



Obstacles to Dataset Citation Using Bibliographic Management Software

RESEARCH PAPER

KRISTINA VROUWENVELDER** 

NATALIE H. RAIA** 

ANDREA K. THOMER 

**Author affiliations can be found in the back matter of this article*

***Contributed equally to the manuscript*

 ubiquity press

ABSTRACT

Governmental, funder, and scholarly publisher mandates for FAIR and open data are pushing researchers to archive data with persistent identifiers in repositories and link datasets in journal articles. Data citations enable transparency in research and credit and impact metrics for data reuse. However, numerous adoption barriers still exist, including that bibliographic reference management software commonly used by researchers to ease the referencing process may not yet be equipped to handle datasets. This paper examines the readiness of commonly used reference management software to support researchers in importing bibliographic metadata for datasets and generating references that comply with leading practices for data citation. Using seven major reference managers and datasets sampled across 14 Earth, space, and environmental sciences repositories, we identify and analyze common errors in reference-manager-facilitated metadata capture, storage, and citation export, using quantitative content analysis to compare repository-provided recommended citations, reference manager results, and DataCite metadata records. We find that a majority of frequently used reference managers do not adequately support data citation, obstructing uptake of data citation by researchers and thereby limiting the growth of credit and incentives for data sharing and reuse. The range and scale of issues uncovered are broadly extensible and relevant to data citation across disciplines. We present actionable recommendations for reference manager, data repository, scholarly publisher, and researcher stakeholders for increasing the ease, efficiency, and accuracy of bibliographic management software-facilitated data citation.

CORRESPONDING AUTHORS:

Kristina Vrouwenvelder

American Geophysical Union,
Washington, DC, United States
kvrouwenvelder@agu.org

Natalie H. Raia

College of Information
Science, University of Arizona,
Tucson, AZ, United States
nraia@arizona.edu

KEYWORDS:

Data citation; data reuse;
scholarly communication;
bibliographic reference
management software;
information infrastructure

TO CITE THIS ARTICLE:

Vrouwenvelder, K., Raia,
N.H. and Thomer, A.K.
2025 Obstacles to Dataset
Citation Using Bibliographic
Management Software.
Data Science Journal, 24: 17,
pp. 1–16. DOI: <https://doi.org/10.5334/dsj-2025-017>

1. INTRODUCTION

In the last decade, data has increasingly become a valuable standalone research product for stakeholders in the research ecosystem including governments, funders, and publishers of scholarly journals (Nelson, 2022; Parsons et al., 2019; Piwowar, 2013; Stall et al., 2018). To encourage transparency in research and credit for reuse of shared data, several groups have released recommendations on data citation in peer-reviewed publications (e.g., Altman and King, 2007; Data Citation Synthesis Group (DCSG): Joint Declaration of Data Citation Principles, 2014; Earth Science Information Partners (ESIP) Data Preservation and Stewardship Committee (DPSC), 2019; Rauber et al., 2021). Citation of data in the Earth and space sciences has been increasing through time (Wooden et al., 2023), supported by adoption of data citation recommendations from some publishers (e.g., Copernicus Publications, n.d.; Fox et al., 2021; PLOS One, 2014; Springer Nature, n.d.). In addition to supporting transparency and reproducibility in scholarly publishing, dataset citation may lead to increased traffic to the citing journal publication (Vines, 2018). Yet barriers to data sharing still exist, including lack of incentives for sharing research beyond the peer-reviewed article, concerns over ‘scooping’, and lack of appropriate credit or attribution for reuse of shared data (Digital Science et al., 2019; Tenopir et al., 2018). Though data citation guidance exists, it remains sporadic, inconsistent, and researcher education and uptake is slow. To enable a culture of credit and attribution for data sharing, both researcher education and research infrastructure must be improved.

Practices implemented by reference management software and by data repositories can support researchers in data citation. Reference managers have great power to ‘shape’ the scholarly record by the nature of the resource types they support and the sources from which they can import metadata, but support for data citation by reference managers remains underdeveloped. Meanwhile, data repositories serve as sources of bibliographic metadata for both researchers and reference managers, but leading practices and standards supporting data citation are haphazardly implemented, hindering the cultural shift towards valuing data as a research output. In this study, we assess how well-equipped major reference managers are to support data citation. We ask:

- Do data repositories provide recommended data citations or downloadable bibliographic reference files? Is this information accurate and compliant with leading practices?
- How accurate are major reference managers in importing and exporting elements of a complete data citation from data repository web pages?
- What changes should data repositories or reference managers make to improve reference manager-produced data citations?

Through quantitative content analysis, we find that a majority of frequently used reference managers do not adequately support data citation, obstructing uptake of data citation by researchers and thereby limiting the growth of credit and incentives for data sharing and reuse. Only Zotero, Paperpile and Sciwheel accurately import at least 50% of dataset citation metadata. While we focus on reference managers’ import of datasets from the Earth, space, and environmental sciences, the range and scale of issues uncovered are broadly extensible and relevant to data citation across disciplines, particularly given the inclusion of generalist repositories in our analysis. We conclude with actionable insights for producers of reference manager software, data repositories, scholarly publishers, and researchers.

2. BACKGROUND AND PRIOR WORK

Key features of modern reference managers, in particular the pathway(s) by which bibliographic information is imported, influence the successful import of dataset metadata for citation. In the 1980s, reference managers began replacing the use of analog index cards as a method of organizing bibliographic references (Fenner et al., 2019). The first desktop reference management software required users to manually add bibliographic information into their reference library (Proske et al., 2023; Steele, 2008), but in the 2000s, web-based applications made it possible to import citation information directly from online databases. Most modern reference managers still allow users to manually add citations, but citations are primarily imported through one of two methods. The first, referred to as the ‘Plugin’ method

throughout, is a web browser plugin that enables import of bibliographic information from journal webpages, data repositories, and other online sources. The second, referred to as the ‘App’ method throughout, uses the reference manager’s desktop application to import dataset metadata via a persistent identifier lookup tool in the user interface.

As convenient and ubiquitous research tools, reference managers can directly influence citation practices. They generate citations and bibliographies in a wide range of citation styles and offer direct integration with word processors to manage reference lists automatically. Prior work on reference managers has examined usage patterns, selection priorities, and user needs (e.g., [Antonijević and Cahoy, 2014](#); [Berry et al., 2020](#); [Emanuel, 2013](#); [Hristova, 2012](#)). These and other studies ([Francese, 2013](#); [Lonergan, 2017](#); [Madhusudan, 2016](#); [McColl, 2018](#); [Speare, 2018](#)) find that the primary drivers for user selection of a reference manager are availability (e.g., institutional subscription), system compatibility, and the recommendations of colleagues, supervisors, and librarians. The ability to export citations from commonly used databases ([Lorenzetti and Ghali, 2013](#)) and the accuracy of generated citations ([Rempel and Mellinger, 2015](#)) can also influence choice; as will be examined in this work, in the case of data citation these vary widely across reference managers. Few to no studies of data citation practices using reference managers exist, a gap that is important in light of work examining disconnects between data citation principles and practices ([Borgman, 2016](#); [Parsons et al., 2019](#); [Silvello, 2018](#)).

Standardized guidelines for data citation in research publications already exist, in particular the domain-agnostic FORCE11 Joint Declaration of Data Citation Principles in 2014 ([DCSG: Joint Declaration of Data Citation Principles, 2014](#)), providing a basis for necessary bibliographic information. The Coalition on Publishing Data in the Earth and Space Sciences (COPDESS), which counts fifteen publishers among the signatories to its best practices for open data, recommends that data citations align with established guidelines from the Earth and space science community ([ESIP DPSC, 2019](#); [COPDESS, n.d.](#)). These guidelines define needed elements (metadata fields) of a complete data citation, including ‘dataset author’ or ‘creator’, ‘title’, ‘public release date’, ‘version ID’, ‘publisher’ or ‘repository’, ‘resolvable persistent identifier’, and ‘access date.’ Digital Object Identifier (DOI) authority DataCite also provides a recommendation for human-readable formatting of data citations in their metadata schema (see [Table 1](#)). We note that the ‘resource type’ field is recommended by DataCite but not included in ESIP’s data citation guidelines. ‘Resource Type’ is also identified as an important citation element to denote data and software citations for several infrastructural reasons ([Stall et al., 2023](#)). Additionally, within bibliographic management software, ‘Resource Type’ influences the types of metadata imported (described further below). Notwithstanding the broad commitment from publishers to support data citation, individual publisher recommendations and style guides for data citation still vary. Additional development of recommendations related to specialized types of data citation, such as citation of dynamic datasets, is ongoing through international groups such as FORCE11 and the Research Data Alliance, though a comprehensive overview of this work is beyond the scope of this study (e.g., [Rauber et al., 2021](#)).

THIS STUDY	DATA CITATION GUIDELINES FOR EARTH SCIENCE DATA	FORCE11, DATA CITATION IMPLEMENTATION PILOT	DATA CITE PROPERTIES
Resource Type	[N/A; ‘resource type’ is not a required concept in these guidelines]	Type	ResourceTypeGeneral
Creator	Author or Creator	Creator	Creator(s)
Title	Title	Title	Title(s)
Publication Year	Public Release Date	Publication Date	PublicationYear
Version	Version ID	Version	Version
Publisher	Repository	Data repository or Archive	Publisher
DOI	Resolvable Persistent Identifier	Dataset Identifier	DOI
Access Date	Access Date	N/A	N/A

Table 1 Mapping of metadata fields in this study to core required elements of a data citation ([ESIP DPSC, 2019](#); FORCE11’s Data Citation Implementation Pilot Project Repository Expert Group, [Fenner et al., 2019](#)) and properties available in DataCite Metadata Schema Version 4.5 ([DataCite Metadata Working Group, 2024](#)).

As providers of unique identifiers and metadata for datasets, data repositories are important actors in enabling accurate data citation. In 2019, recommendations for scholarly data repositories to enable actionable data citation, including explicit specifications for machine-readable metadata on dataset landing pages were developed by the Repositories Expert Group (part of FORCE11's Data Citation Implementation Project) and the National Institutes of Health-funded BioCADDIE project (<https://biocaddie.org>) (Fenner et al., 2019). However, recent analysis of the use of metadata standards by Earth science data repositories revealed that no single metadata standard was used by more than 42% of the 55 data facilities examined and that there is significant variation in metadata structuring and standards (Mayernik, 2016; Mayernik and Liapich, 2022). This heterogeneous landscape is a barrier to integration of data repositories and reference managers in providing standardized, accurate metadata for data citation.

Data repositories can also provide structural elements to help enable accurate data citation by users. Dataset metadata files generated by repositories, such as downloadable structured bibliographic information files (e.g., .bib files), can make it possible for users to import references into reference managers, LaTeX editors, and other software. In providing these, repositories can (in theory) ensure that the correct metadata for data citation are encoded in a standardized format recognized across common reference managers and independent of citation style. One such widely used format is the standard associated with LaTeX, usually encoded as .bib files. Using a dataset-specific template (such as the one provided in the BibLaTeX package (Kime and Wemheuer, 2024)) is essential to ensuring correct data citations are generated from .bib files. Given the extensibility of this method of standardizing dataset metadata across varied user citation tools, including both reference managers and LaTeX editors, providing these standardized templates for datasets is one action repositories can take to ensure data citation is easy and accurate (recommendation 9 in Fenner et al. 2019).

3. METHODS

To better understand how reference managers shape the scholarly record with respect to data citation, we systematically tested the dataset metadata import and export capabilities of seven widely used reference managers (Table 2). Our five-step approach is outlined below.

3.1 STEP 1: REFERENCE MANAGER, DATA REPOSITORY, AND DATASET SELECTION

All 'App' and 'Plugin' import options were tested and compared. DOIs were used as persistent identifiers in the 'App' method. We refer to 'DOI lookup' in the text when describing this functionality associated with the App method.

REFERENCE MANAGER	DESCRIPTION
EndNote	A proprietary reference manager released in 1989 and now produced by Clarivate for Windows and macOS.
Mendeley	A proprietary reference manager released in 2007 and acquired by Elsevier in 2013 for Windows, macOS, and Linux.
Paperpile	A web-based proprietary reference manager released in 2012 and produced by Paperpile, LLC.
Papers	A proprietary reference manager released in 2007 and produced by ReadCube for Windows and macOS.
RefWorks	A web-based proprietary reference manager released in 2001 and acquired by ProQuest in 2008.
Sciwheel	A web-based proprietary reference manager. Formerly called F1000 Workspace; acquired by SAGE Publishing in 2022.
Zotero	An open-source reference manager released in 2006 and managed by the non-profit Corporation for Digital Scholarship as of 2021.

Table 2 Reference manager software examined in this study.

To test reference manager-repository compatibility, we selected fourteen widely-used data repositories serving the Earth, space, and environmental sciences (Table 3). One dataset from each repository was randomly selected for use in the analysis. Datasets had to have a DOI and to be published after the ESIP data citation guidelines in 2019. Four of the selected

datasets (Dataverse, Figshare, Mendeley Data, Zenodo) were from repositories that support ‘versioning’— the ability to release new updates or versions of a dataset. A full list of datasets and data repositories considered is available at <https://doi.org/10.5281/zenodo.15179177> (Vrouwenvelder and Raia, 2025a). The web-based citation formatter CrossCite produces citations via content negotiation, allowing a researcher to input a DOI, select a desired citation style, and generate a citation. We examined the efficacy of CrossCite in producing citations for our sampled datasets and include this data in our dataset, though further discussion of it here fell outside the scope of this reference manager-focused analysis.

Table 3 Data repositories examined in this study and their scope.

DATA REPOSITORY	DESCRIPTION
Climate Data Store	Disciplinary climate data repository for the Copernicus Climate Change Service (C3S).
DataverseNO	Generalist repository for data produced by researchers at Norwegian institutions.
Dryad	Generalist repository for research data from all disciplines.
Figshare	Generalist repository for research data from all disciplines.
EarthChem (Interdisciplinary Earth Data Alliance; IEDA)	Disciplinary data repository for geoscience research (analytical data, data syntheses, models, and technical reports).
Environmental Data Initiative Data Portal (EDI)	Disciplinary data repository for environmental and ecological data.
International Federation of Digital Seismograph Networks (FDSN)	Disciplinary organization and data repository exposing seismological data from member organizations for free and open use.
Mendeley Data	Generalist repository for research data from all disciplines.
NASA Goddard Earth Sciences Data and Information Services Center (NASA GES DISC)	Disciplinary repository serving NASA's Atmospheric Composition, Water & Energy Cycles, and Climate Variability Focus Areas.
NSF National Center for Atmospheric Research Research Data Archive (NCAR)	Disciplinary data repository containing meteorological, atmospheric compositions, and oceanographic observations.
Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC)	One of NASA's Earth Observing System Data and Information System data centers, containing data on biogeochemical dynamics, ecology, and environmental processes.
PANGAEA	Data repository publishing georeferenced data from Earth systems research.
Planetary Data System (PDS)	Data repository hosting data from NASA's planetary missions, astronomical observations, and laboratory measurements.
Zenodo	Generalist repository for research data from all disciplines.

3.2 STEP 2: DESIGNATING DATA CITATION METADATA FIELDS AND RETRIEVING DATACITE METADATA

Drawing from the ESIP data citation guidelines and DataCite properties, we created a list of metadata fields needed for complete data citation and assessed the completeness and accuracy of these fields for each dataset with each reference manager import method (Table 1). For each dataset, the DataCite JSON metadata files were taken as the authoritative source against which imported metadata was compared. These files were retrieved by resolving eachDOI in DataCite Commons. The full sampling frame, data table, and DataCite JSON metadata used for analysis are shared (Vrouwenvelder and Raia, 2025a).

3.3 STEP 3: ESTABLISHING A FRAMEWORK FOR QUANTITATIVE CONTENT ANALYSIS

We established a numeric coding framework designed to quantitatively track the completeness and accuracy of (1) repository recommended citations, (2) repository-provided .bib files, (3) reference manager-facilitated metadata import, and (4) reference manager-facilitated citation export. The codes captured the completeness and accuracy of data citation metadata

fields with respect to the dataset's DataCite JSON metadata: a 0 was assigned if metadata were correct, a 1 was assigned if metadata were incorrect or incomplete, and the field was left blank ('NaN') if a metadata field was altogether missing or if no metadata were imported (Vrouwenvelder and Raia, 2025a). Errors in capitalization for author names (i.e., an author name appearing with all capital letters) were not marked as incorrect; however, errors in capitalization for dataset titles were considered incorrect, because they can meaningfully change the word (for instance, 'Compiled Mg/Ca data...' versus 'Compiled mg/ca data...' for the PANGAEA dataset in this study). For versioned datasets, the following two scenarios were accepted as 'correct' (i.e., assigned a '0' in the data table) for the 'version' metadata field: (1) no version number imported and canonical/concept (primary) DOI imported or (2) a version number and the DOI of the most current version imported. The latter option is the better practice for data citation because it unambiguously provides a link to the exact version of the dataset cited by the authors.

3.4 STEP 4: APPLYING THE CODING FRAMEWORK

Using the fourteen selected datasets and the seven selected reference managers, we applied the coding framework to analyze the completeness and accuracy of four data citation mechanisms as follows: (1) In cases where repositories provided a recommended data citation on dataset landing pages, we applied the coding framework described in Step 3. (2) In cases where repositories provided a downloadable .bib file, we downloaded the file and examined its contents, applying the coding framework described in Step 3. (3) For testing reference manager import functionality, each dataset was imported into the seven designated reference managers using the 'App' and 'Plugin' methods. Quantitative coding was applied as described in Step 3. Two conditions resulted in the assignment of a blank ('NaN') code: if a metadata field was altogether missing in a reference manager or if no metadata were imported into a given field. (4) For examining accuracy of reference manager citation exports, each dataset record in the reference manager was exported as a citation using APA 7th edition style and assigned codes as described in Step 3 (Vrouwenvelder and Raia, 2025a). To ensure inter-coder reliability, the two coding authors met weekly to review coding progress, discuss questions, and spot-check coding work for accuracy. The completed dataset was imported into Python for analysis (Notebook available in Vrouwenvelder and Raia, 2025b). The data were plotted to show counts of missing, inaccurate, and accurate data citation metadata on a repository-by-repository basis and across different reference managers. The mean correct metadata fields and standard error of the mean were calculated for each reference manager import method, averaging across all repository datasets, and for each repository, averaging across reference manager import methods.

4. RESULTS

We found significant obstacles to data citation at nearly all points along a researcher's data citation workflow. Errors or problems exist from the moment the researcher encounters a dataset on a repository webpage, during the metadata import stage (regardless of import method), and through the export of the citation from a reference manager for inclusion in a scholarly publication. Below, we organize results with this generalized data citation workflow in mind.

4.1 DO DATA REPOSITORIES PROVIDE RECOMMENDED DATA CITATIONS OR DOWNLOADABLE BIBLIOGRAPHIC REFERENCE FILES? IS THIS INFORMATION ACCURATE AND COMPLIANT WITH DATA CITATION BEST PRACTICES?

4.1.1 Data repository challenge 1: issues with repository-provided citations

Most users first interact with recommendations for appropriate data citation when downloading data from a data repository. Many data repositories, including all surveyed in this study except for NASA's PDS, already provide a recommended citation, formatted for inclusion in peer-reviewed journal articles, on dataset landing pages. Though PDS dataset pages have a 'Citation' section that contains bibliographic metadata in a table, the information does not appear as a fully formatted citation.

However, there were several issues with the repository-provided citations. First, all repository-provided citations were missing two or more recommended data citation elements (Figure 1; ESIP DPSC, 2019). No repository-provided citation contained all recommended elements; over 50% were missing more than three elements. Moreover, five of 14 repository-provided citations contained ‘incorrect’ elements that did not match DataCite metadata (Table 1; Vrouwenvelde and Raia, 2025a). The proportions of correct, missing, and incorrect metadata fields in repository-provided citations are displayed in Figure 1.

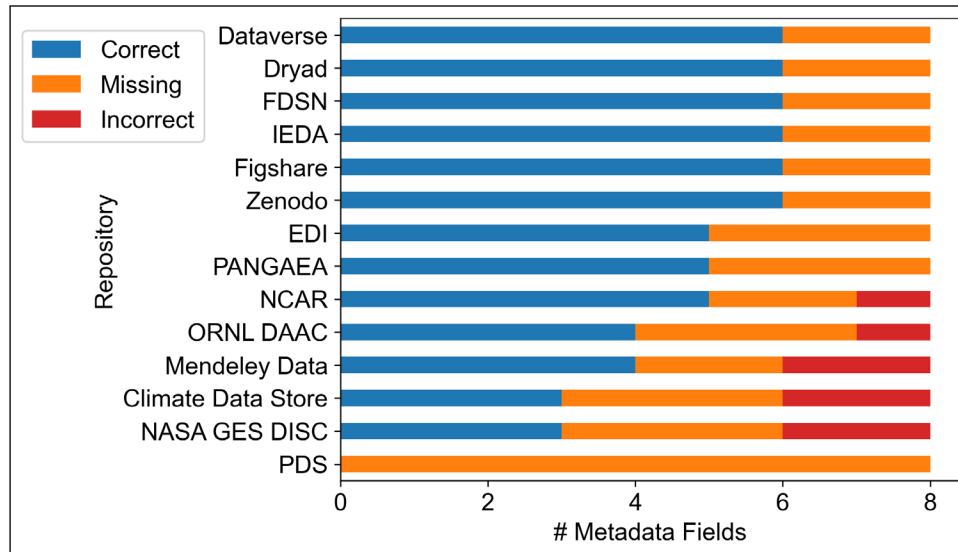


Figure 1 Counts of correct (blue), missing (orange), and incorrect (red) information in repository-provided citations available on dataset landing pages for the 14 repositories surveyed.

Discrepancies between repository names appearing in repository-provided citations and DataCite metadata were common (Table 4). In some cases, the repository-provided citation contained an acronym or abbreviation for the repository (e.g., ‘NASA GES DISC’), whereas DataCite metadata contained the full name of record (‘NASA Goddard Earth Sciences Data and Information Services Center’) and vice versa. In other cases, the name appearing in the repository-provided citation versus the DataCite metadata differed completely: for the Climate Data Store dataset, DataCite metadata listed the publisher as ‘ECMWF’ whereas the recommended citation on the Climate Data Store dataset page listed ‘Copernicus Climate Change Service (C3S) Climate Data Store (CDS)’ as the publisher.

More complicated discrepancies arose for author names (Table 5). Some cases were similar to repository names—for instance, the discrepancy between the author’s name in the repository-provided citation for Copernicus’ Climate Data Store (‘Copernicus Climate Change Service, Climate Data Store’) and the listed DataCite contributor name (‘Copernicus Climate Change Service’). In other cases, parsing of a dataset’s author/contributor names into DataCite’s ‘creator’ and ‘contributor’ properties did not reflect authorship indicated in the repository-provided citation (e.g., Mendeley Data, NASA GES DISC, Table 5, Vrouwenvelde and Raia, 2025a). This was further complicated by inconsistent usage of the ‘givenName’ and ‘familyName’ DataCite sub-properties, which contributed to reversed first and last name ordering in exported citations (Mendeley Data, PDS, Table 5, Vrouwenvelde and Raia, 2025a). In one case, two author names were convolved into one ‘creator’ property in DataCite metadata, with one author listed under the ‘givenName’ sub-property and another under ‘familyName’ (NASA GES DISC, Table 5).

DATASET DOI	PUBLISHER NAME (REPOSITORY-PROVIDED CITATION)	PUBLISHER NAME (DATACITE)
10.24381/cds.ce973f02	‘Copernicus Climate Change Service (C3S) Climate Data Store (CDS)’	‘ECMWF’
10.5065/MM6J-9282	‘Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory’	‘UCAR/NCAR- Research Data Archive’
10.3334/ORNLDAAC/1868	‘ORNL DAAC’	‘ORNL Distributed Active Archive Center’
10.5067/OMPS/OMPS_N20_NMSO2_PCA_L2_Step1.1	‘NASA GES DISC’	‘NASA Goddard Earth Sciences Data and Information Services Center’
10.17632/4dyn8f8srx.2	‘Mendeley Data’	‘Mendeley’

Table 4 Discrepancies in data publisher names provided in repository-provided citations versus DataCite metadata.

DATASET DOI	AUTHOR NAME(S) (VERBATIM, REPOSITORY- PROVIDED CITATION)	AUTHOR ('CREATOR') NAME(S) (DATACITE, ONLY RELEVANT COMPONENTS SHOWN)	ISSUE
10.24381/cds.ce973f02	'Copernicus Climate Change Service, Climate Data Store'	'creators': 'name': 'Copernicus Climate Change Service', 'affiliation': [], 'nameIdentifiers': []	Repository-provided citation and DataCite 'creator' fields not aligned.
10.17632/4dyn8f8srx.2	'Xiong, Wei; Mei, Xi; Huang, Long'	'creators': 'name': 'Wei Xiong' ... 'contributors': 'name': 'Xi Mei', 'contributorType': 'Other' 'name': 'Wei Xiong', 'contributorType': 'Other' 'name': 'Long Huang', 'contributorType': 'Other'	DataCite metadata lists a sole author and three contributors, one of which is also the author. No 'givenName' and 'familyName' sub-properties in DataCite metadata files to disambiguate first and last names.
10.5067/OMPS/OMPS_N20_NMSO2_PCA_L2_Step1.1	'Can Li, Nickolay A. Krotkov, Peter Leonard, et al.'	'creators': 'name': 'Can Li, Nickolay A. Krotkov', 'nameType': 'Personal', 'givenName': 'Nickolay A. Krotkov', 'familyName': 'Can Li', 'nameIdentifiers': [] ... 'contributors': []	DataCite metadata lists two author's names under a single creator property. Author order mis-aligned between repository-provided citation and DataCite metadata. Repository-provided citation has more authors than DataCite metadata, but no identifying names (...et al).'
10.17189/1522849	'Rodriguez-Manfredi, Jose A; de la Torre Juarez, Manuel' + 11 editor names	'creators': ['name': 'Manuel de la Torre Juarez', 'nameType': 'Personal', 'nameIdentifiers': [] 'name': 'Jose A Rodriguez-Manfredi', 'nameType': 'Personal', 'nameIdentifiers': [] 'contributors': ...'contributorType': 'Editor' ... (eleven editors from PDS site are listed, see Vrouwenvelder and Raia, 2025a)	First and second author are switched. PDS 'citation' table leaves ambiguity as to whether editors should be included in data citation.

Table 5 Discrepancies in author names in repository-provided citations versus DataCite metadata.

4.1.2 Data repository challenge 2: issues with bibliographic metadata files

We also identified issues in the bibliographic metadata files provided by some repositories for data citation. We surveyed whether the 14 repositories in this work provide .bib files for archived datasets, and if so, whether these files use a dataset-specific template (Figure 2). Because .bib files can be used in a variety of ways by the user (for instance, imported into a reference manager, or used with a style guide in a LaTeX template for a peer-reviewed publication to directly insert references), we only examined the presence or absence of necessary fields for data citation in the .bib files.

Half of the 14 repositories provided a .bib file with bibliographic information. For these seven repositories, we evaluated how many of the data citation elements were correct, missing, or incorrect with respect to DataCite metadata (Figure 3). Missing fields, followed by incorrect 'resource types', were the most common issues. No repository included a metadata field for 'Access Date' in the .bib file export. 'Version' information was missing from six of the seven repository .bib files. A 'DOI' field was missing from the .bib export for the NCAR dataset, but present in most repository .bib exports.

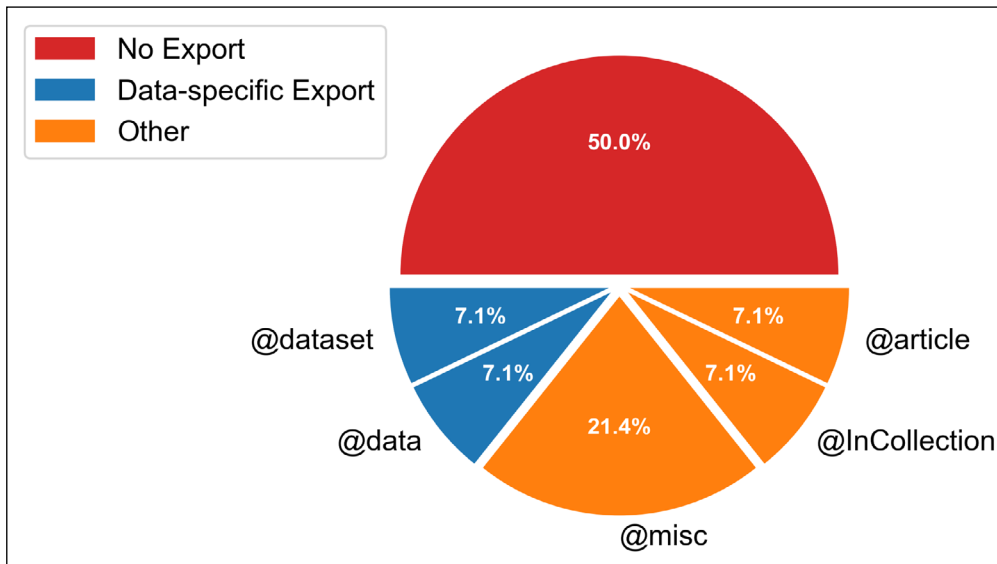


Figure 2 Percentage of surveyed repositories that provide downloadable .bib files for datasets and the types of templates employed.

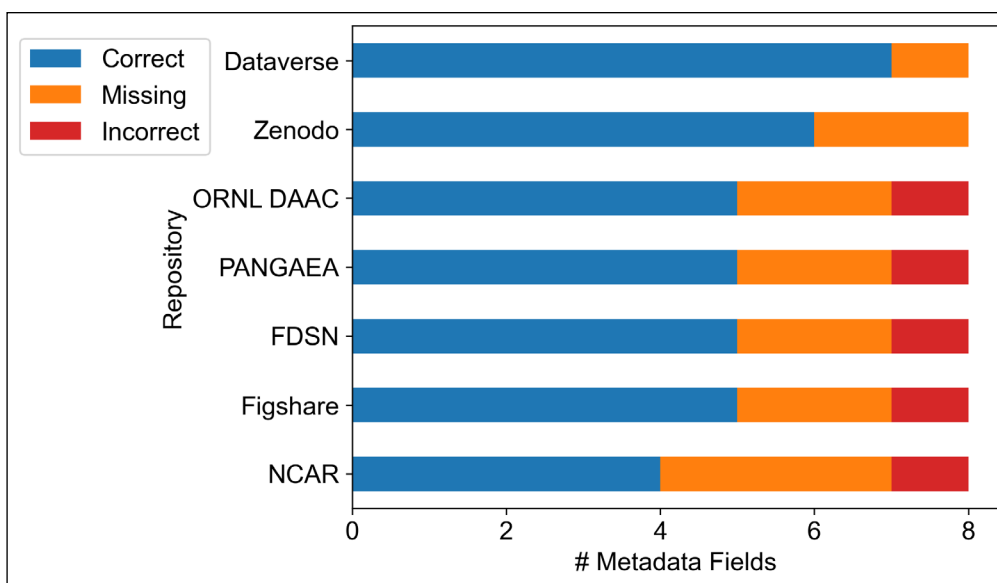


Figure 3 Counts of correct (blue), missing (orange), and incorrect (red) information present in .bib files available for download on dataset landing pages for the seven repositories providing this download option to users.

4.2 HOW ACCURATE ARE MAJOR REFERENCE MANAGERS IN IMPORTING AND EXPORTING ELEMENTS OF A COMPLETE DATA CITATION FROM DATA REPOSITORY WEB PAGES?

4.2.1 Reference manager challenge 1: insufficient reference manager resource types

We identified several barriers to data citation within reference managers themselves. First, several reference managers simply do not allow users to specify that the source is a dataset versus another resource type. All tested reference managers contained a top-level metadata field designating the type of resource or item being cited (variably appearing as an unnamed drop-down menu or called 'Item Type,' 'Reference Type,' 'Type'). Resource types include 'journal article,' 'artwork,' 'book,' 'film,' 'forum post,' 'thesis,' 'video recording,' and 'software,' for example. Four of the seven reference managers (Mendeley, RefWorks, Sciwheel, and Papers) did not have a data-specific resource type. Each resource type selection is associated with a bibliographic template containing type-specific metadata fields. Thus, the template is important, to a first order, in determining what information is ingested into the reference record and influences the final information provided in a data citation.

4.2.2 Reference manager challenge 2: missing metadata during citation import

A second challenge arose during metadata import: significant metadata was often missing or incorrect. Using both App and Plugin methods, we imported each of the 14 datasets into the seven reference managers, and surveyed the number of citation fields with correct metadata (Figure 4A). Using App methods, four of seven reference managers returned >50%

missing or incorrect citation fields; using Plugin methods, six of seven returned >50% missing or incorrect citation fields. App import methods employing a DOI identifier lookup tool were more successful than Plugin methods (Figure 4A), with two exceptions: neither Mendeley nor EndNote's App functions recognized dataset DOIs. While Zotero was the only reference manager to successfully populate >50% correct citation fields using both the App and Plugin methods, no reference manager included all recommended elements of the data citation according to current guidance (ESIP DPSC, 2019). In particular, the 'version' and 'access date' fields were commonly missed, as detailed later. Figure 4B shows the percentage of successfully imported metadata fields that are dropped during the export stage for each reference manager. Overall, most reference managers retained imported metadata through export (note that reference managers that imported more metadata successfully, such as Zotero and Paperpile, appear skewed in the export percentages). Figure 4C shows the mean correct metadata fields for each reference manager, with error bars displaying the standard error of the mean, illustrating the variance in the reference manager's ability to import correct values across the 14 repository datasets. SciWheel and Papers App methods were consistent, with the same number of correct metadata fields imported for each repository, while Mendeley and Endnote App were agnostic in that all fields were missing. To assess whether there are reference manager-agnostic trends in imported metadata completeness and accuracy, we plotted percentages of correct, incorrect, and missing imported metadata by repository and mean correct metadata counts with standard error in SI Figure 1. No clear standouts emerged. Variance between mean correct metadata fields for each repository was larger than for reference manager import methods (Figure 4C), as expected given the wide range of success rates in importing metadata of the reference managers surveyed.

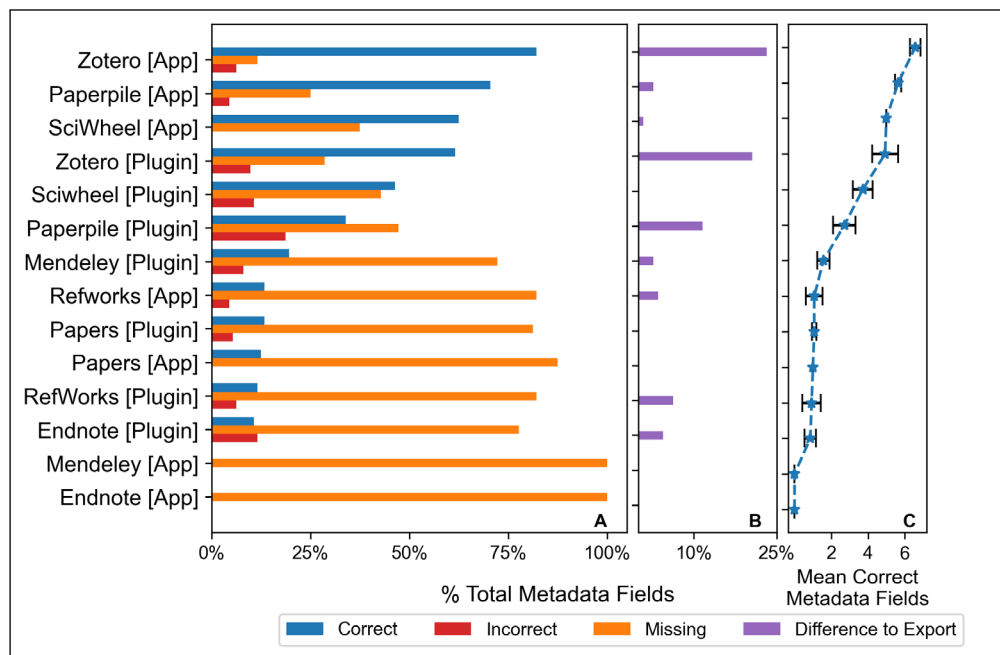


Figure 4 (A) Percentages of correct (blue), missing (orange), and incorrect (red) data citation metadata across all datasets, categorized for each reference manager and metadata import method. **(B)** Percentage of successfully imported metadata fields (correct or incorrect) that are dropped during export. **(C)** Mean correct metadata fields with standard error.

4.2.3 What information is most commonly lost in data citations?

To establish which of the eight fields of a data citation are most often erroneous or missed by reference managers, we assessed all datasets and reference managers surveyed and counted the number of correct, incorrect, and missing instances for each field in the bibliographic entries imported into each reference manager (Figure 5). The same exercise was repeated for exported data citations to examine any fields a reference manager imported, but did not include in the final citation (Figure 6). For both imported bibliographic entries and exported APA-formatted data citations, 'version' was most likely to be missing, while 'DOI' was most likely to be both present and correct; nevertheless, only 50% of imported bibliographic entries and exported APA-formatted data citations correctly included the dataset DOI. We counted an exported field as 'correct' if it agreed with the imported data citation, so while 'resource type' was the most frequently incorrect field in imported bibliographic entries, it was more often simply missing rather than incorrect in the exported APA data citations.

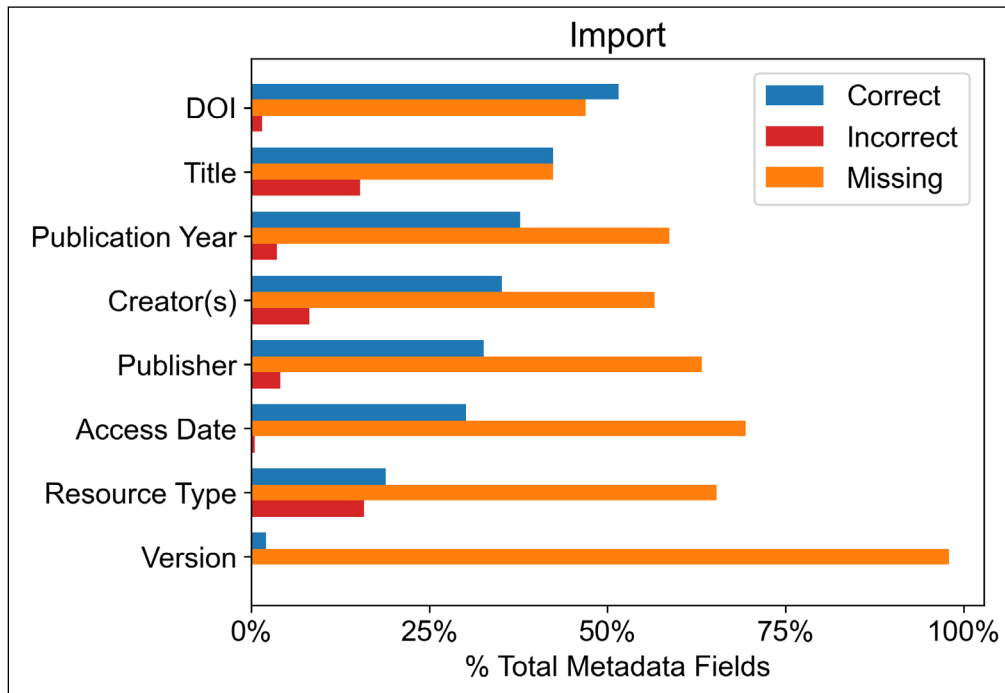


Figure 5 Percent of correct (blue), missing (orange), and incorrect (red) metadata fields for each type of metadata field across all repositories, reference managers, and import methods.

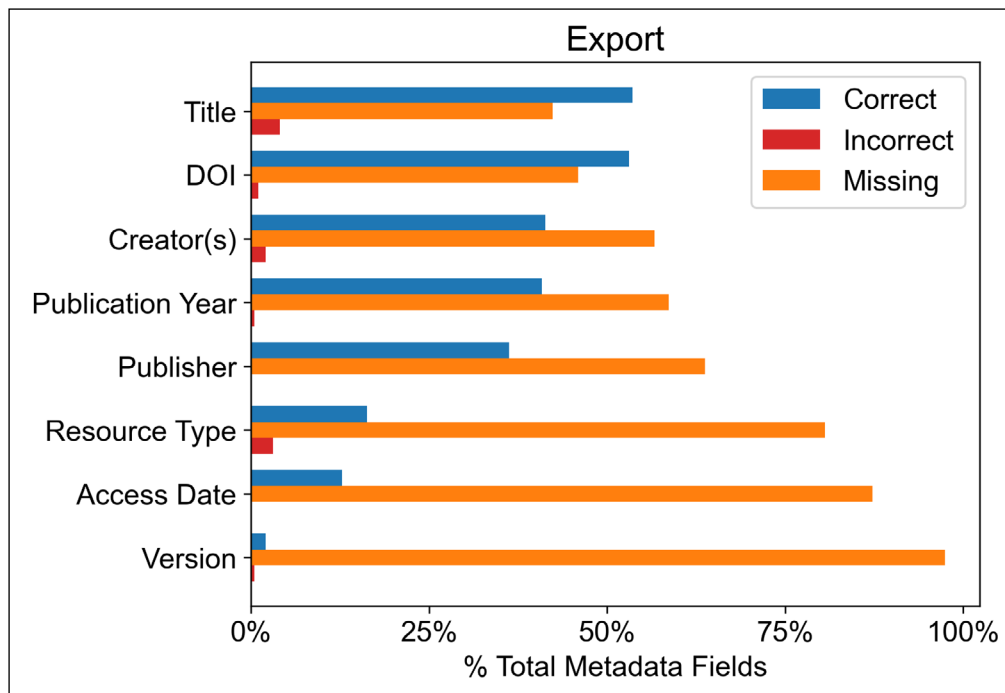


Figure 6 Percent of correct (blue), missing (orange), and incorrect (red) metadata fields for each type of metadata field across all repositories, reference managers, and export methods.

5. DISCUSSION

Collective action is required to enable a functioning data citation ecosystem. Below, we synthesize key challenges revealed through our work and provide actionable recommendations for reference manager software developers, data repositories, scholarly publishers, and researchers.

5.1 CHALLENGE 1: ACCURACY, COMPLETENESS, AND FORMAT OF REPOSITORY-PROVIDED CITATIONS

We found multiple problems with the citations offered by repositories. First, they often differ substantially from metadata in DataCite records and even from repository-provided .bib files, making consistent citation challenging or impossible for users. We also observed that repositories often presented their recommended citations in a single reference style, without explicit indication of which style is being used, and without the ability for users to choose a style appropriate to their application. While users can manually reformat citations, this still presents a barrier to use.

Another repository-provided source and possible tool for standardization of dataset bibliographic information, the .bib file, was lacking in more than 50% of the repositories surveyed. When repositories offered .bib exports, they did not always identify bibliographic information as pertaining to a dataset (e.g., @dataset entry type, [Kime and Wemheuer, 2024](#)). Instead, they often used the @misc, @article, or other .bib export types. To enable widespread use of a standardized format for sharing dataset bibliographic information, repositories should adopt dataset-specific entry types, publishers must update LaTeX style guides to correctly interpret these types, and reference managers should use .bib templates that correctly identify dataset types. Complementary work is needed to ensure that .bib templates for @dataset types include metadata recommended by groups like ESIP and FORCE11. Comparing the accuracy of metadata shared by different .bib dataset templates would be a promising area for future work.

5.2 CHALLENGE 2: REFERENCE MANAGER ACCURACY

Our analysis reveals that most reference managers have not been updated to support best practices in data citation. For instance, Paperpile has no metadata field for ‘version,’ and Mendeley, RefWorks, Sciwheel, and Papers do not include ‘data’ as a resource type. As we show in [Figures 5 and 6](#), even the field most often successfully imported—the DOI—was still missing or incorrect in 50% of the reference manager-repository pairs surveyed. As the DOI is arguably the single most important element of a data citation, this success rate is inadequate.

Other fields were more frequently missed than the DOI field, in particular ‘access date’ and ‘version.’ Though a data citation without either of these elements may be sufficient for some purposes, datasets can be dynamic in ways that publications typically are not, in particular in the case of Earth observation data and other longitudinal datasets iteratively updated over time. Versioning and access date information are essential for reproducibility and transparency for dynamic datasets or in case a dataset is later updated by the data provider. Repositories and reference managers should incorporate access date and/or version as standard elements of data citations and publishers should incorporate these elements into style guides for data citations in publications.

‘Resource type’ was the next most commonly missed field. While resource type is not an element described in the ESIP data citation guidelines ([ESIP DPSC, 2019](#)), bracketed description flags denoting dataset (and software and preprint) citations are used by many publishers and citation styles to distinguish from other scholarly content in the References section, improving both human- and machine-readability of different citation types ([Fox et al., 2021](#); [Stall et al., 2023](#)). In the APA style commonly employed in scholarly publications in the Earth and space sciences, ‘resource type’ is included as a bracketed description flag directly following the dataset title. Noting the ‘type’ of a resource is also crucial in telling software how to structure and format data citations.

5.3 WHAT REFERENCE MANAGERS BEST SUPPORT DATA CITATION?

Zotero, Paperpile and Sciwheel provided the most reliable data citations. For each of these, the App method imported bibliographic information more accurately, and with fewer missing fields, than the Plugin method. We speculate that reference manager Apps are able to pull metadata through content negotiation via DOI lookup from sources like DataCite, whereas Plugins instead scrape information from websites; however, inspection of the tools’ source code would be necessary to confirm this.

The selection of a data-specific resource type upon import is crucial to successful data citation formatting; Mendeley, Refworks, and Papers lack a data-specific resource type altogether ([Figure 4](#)). Sciwheel, too, lacks a data-specific type, but was an exception, performing in the top five. Mendeley was the reference manager with the largest discrepancy between App and Plugin methods. We attribute this to Mendeley’s lack of a data-specific resource type and the App’s inability to recognize any DataCite DOI tested. Mendeley’s Plugin successfully imported some bibliographic fields and performed best for the three datasets that were imported as a ‘journal article’ type ([Vrouwenvelder and Raia, 2025a](#)). While EndNote does have a data-specific type, the EndNote App similarly did not recognize DataCite DOIs, preventing import of bibliographic information. The Plugin was able to import some information across more fields, but fewer repository landing pages.

In addition to performing accurately, reference manager import tools must be easy to use by the researcher. Plugin methods, while performing more poorly overall, are easier to use in the course of browsing scientific literature. Plugin performance varied by repository as well as by reference manager, suggesting that both stakeholders have areas to address to improve performance.

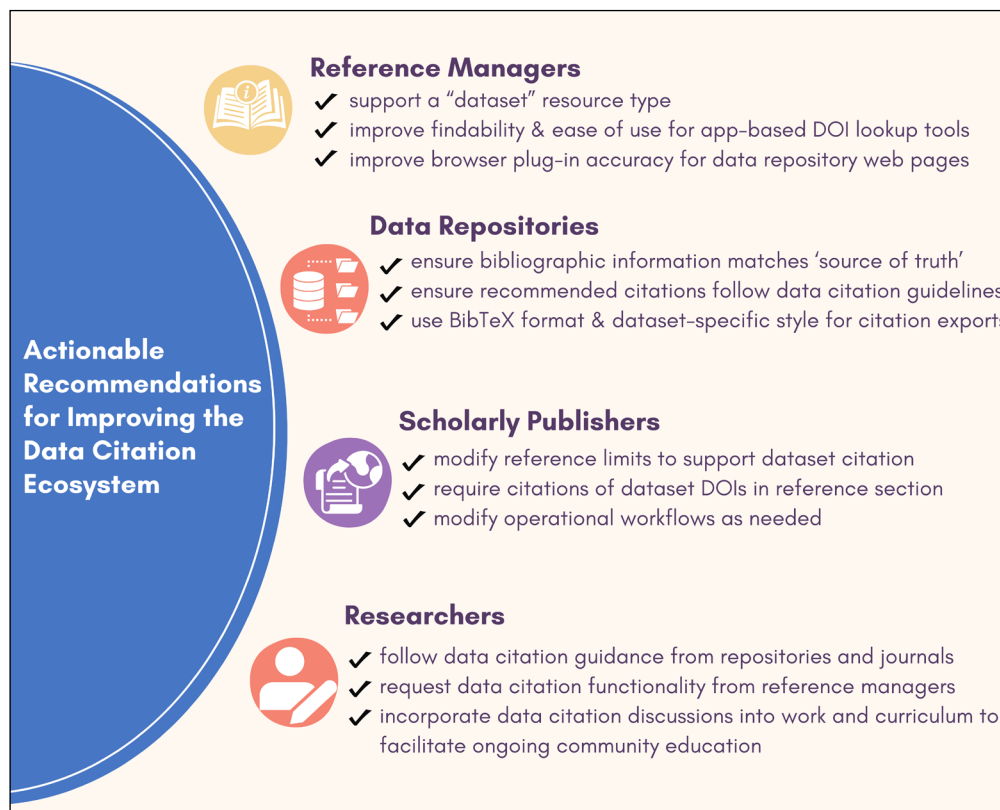


Figure 7 Actionable recommendations for improving data citation by stakeholder.

6. CONCLUSION: RECOMMENDATIONS FOR IMPROVING THE DATA CITATION ECOSYSTEM

Collaboration between reference managers, data repositories, scholarly publishers, and researchers is needed to improve the data citation ecosystem (Figure 7). While this study focuses on readiness of major reference managers for citation of datasets, similar infrastructure challenges likely exist for software citation, which in many disciplines is a newer practice than data citation.

6.1 REFERENCE MANAGERS

Significant development is needed from reference managers to support data citation. For four of the seven surveyed (Mendeley, RefWorks, Sciwheel, and Papers), a key first step is supporting data-specific resource types. Given that six of seven reference managers returned >50% missing or incorrect citation fields, Plugins need further development to accurately interpret data repository landing pages and populate bibliographic information for data citation (Figure 4).

6.2 DATA REPOSITORIES

Data repositories should ensure dataset landing pages follow best practices to enable interoperability (Fenner et al, 2019; Mayernik and Liapich, 2022). When providing recommended citations on dataset landing pages, repositories should ensure these citations follow community guidelines (ESIP DPSC, 2019). Recommended citations should state the citation style used (e.g., APA) and should ideally accommodate selection of popular citation styles. If possible, repositories should provide a standard bibliographic export (e.g., .bib) in a data-specific style, such as the @dataset type maintained by BibLaTeX. Above all, data repositories providing recommended citations and citation export metadata should ensure that these are up-to-date with the ‘source of truth’ for bibliographic metadata (i.e., DataCite or other DOI authority; see Figures 1 and 3, Table 4). Repository capacity to implement these recommendations may vary according to their website design and management.

6.3 SCHOLARLY PUBLISHERS

The addition of datasets to article reference lists is necessary to ensure data citations are indexed. It is well-documented that references in supplemental materials, including datasets, have a low chance of being indexed (Fricke et al., 2021; Seeber, 2008). For some journals, this new paradigm requires re-evaluation of reference limits and revision of author guidance and may require modifications to internal publication workflows.

6.4 RESEARCHERS

Though individual researchers play a critical role in driving data citation, adopting leading practices when using in-development infrastructure is challenging. Researchers should request data citation guidance and support from repositories, journals, and producers of bibliographic management software and be vocal about their struggles and successes with colleagues and mentees. Educators should, where possible, incorporate data citation practices into curricula (for instance, in expectations for course papers). The results of this work may serve as a guide to researchers in choosing bibliographic management software best equipped to support their data citation needs.

6.5 SOCIETAL IMPACT OF A FUNCTIONING DATA CITATION ECOSYSTEM

The potential societal benefits of a functioning citation ecosystem demand urgency with respect to progress on the above recommendations (and the numerous ones that precede it, including Fenner et al., 2019 and Mayernik and Liapich, 2022). Accurate data citation increases research transparency and reproducibility, enhances impact assessment, decreases the ‘time to science,’ and is essential for realizing the vision of FAIR data. This is particularly critical for research addressing societal grand challenges such as climate change, natural disaster response, and public health. New technologies (e.g., AI tools) may help address challenges like the ones surfaced in this work and enhance research efficiency, but ultimately the recognition of data as a high-value, standalone research product requires a transformation of existing scholarly communication infrastructure. Our study shows this transformation is incomplete. Given the urgency of societal and academic community needs and benefits, and the actionable recommendations we identify, we call for continued efforts to incentivize and support the community in making forward progress.

DATA ACCESSIBILITY STATEMENT

The full sampling frame, data table, .bib files, and DataCite JSON metadata and the computational notebook used for analysis are available at <https://doi.org/10.5281/zenodo.15179177> and <https://doi.org/10.5281/zenodo.15179215>.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary file 1.** Repository comparison across reference managers. DOI: <https://doi.org/10.5334/dsj-2025-017.s1>

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

K.V. and N.R. conceptualized the work, collected data, discussed methodology, and wrote and edited all drafts. K.V. developed the code and analysis and produced figures. A.T. contributed to conceptualization, methodology and reviewed drafts.

K.V. and N.R. contributed equally to the manuscript.

Kristina Vrouwenvelder  orcid.org/0000-0002-5862-2502

American Geophysical Union, Washington, DC, United States

Natalie H. Raia  orcid.org/0000-0003-4939-3282

College of Information Science, University of Arizona, Tucson, AZ, United States

Andrea K. Thomer  orcid.org/0000-0001-6238-3498

College of Information Science, University of Arizona, Tucson, AZ, United States

REFERENCES

- Altman, M.** and **King, G.** (2007) 'A proposed standard for the scholarly citation of quantitative data', *D-lib Magazine*, 13(3/4).
- Antonijević, S.** and **Cahoy, E.S.** (2014) 'Personal library curation: An ethnographic study of scholars' information practices', *portal: Libraries and the Academy*, 14(2), pp. 287–306. Available at: <https://doi.org/10.1353/pla.2014.0010>
- Berry, T.U.** et al. (2020) 'Understanding user needs for citation management program support', *Public Services Quarterly*, 16(2), pp. 71–82. Available at: <https://doi.org/10.1080/15228959.2019.1687070>
- Borgman, C.L.** (2016) 'Data citation as a bibliometric oxymoron', *Theories of informetrics and scholarly communication*, 93, p. 94. Available at: <https://doi.org/10.1515/9783110308464-008>
- Coalition for Publishing Data in the Earth and Space Sciences** (n.d.) 'COPDESS Suggested Author Instructions and Best Practices for Journals'. Available at: <https://copdess.org/copdess-suggested-author-instructions-and-best-practices-for-journals/>.
- Copernicus Publications** (no date) *Data Policy*. Available at: https://publications.copernicus.org/services/data_policy.html (Accessed: 28 June 2024).
- Data Citation Synthesis Group** (2014) *Joint Declaration of Data Citation Principles*. Force11. Available at: <https://doi.org/10.25490/A97F-EGYK>
- DataCite Metadata Working Group** (2024) 'DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.5'. Available at: <https://doi.org/10.14454/G8E5-6293>
- Digital Science** et al. (2019) *The State of Open Data Report 2019*. report. Digital Science. Available at: <https://doi.org/10.6084/m9.figshare.9980783.v2>
- Emanuel, J.** (2013) 'Users and citation management tools: Use and support', *Reference Services Review*, 41(4), pp. 639–659. Available at: <https://doi.org/10.1108/RSR-02-2013-0007>
- ESIP Data Preservation and Stewardship Committee** (2019) 'Data Citation Guidelines for Earth Science Data, Version 2'. Earth Science Information Partners. Available at: <https://doi.org/10.6084/m9.figshare.8441816.v1>
- Fenner, M.** et al. (2019) 'A data citation roadmap for scholarly data repositories', *Scientific Data*, 6(1), p. 28. Available at: <https://doi.org/10.1038/s41597-019-0031-8>
- Fox, P.** et al. (2021) 'Data and Software Sharing Guidance for Authors Submitting to AGU Journals'. Zenodo. Available at: <https://doi.org/10.5281/ZENODO.5124741>
- Francesse, E.** (2013) 'Usage of reference management software at the University of Torino', *JLIS* [Preprint], (2). Available at: <https://doi.org/10.4403/jlis.it-8679>
- Fricke, S., Enslow, E.** and **Shipman, S.** (2021) 'Access to Supplemental Journal Article Materials', *The Serials Librarian*, 80(1–4), pp. 85–96. Available at: <https://doi.org/10.1080/0361526X.2021.1883596>
- Hristova, M.** (2012) 'RefWorks usage patterns: Exploring the first four semesters of use by faculty, graduate students, and undergraduates', *Internet Reference Services Quarterly*, 17(2), pp. 45–64. Available at: <https://doi.org/10.1080/10875301.2012.720654>
- Kime, P.** and **Wemheuer, M.** (2024) 'BibLaTeX'. The LaTeX Project. Available at: <https://ctan.org/pkg/biblatex>
- Lonergan, N.** (2017) 'Reference management software preferences among liberal arts faculty', *Reference Services Review*, 45(4), pp. 584–595. Available at: <https://doi.org/10.1108/RSR-06-2017-0024>
- Lorenzetti, D.L.** and **Ghali, W.A.** (2013) 'Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability', *BMC Medical Research Methodology*, 13(1), p. 141. Available at: <https://doi.org/10.1186/1471-2288-13-141>
- Madhusudan, M.** (2016) 'Use of online citation tools by students and research scholars of Department of Library and Information Science, University of Delhi', *DESIDOC Journal of Library & Information Technology*, 36(3). Available at: <https://doi.org/10.14429/djlit.36.3.9428>
- Mayernik, M.S.** (2016) 'Research data and metadata curation as institutional issues', *Journal of the Association for Information Science and Technology*, 67(4), pp. 973–993. Available at: <https://doi.org/10.1002/asi.23425>
- Mayernik, M.S.** and **Liapich, Y.** (2022) 'The role of metadata and vocabulary standards in enabling scientific data interoperability: A study of Earth System Science Data Facilities', *Journal of eScience Librarianship*, 11(2). Available at: <https://doi.org/10.7191/jeslib.619>

- McColl, R.** (2018) 'Leveraging Citation Management Software (CMS) for broad library impact', *Pennsylvania Libraries: Research & Practice*, 6(2), pp. 93–105. Available at: <https://doi.org/10.5195/palrap.2018.189>
- Nelson, A.** (2022) 'Memorandum for the heads of executive departments and agencies: Ensuring free, immediate, and equitable access to federally funded research'. Available at: <https://rosap.nsl.bts.gov/view/dot/65799>
- Parsons, M.A., Duerr, R.E. and Jones, M.B.** (2019) 'The history and future of data citation in practice', *Data Science Journal*, 18(1). Available at: <https://doi.org/10.5334/dsj-2019-052>
- Piwowar, H.** (2013) 'Value all research products', *Nature*, 493(7431), pp. 159–159. Available at: <https://doi.org/10.1038/493159a>
- PLOS One** (2014) 'PLOS One – data availability', 3 March. Available at: <https://journals.plos.org/plosone/s/data-availability>.
- Proske, A., Wenzel, C. and Queitsch, M.B.** (2023) 'Reference Management Systems', in O. Kruse et al. (eds) *Digital Writing Technologies in Higher Education: Theory, Research, and Practice*. Cham: Springer International Publishing, pp. 215–230. Available at: https://doi.org/10.1007/978-3-031-36033-6_14
- Rauber, A. et al.** (2021) 'Precisely and persistently identifying and citing arbitrary subsets of dynamic data', *Harvard Data Science Review*, 3(4). Available at: <https://doi.org/10.1162/99608f92.be565013>
- Rempel, H.G. and Mellinger, M.** (2015) 'Bibliographic management tool adoption and use a qualitative research study using the UTAUT model', *Reference & User Services Quarterly*, 54(4), p. 43. Available at: <https://doi.org/10.5860/rusq.54n4.43>
- Seeber, F.** (2008) 'Citations in supplementary information are invisible', *Nature*, 451(7181), pp. 887–887. Available at: <https://doi.org/10.1038/451887d>
- Silvello, G.** (2018) 'Theory and practice of data citation', *Journal of the Association for Information Science and Technology*, 69(1), pp. 6–20. Available at: <https://doi.org/10.1002/asi.23917>
- Speare, M.** (2018) 'Graduate student use and non-use of reference and PDF management software: An exploratory study', *The Journal of Academic Librarianship*, 44(6), pp. 762–774. Available at: <https://doi.org/10.1016/j.acalib.2018.09.019>
- Springer Nature** (no date) 'Research Data Policy'. Available at: <https://www.springernature.com/gp/authors/research-data-policy/data-policy-faqs> (Accessed: 28 June 2024).
- Stall, S. et al.** (2018) 'Enabling FAIR and open data in Earth and space sciences publications', in *EGU General Assembly Conference Abstracts*. (EGU General Assembly Conference Abstracts), p. 12238.
- Stall, S. et al.** (2023) 'Journal production guidance for software and data citations', *Scientific Data*, 10(1), p. 656. Available at: <https://doi.org/10.1038/s41597-023-02491-7>
- Steele, S.E.** (2008) 'Bibliographic citation management software as a tool for building knowledge', *Journal of Wound Ostomy & Continence Nursing*, 35(5), p. 463. Available at: <https://doi.org/10.1097/01.WON.0000335956.45311.69>
- Tenopir, C. et al.** (2018) 'Research data sharing: Practices and attitudes of geophysicists', *Earth and Space Science*, 5(12), pp. 891–902. Available at: <https://doi.org/10.1029/2018EA000461>
- Vines, T.** (2018) *What's Up with Data Citations? – The Scholarly Kitchen*. Available at: <https://scholarlykitchen.sspnet.org/2018/05/28/whats-up-with-data-citations/> (Accessed: 8 July 2024).
- Vrouwenvellder, K. and Raia, N.** (2025a) 'Reference manager data citation analysis'. Zenodo. Available at: <https://doi.org/10.5281/zenodo.15179177>
- Vrouwenvellder, K. and Raia, N.** (2025b) 'Reference Manager Data Citation Analysis Workbook'. Zenodo. Available at: <https://doi.org/10.5281/zenodo.15179215>
- Wooden, P. et al.** (2023) *AGU Publications Opens Science: Making Science Accessible and Equitable*, Eos, 104. Available at: <https://doi.org/10.1029/2023EO235035>

TO CITE THIS ARTICLE:

Vrouwenvellder, K., Raia, N.H. and Thomer, A.K.
 2025 Obstacles to Dataset Citation Using Bibliographic Management Software.
Data Science Journal, 24: 17, pp. 1–16. DOI: <https://doi.org/10.5334/dsj-2025-017>

Submitted: 14 December 2024

Accepted: 24 April 2025

Published: 20 May 2025

COPYRIGHT:

© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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