Tracing Data Footprints: Formal and Informal Data Citations in the Scientific Literature

Ornella Irrera[1,2][0000−0003−2284−5699], Andrea Mannocci[2][0000−0002−5193−7851], Paolo Manghi[2][0000−0001−7291−3210], and Gianmaria Silvello[1][0000−0003−4970−4554]

[1] Department of Information Engineering, University of Padova, Italy
{ornella.irrera,gianmaria.silvello}@unipd.it
[2] National Research Council (CNR-ISTI), Pisa, Italy
{andrea.mannocci,paolo.manghi}@isti.cnr.it

Abstract. Data citation has become a prevalent practice within the scientific community, serving the purpose of facilitating data discovery, reproducibility, and credit attribution. Consequently, data has gained significant importance in the scholarly process. Despite its growing prominence, data citation is still at an early stage, with considerable variations in practices observed across scientific domains. Such diversity hampers the ability to consistently analyze, detect, and quantify data citations. We focus on the European Marine Science (MES) community to examine how data is cited in this specific context. We identify four types of data citations: formal, informal, complete, and incomplete. By analyzing the usage of these diverse data citation modalities, we investigate their impact on the widespread adoption of data citation practices.

Keywords: Data Citation · Scholarly Graph.

1 Introduction

In recent years, there has been a growing recognition of the significance of data within the scholarly communication ecosystem. Data is no longer considered mere byproducts of research but is acknowledged as a valuable resource that can accelerate research, validate experiments, and generate new knowledge. This shift in perception is leading to a transformation in the traditional research ecosystem, in which textual publications were the sole measure of a researcher's work, to a new paradigm where data and publications hold equal importance.

In this evolving landscape, crediting data authors for their released and reused datasets is essential, akin to the recognition given to authors of textual publications [16]. However, citing data presents a significant challenge that must be addressed to ensure that data authors receive the appropriate credit and enable the scientific community to discover and reuse data effectively.

Several international efforts have been made to define how data should be cited in the literature and which information a data citation should contain to properly identify the data and its authors. Nevertheless, until recently, data has

rarely been cited in the literature, and when it was, the citation was inconsistent, leading to the existence of multiple methods of data citation that are often contradictory [21]. For instance, [2] found that more than 370 different citation variants have been used to cite a dataset in the oceanographic community. In addition, [20] showed that *formal* data citations are less common than *informal* citations occurring in the full text of a publication. A universally accepted standard has not been established yet, and some barriers still prevent researchers from sharing their data; the lack of a robust reward system is the most notable [25].

This work delineates the key distinctions between *formal* and *informal* data citations. Our primary goal is to identify the current patterns of data citation and explore the potential ramifications of different citation styles and methods. To address this challenge effectively, we concentrate on a substantial scholarly graph encompassing textual and data citations within the European Marine Science (MES) research community. The MES community was chosen due to its size, active engagement, and well-established data publication and citation practices, as documented in a previous study [14]. Furthermore, in this research, we enhance the existing scholarly graph by incorporating the PDFs of the publications and employing NLP techniques to extract mentions of datasets and software.

Our analysis encompasses the following aspects: (i) identification of prevalent citation practices; (ii) examination of the sections in which data citations are found within the papers; (iii) investigation of the attributes utilized for data identification in citations; (iv) exploration of the publication and data authors to gain insights into data reusability. Our findings demonstrate that only 24.12% of the identified data citations adhere to formal practices, ensuring proper attribution to the data author, unique identification, and persistent access to the dataset. In contrast, most citations are informal, merely mentioning the dataset DOI or title within the publication's full text, without a comprehensive entry in the reference list. Additionally, we have identified the DOI as the most frequently used attribute for referencing datasets and software. Surprisingly, we have observed that citing data is less prevalent than anticipated, as 83% of the data accompanying the publications is not mentioned in the full text. This suggests a significant gap in acknowledging the data used in scholarly research. Furthermore, our analysis reveals that within the MES community, data re-use is not a common practice because creating new datasets specific to the studied use cases is more common than reusing already published and available datasets. As a result of our work, we publicly release a new scholarly graph where publications and cited data are interconnected and whose edges are enriched with information about dataset mentions – e.g., the position of the mention or whether the citation is formal or informal.

The rest of the paper is organized as follows. Section 2 presents related work focusing on analyzing formal and informal data citations; moreover, it provides the key definitions of the terms employed in this work. Section 3 describes the scholarly graph we used for the analysis, how it was built and enriched to analyze

data citations. Section 4 reports the main finding of our analyses. Section 5 discusses the main findings of this study, and Section 6 draws some final remarks.

2 Background

Related work. Numerous studies have been conducted to analyze the most common data citation practices, examine the advantages and disadvantages of each practice and its diffusion, and explore how these practices vary across the scientific domains in which they are employed. Despite many efforts to define universally accepted and shared standards for data citation – e.g., [10, 7, 26, 1, 8] – there is still no convergence on a common strategy.

The lack of a universally adopted citation standard has resulted in the coexistence of various citation practices both within and across scientific domains [18]. Hence, when studying data citation practices, a very broad definition is often used, which considers not only the citations of a dataset included in a references list but also all its mentions in the text of an article [27, 20]. [20] distinguishes between *formal* and *informal* data citations; the former consists of adding an entry about the dataset in the references list of a publication, plus mentioning the entry in the full text. The latter, instead, consists in mentioning the dataset in the full text of a publication without adding a relative entry in the references list. Some works analyze the articles' full text to detect data citation practices. In [28], for example, authors analyzed data citation practices in 600 articles of *PloS One*. [24] proposed a cross-disciplinary study of data citation practices based on the Data Citation Index (DCI). Other studies have analyzed data citation and sharing practices adopted within some scientific domains. In [28], the authors conducted an analysis involving 12 disciplines and studied their data citation, collection, and sharing practices. They found that URL is the most common attribute used to cite datasets in almost all the disciplines; in addition, the 74% of examined publication that used data contains datasets created by the same authors, indicating the tendency to create new datasets instead of re-using the available ones. Similar results have also been found in [11]. Some works investigate data citation practices in disciplines such as earth science [5], bioinformatics [6], social science [17], genetics [19], and astronomy [22]. Almost all the studies detected a high heterogeneity in the citation practices in terms of the dataset attribute cited – e.g., the dataset DOI or its title, the position in the publication's full text of the dataset mention, and the presence of a reference entry related to the dataset in the references list. [20] detected the prevalence of informal citations compared to formal ones. In addition, in [4, 17, 22, 27], the authors detected a high variety of citation behaviors; in particular, [2] detected 377 variant citation formats. Another finding from the cited studies is that the URLs mentioning a dataset does not always guarantee the accessibility to the dataset [27]. Finally, [12, 28] found that a common practice is citing data papers instead of the datasets. Although this practice guarantees credit attribution, it does not guarantee access and findability of the dataset.

Definition of terms. A *scholarly graph* is a heterogeneous, directed, and labeled graph whose nodes represent entities involved in the scholarly domain, while edges define the semantics of the relation between two nodes. *Metadata*, defined as *data about data*, are structured descriptive information about an entity [9]. Metadata sets are associated with the nodes and relations in scholarly graphs and are used to describe the research entities' nodes and the connections between them; the set of metadata associated with a node usually contains information such as the title, abstract, and date of publication of a product. In this work, we considered scholarly graphs representing the following entities: (i) *Publication*: a digital document documenting a research activity; (ii) *Dataset*: a digital research product including measures, or results – datasets are usually archives, figures, tables, CSV files; (iii) *Software*: code generated from a research activity; (iv) *Author*: a person who contributed to the generation of a research product (be it Publication, Dataset or Software).

The scholarly graph created and analyzed in this work contains the following semantics assigned to edges connecting a publication to a dataset (or software): `IsSupplementedBy`, `Cites`, `References`, `HasAuthor`. `IsSupplementedBy` is assigned when a dataset serves as a supplement for a publication, more specifically, the dataset includes additional relevant material that supports the publication [14]; `Cites` is assigned when a publication mentions the datasets in its full text, or when the publication includes the reference to the dataset; `References` when the publication includes the reference of a dataset in the references list; and, `HasAuthor` when an author contributed to a publication or dataset.

According to [3], we consider a *reference* as a short text describing a research entity included in the references list of a publication (i.e., a citation snippet), and a *citation* as the mention of that reference in the full text of a publication. Hence, a dataset can be referenced at most once by a publication, but it can be cited (mentioned) many times. Furthermore, in the following, we introduce the distinction between *formal* and *informal* data citations. *Formal dataset citations* take place when a dataset is mentioned in the publication full text referring to a reference entry in the reference list of the publication [14, 20], while *informal dataset citations* take place when the dataset – i.e., its URL, DOI, or title – is mentioned in the publication's full text, but there is not a reference entry of the dataset in the references list of the publication [20, 14]. In this work, informal citations comprise also all the datasets included in the references list of a publication but never mentioned in the full text.

We consider formal and informal dataset citation as *incomplete* when it is impossible to determine whether the citation or reference refers to a dataset, a data paper, or none of them. This occurs when there is a lack of URLs or DOIs that allow for the unique identification of the dataset. All the dataset mentions which include the DOI (or URL) are referred to as *complete* citations. In Figure 1, we illustrate formal and informal data citations in a publication and their representation in the scholarly graph. Datasets A and F are formally cited: they are included in the article's references list and the full text contains a pointer to the reference entry. The formal citation of F is *incomplete* since the reference

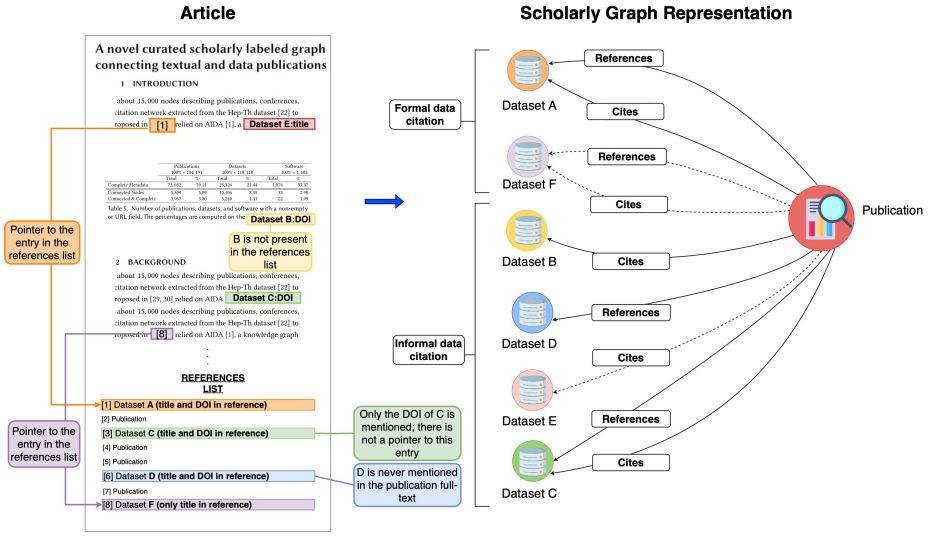

Fig. 1. Representation of formal and informal data citation in literature and in the scholarly graph. Dashed edges represent incomplete data citations. Datasets A and F are formally cited since they are reported in the references list, and there is a pointer to that reference in the full text. Dataset B is mentioned in the full text and not in the references list; Dataset C is mentioned in the references list, and its DOI is mentioned in the full text; Dataset D is mentioned in the references list, but it is never cited; the title of Dataset E is reported in the full text. The citations of E and F are incomplete due to the lack of a DOI or URL able to uniquely identify the datasets.

contains only the title and it is not possible to uniquely identify the dataset. Datasets B, C, D, and E are informally cited: the DOI of B is mentioned in the full text; the DOI of C is mentioned in the full text, it has a reference entry in the references list, but there is no pointer from the mention to the reference list entry; D is mentioned in the references list but not in the full text; the sole title of E is mentioned in the full text: in this case, the citation is incomplete since it is impossible to uniquely identify the dataset.

3 Data and Methods

The scholarly graph considered in this study is described in [14]. It is a curated scholarly graph representing the MES community of OpenAIRE[3]; it comprises 4,047 publications, 5,488 datasets, 22 software, and 21,561 disambiguated authors. It counts 9,692 edges interconnecting publications to datasets and software; edges are labeled with semantics that outline whether a publication is citing, referencing, documenting, or supplementing another research product. Publication, datasets, and software nodes contain the following metadata: title,

[3] `https://graph.openaire.eu`

abstract, date of acceptance, id, URL(s) – a list of one or more URLs pointing to the repositories where the research product has been deposited. The graph was generated through a semi-automatic curation procedure that utilized multiple sources of information, including the metadata of nodes and edges, full text publications, and web pages of datasets and software repositories. The curation process aimed to add new relationships while removing inaccurate ones, enrich the nodes' metadata, and disambiguate authors.

From the curated scholarly graph, we extracted the subgraph including publications, datasets, and software – and their authors, connected with edges whose semantics were `IsSupplementedBy`, `Cites` and `References`. For each pair of connected publication and dataset (or software) nodes, we downloaded the publication's PDF, and we extracted the mentions to the connected datasets. To this aim, having the PDF of each publication, we processed it with GROBID [15], an open-source software that uses machine learning techniques to extract structured data from scientific articles. GROBID processes the PDF and returns an XML file representing the textual content of the PDF, its sections, as well as the references list. We parsed the generated file to identify mentions of the connected dataset, specifically focusing on mentions of the title, URL, and DOI. If the mention occurred in the references list of the publication, hence the dataset had the related references entry, we assigned the `References` semantics; if the DOI or the titles were mentioned in the full text or the dataset's references entry was cited in the full text, we assigned `Cites`. For each new mention found, we added a new edge. We enriched each edge with the following information: the position of the dataset mentioned in the full text – e.g., the title of the section; additional information about the section – i.e., we assigned *main* if the mention occurred in the full text, *references* if it occurred in the references list, *secondary* if it occurred in footnotes or endnotes, and *captions* if it occurred in figures or tables captions; the attribute mentioned – e.g., whether it was mentioned the DOI or the title; the citation type – e.g., formal, informal, formal incomplete, informal incomplete. As said, we considered a formal citation *incomplete* when the dataset entry in the references list did not include the dataset DOI or it was different from the one provided in the graph. In the resulting graph, if a publication formally cites a dataset, they are connected with a `References` and a `Cites` edges. If the dataset is informally cited in the publication, and the mention occurs in the full text, the dataset will be connected to the publication by a `Cites` edge; the `References` edge is added when the mention occurs in the references list of the publication. In addition, some papers reported a separate list of references dedicated to datasets: also, in this case, we marked these mentions as formal and incomplete since the datasets reference entries were not included in the main list of references. Informal data citations were marked as *incomplete* when only the dataset title was mentioned in the publication's full text. The `IsSupplementedBy` edges have not been modified, as well as the edges connecting research outputs to their authors. The data model of the resulting graph is reported in Figure 2. Publication, dataset and software nodes share the same set of properties. Edges connect publications to datasets and software, and publications, datasets, and

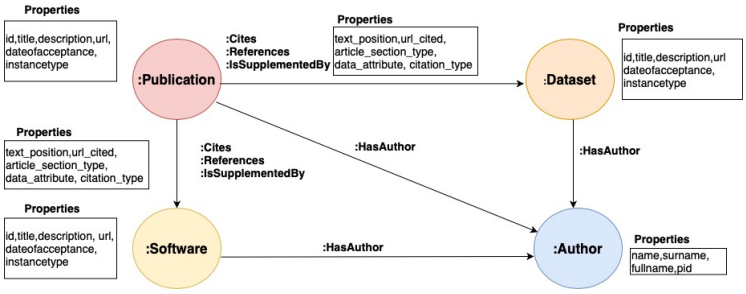

Fig. 2. Graph data model. Inside the rectangles, there are the properties of nodes and relationships. Publications, datasets, and software share the same set of properties. `Cites` and `References` semantics share the same set of properties, the `IsSupplementedBy` semantics, instead, has no properties.

Table 1. Attributes used to mention datasets and software in the references list of an article – `References` labeled edges, and in its full text – `Cites` labeled edges.

	Datasets mentions				Software mentions			
	Title	DOI	Title & DOI	URL	Title	DOI	Title & DOI	URL
References	111	291	480	0	2	7	3	0
Cites	132	761	38	1	0	11	0	0

software to their authors. Edges highlighting authorship relationships have the `HasAuthor` semantics. Edges connecting publications to datasets and software have `Cites, References` or `IsSupplementedBy` semantics. The resulting graph [13] is publicly available at `https://doi.org/10.5281/zenodo.8006578`.

4 Results

This section presents some analysis we performed on the resulting graph. We analyzed all pairs of papers and datasets (or software) connected by at least one edge with the semantics `Cites` or `References` to investigate how they are cited in the literature. The resulting graph counts – 4,497 datasets, 2,636 publications, 21 software and 894 `References` labeled edges, 1,890 `Cites` labeled edges, and 4,287 `IsSupplementedBy` labeled edges.

To cite a dataset in the literature, attributes such as the title, the DOI, the URL – or a combination of them are commonly used. In Table 1, we report the results of this analysis. The most commonly used attribute to mention a dataset in the references list of a paper – i.e., `References` labeled edge – is the combination of the title and the DOI, used in 480 datasets and 3 software mentions. The DOI without the title has been used in 291 dataset, 7 software mentions, while the title in 111 datasets and 2 software mentions. The URL – intended to link to the data repository and different from the DOI, has never

Table 2. Analysis of the detected citation practices in terms of 5 out of 8 Data Citation Principles. The lack of a checkmark means that the principle is not satisfied.

	Attribution	Evidence	Unique Identification	Access	Importance
Formal Citation					
Reference & Citation	✓	✓	✓	✓	✓
Informal Citation					
Dataset reference	✓		✓	✓	✓
Reference without DOI	✓				✓
Dataset DOI		✓	✓	✓	
Dataset Title		✓			

been used. To mention a dataset in the full text – i.e., the `Cites` labeled edges – the most frequent attribute is the DOI, used 761 times. The title has been used 132 to mention datasets, while the title and the DOI have been used only 38 times. Only 1 dataset URL has been detected. Finally, 11 software DOIs mentions have been detected.

We analyzed how the detected practices comply with 5 of 8 FORCE 11 Data Citation Principles [1]: (i)*Importance*: Data should be considered legitimate, citable products of research; (ii) *Attribution*: data citations should facilitate giving scholarly credit; (iii) *Evidence*: if claim relies upon data, the corresponding data should be cited; (iv) *Unique Identification*: a data citation should include a persistent method for identification; (v) *Access*: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials. The results are depicted in Table 2. Formal citations comply with all the principles. Mentioning a dataset in the references without citing it in the full text complies with the selected principles except for *Evidence* because it does not support any claim in the full text. If the reference lacks the DOI or the provided DOI is wrong, only *Attribution* and *Importance* are satisfied. Mentioning the DOI of a dataset in the full text complies with *Unique Identification* and *Access*, but it is not possible to give credits to contributors; in addition, *Importance* is not satisfied since the dataset is not included in the references section. Finally, mentioning the title of a dataset complies only with *Evidence*; there is not enough information to give credit to contributors and uniquely identify the dataset.

In Table 3, we analyzed the dataset and software citations, distinguishing between *formal* and *informal*, *complete*, and *incomplete* data citations. We found a total of 2,147 dataset citations – this value includes also all the datasets cited more than once in a publication's full text. Only the 24.12% of citations are formal and complete, containing enough information to uniquely identify the cited dataset and attribute it to its authors. The 19.70% is represented by incomplete formal citations: in this case, the lack of DOI prevents accessing and identifying the correct instance of the dataset. Formal dataset citations are

Table 3. Overview of formal and informal data citations. *Citation only* means that there is not a dataset entry in the references list of the publication. In contrast, *references only* means that there is a reference entry but is never cited. *Complete* citations refer to all the mentions that comprise the DOI of the dataset, *incomplete* citations include only the title of the dataset.

			$p \rightarrow d$ edges (2,147 citations)		$p \rightarrow s$ edges (23 citations)	
			count	%	count	%
Formal	Reference & citation	Complete	518	24.12	5	21.74
		Incomplete	423	19.70	1	4.35
Informal	Citation only	Complete	800	37.26	11	47.82
		Incomplete	132	6.15	0	0
	Reference only	Complete	216	10.06	5	21.74
		Incomplete	58	2.70	1	4.35

the 44% of the entire count of citations. The remaining 56% of the citations are informal. The largest portion of informal citations is DOI mentions in the full text without a dataset reference entry – i.e., 37.26%. The datasets' reference entries not mentioned in the full text represent only the 10% of the total. Incomplete informal dataset citations occurred in less than the 10% of cases. Regarding software citations, only one formal and one informal citation are incomplete; the 47.82% is informal – cited in full text without a reference, and the remaining part is equally split between formal and informal citations.

Furthermore, among the pairs of publications and datasets connected with a `References` or `Cites` edge, we found that in 144 publications the connected dataset is both formally and informally cited: the DOI (or the title) of the dataset is mentioned in the full text, and, at the same time, the dataset is present in the reference list of the publication and the related entry is formally cited; this aspect has been noticed in three pairs of publications and software instead.

We investigated the number of formal and informal citations in six date ranges.Our findings indicate that most citations were recorded after 2010, with fewer than 30 citations observed before that year. Additionally, informal complete citations were the prevailing type throughout all the periods starting from 2010. Regarding formal citations, between 2010 and 2014, there is a greater frequency of formal incomplete citations (69) compared to formal complete ones (39). From 2015 to 2019, formal complete and incomplete citations were nearly equal (318 formal complete and 299 incomplete), while after 2020, formal complete citations prevail over incomplete ones – 159 formal complete and 54 incomplete. Informal, incomplete citations are always the least common type.

We studied how many datasets supplementing the publications – connected with a `IsSupplementedBy` labeled edge, are also cited in the full text: 57 pairs of connected publications and datasets are formally cited in full text, 531 are informally cited, and $3,579$ are not cited. No software is formally cited, 6 are informally cited, and 12 are not cited.

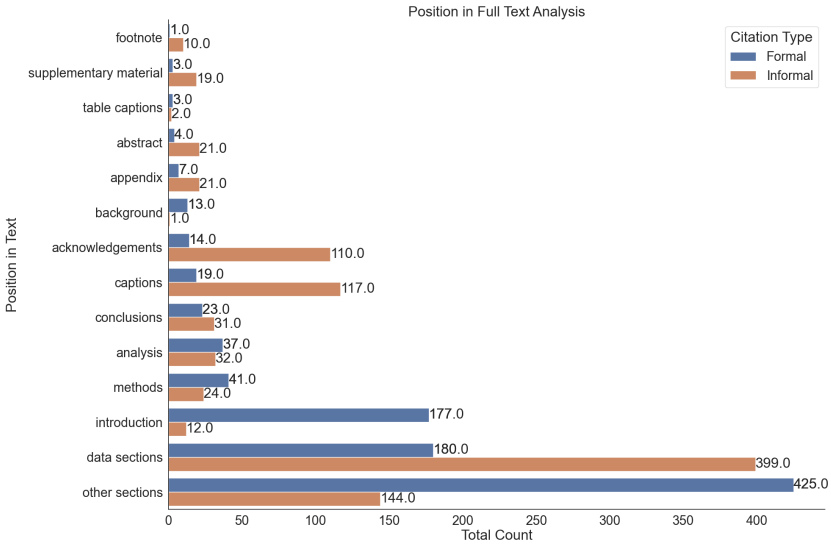

Fig. 3. The bar plot illustrates the positions in the full text of formal and informal citations. In the y-axis there are the possible sections, in the x-axis there is the total count of citations per position.

Table 4. Analysis of authors who contributed to the publication and the connected dataset. We analyzed the pairs of nodes having no authors in common, those having at least one author, and those sharing the entire list of authors.

	No authors	At least one author	All authors
IsSupplementedBy	133	1,612	**2,542**
Cites	300	**475**	398
References	348	**374**	111

In the bar plot reported in Figure 3, we illustrate the positions of a dataset (or software) formal and informal citations in the full text. The largest part of dataset citations is in the introduction, in sections that contain descriptions about the used and generated data – *data sections*, and in one of the sections composing the textual article – *other sections*. Most of the informal citations, instead, are in *data sections*, *other sections*, *acknowledgments*, and *captions*.

We analyzed the authors of the connected research outputs. In Table 4, we show for each semantics how many pairs of nodes do have not any author in common, share at least one author, and share all the authors respectively. The largest part of nodes connected with a `IsSupplementedBy` labeled edge, share all the authors – 2,542 pairs share all the authors, 1,612 share at least one author (but not all), and 133 pairs have no authors in common. The majority of pairs of nodes connected with `Cites` and `References` semantics – 475 and 374 pairs respectively, have at least one author in common; 111 `References` and 398

`Cites` labeled pairs have all the authors in common, and 348 `References` and 300 `Cites` labeled pairs have disjointed lists of authors.

We analyzed publications, datasets, and software to examine whether there exists a difference among the authors of these three research products. We found 13,608 distinct publication authors, 9,804 dataset authors, and 59 software authors. Only 30 authors contributed to publications, datasets, and software; the largest part of authors contributed both to publication and datasets – 8,759 authors, while 4,796, 1,104 and 7 authors contributed only to publications, datasets, and software, respectively. 21 authors contributed to publications and software, while only 1 author to datasets and software.

5 Discussion

About referencing and citing data – `References` and `Cites` labeled relationships – we examined the formal and informal data citations showing that there is not a significant gap between them, accounting for 44% and 56% of the identified citations, respectively. Such a small difference may be related to where the examined datasets are deposited because they mostly belong to Pangaea, Zenodo, Dryad, and Figshare, which promote data citation and provide guidelines that adhere to the 11 data citation principles. However, the lack of a universal way to cite data promotes the coexistence of multiple approaches adopted for data citation. For example, a dataset may be mentioned only in the references list of the publication and be absent from the full text, or vice versa, it may be present only in the full text and not in the references list. It is worth noting that informal citations to datasets are not considered by infrastructures such as OpenCitations [23], which captures formal citations instead and would consequently miss more than half of the detected citations.

Furthermore, there are several different approaches to referring to a dataset – e.g., relying on its DOI, URL, or title. Based on our results, datasets are most commonly cited by including their DOI in the publication, sometimes accompanied by the dataset title. However, there are instances where only the dataset title is provided, resulting in the inability to access the dataset itself. Additionally, it is often observed that when the DOI associated with a dataset is not the one pointing to the dataset repository, the publication is referring not to the dataset itself but to a data paper. This occurs in incomplete formal citations, where an element with the same title as the dataset is cited, but the DOI to the dataset repository is not provided.

The analysis of data citations in different date ranges revealed that citing data is a common practice only in the last decade. In recent years, complete formal citations are becoming more frequent than incomplete ones suggesting a growing consensus on the importance of citing data, a greater interest in following suggested citation practices, and the use of DOIs and persistent URLs.

About supplementary data – i.e., `IsSupplementedBy` labeled edges – we emphasize that despite the close relationship between a publication and its supplementary material, it is rarely cited in the literature. The analysis reveals that out

of 4,287 datasets, 3,579 (83%) are not cited in the literature. This not only hinders the ability of data authors to diversify their contributions [16] and receive credit but also hinders experiment reproducibility, discovery, and re-use.

Finally, the authors' analysis has allowed us to draw important conclusions regarding the trends of authors in discovering and reusing existing data in the literature. Our results show that when citing data within a publication, there is a tendency to cite datasets produced by the same authors instead of taking advantage of already released datasets. For instance, in pairs of nodes connected by a `Cites` edge, the number of publications and datasets sharing more than one author is more than twice the number of pairs without any common authors. This becomes even more evident when examining supplementary materials: in this case, only 133 pairs do not have any common authors, while more than 2,500 pairs share the entire list of authors. Similar results have been achieved in [11] and [28]. This finding can be related to the difficulties in re-using datasets and software released by other researchers. Using already released datasets requires a deep understanding of them, which can be acquired by relying on detailed documentation associated with the dataset. However, it is not guaranteed to find good and precise documentation, as its creation is at the discretion of the author. Additionally, most of the time, existing datasets may need to be selected and adapted for the specific use case. These conditions often result in a significant time loss, making it more convenient to create new datasets instead of re-using the already available ones. Furthermore, while there is a high number of authors working on publications, datasets, and software authors often contribute also to publications. This result is related to the lack of a universally adopted approach to citing data and software and a stable and established rewarding mechanism, such as the one for publications, for assigning credits to authors.

6 Conclusions

In this study, we utilized a curated scholarly graph that establishes connections between publications and research data to investigate how datasets and software are cited in the MES scientific literature. To identify dataset (and software) citations, we conducted an analysis of the PDFs of the publications. We focused on several key aspects, including the location of the citation within the full text, the attribute employed to reference datasets, and the categorization of citations as either formal or informal. Our findings confirmed the absence of a standardized approach to data citation. The results indicated a prevalence of informal citations compared to formal citations. The majority of dataset references included both the DOI and the title of the dataset. In cases where dataset mentions occurred within the full text, the dataset DOI emerged as the most frequently used attribute. We discovered that a small fraction of datasets accompanying the literature were cited within full texts, hindering dataset discovery, reuse, and the reproducibility of experiments. Additionally, our analyses revealed that generating new datasets was more prevalent than relying on previously released ones.

References

1. Altman, M., Borgman, C., Crosas, M., Matone, M.: An introduction to the joint principles for data citation. Bulletin of the Association for Information Science and Technology **41**(3), 43–45 (2015)
2. Belter, C.W.: Measuring the value of research data: A citation analysis of oceanographic data sets. PLoS One **9**(3), e92590 (2014)
3. Buneman, P., Dosso, D., Lissandrini, M., Silvello, G.: Data citation and the citation graph. Quantitative Science Studies **2**(4), 1399–1422 (2021)
4. Callaghan, S.: Preserving the integrity of the scientific record: data citation and linking. Learned Publishing **27**(5), S15–S24 (2014)
5. Chao, T.C.: Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. Proceedings of the American Society for Information Science and Technology **48**(1), 1–8 (2011)
6. Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.Q., Bourne, P.E.: Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution **28**(8), 454–461 (2013)
7. Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., et al.: A data citation roadmap for scientific publishers. Scientific data **5**(1), 1–11 (2018)
8. Crosas, M.: The evolution of data citation: from principles to implementation. IAssist quarterly **37**(1-4), 62–62 (2014)
9. Duval, E.: Metadata Standards: What, Who & Why. J. Univers. Comput. Sci. **7**(7), 591–601 (2001)
10. Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., et al.: A data citation roadmap for scholarly data repositories. Scientific data **6**(1), 28 (2019)
11. He, L., Nahar, V.: Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository. Aslib Journal of Information Management (2016)
12. Huang, Y.H., Rose, P.W., Hsu, C.N.: Citing a data repository: a case study of the protein data bank. PloS one **10**(8), e0136631 (2015)
13. Irrera, O.: Mes citations scholarly graph (2023). https://doi.org/10.5281/zenodo.8006578, `https://doi.org/10.5281/zenodo.8006578`
14. Irrera, O., Mannocci, A., Manghi, P., Silvello, G.: A novel curated scholarly graph connecting textual and data publications. J. Data and Information Quality (2023). https://doi.org/10.1145/3597310, `https://doi.org/10.1145/3597310`
15. Lopez, P.: GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13. pp. 473–474. Springer (2009)
16. Mannocci, A., Irrera, O., Manghi, P.: Open science and authorship of supplementary material. evidence from a research community. arXiv preprint arXiv:2207.02775 (2022)
17. Mooney, H.: Citing data sources in the social sciences: do authors do it? Learned Publishing **24**(2), 99–108 (2011)
18. Mooney, H., Newton, M.P.: The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication **1**(1) (2012)

19. Park, H., Wolfram, D.: An examination of research data sharing and re-use: implications for data citation practice. Scientometrics **111**, 443–461 (2017)
20. Park, H., You, S., Wolfram, D.: 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 (2018)
21. Parsons, M.A., Duerr, R., Minster, J.B.: Data citation and peer review. Eos, Transactions American Geophysical Union **91**(34), 297–298 (2010)
22. Pepe, A., Goodman, A., Muench, A., Crosas, M., Erdmann, C.: How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. PloS one **9**(8), e104798 (2014)
23. Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020)
24. Robinson-García, N., Jiménez-Contreras, E., Torres-Salinas, D.: Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology **67**(12), 2964–2975 (2016)
25. Silvello, G.: Theory and practice of data citation. Journal of the Association for Information Science and Technology **69**(1), 6–20 (2018)
26. Walton, D.W.: Data citation-Moving to new norms. Antarctic Science **22**(4), 333–333 (2010)
27. Yoon, J., Chung, E., Lee, J.Y., Kim, J.: How research data is cited in scholarly literature: A case study of HINTS. Learned Publishing **32**(3), 199–206 (2019)
28. Zhao, M., Yan, E., Li, K.: 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 (2018)