A Novel Curated Scholarly Graph Connecting Textual and Data Publications

ORNELLA IRRERA*, Department of Information Engineering, University of Padua, Italy and CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Italy
ANDREA MANNOCCI, CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Italy
PAOLO MANGHI, CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Italy
GIANMARIA SILVELLO, Department of Information Engineering, University of Padua, Italy

In the last decade, scholarly graphs became fundamental to storing and managing scholarly knowledge in a structured and machine-readable way. Methods and tools for discovery and impact assessment of science rely on such graphs and their quality to serve scientists, policymakers, and publishers. Since research data became very important in scholarly communication, scholarly graphs started including dataset metadata and their relationships to publications. Such graphs are the foundations for Open Science investigations, data-article publishing workflows, discovery, and assessment indicators. However, due to the heterogeneity of practices (FAIRness is indeed in the making), they often lack the complete and reliable metadata necessary to perform accurate data analysis; e.g., dataset metadata is inaccurate, author names are not uniform, and the semantics of the relationships is unknown, ambiguous or incomplete.

This work describes an open and curated scholarly graph we built and published as a training and test set for data discovery, data connection, author disambiguation, and link prediction tasks. Overall the graph contains 4,047 publications, 5,488 datasets, 22 software, 21,561 authors; 9,692 edges interconnect publications to datasets and software and are labeled with semantics that outline whether a publication is *citing, referencing, documenting*, *supplementing* another product.

To ensure high-quality metadata and semantics, we relied on the information extracted from PDFs of the publications and the datasets and software webpages to curate and enrich nodes metadata and edges semantics. To the best of our knowledge, this is the first ever published resource, including publications and datasets with manually validated and curated metadata.

CCS Concepts: • **Information systems → Digital libraries and archives**; **Data cleaning**; Incomplete data; Inconsistent data.

Additional Key Words and Phrases: Scholarly Knowledge Graphs, Data Curation, Data Enrichment, Datasets, Open Science

Authors' addresses: Ornella Irrera, ornella.irrera@unipd.it, Department of Information Engineering, University of Padua, Via Gradenigo, 6/B, Padua, Italy, 35131 and CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Via Giuseppe Moruzzi, 1, Pisa, Italy; Andrea Mannocci, CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Via Giuseppe Moruzzi, 1, Pisa, Italy, andrea.mannocci@isti.cnr.it; Paolo Manghi, CNR-ISTI – National Research Council, Institute of Information Science and Technologies, Via Giuseppe Moruzzi, 1, Pisa, Italy, paolo.manghi@isti.cnr.it; Gianmaria Silvello, gianmaria.silvello@unipd.it, Department of Information Engineering, University of Padua, Via Gradenigo, 6/B, Padua, Italy, 35131.

1936-1955/2022/5-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Ornella Irrera, Andrea Mannocci, Paolo Manghi, and Gianmaria Silvello. 2022. A Novel Curated Scholarly Graph Connecting Textual and Data Publications. *ACM J. Data Inform. Quality* 1, 1 (May 2022), 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Releasing research datasets is crucial to amplify data exposure and promote the reproducibility of scientific experiments. Over the past two decades, scholarly communication moved from a publication-centered ecosystem to one where datasets have been elevated to artifacts with similar curation attention, at least in principle, as publications [8]. The scholarly communication community agreed that datasets should always be available, relying on persistent identifiers, well documented to facilitate re-use, and formally cited [24]. Usually, the literature around research datasets focuses on three main concerns: (i) how datasets are published, (ii) how datasets are cited, and (iii) how articles and datasets are syntactically and semantically connected.

About the first aspect, we need to consider that research datasets profoundly differ from traditional research publications and have their own specificities. Researchers need to put a significant effort into depositing a research dataset because when it is published, the quality of the associated metadata is central to discovering, understanding, and re-using the dataset in the future. Nevertheless, if this task is wholly delegated to the authors, no assurance can be made about the quality and completeness of the metadata.

For the second aspect, we note that datasets are not always formally cited. For instance, dataset references can appear in the text, in the footnotes, or an article's reference list. Moreover, the references may only contain a URL to the dataset, sometimes a textual description, and rarely follow a consistent citation style [44]. Data citation is another task that should be computer-aided and not wholly entrusted to the authors of an article, as automation can enable better consistency and completeness in the citation practices and styles, both intra- and inter-domain.

Finally, even though linking datasets facilitates scientific progress, discoverability, and credit attribution, it is rarely done. One reason is that publishers and data service providers have not developed agreements to standardize dataset linking. Although the automation of the dataset-paper links would improve the consistency and quality of links, it is difficult to achieve [6].

Automatically creating typed links – i.e., labeled links or links enriched with a predicate explaining the relationship between source and target node – between published papers and datasets is a challenging task with no ready-to-use solution.

For such reasons, publishing and describing a dataset, citing it, and linking the dataset to papers are burdensome tasks that researchers perceive as onerous [39], and that lack incentivization for scientists. A major barrier to automated linking, data citation, and automated description is the lack of datasets to train and test computational methods.

In this respect, the OpenAIRE Graph (OAG) [29], developed and maintained by OpenAIRE[1], is an Open Science Graph where the aggregated metadata of research products as publications, datasets and software, and about organizations, projects, funding agencies, authors are semantically interlinked. The OAG, in 2021, counted $140M$ publications, $50M$ datasets, $256K$ software, and about $3.5B$ relationships. The OAG's open availability and domain-agnostic coverage of science make it a resource with the crucial potential to understand more about the scholarly communication ecosystem and improve data publication, citation, and linking. The OAG is a resource aggregating data from many heterogeneous sources. As such, the OAG cannot be used as is because its metadata seldom describe a research product in enough detail. The OAG is not curated and therefore the

[1] OpenAIRE – https://www.openaire.eu

correctness of metadata, data-paper links, and their semantics is not guaranteed. Moreover, it is challenging to study the OAG in details due to its scale.

Hence, in this work, we focused on a specific OAG's subgraph to provide a curated resource for studying scholarly communications. The goal is to release a curated, reliable, and sufficiently sizeable scholarly graph with verified rich metadata and semantic connections, which can aid researchers in training and testing link prediction, data search and recommendation, and author disambiguation algorithms in the scholarly communication domain. To this end, we work on the European Marine Science (MES) subgraph included in the OAG because it represents a large and active community with well-established data publication and citation practices. Moreover, in MES there is a good balance between publications and research datasets collected in data repositories whose sharing, re-use, citation recommendations, and guidelines are coherent with the most recent and recommended practices.[2]

Overall, the curated MES graph we release contains 31,118 nodes, 21,561 disambiguated authors, 4,047 publications, 5,488 datasets, and 22 software products. There are 9,649 direct labeled edges that connect publications (source) and datasets (target), and 43 publications and software; edges are labeled with semantics that outline whether the publication or the dataset is *citing*, *referencing*, *documenting*, or *supplementing* the linked products. 69,053 edges denote authorship and connect a research product to the authors. To verify the correctness of metadata and enrich them with new additional information, we parsed the full-text of all the publications and scraped the dataset web pages to collect as much information as possible; the same sources have been used to validate and augment edges and semantics. Part of the metadata was manually curated when it was impossible to apply (semi-)automatic methods.

Moreover, the curation of the MES subgraph allowed us to draw some conclusions about the current state of the OAG, the current scholarly practices about dataset description, citation, and linkage, and the quality of metadata and semantics.

Before the curation process, we verify that (i) the metadata representing research products in MES are heterogeneous and usually do not contain enough information to describe a research product in detail; (ii) the authors' metadata are usually incomplete and challenging to disambiguate; (iii) more than the 90% of datasets and software are not connected to any publications; (iv) the article full-text does not provide enough information to track all the publication-dataset links because less than the 20% of research datasets is mentioned by the connected publication; and, (v) the edge semantics are often imprecise or ambiguous; more than 30% of the edges semantics we analyzed were incorrect.

On the one hand, it is surprising to detect so many under-described or inaccurate aspects within the MES subgraph, given that it is built by actively sharing and describing research data. On the other hand, we see the necessity of a high-quality resource to study scholarly communication practices and validate algorithms for automating aspects of scholarly communication.

The manuscript is organized as follows: in Section 2, we focus on the available resources in scholarly publishing and the current state of the art; in Section 3, we describe the pipeline to create and curate the resource we release; in Section 4, we provide some statistics about the obtained curated MES graph, and we analyze the main differences between it and the original OAG before the curation; in Section 5, we discuss the results obtained, and the scenario where this graph can be applied; in Section 6, we summarise the most critical aspects of the work; in Section 7 we discuss the limitations of the proposed approach, and we summarise the future works.

[2] See Pangaea https://www.pangaea.de, Dryad https://datadryad.org/stash, and Mendeley https://www.mendeley.com.

2 BACKGROUND

In the last decade, we experienced an exponential increase in research products; these are not limited to conventional publications, but also encompass datasets, code, software, and related metadata. One of the significant challenges is how to effectively and efficiently make scholarly data available in a persistent, accessible, flexible, machine-readable fashion such that the scientific community can benefit from them [11, 35]. Several solutions propose to describe scholarly knowledge by means of network representations where actors, documents, research products, and organisations are all interconnected and form the "scholarly graphs": the Microsoft Academic Graph (MAG) [49] and the Microsoft Academic Knowledge Graph (MAKG) [10], the Open Academic Graph[3], the Open Research Knowledge Graph (ORKG)[4] [20], AMiner [46, 48], the OpenCitations corpus[5] [38], SciGraph[6], and the OAG [2] are some notable examples. Other examples of scholarly graphs focused more on datasets are: the Google Dataset Search dataset [4], DataMed [36], and the Data Set Knowledge Graph (DSKG) [11].

SoMeSci, instead, is a knowledge graph which interconnects software to the publications that mention them [41].

Among the graphs mentioned above, the OAG is the unique openly available, large-scale scholarly graph, with publications interconnected to datasets and software.

Despite the ever-increasingly recognized value of data publication, most scholarly graphs are still publication-centered. A common and universally adopted approach for interlinking data and literature is missing, raising barriers to the communication and interoperability between literature publishers and datasets ones [6]. Moreover, research data and software lack a common and globally adopted standard to be cited in the literature [31], and a wide variety of citation practices still exist [3, 37]. These aspects hinder the creation of large networks where publications and research data co-exist and are interconnected.

A large amount of available scholarly data has the potential to improve the scholarly communication ecosystem, supporting the development of algorithms to perform fundamental tasks such as author disambiguation, link prediction, and paper recommendations. Large scholarly graphs, such as the MAG or the OAG, are unsuitable for these tasks because they are incomplete and inaccurate, being created from data crawled from the Web without any control or curation [4, 9, 16] and most of them do not explicitly account for datasets and their connections [5]. No other available, curated scholarly graphs include publications, datasets, software, and their connections.

In the context of link prediction, some works leveraged ad-hoc scholarly graphs; an example is the graph created in [15] and also used in [35]. They created a graph with about 15,000 nodes describing publications, conferences, authors, and departments. In [26], the authors relied on a citation network extracted from the Hep-Th dataset [25] to construct a paper correlation graph; the methodologies proposed in [33, 34] relied on AIDA [1], a knowledge graph which includes both academic and knowledge entities (e.g., publications, patents). In [32], the authors relied on the citation networks of five sections (astrophysics, condensed matter, general relativity and quantum cosmology, high energy physics–phenomenology, and high energy physics–theory) of the physics e-Print arXiv. However, the aforementioned resources never consider datasets or software, they are primarily focused on papers (and the related entities such as venues, and journals), authors, and patents.

[3] OAG – https://www.aminer.org/open-academic-graph
[4] ORKG – https://orkg.org
[5] OpenCitations – https://opencitations.net
[6] SciGraph – https://www.springernature.com/gp/researchers/scigraph

Similarly, several solutions to the authors disambiguation problem involve scholarly graphs and citation networks. An example is Aminer, adopted in [17, 28, 40, 45]; it has sometimes been used in conjunction with Semantic Scholar [45]. Other solutions, instead, relied on combining more than two data sources. In [23], the authors relied primarily on MEDLINE, the MAG, and DBLP datasets. Other solutions relied on citation networks such as Citeseer [30] and WoS [42]. Other approaches relied on custom-made datasets, such as the Vietnamese dataset [18, 47] or a dataset on Korean scholarly data [43]. All the proposed solutions are based on the authors of textual publications, but the authors/curators of datasets are not considered. The absence of resources and methods to perform data authors' disambiguation prevents accurately computing data authors' impact and giving them credit.

There are only a few resources connecting publications and datasets (or software) that are available in the scholarly ecosystem, and whose purpose is to provide a valuable ground-truth for tasks such as link prediction or authors disambiguation. The Open Research Knowledge Graph (ORKG) has been conceived as "an infrastructure for the acquisition, curation, publication, and processing of semantic scholarly knowledge" [20]. According to the website[7], the ORKG contains $161K$ resources and more than $2.5M$ statements. In this context, one of the most important aspects is the curation performed via crowdsourcing; users are free to modify or add new scholarly contributions. This graph has played a crucial role in tasks such as question answering [22], data enrichment [14], and triples classification [21]. Nevertheless, it contains no information about datasets, software, and their connections to publications.

One of the few resources where datasets and publications are linked is described in [11]. Authors released a dataset knowledge graph, the Data Set Knowledge Graph (DSKG), with $2K$ datasets interconnected to $635K$ publications using $835K$ edges. The provided graph is linked to other Linked Data sources such as the MAKG, ORCID, and Wikidata; the goal of this resource is to facilitate the exchange of knowledge, scholarly search systems, trend detection algorithms, and impact quantification. Conversely to the other resources mentioned above, it includes datasets connected to publications. Nevertheless, this solution does not propose a curation procedure, which is crucial to resolving metadata inconsistencies. In addition, this solution includes the datasets mentioned by the publications and does not consider the datasets that are not present in the publications. Finally, the relationship between the publication and the dataset is not modeled. It is impossible to detect whether the dataset is a supplement for the publication or is only cited or referenced in its text.

In [51], the authors relied on two citation networks for assigning credits to datasets assuming the presence of biases in dataset citations. The nodes of the networks are papers and datasets and the edges are used to describe paper-paper and paper-datasets citations. To create paper-paper edges, the OpenCitations Index (COCI) and the MAG were utilized, while paper-dataset edges were generated using GenBank and Figshare. It's worth noting that while these networks contain a large number of nodes, they do not include software, and their edges only represent citation relationships, rather than other types of connections.

The detection of software and dataset mentions in the textual publications is crucial to infer new links between publications and datasets. In [41] for example, authors manually annotated the software mentions in the textual documents. However, manual annotation is usually a time consuming task. The model proposed in [50] uses a sequence-to-sequence recurrent neural network that returns the probability of a token being part of a dataset mention. This model obtained relatively high performances in terms of F_1 measure. In [12] authors proposed a semi-automatic approach that relies on TF-IDF and cosine similarity to detect dataset references.

[7] https://orkg.org

Table 1. Overview of the edge semantics connecting a publication p to a dataset or a software d. The last two rows involve the semantics used to interconnect a research product ro to an author a, and a dataset d_sub to another dataset d_super which includes d_sub.

Edge labels	Definition
p −`References`→ d	A publication p mentions a dataset d in the references list.
p −`IsReferencedBy`→ d	A dataset d includes the reference entry of a publication p in its webpage.
p −`Cites`→ d	A publication p mentions the dataset d in its full-text, or p formally cites the reference entry of d. The `cites` edge semantics is used both for formal and informal citations.
p −`IsCitedBy`→ d	A dataset d mentions the identifier of a publication p in its webpage.
p −`Documents`→ d	A publication p is a report of a dataset d. The webpage of d reports the URLs or DOIs of the publication p that documents d.
p −`IsSupplementTo`→ d	A publication p is supplementary material for a dataset d. This information might be included in the webpage of d and not in the text of p.
p −`IsSupplementedBy`→ d	A dataset d includes in its webpage the URL or the DOI of the publication p it is supplementing. Similarly, the full-text of p includes a mention of a supplementary dataset d.
$\{p\|d\}$ −`HasAuthor`→ a	The author a appears in the list of authors of the publication p or of the dataset d.
d_sub −`IsPartOf`→ d_super	The dataset d_super includes several datasets: the dataset d_sub is one of the datasets in d_super. The datasets included in d_super are listed in the webpage of d_super.

SoMeSci [41] is an example of open knowledge graph connecting software to the scholarly articles that mentions them. It counts $400K$ triples describing 3,756 software mentions in 1,367 articles.

The manual annotation of research articles to discover new software mentions, the disambiguation of spelling variations, and the enrichment with new additional information, are all aspects that make this knowledge graph a valuable, and trustworthy resource in the scholarly domain; it is a gold standard essential in the training and evaluation of several tasks such as: Entity Disambiguation, and Relation Extraction. This resource, similarly to the previous one, considers only mentions in the publication full-text, and never considers different types of relations.

Most existing resources for scholarly communication listed above primarily focus on publications and authors. Research products such as research datasets and software are left aside. Data are complex objects and completely differ from publications: the lack of resources comprising research datasets has a strong implication in the scholarly ecosystem, hindering the possibility of developing new methods involving datasets.

2.1 Definitions of Terms

In this section, we define some concepts relevant to our work.

Scholarly graphs are labeled and directed graphs, where the nodes are the entities involved in the scholarly domain, while edge labels define the semantics of the relation between two nodes. In this work, we consider the following node types:

- *Publication*: digital research document describing a research activity or product;
- *Dataset*: digital artifacts encoding observations, measures, results. Examples of datasets can be CSVs files, compressed archives, figures, and tables;
- *Software*: code produced within a research activity – e.g., web applications, scripts, libraries;
- *Author*: a person who contributed to a research product (publication, dataset or software).

An edge label describes the relationship existing between two scholarly products. In Table 1, we summarize the edge semantics we employ in the curated scholarly graph we release. Every definition involving datasets holds also for software products, except for the last row (`IsPartOf`), which exclusively concerns datasets. The first 7 semantics belong to the original DataCite metadata

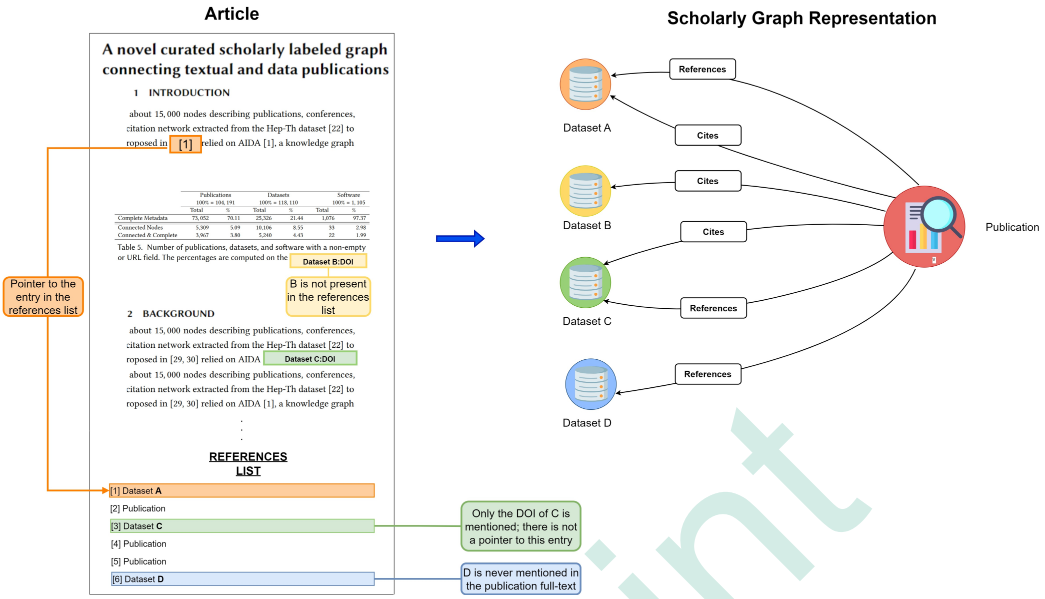
Article
A novel curated scholarly labeled graph connecting textual and data publications
1 INTRODUCTION
[1]
Pointer to the entry in the references list
Dataset B:DOI
B is not present in the references list
2 BACKGROUND
Dataset C:DOI
REFERENCES LIST
[1] Dataset A
[2] Publication
[3] Dataset C
[4] Publication
[5] Publication
[6] Dataset D
Only the DOI of C is mentioned; there is not a pointer to this entry
D is never mentioned in the publication full-text
Scholarly Graph Representation
Dataset A
Dataset B
Dataset C
Dataset D
References
Cites
Cites
Cites
References
References
Publication

Fig. 1. Representation of data citation and data reference in literature and in the scholarly graph we propose. The "Dataset A" is the only one that is formally cited: it has a related entry in the references list of the publication, and the full-text contains a pointer to the entry; the "Dataset B" is mentioned only in the full-text; the "Dataset C" is mentioned in the references list, and in the full-text it is reported its DOI; the "Dataset D" is mentioned only in the references list.

schema [7]; the last two semantics instead, are used to highlight the authorship relationship and to describe whether a dataset is contained in another one. With the term *supplementary material* we refer to additional relevant material used or supporting a publication. It is deposited together with the publication, but it is not contained in the publication text. An example can be a CSV file used for some experiments, but not reported in a table of the publication.

DataCite provides a very high-level distinction between a "reference" and "citation", not enabling a clear distinction between the two concepts. These two terms, in fact, are commonly used interchangeably, and a common agreement on their definition is still missing [5]. In this work, we model `Cites` and `References` edge semantics following the definition given in [13], where a "reference" is a work reported in the references list of a publication. A "citation" is the mention of a reference in the full-text of a publication. In the following, we define the dataset reference, and the dataset formal and informal citations. Please note that our main concern is datasets, but the definitions below, take also software into account.

- *Dataset reference*: an entry in the references list of a publication representing a dataset [5]; a reference entry containing author(s), a title, a date, and a publisher is considered to be complete;
- *Formal dataset citation*: a dataset mention occurring in the full-text of a publication and referring to a reference entry in the references list of the publication (see [5], and "formal citation" in [37]);
- *Informal dataset citation*: a dataset mention occurring in the content of a publication, but not tied to a corresponding reference entry in the references list of the publication (see "informal citation" in [37]);

Table 2. Communities quantitative analysis. For each community, the number of publications, datasets, and software are reported. In boldface, we marked the MES community we selected.

	Publications	Datasets	Software
DH-CH	4.3M	492K	1,085
Enermaps	2.6M	47K	360
RDE	1.5M	30K	656
Neuorinformatics	790K	12K	522
Covid-19	510K	23K	1,357
SIPS	157K	240K	3,513
MES	**104K**	**118K**	**1,105**
NEANIAS	25K	2,825	12

In Figure 1, we illustrate how we represent datasets references and formal and informal citations from a publication to a dataset. "Dataset A" (in orange) is *formally* cited in the article because there is a pointer in the full-text to the entry of the dataset in the references list of the citing publication. We model this formal citation by connecting the publication and the "Dataset A" with two edges, one labeled with the `References` semantics outlining the dataset entry in the references list, and another one labeled with `Cites` outlining the presence of a pointer in the full-text to the entry in the references list of the publication. Another case is represented by "Dataset B" (in yellow), whose DOI is mentioned in the full-text of the publication, but it is not tied to an entry in the references list. We model this case by connecting the publication to the "Dataset B" with a single `Cites` labeled edge. "Dataset C" (in green) is reported in the references list of the paper, and its DOI is mentioned in the full-text, but there is no pointer going from the text to the references list. Similarly to formal citations, we model this case by connecting the publication and "Dataset C" with a `References` labeled edge, and a `Cites` labeled edge. Despite the "Dataset A" and the "Dataset C" are connected with the publication through the same labeled edges, the former is a formal citation, whereas the latter is an informal one. Finally, "Dataset D" (in blue) is reported only in the references list, without any mention in the full-text of the publication. In this case, we connect the publication and "Dataset D" only with a `References` edge.

2.2 Community Detection

Research communities are intended as communities of practice in a research field, willing to share and discover scientific results among the community itself and beyond [2]. We detected eight communities in the OAG relevant to our task: (i) Digital Humanities and Cultural Heritage (DH-CH)[8], (ii) EnerMaps[9], (iii) Rural Digital Europe (RDE)[10], (iv) Neuroinformatics[11], (v) Covid-19[12], (vi) Science and Innovation Policy Studies (SIPS), (vii) European Marine Science (MES)[13], (viii) NEANIAS Underwater Research Community[14].

We analyzed the communities from a quantitative point of view to detect those with a good balance between the number of publications, datasets, and software. This analysis allowed us to

[8] https://dh-ch.openaire.eu/
[9] https://enermaps.openaire.eu/
[10] https://rural-digital-europe.openaire.eu/
[11] https://ni.openaire.eu/
[12] https://covid-19.openaire.eu/
[13] https://mes.openaire.eu/
[14] https://neanias-underwater.openaire.eu/

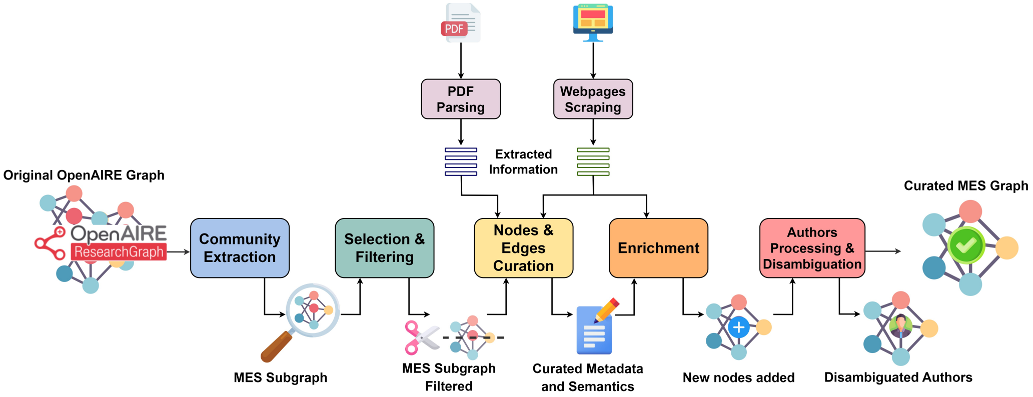

Fig. 2. Curation pipeline. The output of each phase is the input of the subsequent phase. There are five main phases: (i) Community Extraction, (ii) Selection & Filtering, (iii) Nodes & Semantics Curation, (iv) Enrichment, (v) Authors Processing & Disambiguation. The input of the entire pipeline is the OpenAIRE Graph, while the final output is the curated MES graph.

select the community to curate. In Table 2, for each community, we provide the total count of publications, datasets, and software.

In all the communities, the number of software products is considerably lower than the number of datasets and publications. Additionally, in all the communities except for SIPS and MES, there is a significant imbalance between the number of publications and datasets. For instance, the Neuroinformatics community has 790K publications and only 12K datasets. SIPS and MES are the only two communities where publications and datasets are balanced.

DH-CH and NEANIAS communities were unsuitable for our task, due to their size, with DH-CH having the most nodes and NEANIAS the fewest. Therefore, we excluded them. Enermaps, RDE, Neuroinformatics, and Covid-19 communities had a high imbalance between datasets and publications, and we excluded them for this reason. We selected SIPS and MES as the only suitable communities for our task. We chose to work with the MES community because it is more manageable and it is known as an active community with well-established citation practices.

3 CURATION PIPELINE

In Figure 2, we report the five-phases curation pipeline we adopted for this work.

First and foremost, we take the original OAG as input[15]. The OAG is a directed and labeled graph aggregating the metadata about research products and their links from over 96K sources worldwide. OAG's nodes are research products (e.g., publications, datasets, software), research organizations, projects, and data sources. Each node is described by a set of metadata records. The typed and directed edges represent the links between the research products, labeled with semantics from the DataCite Metadata Schema [7].

The OAG counts a total of 140M publications, 50M datasets, 256K software, and 3,8B relationships; we consider relationships connecting publications to datasets and software, and whose semantics are: `References`, `Cites`, `Documents`, `IsSupplementedBy`, and their inverse. All the edge semantics consider the publication as the source node and the dataset (or software) as the target node.

The first phase of the pipeline, *Community Extraction*, selects from the original OAG the nodes and the links belonging to the MES community – there are 104K publications, 118K datasets, and

[15] Release of December 2021

1,105 software. The metadata set of a node contains information about the community to which it belongs, if any. To identify the MES community, we examined the nodes in the OAG and selected only those whose metadata indicated the participation in the MES community. In this phase, we connected each node to its authors. For each publication, dataset, or software, we extracted the authors list from its metadata, and we created a distinct node for each author containing the name and surname, the full name (intended as the concatenation of the first name and the surname), the PID (e.g., ORCID), and the rank (i.e., the position of the author in the original authors' list of the publication/dataset/software). The output of this phase is the MES subgraph: a graph extracted from the original OAG representing the MES community, and where publications, datasets and software are interconnected, and each node is connected to its authors.

The MES subgraph is the input of the *Selection & Filtering* phase, where we removed the nodes with incomplete metadata. A node with *complete* metadata must include authors (at least one author must be defined), title, description (e.g., the abstract), date of publication, a pointer to a repository, and a list of keywords describing the product research areas. Then, we filtered out the isolated nodes with zero degree (i.e., no outgoing or ingoing edges). The lack of associations of a publication (dataset or software) could be due to either the non-existence of linked research products in the OAG or the removal of associated research products due to incomplete metadata or non-inclusion in MES. Finally, we filtered out the nodes with metadata presenting multiple URLs pointing to different datasets. This aspect hinders the curation because we cannot uniquely associate a dataset to a repository webpage, and there might be contrasting, or incoherent information reported on different webpages.

In the *Nodes and Edges Curation* phase, we downloaded the PDFs of the publications and automatically extracted title, abstract, the publication sections, authors, keywords, references list, footnotes, and figure and table captions. To parse the PDF we primarily relied on GROBID [27], a machine learning library to extract structured information from scientific documents in PDF format. Specifically, GROBID was crucial in extracting the list of references, author information, and identifying the various sections within the publication. We scraped the repository webpages of the datasets and software and we extracted title, description, authors, keywords, and a set of related research products; each research product is often reported together with the semantics describing the relation between the product and the dataset in the webpage. Publications' metadata have been compared to the information extracted from PDFs, while datasets and software metadata to the information extracted from the repository webpage. If the original metadata were coherent with the extracted information, we kept the metadata information as it is, otherwise we replaced them with the information extracted from the PDFs or the webpages.

The research products listed in the webpage of a dataset allowed us to collect information about referenced or cited publications (`IsReferencedBy`, `IsCitedBy` semantics), supplementary material (`IsSupplementedBy` semantics), and documentation (`Documents` semantics) related to the dataset examined. The semantics associated to each related research product determines the appropriate semantics to assign to the edge connecting the dataset to the respective product. The information extracted from the PDFs allowed us to determine if a publication referenced or/and cited a dataset. If the dataset was mentioned in the references list of a publication, we labeled the edge with `References` semantics; if the mention to a dataset (intended as the pointer to a reference entry, or the mention of the dataset DOI/URL) occurred in the full-text instead, the assigned semantics was `Cites`. Anytime that a dataset was mentioned in the full-text, we stored its position (e.g., captions, footnotes, publication sections, endnotes).

For each pair of connected nodes, we created a new list of semantics, collected processing the PDF of the publication and the webpage of the connected product. We first processed the publication PDF to see if and where the dataset was mentioned; this allowed us to detect the `References`, `Cites` semantics. Then, we looked into the dataset webpage to see if it mentioned the related publication and the semantics able to describe the type of relationship between the two research products. We compared the new set of semantics collected processing the publication and the webpages with those already existing in the MES subgraph. If a semantics of the new generated list did not belong to the original MES subgraph, we added a new edge, and we labeled it with that semantics; if a semantics in the MES subgraph did not belong to the new generated list, we removed the corresponding edge; finally, if one semantics assigned in the MES subgraph belong also to the new list, we left the edge as it was. As for software, since they are limited in number, we manually curated them. Notably, for each pair of publication and software, we parsed the publication to determine citations and references to the software, and analyzed the software webpages to determine the related research products and the associated semantics.

In the *Enrichment* phase, we further enriched the curated graph obtained in the previous phase with new nodes. Since our interest was enriching datasets' connections, we extracted from the dataset webpages all the related research products (e.g., datasets, publications), each one with the associated semantics that identifies the relation between the product and the dataset. For each product not yet in MES we created a new node, and we relied on the provided semantics to connect it to the dataset described in the webpage. The metadata of the new publications included the information extracted from the PDFs, while the metadata of new datasets included the information extracted from their repository webpages. In this phase, we added some connections having as source and target two datasets: connections involving only datasets are used only to indicate that the source dataset is contained in the target one. Edges connecting pairs of datasets have semantics `IsPartOf`. Finally, we connected each new node to its authors; if the authors already exist, we inserted a new edge between the author and the new product; if one or more authors were not in the graph, we created the related nodes.

The last phase of the curation pipeline is *Authors Processing and Disambiguation*. It is important to notice that multiple nodes may represent the same author relying on different metadata; for example, a node may report the full first name, while another one only has the initials. An author name disambiguation procedure is needed to recognize the nodes representing the same person. To achieve this, we relied on two pieces of information: the PID and the full name (i.e., the concatenation of name and surname, or vice versa). A couple of authors represented the same person if one of the following conditions occurred: (i) they shared the same PID, or (ii) the Jaro-Winkler similarity [20] measure applied to the full names exceeded a given threshold equal to 0.95. This similarity measure allowed us to disambiguate authors also when their full names were not exactly the same. For example, the Jaro-Winkler similarity for the authors: *Armand, Leanne* and *Armand, Leanne K* is higher than 0.96, pointing out that they probably refer to the same person. In some cases, the Jaro-Winkler similarity measure failed and returned a high level of similarity despite the authors corresponding to different persons. This happens when both the surname and name are short, or in the case of homonyms. As a consequence, when one of the conditions above occurred, our method failed and we performed manual disambiguation. Despite these limitations, the conditions above rarely occurred since the size of the set examined was limited.

We merged all the nodes representing the same person in a single node whose metadata were the union of the merged authors.

The result of the curation pipeline is a new curated research graph where publications, datasets, and software are interconnected, and each product is connected to its authors. After nodes and edges curation, and authors disambiguation, the graph not only is a trustable representation of the actual MES community but also accurately describes the data publication and most common citation practices, which are useful information to understand the role of data scholarly publication ecosystem.

Table 3. Nodes and edges count after each phase of the curation pipeline. $p \longrightarrow d$ refers to edges connecting publications to datasets, $p \longrightarrow s$ refers to edges connecting publications to software.

	Publications	Datasets	Software	Authors	$p \longrightarrow d$	$p \longrightarrow s$
Community Extraction	104,191	118,110	1,105	924,168	13,703	34
Selection and Filtering	3,793	5,163	22	45,023	7,527	26
Nodes and Edges Curation	3,793	5,163	22	45,023	8,436	43
Enrichment	4,047	5,488	22	50,065	9,649	43
Authors Disambiguation	4,047	5,488	22	21,561	9,649	43

4 ANALYSIS AND RESULTS

In this section, we present some statistics to show how the "original MES subgraph" – i.e., the graph about the MES community originally extracted from the OAG – differs from the "curated MES graph" obtained after the curation pipeline described in Section 3.

The original MES subgraph counts more than 104K publications, 118K datasets, and 1,105 software; while 13,703 edges connect publications to datasets, and 34 publications to software (see *Community Extraction* row in Table 3). After the curation pipeline, the curated MES graph counts 4,047 publications, 5,488 datasets, 22 software, and 21,561 authors; while 9,649 edges connect publications to datasets, and 43 edges connect publications to software.

In Table 3, we show how the original MES subgraph changes after each phase of the curation pipeline, leading to the curated MES graph. The *Selection and Filtering* is the phase removing nodes with incomplete metadata and the isolated ones, and thus incurs in the largest size reduction. The *Nodes and Edges Curation* and the *Enrichment* are the only phases where the number of nodes and edges change. In the *Nodes and Edges Curation* phase, we added less than 1K new edges; the limited increase is related to how curation was performed: we curated only the pairs of nodes that were already connected in the OAG subgraph we considered. In addition, in the majority of the cases, the semantics of the edges has been validated, or replaced with the most appropriate one. Only in few cases we added a new edge between two nodes. Whereas, in the *Enrichment* phase, for almost all the new pairs of nodes we added a single edge, and also in this case, the increase is marginal with respect to the total count of edges. In Table 4, we analyze the most filled fields in the metadata of publications, datasets, and software. Title, authors, date of acceptance, and URL are always filled with only a few exceptions, whereas description and keywords fields are less used. The description field is filled in 85.90% of publications, but only in 21.80% of the datasets. For what concerns the keywords field, instead, the 73.80% of publications and 64.19% of datasets have at least one keyword. In Table 5, we show how many nodes of the original MES subgraph have complete metadata (i.e., title, description, keywords, authors, date of acceptance, and URL fields are all defined), also showing how many nodes are both not isolated and complete. There is a gap between publications and datasets because 70.11% of publications have complete metadata versus only 21.44% of the datasets. Only 5% of publications are connected to one or more research products, 8.55% of datasets and 2.98% of software are connected to one or more publications. The least common case is having both complete metadata and not isolated nodes: this condition characterises only the 3.80% of publications, the 4.43% of datasets, and the 1.99% of software.

In Table 6, we report the number of metadata fields that we updated/corrected in the *Nodes and Edges Curation* phase. We corrected 14 publication titles containing some parsing errors of the PDFs and 796 dataset titles because they contained some extra information erroneously scraped from the dataset webpage. Besides, we modified 587 publication descriptions (i.e., the abstracts), due to errors in parsing which led to the extraction of the wrong portion of text from the PDFs, and errors

Table 4. Number of publications, datasets, and software with a non-empty title, description, authors, date of acceptance, keywords, or URL field. The percentages are computed on the original MES subgraph.

	Publications		Datasets		Software	
	100% = 104,191		100% = 118,110		100% = 1,105	
	Total	%	Total	%	Total	%
Title	104,191	100.00	118,110	100.00	1,105	100.00
Description	89,510	85.90	25,757	21.80	1,100	99.54
Authors	102,018	97.91	117,842	99.77	1,105	100.00
Date of Acceptance	102,677	98.54	118,071	99.96	1,105	100.00
Keywords	76,893	73.80	75,820	64.19	1,077	97.46
URL	104,149	99.95	118,110	100.00	1,105	100.00

Table 5. Analysis of nodes count in the original MES subgraph. *Complete Metadata* reports the number of nodes whose metadata include authors, title, description, keywords, date of acceptance, and URL; *Connected Nodes* reports the number of non-isolated nodes; *Connected & Complete* is the number of connected nodes with complete metadata.

	Publications		Datasets		Software	
	100% = 104,191		100% = 118,110		100% = 1,105	
	Total	%	Total	%	Total	%
Complete Metadata	73,052	70.11	25,326	21.44	1,076	97.37
Connected Nodes	5,309	5.09	10,106	8.55	33	2.98
Connected & Complete	3,967	3.80	5,240	4.43	22	1.99

Table 6. Overview of titles, descriptions, and keywords fields we curated/modified.

	Publications		Datasets		Software	
	100% = 3,793		100,% = 5,163		100% = 22	
	Total	%	Total	%	Total	%
Title	14	0.37	796	15.41	1	4.54
Description	587	15.47	30	0.6	2	9.09
Keywords	738	19.45	468	9.1	0	0.00

in scraping the repository webpage of the publication, which led to wrongly formatted textual content. In 738 publications, and in 468 datasets, we enriched the original set of keywords with those extracted from the PDFs and the webpages, respectively. The URL and the date of acceptance have not been curated. Almost all the software metadata did not need any modification/enrichment.

Nodes curation had a limited impact on authors' nodes: among the 45,023 authors considered in the *Nodes and Edges Curation* phase, we enriched/modified 261 PIDs, 132 full names, 99 first names, and 88 surnames. The authors scraped from the webpages have been crucial to enrich PIDs. The authors extracted from the PDFs and the webpages often include the entire first name of a person, allowing us to enrich the metadata of the authors, whose first name and full name usually reported only their initials.

In Table 7, we overview the edges' semantics in the curated MES graph. We distinguished between (i) the edges that we validated to check if their semantics was correct (VAL columns in green), (ii) the edges we added in the *Nodes and Edges Curation* and *Enrichment* phases of the pipeline

Table 7. Overview of edges semantics in the curated MES graph. $p \longrightarrow d$ and $p \longrightarrow s$ refer to all the publication-dataset and publication-software connections respectively. VAL (*Validated*): the edge correctly describes the correlation between two nodes. ADD (*Added*): a new edge is added; the new edge is present only in the curated MES graph. DEL (*Removed*): the curation revealed that the edge in the original MES subgraph improperly describes the relation between two products; hence the edge is removed from the final graph.

	$p \longrightarrow d$			$p \longrightarrow s$		
Semantics	VAL	ADD	DEL	VAL	ADD	DEL
IsSupplementedBy	2,681	1,585	3	17	4	0
IsSupplementTo	0	0	64	0	0	3
IsReferencedBy	1,466	1,407	1,054	0	0	0
References	483	387	209	0	9	0
IsCitedBy	3	0	96	1	0	0
Cites	349	1,007	926	5	7	0
IsDocumentedBy	0	0	0	0	0	0
Documents	191	90	2	0	0	0
Total	5,173	4,476	2,354	23	20	3
Total (VAL + ADD)	9,649			43		

(ADD columns in yellow), (iii) the edges we removed due to ambiguous or incorrect semantics (DEL columns in red). The edges present in the curated MES graph are those reported in the green and yellow columns. We added a total of 4,476 edges between publications and datasets and 20 between publications and software; we validated 5,173 edges between publications and datasets, and 23 between publications and software – i.e., the 68.72% of the $p \longrightarrow d$, and the 88.46% of the $p \longrightarrow s$ edges at the beginning of the *Nodes & Edges Curation* phase; we removed 2,354 edges between publications and datasets, and 3 between publications and software –i.e., the 31.27% of the $p \longrightarrow d$, and 11.53% of the $p \longrightarrow s$ edges at the beginning of the *Nodes & Edges Curation* phase.

We can observe that almost all the `IsSupplementedBy` labeled edges between publications and datasets in the original MES subgraph were correct. We inserted 1,585 new edges between publications and datasets, and 4 between publications and software. The largest part of them was added in place of `IsCitedBy`, `IsSupplementTo`, `IsReferencedBy` labeled edges we removed due to their incorrect semantics.

The vast majority of the `IsSupplementTo` edges is about datasets deposited on Zenodo. In Zenodo's webpages describing the datasets, the publication connected to a dataset is often specified in the *"Supplementary material"* section, which may lead to ambiguities in defining the supplement or the supplemented products. After manual validation, where we checked the semantics of the relation, we decided to remove all the `IsSupplementTo` labeled edges and replace them with `IsSupplementedBy`semantics. This decision is because we found out that the publication is supplemented by the dataset and not vice versa, as specified in Zenodo's web pages.

We also validated 1,466 `IsReferencedBy` labeled edges between publications and datasets. We discovered that all the removed edges involved datasets deposited in PANGAEA where their descriptive webpages specified a `IsSupplementedBy` semantics instead of the `IsReferencedBy` semantics assigned by the OAG. Hence, we manually verified that the PANGAEA semantics was the correct one and replaced 1,054 `IsReferencedBy` edges with `IsSupplementedBy` ones. By using the information scraped from the dataset repositories webpages we also added 1,407 new `IsReferencedBy` labeled edges. No `IsReferencedBy` labeled edges between publications and software exist.

Table 8. Description of formal and informal citations from the graph and article points of view. We report the occurrences of formal and informal citations publication-datasets pairs ($p \longrightarrow d$) and publication and software nodes ($p \longrightarrow s$) respectively connected with edges whose label is `References` or `Cites`.

Type	Research graph	$p \longrightarrow d, p \longrightarrow s$ 100% = 1,383	
		Total	%
Formal	*References* and *Cites* edges occur together	588	42.52
Informal	*Cites* occurs without *References*	504	36.44
	References occurs without *Cites*	15	1.09
	References and *Cites* edges occur together	276	19.95

Furthermore, we validated 483 `References` labeled edges connecting publications and datasets. The 209 edges we removed were used to connect (publication, dataset) pairs where the publication PDF did not reference the connected dataset or the reference did not occur in the references list. We added 387 new edges we obtained by parsing the publications PDFs and searching for the dataset mentions in the publications references list. We added 9 `References` edges connecting publications to software.

We manually curated the `IsCitedBy` labeled edges, and we found out that most were `IsSupplementedBy` edges. Almost all these edges involved datasets deposited in Dryad repository whose title reported the title of the publication preceded by *"Data from: "*, or *"Supplemental data"*. We updated all these edges, replacing the `IsCitedBy` with the `IsSupplementedBy` semantics. Taking a closer look at `Cites` labeled edges between publications and datasets, we removed 926 edges, while we validated only 349 edges. The largest part of the removed edges connected a publication that did not mention the connected dataset in the full-text. 1,007 edges instead have been added; this result indicates that dataset citations and mentions occurring in the publication full-text are rarely captured by the edges in the original MES subgraph. Only 12 `Cites` edges connect publications to software: 5 have been validated, while 7 have been added.

It is worth noting that `References` and `Cites` are the only semantics, pointing out if a publication exhibits a dataset mention anywhere in the full-text. The total count of `References` and `Cites` edges, connecting publications to datasets or to software, is 2,247 (obtained by adding the values reported for these semantics in green and yellow columns, respectively), which corresponds to the 23.18% of the total count of edges. Furthermore, the number of pairs connected by at least one between the `References` and `Cites` edge semantics is 1,383, which corresponds to the 14.27% of the pairs; this result shows that in the 85.73% of the pairs, the only perusal of the publication PDF is not enough to acknowledge the connected datasets and/or software. We analyzed these pairs and found that they involved only 1,063 publications – the 26.26%, 1,012 datasets – the 18.51%, and 11 software – the 50%. In the portion of the original MES subgraph we curated, there were no `IsDocumentedBy` labeled edges. `Documents` labeled edges connect only publications to datasets; we validated almost all of them and added 90 new edges with such semantics.

It is possible to rely on the curated MES graph to detect the presence of formal and informal citations. Formal citations require the presence of a dataset reference entry in the references list and at least a citation of that entry in the full-text of the article; as a consequence, to detect formal citations in the curated MES graph, it is sufficient to find the pairs of nodes connected by a `Cites` and a `References` labeled edges such that the dataset was included in the references list of the article and the full-text contained at least a pointer to the dataset reference entry. Informal citations, instead, do not require the presence of a reference entry in the references list. Therefore, in the curated MES graph informal citations are all the node pairs where `Cites` and `References` labeled edges do not co-occur. Another example of informal citations is when `Cites` and `References` labeled edges co-occur and the dataset reference is not well-formed, or there is not any formal pointer to the related reference entry in the full-text. The 42.52% of citations are formal, while the remaining 57.48% are informal; we can see no significant gap between formal and informal citation counts. The 36.44% of citations involve publications that mention a dataset or software in the full-text without including it in the references list. Only 1.09% of datasets are referenced (hence they appear in the references list) without being cited in the full-text; we considered a reference without a pointer in the full-text an informal citation. The 19.95% of node pairs have been considered informal citations, despite being connected by a `References` and a `Cites` edges. A limited portion of these pairs involves datasets that are included in the references list of the associated publications, but the related reference entries are never formally cited in the full-text; at the same time, the DOI or the title of the datasets are reported in the full-text of the publications. The largest portion of these pairs, instead, involve publications that include the dataset in the references list, and the dataset reference entry is not well-formed: in these cases, the DOI is missing or not the one provided in the metadata of the linked dataset. One possible reason for not including the DOI in a reference entry is publishing the paper before the dataset. In this case, the DOI of the dataset has not been provided at the time of the article's publishing. A difference in the DOIs might be related to the citation of a paper describing the dataset instead of citing the dataset itself.

Supplementary datasets are research products derived by and/or essential to perform the experiments described in a publication. As a consequence, the publication and its supplementary materials are strictly correlated. We analyzed the 4,287 node pairs connected by a `IsSupplementedBy` labeled edge to detect if the curated metadata of the connected products could evidence this strict correlation. We found out that in 31.53% node pairs, the publication and the dataset share the same title (or part of it), 31.28% of the pairs share the same description (or part of it), 96.38% share one or more authors, while 70.81% have the same publication year. In this respect, the title and the description are the most informative field to infer the `IsSupplementedBy` semantics between a publication and a dataset. On the contrary, sharing the authors or the publication year is a less distinctive feature because these aspects are characteristic also of other semantics such as `IsReferencedBy` or `Documents`.

Moving forward to the next phase of the pipeline, the *Enrichment* phase, we found new valuable data leading to the addition of 254 publications, and 325 datasets.

Finally, in the *Authors Processing and Disambiguation* phase, we eliminated the duplicated authors, reducing their number from $50K$ to $21K$ – a 56% decrease. In Table 9, we show how the filled fields changed in authors' metadata before and after disambiguation. The percentage of filled name and surname fields increased from 79.80% to 90%. For what concerns PIDs instead, the percentage increased from 37.15% to 39.40%. The full name is the only field always filled.

In Table 10, we analyzed the maximum, minimum, and average number of publications, datasets, and software produced by the authors of publications, datasets, and software, respectively. Considering the authors who contributed to at least one dataset, we see that after the disambiguation, the maximum and the average number of datasets per author increased from 103 to 226 and from 2.76 to 3.22, respectively. Similarly, for the publication authors, we found out that the maximum number of publications per author increased from 7 to 107 and the average from 1.05 to 1.75. For what concerns software, the maximum and the average software per author moderately increased

Table 9. Authors metadata filling before and after disambiguation. The percentages are computed concerning the number of authors before and after disambiguation.

	Before Disambiguation 100% = 50,065		After Disambiguation 100% = 21,561	
	Total	%	Total	%
Fullname	50,065	100.00	21,561	100.00
Name	39,952	79.80	19,418	90.00
Surname	39,951	79.80	19,422	90.00
PID	18,601	37.15	8,494	39.40

Table 10. Maximum, minimum, average number of publications, datasets, software per author.

	Before disambiguation			After disambiguation		
	Max	Min	Avg	Max	Min	Avg
Publications	7	1	1.05	107	1	1.75
Datasets	103	1	2.76	226	1	3.22
Software	2	1	1.17	4	1	1.38

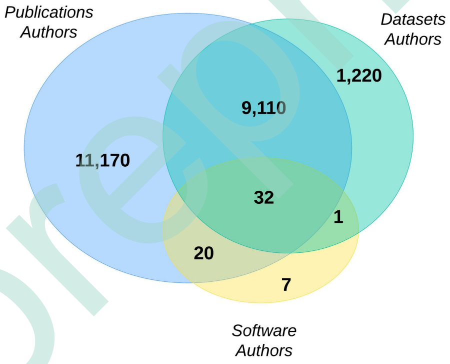

Fig. 3. Venn diagram of authors of publications, datasets, and software. Intersections identify authors who contributed to two or three different types of research products.

the disambiguation. These results highlight that a disambiguation phase is needed to associate the research products with the correct author. The duplication of authors and the ambiguities lead to underestimating authors' contributions and, consequently, their impact.

We conducted an analysis of authors to check if there is a clear distinction between the authors of publications, datasets, and software. We illustrate this analysis in a Venn diagram in Figure 3. We found 20,332 publication authors, 10,363 dataset authors, and 60 software authors. 11,170 publications authors – 54.93% of the total – did not contribute to any dataset or software. 1,220 datasets authors – 11.77% of the total – and 7 software authors contributed exclusively to datasets and software, respectively. The largest part of datasets and software authors also contributed to publications, in particular, 32 authors contributed to all three types of products, 9,110 authors contributed both to datasets and publications, and 20 to publications and software. These results

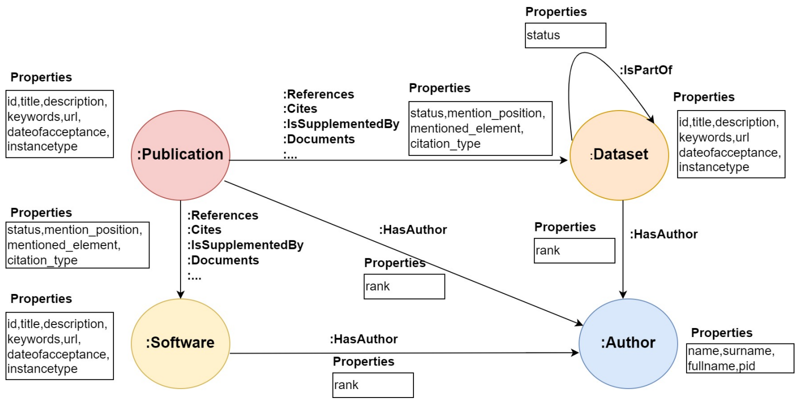

Fig. 4. Graph data model of the released property graph. Inside the rectangles, we put the properties of nodes and relationships. Publications, datasets, and software share the same set of properties, as well as edges from publications to datasets or software.

show that there is a large part of people exclusively work on publications, whereas dataset and software authors are keener to contribute also to publications.

5 GRAPH AVAILABILITY

We release three graphs: (i) the curated MES graph obtained from the curation; (ii) the curated MES graph including also removed edges; (iii) the original MES subgraph extracted from the OAG, not curated, and without the isolated nodes.

The graphs are available at [19]. We provide the resources as property graphs that can be imported in *Neo4j*. We also provide the resources as JSON files. In Figure 4, we report the model of the property graph we release. In order to query and manipulate the graph, Cypher[16] query language must be used. The *instancetype* property of publications, datasets and software indicates the research product type (e.g., the publication can be a journal article, book, book chapter); the edge *status* property indicates whether the edge has been added, verified, or removed from the original graph; *mention_position*, *citation_type*, and *mentioned_element* occur when the label is `References` or `Cites` and refer to the place in the full-text where the reference or citation occurred (e.g., references, footnotes, introduction), the type of citation (i.e., informal, formal) and what statement was used to refer to the dataset (e.g., title, DOI, URL).

For each graph, we provide 5 JSON files including publications, datasets, software, authors, and relationships respectively. Each line of the files contains a JSON representing a research product (or relationship). Each JSON line representing a product includes the properties associated with that product, depicted in Figure 4. For each relationship, we also include the IDs of the source and target research products.

6 DISCUSSION

The study on the MES subgraph extracted from the OAG, and the definition and application of the curation pipeline, pointed out some critical aspects concerning the original OAG, and the curated MES graph we release.

Our results evidenced that nodes with incomplete metadata are more likely to be isolated in the graph than nodes with complete sets. The information of nodes' metadata is useful to infer new

[16] Cypher – https://neo4j.com/developer/cypher

labeled edges between two research products (e.g., metadata are valuable to infer the presence of `IsSupplementedBy` semantics). Metadata can be automatically inferred or deposited by researchers in repositories. It is worth noting that, in the second case, the deposition of complete metadata is a time-demanding task for which researchers are seldom rewarded. This might be a key reason why metadata are often inaccurate and imprecise. We detected a high heterogeneity in metadata, and publications and datasets reported two contrasting shreds of evidence. In particular, most publications have complete metadata; conversely, the largest part of the datasets has incomplete sets. Publications traditionally have a primary role in the scholarly ecosystem, and they have always been the unique yardstick for evaluating the work of researchers. Consequently, there is usually a greater interest in increasing the visibility of publications rather than datasets or software. The lack of complete metadata about datasets hinders the creation of a large network where publications, datasets, and other research outcomes are interconnected; this affects not only the discoverability and the reproducibility of the experiments but also the visibility of datasets authors, who barely achieve the credits they deserve. Given the importance and the potential of metadata in the scholarly communication ecosystem, providing complete and detailed metadata is an essential task that should be computed for all the deposited research products, independently of their type.

In the OAG, the metadata about a research product also includes the metadata about its authors. Given that authors' metadata are deposited along with the research product metadata, it is common for different products sharing the same author to describe the same author with different metadata. Several aspects correlate to the presence of multiple descriptions for the same person. First, different researchers may provide different metadata for the same author. In addition, the type of research product influences the completeness of the author's metadata; the author who contributed both to a publication and a dataset may have more complete information in the metadata of the former. Finally, different metadata are also related to how they are collected; manually deposited metadata may differ from the inferred ones. Capturing and disambiguating all the metadata representing the same author is crucial to improve the computation of author-based statistics essential to understanding an author's impact, attributing credits, and monitoring research collaborations and topics of interest. However, author disambiguation is an open problem, currently unsolved for authors in the OAG, due to the vast amount of authors and the presence of homonyms, synonyms, and ambiguities. In the analyses we performed, we found that the largest part of datasets and software authors are also authors of publications. As of the time of writing, many barriers still prevent authors from achieving credits for datasets and software outcomes. They are usually included in the related publication's authors list. Contributing also to publications allows them to benefit from the well-established credits attributions mechanisms proper of the publications.

Another aspect we discuss concerns semantics assignment. New labeled edges between nodes can be automatically created by applying inference algorithms or manually deposited by the researchers. The assignment of the most appropriate semantics is a challenging task that requires a deep understanding of the wide range of the semantics provided by DataCite; this is why both the automatically inferred semantics and those manually assigned may be inaccurate and imprecise. Examples of ambiguities in semantics assignments are the `IsCitedBy, IsReferencedBy` and `IsSupplementTo` semantics that we replaced with `IsSupplementedBy`. The `Cites` labeled edges we removed show that this semantics has been used to highlight the correlation between a publication and a dataset, ignoring the actual definition of citation. The analysis performed on `Cites` and `References` labeled edges pointed out that dataset and software mentions occurring in the full-text of a publication are rarely represented in the OAG; accurate processing of the original PDF documents is important to create new connections and consequently promoting the discoverability of the datasets and software and their authors. These aspects inevitably affect the reliability of the ORG; the amount of relationships existing in the OAG (more than 3B) makes it impossible to control the correctness of automatically and manually assigned semantics, which are often imprecise or inaccurate.

About References and Cites edges, we analyzed the occurrences of formal and informal citations. Informal citations tend to prevail over formal ones, but there is no deep gap between them. This points out that within the MES community, there is not a data citation practice commonly adopted: dataset may be cited formally, adding the dataset to the references list of a paper and inserting in the paper full-text a pointer to it, or informally, hence mentioning the datasets (its DOI or title) in the full-text without mentioning it in the references list. It is worth mentioning that the most common data repositories in MES (e.g., Pangaea, Zenodo, Figshare) already provide some data citation guidelines, and all recommend formally citing data; in addition, these repositories all provide a DOI for the deposited dataset. Despite these facts, researchers still adopt different practices. This is related to the essential role datasets achieved in the last decade and, at the same time, the lack of a well-established data citation reward system and credit attribution mechanism: as a consequence, researchers who agree on the importance of data in the scholarly ecosystem formally cite datasets, otherwise they cite them informally.

The metadata (in)completeness, the presence of duplicated authors, and the partial reliability of edges semantics are all aspects that make the OAG unsuitable for developing and testing computational methods to perform link prediction, author name disambiguation, data search, and data enrichment strategies based not only on publications and their authors but also on datasets and software. The curated MES graph proposed in this work, generated by applying the curation pipeline described in Section 3, tackles the problems above by providing disambiguated authors and curated and enriched nodes metadata and edges semantics. Our results proved that curation is essential to fix improper metadata and semantics and provide a more reliable representation of the connections between publications and datasets or software.

7 LIMITATIONS AND FUTURE WORKS

Despite the improvements the resource we share can bring to the scholarly ecosystem, there are some limitations to be considered. Firstly, the proposed pipeline cannot be fully automatized: the majority of nodes and edges have been automatically curated, but in some cases manual curation was necessary. Secondly, our curation pipeline has been created to work on graphs whose dimensions are similar to those of MES. These considerations limit the application of the proposed pipeline to sizeable graphs containing millions of nodes and relationships since it would be time-demanding. Furthermore, the authors' disambiguation procedure we proposed might have some issues with very short names when the authors have similar surnames and similar first names, and in some cases of homonymy. As a consequence, in all these cases manual curation would be required. Despite these limitations, the curation pipeline could be applied to other different communities of the OAG having similar dimensions.

We see that introducing a curation procedure in the OAG is crucial to improve its quality and trustworthiness. In this respect, having large curated scholarly graphs would substantially improve the scholarly ecosystem, providing valuable ground-truths for link prediction, recommendation, and authors disambiguation tasks. A potential improvement in this direction would be the creation of a comprehensive framework, to facilitate and speed up the curation process. Given that curation is a time-demanding task, it would be of great help to have a framework able to operate on limited portions of the OAG, selected by the users, and not strictly related to a community of interest. This would offer not only the possibility to curate different portions of graphs according to the user's need, but also to allow the user to manually curate metadata and semantics, improving the overall framework reliability. To this end, some modifications should be applied to our pipeline. First of all, the authors disambiguation pipeline should be changed in order to be faster and better recognize homonyms and identify short names. Moreover, we should provide more general scrapers we can apply to different dataset repositories webpages.

Finally, we plan to extend the analyses concerning formal and informal citations. We will focus on `IsSupplementedBy`, `Cites` and `References` labeled edges to analyze the positions of formal and informal citations in the full-text, the co-occurrence of formal and informal citations, and determine whether the cited datasets are also part of the supplementary material.

8 FINAL REMARKS

In this paper, we presented our work towards the creation of a curated scholarly graph representing the MES community of OpenAIRE. To this end, we defined a five-phase curation pipeline that takes in input the OAG, extracts the subgraph representing the MES community, curates and enriches nodes and edges, and disambiguates authors. The information extracted from the PDFs of the publications and the webpages of datasets and software has been crucial to validate, enrich, and fix the information stored in nodes' metadata and edges semantics. To disambiguate authors, we primarily relied on the Jaro-Winkler similarity measure; the authors representing the same person have been merged, as well as their metadata.

Our analyses pointed out a high heterogeneity in metadata, and metadata completeness was highly dependent on the type of research product: publications metadata were more complete than datasets ones. Furthermore, the presence of connections in the graph is related to metadata completeness.

Moreover, we found out that curation had a strong impact on semantics: approximately half of the edges ($4.5K$), in fact, have been added during curation, and more than $2K$ edges have been removed. Besides, the assignment of the correct semantics is related to the researcher's expertise and understanding of semantics meaning, inevitably leading to inconsistencies in semantics assignments. We analysed the `References` and the `Cites` edges, and we detected the prevalence of informal citations over formal ones: this is related to the lack of a commonly adopted practice on how to cite data: despite data repositories usually provide a set of instructions on how to cite data, the decision on how to include a dataset in a paper is demanded to the researcher, and a wide range of practices currently co-exist.

The graph we provide is a unique resource for the scholarly ecosystem. The released graph is currently available at [19] and it includes 4,047 publications, 5,488 datasets, 22 software, and 21,561 authors. 9,649 relationships connect publication to datasets, and 43 connect publication to software; the semantics used to label the edges are: `Documents`, `Cites`, `References`, `IsSupplementedBy` and their inverse, and they belong to the DataCite metadata schema [7]. It is focused on the interconnections between publications and datasets or software, and this makes it a useful ground truth to evaluate link prediction, data search, and data enrichment methods involving different types of products. The authors are disambiguated, and each author is represented as a distinct node: this is crucial in the definition of authors' disambiguation algorithms, and automatic methods to compute authors' impact, and monitor collaborations.

As the number of publications, datasets, and software in the OAG continues to increase, we plan to periodically update our resource as soon as a new version of the OAG is released; we will extract and curate the new versions of the MES community, integrating into our resource all the missing nodes and edges. The resulting graph will be a faithful and up-to-date snapshot of the MES community subgraph.

ACKNOWLEDGEMENTS

This work was partially funded by the EC H2020 project OpenAIRE-Nexus (Grant agreement 101017452). Moreover, we are deeply grateful to Dr Allyson Lister and the FAIRsharing team for the suggestions they provided us with.

REFERENCES

[1] Simone Angioni, Angelo Salatino, Francesco Osborne, Diego Reforgiato Recupero, and Enrico Motta. 2021. AIDA: A knowledge graph about research dynamics in academia and industry. *Quantitative Science Studies* 2, 4 (2021), 1356–1398.

[2] Miriam Baglioni, Alessia Bardi, Argiro Kokogiannaki, Paolo Manghi, Katerina Iatropoulou, Pedro Principe, André Vieira, Lars Holm Nielsen, Harry Dimitropoulos, Ioannis Foufoulas, et al. 2019. The OpenAIRE research community dashboard: on blending scientific workflows and scientific publishing. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 56–69.

[3] Christopher W Belter. 2014. Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One* 9, 3 (2014), e92590.

[4] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference*. 1365–1375.

[5] Peter Buneman, Dennis Dosso, Matteo Lissandrini, and Gianmaria Silvello. 2021. Data citation and the citation graph. *Quant. Sci. Stud.* 2, 4 (2021), 1399–1422. https://doi.org/10.1162/qss_a_00166

[6] Adrian Burton, Hylke Koers, Paolo Manghi, Markus Stocker, Martin Fenner, Amir Aryani, Sandro La Bruzzo, Michael Diepenbroek, and Uwe Schindler. 2017. The scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine* 23, 1/2 (2017).

[7] DataCite Metadata Working Group. 2021. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. https://doi.org/10.14454/3w3z-sa82 Accessed: 2022-07-11.

[8] Hilary M Davis and John N Vickery. 2007. Datasets, a shift in the currency of scholarly communication: Implications for library collections and acquisitions. *Serials Review* 33, 1 (2007), 26–32.

[9] Suhendry Effendy and Roland HC Yap. 2017. Analysing trends in computer science research: A preliminary study using the microsoft academic graph. In *Proceedings of the 26th international conference on world wide web companion*. 1245–1250.

[10] Michael Färber. 2019. The Microsoft Academic Knowledge Graph: A linked data source with 8 billion triples of scholarly data. In *International semantic web conference*. Springer, 113–129.

[11] Michael Färber and David Lamprecht. 2021. The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies* 2, 4 (2021), 1324–1355.

[12] Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange. 2016. Identifying and improving dataset references in social sciences full texts. *arXiv preprint arXiv:1603.01774* (2016).

[13] G Nigel Gilbert and Steve Woolgar. 1974. Essay Review: The quantitative study of science: an examination of the literature. *Science studies* 4, 3 (1974), 279–294.

[14] Muhammad Haris, Markus Stocker, and Sören Auer. 2022. Enriching Scholarly Knowledge with Context. *arXiv preprint arXiv:2203.14617* (2022).

[15] Veronika Henk, Sahar Vahdati, Mojataba Nayyeri, Mehdi Ali, Hamed Shariat Yazdi, and Jens Lehmann. 2019. Metaresearch recommendations using knowledge graph embeddings. In *RecNLP workshop of AAAI Conference*.

[16] Drahomira Herrmannova and Petr Knoth. 2016. An analysis of the microsoft academic graph. *D-lib Magazine* 22, 9/10 (2016), 37.

[17] Ijaz Hussain and Sohail Asghar. 2018. DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science* 44, 6 (2018), 830–847.

[18] Tin Huynh, Kiem Hoang, Tien Do, and Duc Huynh. 2013. Vietnamese author name disambiguation for integrating publications from heterogeneous sources. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 226–235.

[19] Ornella Irrera, Andrea Mannocci, Paolo Manghi, and Gianmaria Silvello. 2022. *A Novel Curated Scholarly Graph Connecting Textual and Data Publications*. https://doi.org/10.5281/zenodo.7464120

[20] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*. 243–246.

[21] Mohamad Yaser Jaradeh, Kuldeep Singh, Markus Stocker, and Sören Auer. 2021. Triple classification for scholarly knowledge graph completion. In *Proceedings of the 11th on Knowledge Capture Conference*. 225–232.

[22] Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Question Answering on Scholarly Knowledge Graphs. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 19–32.

[23] Jinseok Kim. 2019. Scale-free collaboration networks: An author name disambiguation perspective. *Journal of the Association for Information Science and Technology* 70, 7 (2019), 685–700.

[24] John Kratz and Carly Strasser. 2014. Data publication consensus and controversies. *F1000Research* 3 (2014).

[25] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 177–187.

[26] Hanwen Liu, Huaizhen Kou, Chao Yan, and Lianyong Qi. 2019. Link prediction in paper citation network to construct paper correlation graph. *EURASIP Journal on Wireless Communications and Networking* 2019, 1 (2019), 1–12.

[27] Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*. Springer, 473–474.

[28] Xiao Ma, Ranran Wang, and Yin Zhang. 2019. Author name disambiguation in heterogeneous academic networks. In *International Conference on Web Information Systems and Applications*. Springer, 126–137.

[29] Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, Pedro Principe, Michele Artini, Amelie Becker, Michele De Bonis, et al. 2019. The OpenAIRE research graph data model. *Zenodo* (2019).

[30] Duncan M McRae-Spencer and Nigel R Shadbolt. 2006. Also by the same author: Aktiveauthor, a citation graph approach to name disambiguation. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. 53–54.

[31] Hailey Mooney and Mark P Newton. 2012. The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication* 1, 1 (2012).

[32] Carlos Pedro Muniz, Ronaldo Goldschmidt, and Ricardo Choren. 2018. Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowledge-Based Systems* 156 (2018), 129–137.

[33] Mojtaba Nayyeri, Gökce Müge Cil, Sahar Vahdati, Francesco Osborne, Andrey Kravchenko, Simone Angioni, Angelo Salatino, Diego Reforgiato Recupero, Enrico Motta, and Jens Lehmann. 2021. Link prediction of weighted triples for knowledge graph completion within the scholarly domain. *IEEE Access* 9 (2021), 116002–116014.

[34] Mojtaba Nayyeri, Gokce Muge Cil, Sahar Vahdati, Francesco Osborne, Mahfuzur Rahman, Simone Angioni, Angelo Salatino, Diego Reforgiato Recupero, Nadezhda Vassilyeva, Enrico Motta, et al. 2021. Trans4E: Link prediction on scholarly knowledge graphs. *Neurocomputing* 461 (2021), 530–542.

[35] Mojtaba Nayyeri, Sahar Vahdati, Jens Lehmann, and Hamed Shariat Yazdi. 2019. Soft marginal transe for scholarly knowledge graph completion. *arXiv preprint arXiv:1904.12211* (2019).

[36] Lucila Ohno-Machado, Susanna-Assunta Sansone, George Alter, Ian Fore, Jeffrey Grethe, Hua Xu, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Anupama E Gururaj, Elizabeth Bell, et al. 2017. Finding useful data across multiple biomedical data repositories using DataMed. *Nature genetics* 49, 6 (2017), 816–819.

[37] Hyoungjoo Park, Sukjin You, and Dietmar Wolfram. 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 (2018), 1346–1354.

[38] Silvio Peroni and David Shotton. 2020. OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies* 1, 1 (2020), 428–444.

[39] Nicolas Robinson-García, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 2016. Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology* 67, 12 (2016), 2964–2975.

[40] Tanay Kumar Saha, Baichuan Zhang, and Mohammad Al Hasan. 2015. Name disambiguation from link data in a collaboration graph using temporal and topological features. *Social Network Analysis and Mining* 5, 1 (2015), 1–14.

[41] David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. Somesci-A 5 star open data gold standard knowledge graph of software mentions in scientific articles. *arXiv preprint arXiv:2108.09070* (2021).

[42] Christian Schulz, Amin Mazloumian, Alexander M Petersen, Orion Penner, and Dirk Helbing. 2014. Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Science* 3 (2014), 1–14.

[43] Jae-Wook Seol, Seok-Hyoung Lee, and Kwang-Young Kim. 2016. Author disambiguation using co-author network and supervised learning approach in scholarly data. *International Journal of Software Engineering and Its Applications* 10, 4 (2016), 73–82.

[44] Gianmaria Silvello. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology* 69, 1 (2018), 6–20.

[45] Qingyun Sun, Hao Peng, Jianxin Li, Senzhang Wang, Xiangyun Dong, Liangxuan Zhao, S Yu Philip, and Lifang He. 2020. Pairwise learning for name disambiguation in large-scale heterogeneous academic networks. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 511–520.

[46] Jie Tang. 2016. AMiner: Toward understanding big scholar data. In *Proceedings of the ninth ACM international conference on web search and data mining*. 467–467.

[47] Hung Nghiep Tran, Tin Huynh, and Tien Do. 2014. Author name disambiguation by using deep neural network. In *Asian conference on intelligent information and database systems*. Springer, 123–132.

[48] Huaiyu Wan, Yutao Zhang, Jing Zhang, and Jie Tang. 2019. Aminer: Search and mining of academic social networks. *Data Intelligence* 1, 1 (2019), 58–76.

[49] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.

[50] Tong Zeng and Daniel E. Acuna. 2020. Finding datasets in publications: the Syracuse University approach. https://doi.org/10.5281/zenodo.4402304 Tong Zeng was funded by the China Scholarship Council #201706190067. Daniel E. Acuna was funded by the National Science Foundation awards #1646763 and #1800956.

[51] Tong Zeng, Longfeng Wu, Sarah Bratt, and Daniel E Acuna. 2020. Assigning credit to scientific datasets using article citation networks. *Journal of Informetrics* 14, 2 (2020), 101013.