-fundamental-science-practices-review-and-approval-scientific-data-release [Last accessed on 14 September 2022].

U.S. Geological Survey Office of Science Quality and Integrity (USGS OSQI) 2021a *Fundamental Science Practices (FSP) Guide to Data Releases with or Without a Companion Publication*. Available at https://www.usgs.gov/office-of-science-quality-and-integrity/fundamental-science-practices-fspguide-data-releases-or [Last accessed on 14 September 2022].

U.S. Geological Survey Office of Science Quality and Integrity (USGS OSQI) 2021b *1100.2 – Editorial Review of U.S. Geological Survey Publication Series Information Products*. Available at https://www.usgs.gov/survey-manual/11002-editorial-review-us-geological-survey-publication-series-information-products [Last accessed on 14 September 2022].

U.S. Geological Survey Office of Science Quality and Integrity (USGS OSQI) 2021c *1100.3 – U.S. Geological Survey Publication Series*. Available at https://www.usgs.gov/survey-manual/11003-us-geological-survey-publication-series [Last accessed on 15 June 2023].

Wieferich, D, Serna, B, Langseth, M, et al. 2020 *Publink*. U.S. Geological Survey Software Release. DOI: https://doi.org/10.5066/P92MX1NF

Wilkinson, L 2022 *About us*. Available at https://www.crossref.org/about/ [Last accessed 14 September 2022].

Wilkinson, M, Dumontier, M, Aalbersberg, I, et al. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3(1): 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Zhao, M, Yan, E and **Li, K** 2017 Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology,* 69(1): 32–46. DOI: https://doi.org/10.1002/asi.23919

TO CITE THIS ARTICLE:
Donovan, G C and Langseth, M L 2024 Are Researchers Citing Their Data? A Case Study from The U.S. Geological Survey. *Data Science Journal*, 23: 24, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2024-024

Submitted: 28 September 2023
Accepted: 08 April 2024
Published: 30 April 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.