






# Initial insight into three modes of data sharing: Prevalence of primary reuse, data integration and dataset release in research articles

Yukiko Sakai <sup>1,2\*</sup>, Yosuke Miyata <sup>2</sup>, Keiko Yokoi <sup>3</sup>, Yuqing Wang <sup>4</sup>, and Keiko Kurata <sup>2</sup>

<sup>1</sup>Center for General Education, Teikyo University, Tokyo, Japan

<sup>2</sup>Faculty of Letters, Keio University, Tokyo, Japan

<sup>3</sup>Policy Research, The Tokyo Foundation for Policy Research, Tokyo, Japan

<sup>4</sup>Graduate School of Letters, Keio University, Tokyo, Japan

ORCID:

Y. Sakai: [0000-0003-2795-8297](https://orcid.org/0000-0003-2795-8297)

Y. Miyata: [0000-0002-5239-5396](https://orcid.org/0000-0002-5239-5396)

K. Yokoi: [0000-0002-2546-347X](https://orcid.org/0000-0002-2546-347X)

Y. Wang: [0000-0002-3945-8620](https://orcid.org/0000-0002-3945-8620)

K. Kurata: [0000-0002-8486-2438](https://orcid.org/0000-0002-8486-2438)

\*Corresponding author: Yukiko Sakai, Center for General Education, Teikyo University, 359 Otsuka Hachioji-shi, Tokyo 192-0395, Japan.  
E-mail: [ysakai@main.teikyo-u.ac.jp](mailto:ysakai@main.teikyo-u.ac.jp); [yukiko@a2.keio.jp](mailto:yukiko@a2.keio.jp)

**Abstract:** While data sharing has received research interest in recent times, its real status remains unclear, owing to its ambiguous concept. To understand the current status of data sharing, this study examined primary reuse, data integration, and dataset release as the actual practices of data sharing. A total of 963 articles, chosen from those published in 2018 and registered in the Web of Science global citation database, were manually checked. Existing data were reused in the mode of data integration (13.3%) as frequently as they were for the mode of primary reuse (12.1%). Dataset release was the least common mode (9.0%). The results show the variation in data sharing and indicate the need for standardization of data description in articles based on thorough registration and expansion in public data archives to close the loop that results in the virtuous cycle of research data.

**Keywords:** data reuse, data sharing, research data

## INTRODUCTION

Data sharing has been prioritized in scholarly communication recently and studies investigating its prevalence have been conducted. However, the explicit status of data sharing remains vague because of the ambiguous concept of data (Borgman, 2015, p. 21) and data sharing.

Data sharing actually contains diverse modes, which can be divided into two main aspects: data release and data reuse. Data release is an act documenting research data for one's own study

as well as for that of others for further reuse (Borgman, 2015, p. 13). Data reuse means using existing research data for new research. Whether data reuse includes data use by the same author in subsequent studies depends on researchers.

Although data sharing comprises these two aspects, most studies, especially earlier ones, focused on data release compared with data reuse (Kim, 2022, p. 710). For example, in a worldwide researcher survey study, the DataONE project reported that 90.5% of scientists have already made at least some data available for reuse by other scientists (Tenopir et al., 2015). Some studies, including the DataOne project, have repeated analysis mostly of survey responses to understand the positive and negative factors affecting data release behaviour among researchers (Kim, 2022; Tenopir et al., 2020; The State of Open Data, 2022).

Compared with focus on data releasing, less attention has been paid to data reuse among data sharing research

A preliminary study based on this report was presented as 'Data Integration as the Major Mode of Data Reuse' in a poster session at the 83rd Annual Meeting of the Association for Information Science and Technology on 28 October 2020. doi.org/10.1002/pra2.391.

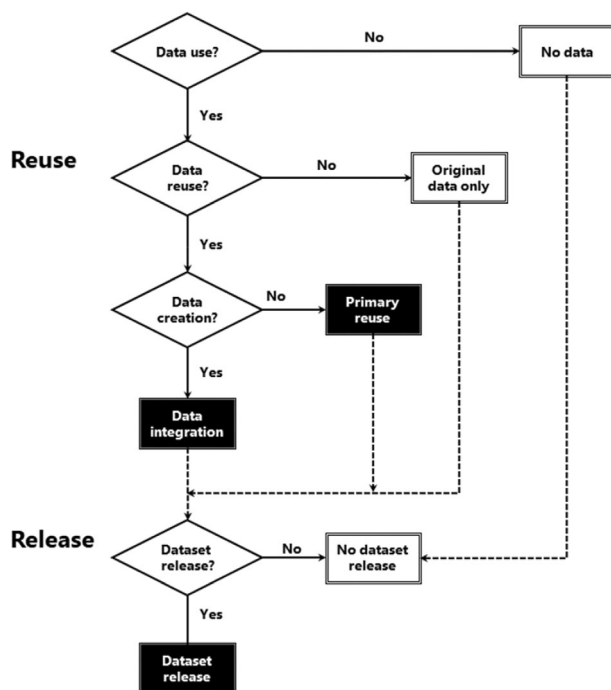
(Curty et al., 2017, p. 2). Although recent studies are gradually pursuing data reuse, the practices of data reuse are complicated and need to be further examined along with practices of data release to obtain clarity on the entire process of data sharing. Otherwise, the data-sharing cycle, such as one suggested by JISC (2021), would not be completed.

## Distinct modes of data sharing used for the present study

To investigate data sharing found in research articles in-depth, the authors suggest three distinct modes consisting of primary reuse, data integration, and dataset release (see Fig. 1).

Figure 1 illustrates the flow used to sort sample research articles for the present study regarding research data. The first step is to exclude studies that use ‘no data’. The second step is to exclude studies with no reused data, which should use ‘only original data’ rather than existing data. The third step is undertaken to further categorize data reuse based on the presence or absence of data creation. Data reuse without data creation is ‘primary reuse’, while data reuse with data creation is ‘data integration’. The remaining mode is ‘dataset release’ rather than data release, as data release could be too broad and difficult to identify in research articles. The authors decided that released dataset should be raw data that can be used for verification of the current study or easily used for another study (i.e., reuse). Dataset release could be found in any articles except those with ‘no data’, since the sharing phase is different from that of data reuse.

Therefore, the definition of the three modes is as follows. ‘Primary reuse’ is reuse without creating or collecting original



**FIGURE 1** Three modes of data sharing.

## Key points

- In the 963 articles chosen from those published in 2018, primary reuse occurred in 117 (12.1%), data integration in 128 (13.3%), and dataset release in 87 (9.0%) articles.
- Primary reuse (i.e., data reuse without creating or collecting original data) articles used two types of existing data: resource data and specific research data.
- Data integration (i.e., data reuse with original data created or collected by the authors for the current study) was categorized into empirical type, introduction/material/research method type, and combined data analysis type.
- Public data archives were less used as platforms of dataset release and sources of reused data.
- Thorough registration and expansion of research data in public data archives are expected to promote more reuse.
- Standardization of reused data description in research articles is also expected.

data, whereas ‘data integration’ is data reuse with original data created or collected by the authors for this study. Dataset release involves making the dataset available for the authors as well as others and documenting it in the identifying form in the study.

Regarding data reuse modes, detailed information on the purpose, creator of reused data, and an alternative definition should be conveyed to clarify the definition. In terms of the purpose of data reuse suggested by Pasquetto (2018), ‘primary reuse’ occurs only for foreground purposes (reuse focusing on the secondary analysis of existing data) and ‘data integration’ occurs for both foreground and background purposes (e.g., ‘to set up experiments, to annotate novel sequences, and to interpret preliminary results’) (Pasquetto, 2018, p. 208).

Note that reused data for both reuse modes included those created or collected by the authors in past studies. In addition, note that the definitions of the two modes of data reuse differ from the primary and secondary reuse used by Yoon et al. (2019). Their definition is described in Section 2.

## LITERATURE REVIEW

The present study seeks to refine data sharing research by paying more attention to data reuse. Therefore, this section reviews data sharing studies that focus on data reuse.

## Purpose and process of data reuse

Some studies pointed out that various purposes of data reuse exist in research. For example, Pasquetto (2018) presented two kinds of purposes from a 2-year participative observation study. In addition

to the foreground purpose (reuse focusing on the secondary analysis of existing data), daily data reuse among scientists to obtain 'small facts' was described as background purpose reuse. In fact, the background reuse is more often readily found (Pasquetto, 2018). From the survey among researchers, Gregory et al. (2020) questioned specific reasons of data reuse corresponding to the purposes for their research. The most frequent reasons were 'basics of new study', 'prepare for new project or proposal', followed by 'verify own data' (Gregory et al., 2020).

The process of data reuse in research has also been investigated along with the purposes of data reuse. For example, Wang et al. (2021) analysed text from 42 data reuse studies using grounded theory and recognized three stages (i.e., initiation, exploration and collection and repurposing) of data reuse.

### Positive and negative factors affecting data reuse

Researcher studies have sought positive and negative factors affecting data reuse. Some internal factors such as perceived usefulness of, attitude towards, and subjective norm of data reuse, and some external factors such as accessibility and interoperability of data and metadata standards have been recognized as positive factors. Some negative factors recognized are perceived concern involved in data reuse and lack of support (Curty et al., 2017; Kim & Yoon, 2017; Tenopir et al., 2020; Wang et al., 2021; Winkler & Berenbon, 2021; Yoon, 2016).

### Data reuse in research articles

Studies examining research articles are relatively few compared to those investigating researchers' attitude and behaviour through surveys, interviews, and observation. However, several studies analysed data reuse in research articles as an indispensable mode of data sharing.

For example, a study investigated 200 articles published in the top 10 journals with higher impact factor in four categories in 2013. Of the 152 (76%) articles that contained or used 'more than trivial' amount of data, 123 (80.6%) used original data, and 29 (19%) reused data (Womack, 2015). A study investigating 600 articles published in PLoSOne in 2014 and 2015 reported 312 (52.0%) that used datasets (Zhao et al., 2018). Data were created or collected by the authors in 231 articles (74.0%) and reused in the other 81 articles (26.0%).

Park et al. (2018) conducted more detailed analysis on data sharing in 313 full-text articles written by authors, who used the most cited datasets at least once. They searched indicator terms of data reuse identified in their previous study, in addition to manual coding. They found 208 articles (66.5%) indicating data reuse. Another finding was related to the location of the indicator term in the article. The terms indicating data reuse tended to be present, not in the official reference section (8.8%), but in the main text (81.9%). They concluded that 'informal' data citation was more common in the articles investigated.

### Different ways to reuse data in research articles

A few article studies have suggested that there are different ways to reuse data. However, few attempts have been made to identify the differences and distinctively analyse them in article studies.

For example, Park et al. (2018) pointed out a specific method of data reuse by examining existing data with authors' original data, in addition to another method of data reuse that involves examining existing data only. However, they did not distinguish between the two methods, nor report the frequency of each.

Yoon et al. (2019) analysed the reuse of HINTS (Health Information National Trends Survey) data based on an alternate definition of two methods. They defined primary reuse as reuse of HINTS data only, and secondary reuse as reuse of HINTS data alongside other data, where HINTS data was prioritized. Therefore, the other data can be any data in secondary reuse, regardless of the creator and time of creation. They analysed 250 articles on primary and secondary reuse of HINTS data without distinguishing between the two methods.

### Framework of the study

The literature review revealed that data reuse captured in research articles has not been fully analysed, although the research article study has an advantage of actually seeing the evidence of data sharing rather than relying on the self-reported attitude or behaviour of researchers. First, existing studies did not examine an extensive number of articles from across research fields. Second, despite having found different methods of data reuse, they did not distinguish them along with different purposes or the process of research. To analyse the whole process of data sharing, this study investigates various items of the three modes of data sharing in research articles from all research fields.

The results of the study could contribute to further data reuse and an organic data-sharing cycle, providing useful implications to stakeholders of data sharing such as policy makers, people involved in public data archives, and publishers.

## METHODS

### Sample

The authors purchased bibliographic data from Clarivate, containing 3000 randomly sampled 'article' type records published in 2018 from Web of Science (WoS). As the WoS categories were unwieldy for our research, similar research fields were organized and aggregated into the study's original 14 categories. A stratified random sample of 1000 articles was selected from the purchased data in the 14 categories. After excluding 37 articles (non-original articles such as commentaries, opinions, protocols, and educational reviews), 963 articles were finally analysed. Each author was assigned approximately 200 articles.

## Coding

The articles in our sample were analysed manually from three perspectives: (1) field of research; (2) target of study and (3) data sharing. The authors describe their coding schema below. The details are included in the Appendix (Table S1), found in the Supporting Information and in Zenodo.

The field of research was used with the 14 categories. In summarizing the results, two categories (i.e., astronomy & astrophysics, and geosciences), which were few in the sample, were merged, and a total of 13 field categories were used for analysis.

In this study, the authors analysed the entity under study, regardless of the field. The scheme of the target of the study was constructed inductively through discussions among the researchers. Four broad categories with 16 entities in total were identified: (1) the biological systems category including six entities such as gene, human, and protein; (2) the material and physical science category including four entities such as compound, material, and physical phenomenon; (3) the mathematical systems category including four entities such as architecture, mathematical model, and protocol; and (4) the other category including the two entities of documents and social infrastructure.

Three distinct modes of data sharing (i.e., primary reuse, data integration and dataset release) were manually coded to find indications in the entire body of articles and supporting information. For primary reuse, the authors further categorized the types of reused data: (a) resource data (created or collected for general purposes other than specific individual studies) or (b) specific research data (created or collected for specific individual studies). For data integration, the authors determined the reuse type based on the purpose: (a) empirical type (comparison/verification type), (b) introduction/material/research methods type, and (c) combined data analysis type. Additionally, in the primary reuse and data integration, the data sources were identified as public data archive, citation (i.e., other articles in which the citation was recorded in the article), self (the authors' own storage), or other (i.e., other researchers or institutions). For dataset release, the methods of storing data as public data archive, on request (i.e., storing data somewhere which can be accessed or supplied on request), and supporting information of articles, were also identified.

## RESULTS

### Overall result

Among the 963 sample articles, the studies reported in 25 articles (2.6%) did not use any data, while those reported in 693 articles (72.0%) utilized original research data only. The indications for each of the three modes of data sharing were found in 117 articles (12.1%) for primary reuse, 128 articles (13.3%) for data integration, and 87 articles (9.0%) for dataset release (Fig. 2).

The distribution numbers of the three modes in the figure include overlap. As data reuse and dataset release occur in different phases, a few articles show mode of data reuse and data

release concurrently: primary reuse and dataset release in six articles (0.6%) and data integration and dataset release in 10 articles (1.0%).

### Primary reuse

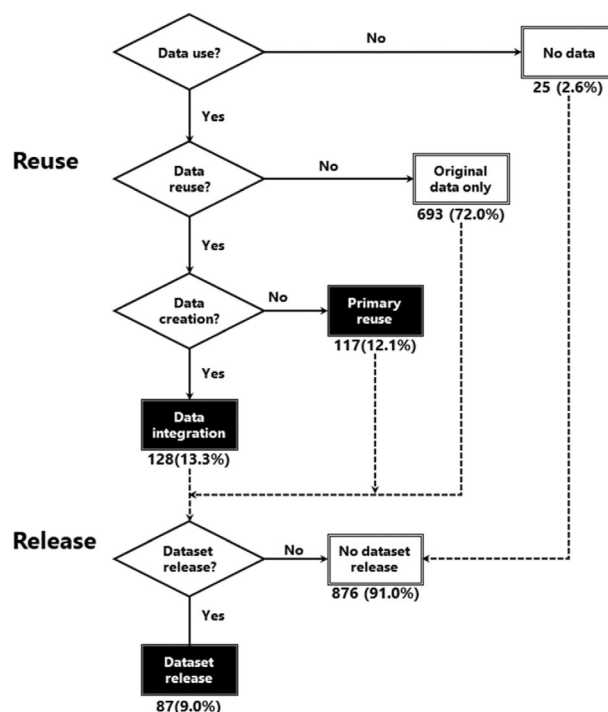
Table S2 shows the rate of primary reuse of articles in each research field and the number of reuses. The highest percentage was found in astronomy & astrophysics + geosciences (26.7%, 12 out of 45), followed by social sciences (25.4%, 15/59), and medicine (24.3%, 52/214).

By target of the study, social infrastructure (33.8%, 25/74), human (29.3%, 41/140), natural phenomenon (24.3%, 9/37) and human behaviour or mind (22.6%, 7/31) showed a rate of more than 20% (Table S3).

Table 1 presents the sources of reused data in primary reuse. Public data archive was the most common source of primary reused data (45.3%, 53/117).

### Two types of primary reused data

Primary reuse data fall into two types: resource data or specific research data. Resource data used by 74.4% (87/117) of primary reuse articles are data created or collected for general purposes other than specific individual studies, while specific research data (39.3%, 46/117) are data created or collected for specific individual studies. Both types of data were used in 13.7% (16/117) of the articles.



**FIGURE 2** Distribution of articles with three modes of data sharing.

**TABLE 1** Sources of reused data in primary reuse.

Source	n	%
Public data archive	53	45.3
Other	43	36.8
Citation	29	24.8
Self	6	5.1

Note: The total is not equivalent to 100% because a single article may reuse data from multiple sources.

Resource data are often large longitudinal observation data such as climate data for forecasting, which cannot be collected for a single study. Compiled practical business data such as electronic health records and general-purpose statistics are also resource data. A typical study in biology obtained longitudinal observation data and statistics resource data (i.e., atmospheric surface temperature records and mortality records) (Méndez-Lázaro et al., 2018). Another example study in the social sciences reused video data placed on general websites to analyse the discourses around LGBTI (Egan, 2018).

Specific research data are also reused when they matched new studies. Most often, only the necessary parts were extracted and incorporated into the research framework for reanalysis. For example, an article in geosciences indicated that data on the stable isotope ratios of oxygen and carbon as well as existing pollen and charcoal data were drawn from multiple specific studies to investigate climate change and its influence (Roberts et al., 2018).

Almost half of primary reuse articles with both types were in medicine (7/16). In the example study, resource data (i.e., electronic health records, housing data and census data) and specific research data (i.e., data from a large cohort study of colorectal cancer screening) were integrated to find indicators of social disadvantage (Hughes et al., 2018).

## Data integration

Looking at the rate of articles with data integration (i.e., while creating the original data and reusing other research data) by research field, only computer science (31.3%, 10/32) exceeded 30%, followed by astronomy & astrophysics + geosciences (20.0%, 9/45), and chemistry (18.6%, 22/118; Table S4).

By target of study, only natural phenomena (32.4%, 12/37) exceeded 30% (Table S5). All other targets were in the 10-figure percent or less.

Table 2 presents the sources of reused data in data integration. Most articles (97.7%, 125/128) recorded citations as a source of reused data.

## Three types of data integration

Data integration studies have been categorized into the following three types: (a) empirical (which was the most common, 60.2%, 77/128); (b) introduction/material/research method (36.7%, 47/128); and (c) combined data analysis (11.7%, 15/128).

The empirical type of data integration is conducted in a study that creates its own new model and uses the data from previous studies for comparison and evaluation, although its own model is based on previous studies. The empirical type of reused data was found in the results section of articles citing previous studies. For example, a study in engineering examined how 3D images can be effectively expressed using their newly proposed algorithm. Data from the previous algorithm were used and compared with their own results for evaluation (Pahwa et al., 2018).

The introduction/material/research method type uses data from existing research as a background or premise, or as a part of the research method or material rather than directly using it to derive the results. An example article in engineering showed figures of the breakdown of the types of incinerators based on information prepared by the Japanese Ministry of Environment (Fujiwara et al., 2018). Another example study in humanities used archived image data to develop its own tests examining the recognition of human visual beauty. In this case, the existing data were used as stated in the methods section (Mayer & Landwehr, 2018).

The combined analysis type includes studies that arrive at new results by reanalysing the newly investigated data and existing data as one dataset from a unified viewpoint. While the empirical type of data integration mainly focuses on the original research data to be verified, the combined data analysis type uses both data equally to arrive at the new results. Most studies combined similar kinds of data. For example, a study in the field of multiple sciences attempted to clarify the ethnic roots of the Uighurs. Researchers combined their original genetic data on Uighurs with existing genetic data from various ethnic groups to conduct a principal component analysis (Zhan et al., 2018). Other studies combined heterogeneous data. For example, a study in medicine combined existing electronic health records of the participants and newly measured their clinical data to perform logistic regression analysis (Zhang et al., 2018).

## Dataset release

Dataset releases were recognized mostly in the interdisciplinary/multidisciplinary field (36.4%, 12/33), followed by chemistry (23.7%, 28/118; Table S6). No dataset release was identified in our sample articles in computer science, social sciences, mathematics and humanities.

By target of the study, dataset release was the most common in cases of research on compounds (30.6%, 26/85), followed by research on genes (26.7%, 12/45; Table S7).

## The methods of dataset release

Regarding the methods of dataset release, more than half (63.2%, 55/87) of the data were released in supporting information on the e-journal platform, whereas only 23.0% (20/87) were published in public data archives (Table 3).



**TABLE 2** Sources of reused data in data integration.

Source	n	%
Citation	125	97.7
Public data archive	14	10.9
Other	14	10.9
Self	5	3.9
N.A.	1	0.8

Note: The total is not equivalent to 100% because a single article may reuse data from multiple sources.

## Examples of dataset release

In the 26 compounds articles, most (24 cases) published data as supporting information (Table S7). For example, data on compounds such as structure and some measurements prepared by experiments were partly released in tables and figures in the main text, and the details were provided in the supporting information in the form of a table (Burilov et al., 2018). The other methods of releasing compound data were a public data archive (Cambridge Crystallographic Data Centre) and on request.

Even in the 12 articles targeting genes, half of the articles published data as supporting information. One example is a study on the DNA characteristics of different ethnic groups in China that published the raw genotypes in a spreadsheet format under the supporting information (Zhan et al., 2018). Another seven cases published a dataset in public data archives as expected and one case published it as on request.

In total, 22 specific public data archive names were found in 20 cases regardless of the field and target of the study (Table S8). Among the individual archives, general-purpose Github was the most common (four cases); however, NCBI Gene Expression Omnibus (GEO; three cases) and other specialized bioinformatic archives (twelve cases in total) accounted for the majority.

## DISCUSSION

The present study revealed initial insights into current data sharing in three mode practices. The rate of dataset release (9.0%) was lower than that of either mode of data reuse. Among data reuse, data integration (13.3%) was a common mode as primary reuse (12.1%).

The gap between dataset release and data reuse infers an incomplete cycle of data-sharing. To encourage further data

**TABLE 3** Methods of dataset release.

Methods	n	%
Supporting information	55	63.2
Public data archive	20	23.0
On request	16	18.4

Note: The total is not equivalent to 100% because a single article may use multiple methods for dataset release.

reuse, issues found by carefully examining indicators in multiple locations in the sample articles should be discussed as external factors affecting data reuse, such as findability and accessibility. This section elaborates on the two kinds of necessities required for further data reuse.

## Necessity of thorough registration and expansion in public data archives

Data platforms other than public data archives oppose FAIR Principles (Wilkinson et al., 2016) and the US OSTP guidance (Nelson, 2022). However, they were conspicuously used, not only for dataset release, but also as sources of reused data. Although nearly half of the primary reuse articles reused data from public data archives (45.3%), relatively few data integration articles reused data from public data archives (10.9%). The reason for less archival data could be that data integration occasionally occurs for background purpose reuse to affirm ‘small facts’ (Pasquetto, 2018).

The minority of archival data indicate the difficulty in validating reused data and encouraging further reuse. For example, to find and access data from citations (24.8% of primary reuse and 97.7% of data integration), researchers must first locate another research article. The data may be buried in the main text, tables, and figures or attached as supporting information. Imker et al. (2021) actually found that any place in an article can be recognized as the largest source of reused data in studies.

Specifically, most supporting information for dataset release (63.2%, 55/87) would cause trouble, as pointed out in other studies (Imker et al., 2021; Jiao & Li, 2022). Supporting information is only available on the electronic journal site and require an access contract. Additionally, they included a wide variety of materials, such as protocol details and additional charts pertaining to the results, and identifying the research data was time-consuming.

To find and access reused data from ‘other’ (36.8% of primary reuse and 10.9% of data integration) is also difficult in cases of special request, as researchers need to make extra effort to request data through a ‘person’. Data, such as electronic health records, questionnaires, and interview records, may not be available due to anonymity and sensitivity, even if contact information is provided in the article. For this reason, Herold’s study did not recognize data ‘upon request’ as data sharing (Herold, 2015).

Ideally, all research data including data reused for background purposes, such as ‘aggregate or summary level data for “small fact”’ (Pasquetto, 2018), and that are currently buried in the main text, tables and figures, or attached as supporting information, should be registered in a public data archive to solve this data platform problem. For sensitive information, a public data archive could restrict access, and provide as much access as possible in a unified platform.

## Necessity of standardization of data description in research articles

Regarding the description and metadata format of the data in the articles, this study confirmed the problem of inconsistency.

For example, released dataset and reused data were frequently described and buried only in the main text. The data availability statement (DAS) section was often found in the articles, but the description varied. In the reference list, citations indicating reused data were not distinguished from bibliographic references. All these problems create findability (F) and accessibility (A) issues, as highlighted by the FAIR principles (Wilkinson et al., 2016).

Regardless of whether the original data are created or collected and published or reused, they should at least be stated in a common specific section in the article, such as DAS linked to the data archive (Colavizza et al., 2020). Additionally, apart from the reference section, a standardized style of metadata should be listed. To this end, the publisher should specify the location and format of the description and metadata of the research data in the article. Ideally, metadata would include 'public data archive identifier + dataset identifier'. Standardization of data citations is an incentive for researchers to share data (Marwick & Birch, 2018), and is essential for promoting reuse in terms of accessibility (Tenopir et al., 2020; Yoon, 2016).

This study has limitations. While manual coding revealed the current state of various data sharing, it was analysed over an extended period of time and the results of the articles in 2018 may not reflect the latest situation. In addition, research article studies can only reveal what is written in the article. Since there is no standard for describing the extent of reused data in the current articles, there is a limit to understanding the actual status of reused data.

## CONCLUSIONS

Data reuse in research articles is an indispensable mode of data sharing, as the purpose of data sharing is to enable data reuse (Duan et al., 2022). In addition to the dataset release, this study provides a clear account of the current practice of two modes of data reuse: the mode of primary reuse that reanalyses the dataset and the mode of data integration that uses existing data with created or collected data for various purposes, including background purposes (i.e., introduction/material/research method). Findings about the two modes of data reuse were made possible by cautious investigation using manual coding at multiple locations in the research articles.

Research can be expected to be activated through widespread reuse. However, this study's findings revealed that the data to be reused are not in an accessible state, and it is difficult to develop further data reuse. Stakeholders of data sharing such as policy makers, people involved in public data archives and publishers need to understand the reality of the reuse situation in depth and take action. For example, publishers could require the inclusion of the DAS and promote standardization of its content, referring to the guidance shown by the US OSTP (Nelson, 2022).

To accelerate promotion of data sharing in the future, all data created or collected in research should be stored in one of the public data archives, and all data created or collected and reused should be mentioned and recorded in a standardized format in the article. This could then close the loop that results in a virtuous cycle of research data-sharing.

## AUTHOR CONTRIBUTIONS

**Yukiko Sakai:** Project administration, Investigation, Writing—original draft, review and editing. **Yosuke Miyata:** Investigation, Data curation, Formal analysis, Visualization Writing—original draft. **Keiko Yokoi:** Investigation, Writing—original draft. **Yuqing Wang:** Investigation, Writing—original draft. **Keiko Kurata:** Funding acquisition, Conceptualization, Methodology, Investigation, Writing—review and editing.

## ACKNOWLEDGMENT

The authors wish to thank Shuichi Ueda, Professor Emeritus of Keio University, Kazuhiro Hayashi, Director, Research Unit for Data Application, the National Institute of Science and Technology Policy, and Mamiko Matsubayashi, Assistant Professor of the University of Tsukuba for their continuous advice on the study. Thanks also to the reviewers for their comments and suggestions to improve the paper.

## FUNDING INFORMATION

This work was supported by JSPS KAKENHI Grant Number JP19H04423.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.7824162>.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

**Appendix S1.** Supplementary Information

## REFERENCES

- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. The MIT Press. <https://doi.org/10.7551/mitpress/9963.001.0001>
- Burilov, V. A., Fatikhova, G. A., Dokuchaeva, M. N., Nugmanov, R. I., Mironova, D. A., Dorovatovskii, P. V., Khrustalev, V. N., Solovieva, S. E., & Antipin, I. S. (2018). Synthesis of new p-tert-butylcalix[4]arene-based polyammonium triazolyl amphiphiles and their binding with nucleoside phosphates. *Beilstein Journal of Organic Chemistry*, 14, 1980–1993. <https://doi.org/10.3762/bjoc.14.173>
- Colavizza, G., Hrynaszkiwicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLoS One*, 15(4), 1–18. <https://doi.org/10.1371/journal.pone.0230416>
- Curry, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLoS One*, 12(12), 1–22. <https://doi.org/10.1371/journal.pone.0189288>
- Duan, Q., Wang, X., & Song, N. (2022). Reuse-oriented data publishing: How to make the shared research data friendlier for

- researchers. *Learned Publishing*, 35(1), 7–15. <https://doi.org/10.1002/leap.1444>
- Egan, M. (2018). LGBTI staff, and diversity within the Australian accounting profession. *Sustainability Accounting, Management and Policy Journal*, 9(5), 595–614. <https://doi.org/10.1108/SAMPJ-07-2017-0069>
- Fujiwara, H., Kuramochi, H., Maeseto, T., Nomura, K., Takeuchi, Y., Kawamoto, K., Yamasaki, S., Kokubun, K., & Osako, M. (2018). Influence of the type of furnace on behavior of radioactive cesium in municipal solid waste thermal treatment. *Waste Management*, 81, 41–52. <https://doi.org/10.1016/j.wasman.2018.09.029>
- Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*, 2, 1–51. <https://doi.org/10.1162/99608f92.e38165eb>
- Herold, P. (2015). Data sharing among ecology, evolution, and natural resources scientists: An analysis of selected publications. *Journal of Librarianship and Scholarly Communication*, 3(2), 1244. <https://doi.org/10.7710/2162-3309.1244>
- Hughes, A. E., Tiro, J. A., Balasubramanian, B. A., Skinner, C. S., & Pruitt, S. L. (2018). Social disadvantage, healthcare utilization, and colorectal cancer screening: Leveraging longitudinal patient address and health records data. *Cancer Epidemiology Biomarkers and Prevention*, 27(12), 1424–1432. <https://doi.org/10.1158/1055-9965.EPI-18-0446>
- Imker, H. J., Luong, H., Mischo, W. H., Schlembach, M. C., & Wiley, C. (2021). An examination of data reuse practices within highly cited articles of faculty at a research university. *Journal of Academic Librarianship*, 47(4), 102369. <https://doi.org/10.1016/j.acalib.2021.102369>
- Jiao, C., & Li, K. (2022). Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications. *Journal of Information Science*, 016555152211018. <https://doi.org/10.1177/01655515221101830>
- JISC. (2021). *Research data management toolkit*.
- Kim, Y. (2022). A sequential route of data and document qualities, satisfaction and motivations on researchers' data reuse intentions. *Journal of Documentation*, 78(3), 709–727. <https://doi.org/10.1108/JD-02-2021-0044>
- Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multi-level analysis. *Journal of the Association for Information Science and Technology*, 68(12), 2709–2719. <https://doi.org/10.1002/asi.23892>
- Marwick, B., & Birch, S. E. P. (2018). A standard for the scholarly citation of archaeological data as an incentive to data sharing. *Advances in Archaeological Practice*, 6(2), 125–143. <https://doi.org/10.1017/aap.2018.3>
- Mayer, S., & Landwehr, J. R. (2018). Quantifying visual aesthetics based on processing fluency theory: Four algorithmic measures for antecedents of aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts*, 12(4), 399–431. <https://doi.org/10.1037/aca0000187>
- Méndez-Lázaro, P. A., Pérez-Cardona, C. M., Rodríguez, E., Martínez, O., Taboas, M., Bocanegra, A., & Méndez-Tejeda, R. (2018). Climate change, heat, and mortality in the tropical urban area of San Juan, Puerto Rico. *International Journal of Biometeorology*, 62(5), 699–707. <https://doi.org/10.1007/s00484-016-1291-z>
- Nelson, A. (2022). Memorandum for the heads of executive departments and agencies. [www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf](https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf)
- Pahwa, R. S., Lu, J., Jiang, N., Ng, T. T., & Do, M. N. (2018). Locating 3D object proposals: A depth-based online approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3), 626–639. <https://doi.org/10.1109/TCSVT.2016.2616143>
- Park, H., You, S., & 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. <https://doi.org/10.1002/asi.24049>
- Pasquetto, I. (2018). *From open data to knowledge production: Biomedical data sharing and unpredictable data reuses*. UCLA: Center for Knowledge Infrastructures <https://escholarship.org/uc/item/1sx7v77r>
- Roberts, N., Woodbridge, J., Bevan, A., Palmisano, A., Shennan, S., & Asouti, E. (2018). Human responses and non-responses to climatic variations during the last glacial-interglacial transition in the eastern Mediterranean. *Quaternary Science Reviews*, 184, 47–67. <https://doi.org/10.1016/j.quascirev.2017.09.011>
- Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One*, 10(8), e0134826. <https://doi.org/10.1371/JOURNAL.PONE.0134826>
- Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS One*, 15(3), e0229003. <https://doi.org/10.1371/journal.pone.0229003>
- The State of Open Data. (2022). The longest-running longitudinal survey and analysis on open data (2022). Digital Science Report. <https://doi.org/10.6084/m9.figshare.21276984.v5>
- Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, 72(9), 1161–1182. <https://doi.org/10.1002/asi.24483>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 1–9. <https://doi.org/10.1038/sdata.2016.18>
- Winkler, C. E., & Berenbon, R. F. (2021). Validation of a survey for measuring scientists' attitudes toward data reuse. *Journal of the Association for Information Science and Technology*, 72(4), 449–453. <https://doi.org/10.1002/asi.24412>
- Womack, R. P. (2015). Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS One*, 10(12), 1–22. <https://doi.org/10.1371/journal.pone.0143460>
- Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–6. <https://doi.org/10.1002/pr2.2016.14505301126>
- Yoon, J. W., Chung, E. K., Lee, J. Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS. *Learned Publishing*, 3, 199–206. <https://doi.org/10.1002/leap.1213>



- Zhan, X., Adnan, A., Zhou, Y., Khan, A., Kasim, K., & McNevin, D. (2018). Forensic characterization of 15 autosomal STRs in four populations from Xinjiang, China, and genetic relationships with neighboring populations. *Scientific Reports*, 8(1), 1–7. <https://doi.org/10.1038/s41598-018-22975-6>
- Zhang, J., Hu, J., Zhang, C., Jiao, Y., Kong, X., & Wang, W. (2018). Analyses of risk factors for polycystic ovary syndrome complicated with non-alcoholic fatty liver disease. *Experimental and Therapeutic Medicine*, 15(5), 4259–4264. <https://doi.org/10.3892/etm.2018.5932>
- Zhao, M., Yan, E., & Li, K. (2018). 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. <https://doi.org/10.1002/asi.23919>