Standing on Shoulders or Feet? The Usage of the MSR Data Papers

Zoe Kotti, Diomidis Spinellis
Department of Management Science and Technology
Athens University of Economics and Business
Athens, Greece
{t8150062, dds}@aueb.gr

Abstract*—*Introduction: **The establishment of the Mining Software Repositories (MSR) Data Showcase conference track has encouraged researchers to provide more data sets as a basis for further empirical studies.**
Objectives: **Examine the usage of the data papers published in the MSR proceedings in terms of use frequency, users, and use purpose.**
Methods: **Data track papers were collected from the MSR Data Showcase and through the manual inspection of older MSR proceedings. The use of data papers was established through citation searching followed by reading the studies that have cited them. Data papers were then clustered based on their content, whereas their citations were classified according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge.**
Results: **We found that 65% of the data papers have been used in other studies, with a long-tail distribution in the number of citations. MSR data papers are cited less than other MSR papers. A considerable number of the citations stem from the teams that authored the data papers. Publications providing repository data and metadata are the most frequent data papers and the most often cited ones. Mobile application data papers are the least common ones, but the second most frequently cited.**
Conclusion: **Data papers have provided the foundation for a significant number of studies, but there is room for improvement in their utilization. This can be done by setting a higher bar for their publication, by encouraging their use, and by providing incentives for the enrichment of existing data collections.**

***Index Terms*—Software engineering data; Bibliometrics; Data paper; Reproducibility; Data showcase track**

To appear in *MSR '19: Proceedings of the 16th Conference on Mining Software Repositories*, 2019

> *"Indeed, one of my major complaints about the computer field is that whereas Newton could say, 'If I have seen a little farther than others, it is because I have stood on the shoulders of giants,' I am forced to say, 'Today we stand on each other's feet.' Perhaps the central problem we face in all of computer science is how we are to get to the situation where we build on top of the work of others rather than redoing so much of it in a trivially different way."*
>
> — Richard Wesley Hamming[1]

I. Introduction

Software engineering data sets are often a key ingredient for performing empirical software engineering by testing a hypothesis through an experiment run on such data [16]. They can be used to empirically evaluate software product quality and development process attributes and also to create or verify estimation models [47]. In addition, publicly available data sets can help researchers perform so-called *exact* replications of existing studies and thus address potential internal validity problems [79]. These, in contrast to *conceptual* replications, which follow an independently developed experimental procedure, attempt to control as many factors of the original study as possible, varying almost no (in *dependent* replications) or only some (in *independent* replications) conditions of the experiment [79].

Yet, at least in the past, data sets for software engineering research were small in size and difficult to obtain [43]. The situation has improved over the past decades with the emergence of open source software [90]. For this reason researchers have collaborated [16] through various initiatives to develop data set repositories, such as the *International Conference on Predictive Models and Data Analytics for Software Engineering* (PROMISE), or to promote the sharing and publication of data, as through the US National Institute of Standards and Technology's "Error, Fault, and Failure Data Collection and Analysis Project" [92], the Mining Software Repositories (MSR) conference data showcase track, or the *awesome-msr* GitHub project.[2]

[1] 1968 ACM Turing Award Lecture [36]
[2] https://github.com/dspinellis/awesome-msr

The MSR data showcase track, established in 2013, aims at encouraging the research community to develop, share, and document software engineering research data sets. In the words of the 2013 MSR conference chairs [105],

> "rather than describing research achievements, data papers describe datasets curated by their authors and made available to others. Such papers provide description of the data, including its source; methodology used to gather it; description of the schema used to store it, and any limitations and/or challenges of this data set."

In the past decade tens of data set papers have been published in the MSR conference. Given the effort that went into creating the data sets and publishing the corresponding papers, it is reasonable to investigate what the outcome has been. This study aims to answer the question by examining the usage of the data papers published in the MSR proceedings in terms of use frequency, users, and use purpose. The study's contributions are:

- the systematic collection of research that has been based on MSR data papers,
- the categorization of the subjects tackled using MSR data papers, and
- the quantitative analysis of the MSR data papers' impact.

In the following Section II we describe our study's methods. We then present our results in Section III, discuss them in Section IV, and outline the associated validity threats in Section V. The study is complemented by an overview of related work in Section VI, followed by our conclusions in Section VII. The data sets associated with our study (data papers, citing papers, categorizations, MSR papers, and citations) are made available online.[3]

II. Methods

We framed our investigation on the usage of MSR data papers in terms of the following research questions.

RQ1 *What data papers have been published?* We answer this by finding all data papers published in the MSR proceedings, and further elaborate by classifying them based on the year of publication and the content.

RQ2 *How are data papers used?* We answer this by collecting all citations to MSR data papers and classifying them according to their subject and authors.

RQ3 *What is the impact of published data papers?* We answer this through the statistical analysis and visualization of the citations and their slicing according to their type.

A. Data Paper Collection and Clustering

To perform the particular research, we first obtained all data papers of the proceedings of the International (Working) Conference on Mining Software Repositories (MSR). By the term *data papers* we refer to all papers included in the data showcase track of the MSR proceedings, as well as other papers from older proceedings that primarily provide a data set (e.g. Conklin et al.'s collection of FLOSS data and analyses [15]).

TABLE I
MSR Data Papers by Year

Year	Data Papers
2005	[80] [55] [15]
2006	[41]
2010	[61]
2012	[40]
2013	[8] [18] [28] [89] [91] [38] [50] [83] [57] [12] [84] [46] [30] [70] [35]
2014	[44] [76] [32] [58] [66] [102] [72] [48] [9] [22] [31] [95] [20] [56] [6]
2015	[88] [77] [62] [45] [25] [4] [81] [94] [53] [7] [39] [65] [68] [101] [34] [29]
2016	[69] [2] [85] [99] [5] [104] [63]
2017	[59] [1] [103] [73] [73] [74] [97]
2018	[52] [100] [60] [24] [96] [75] [64] [98] [82] [23] [13] [51] [78] [26] [19]

To acquire the aforementioned papers, we searched through the programs of the MSR conferences on their respective website. Programs that contained an explicit Data Showcase section immediately informed us of the particular year's data papers. In contrast, programs that did not include the forenamed section, were manually searched for potential research offering data sets. From the gathered studies, those which genuinely offered complete data sets were included in our data paper archive. In total we identified the 81 data papers shown in Table I.

Following the collection, we classified the data papers into distinct clusters. This classification would provide us with a different perspective on the analysis of the papers.

We manually sorted all data research into different categories in the way described further on. The first data paper in order was assigned into the first category. The second paper was semantically compared to the first one; if any conceptual relation was recognized between them, then they were grouped together. Otherwise, the second paper was placed in a new category. The procedure continued accordingly; all papers were classified into existing clusters in case of conceptual relation, or into new clusters when no association with the existing categories was noted. Eventually, a set of seven categories was formed, as presented in Table II.

B. Data Paper Use Identification and Classification

To conduct the analysis on the data paper research usage, we implemented the *Identification of Research* and *Study Selection* processes, as proposed in Kitchenham's work on procedures for performing systematic reviews [42].

The Identification of Research was made through widely used and established platforms that provide citation data: *Google Scholar*,[4] *Scopus*—Elsevier's abstract and citation database[5] and the *ACM Digital Library*.[6] Most research papers

[3]http://doi.org/10.5281/zenodo.2544957

[4]https://scholar.google.com/

[5]https://www.scopus.com/

[6]https://dl.acm.org/

TABLE II
DATA PAPER CATEGORIES AND CITATIONS

Category	Data Papers	Cited DPs	Non-cited DPs	Citation Ratio (%)	Citations
Repository Data & Metadata	26	17	9	65	255
Bugs, Defects, Smells	17	10	7	59	36
Software Evolution	12	8	4	67	18
Software Development Process	9	6	3	67	30
Computing Education, Programming Practices & Skills	7	5	2	70	17
Human-centered Data	6	3	3	50	18
Mobile Application Data & Metadata	4	4	0	100	66
Total	81	53	28	65	440

that were not publicly available were provided to us through personal communication with the authors.

After collecting the citations of a particular data paper, we followed the Study Selection process. Specific criteria were applied to the collected research, in order to ensure quality and validity for our analysis. First, we applied the whitelisting practice and kept studies of conferences' proceedings, articles, master's and doctoral theses, books, and technical reports. Studies published in multiple venues were only listed once. Priority was given sequentially to books, articles, proceedings, reports, and, lastly, theses. We additionally decided to retain studies written in the English language, due to its widespread adoption for scientific communication.

The main criterion for retaining citing studies was their use of the data sets of the papers they had cited. We term these *strong* citations. Research that solely referred to a data paper without using its data set was not taken into account in our study. A representative example of a non-strong citation is the study of repository badges in the npm ecosystem [87], which has cited the collection of social diversity attributes of programmers [88], although it has not used its data.

To determine the most cited data papers, we counted for each research paper its total strong citations and sorted them in descending order (see Table III).

Furthermore, we classified the collected strong citations according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge [11] (SWEBOK—see Table IV).

During the collection of the strong citations, we noticed that some studies shared at least one author with the data paper they had cited. For these data studies, we divided their strong citations into three categories. The first category contains references to the data papers made by their first author. The second category includes citations made by at least one co-author of the respective data paper. The remaining references that were not made by any author of the particular data paper were placed in the third category.

C. Citation Analysis

To assess in an objective manner the impact of data papers compared to other MSR papers we collected all MSR papers and coupled them with citation data provided by Scopus. This process differs from the one described in the preceding Section II-B, because citations are not manually evaluated regarding actual use, and are retrieved only from a single source (Scopus). Consequently, the collected metrics are only appropriate for assessing relative rather than absolute impact.

We first created a data set of all 1267 MSR papers by downloading the complete DBLP computer science bibliography database,[7] and filtering its XML records to obtain only those whose *booktitle* tag contained *MSR*. We split the MSR papers at hand into two sets: data papers (as determined in Section II-A) and the rest. We also split the MSR papers by year to simplify the selection of samples.

As citations and data papers are unevenly distributed over the years (see Figure 1), we created a collection mirroring the yearly distribution of data papers in order to compare in a fair manner citations to data papers against citations to other MSR papers. We created this collection as follows. For each year in which N data papers were published, we randomly chose N non-data papers from the MSR papers published in the same year.

To assess research building on data papers we also created a set of MSR papers that cite MSR data papers. We did this by calculating the intersection between all MSR papers and the papers that use them (as determined in Section II-B). Although this new set of papers citing data papers is not exhaustive (it only contains MSR papers), it allows us to compare the citation metrics of these papers against those of a known tractable population, namely MSR papers as a whole.

We then used the Scopus REST API to obtain the number of times each MSR paper was cited. The citation data obtained in this step are not comparable with those we obtained through the widespread search and manual filtering described in Section II-B, because they may be associated with false positives and false negatives. However, they allow comparisons to be made between different MSR sets, because all citation metrics are obtained through the same methods employed by Scopus and all probably suffer through the same types of bias.

Finally, we joined the Scopus citation data with the sets obtained in the previous steps. We then calculated simple descriptive statistics for the citation counts of the following sets:

- all MSR data papers,
- a sample of MSR non-data papers mirroring the yearly distribution of data papers,

[7] https://dblp.org/

TABLE III
TOP FIVE DATA PAPERS IN NUMBER OF CITATIONS

Title	Data Paper	Year	Category	Citations
The GHTorent Dataset and Tool Suite	[30]	2013	Repository Data & Metadata	165
AndroZoo: Collecting Millions of Android Apps for the Research Community	[2]	2016	Mobile Application Data & Metadata	57
Lean GHTorrent: GitHub Data on Demand	[31]	2014	Repository Data & Metadata	24
Who Does What During a Code Review? Datasets of OSS Peer Review Repositories	[35]	2013	Software Development Process	16
The Maven Repository Dataset of Metrics, Changes, and Dependencies	[70]	2013	Repository Data & Metadata	12
The Eclipse and Mozilla Defect Tracking Dataset: A Genuine Dataset for Mining Bug Information	[46]	2013	Bugs, Defects, Smells	12
The Emotional Side of Software Developers in JIRA	[63]	2016	Human-centered Data	12

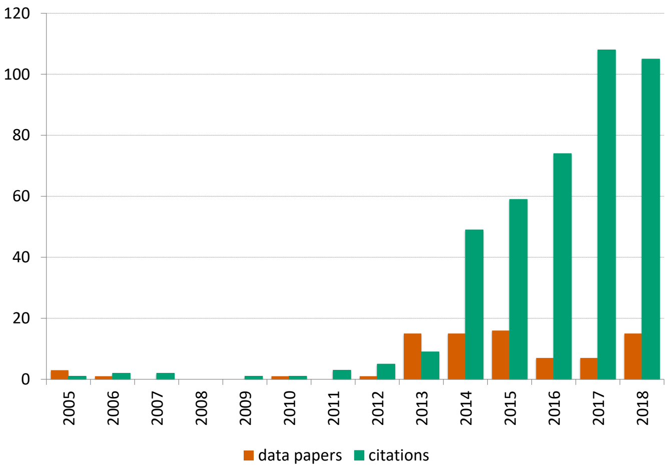

Fig. 1. Timeline of the data papers and the strong citations. Each year depicts the number of data papers published in the particular year and the number of studies published in the particular year that are based on any data paper.

- all MSR non-data papers for years in which data papers were published, and
- MSR papers citing MSR data papers.

III. RESULTS

We examined all 81 identified data papers, which comprise about 15% of the 507 papers published in the MSR conference in the years when data papers appeared. The MSR data papers are associated with 1169 citations to them, coming from 982 distinct studies. Out of the 1169 citations, 440 (419 distinct studies) use the data sets provided by the data papers. The remaining 729 citations (610 distinct studies) refer to data papers without utilizing the particular data sets. Six citations were received from their respective authors through personal communication, as stated in Section II-B, but no access was provided for another three. (These three studies have been excluded from the total citations.)

The timeline of the data studies and the research based on them is depicted in Figure 1. For each year, the number of published data papers is showcased, along with the number of studies published the particular year, which have been based on any data paper. There has been a significant rise in the number of data papers since 2013, which is the year when the data showcase track was founded [105]. Until then, 2005 was the year with the most data studies recognized. The smallest number of data showcase research papers—seven—was published in 2016 and 2017. Nevertheless, 2018 indicates a double increase in data publications—15 (see Table I).

From the classification of the data papers, as described in Section II-A, seven data categories emerged. Table II shows for each category the number of data papers it comprises, the number of strongly cited and non-cited data papers, and the references that have been made to them. We consider as *non-cited* the data papers with either non-strong citations or no citations at all. The categories are sorted in descending order of data papers.

Repository Data & Metadata preponderate. The particular category consists of 26 studies that provide raw or processed data, along with descriptive statistics and analyses. The collection of Java source code of the Merobase Component Finder project [38] is part of this category.

Bugs, Defects, Smells concern security failures, software inconsistencies and unfavorable programming practices detected in a variety of software applications and ecosystems. For instance, VulinOSS offers a data set of security vulnerabilities in open-source systems [26].

Software Evolution involves twelve collections with information on the evolution of artifacts such as operating systems [82], software products [103], or frameworks [91].

Nine data papers were grouped together due to their common intention of assisting developers in ordinary development practices, such as maintenance [18] and verification [35]. These papers constitute the *Software Development Process* category.

Papers that shared records regarding novices' and experts' programming practices and abilities (e.g. the list of Scratch programs of students [1]) were classified in *Computing Education, Programming Practices & Skills*. The aim of this category is to facilitate studies on Computing Education.

The class of *Human-centered Data* is composed of data papers that concentrate on the social aspect [88] and the emotional side of developers [63].

The last category we defined is the *Mobile Application Data & Metadata*, which shares collections of Android applications and meta-information [45]. Only four papers represent this category, however their second-in-order number of strong citations attests their significance and their differentiation from the other classes.

According to our analysis on the strong citations to data papers (see Table III), Gousios's collection of GitHub repository data [30] is the most cited study with a total of 165 uses, followed by the AndroZoo collection of Android ap-

TABLE IV
AREAS OF CITING STUDIES

SWEBOK Knowledge Area	Studies	Percentage
Software Quality	119	28.4
Software Maintenance	65	15.5
Software Engineering Process	41	9.8
Software Configuration Management	38	9.1
Software Engineering Management	37	8.8
Software Construction	33	7.9
Software Engineering Professional Practice	33	7.9
Software Engineering Models and Methods	21	5.0
Software Testing	16	3.8
Software Design	11	2.6
Software Requirements	4	1.0
Software Engineering Economics	1	0.2

plications [2] with 57 citations. GHTorrent's complementing work, Lean GHTorrent [31], and the peer review data set [35] have also attracted the attention of the research community. Finally, the repository of Maven meta-information [70], along with the collections of Eclipse and Mozilla defects [46] and developers' sentiments [63] share the same number of uses—twelve.

Overall, 53 MSR data papers (65%) have been utilized (by their authors or others), while 28 papers' data sets have never been used. The majority of them belong to the categories of Repository Data & Metadata and Bugs, Defects, Smells. The unused number of data sets is noteworthy, considering the effort required to produce them.

The categorization of the studies based on data papers according to the knowledge areas of the SWEBOK (see Table IV) suggests that research on *Software Quality* and *Software Maintenance* uses data papers to a considerable extent. On the other hand, only a slight portion of research on *Software Requirements* and *Software Engineering Economics* uses data showcase papers.

Furthermore, concerning the use of data papers by their respective authors, our findings show that 37 papers have been referenced by the teams that authored them. Specifically, 15 studies have been solely deployed either by their first author or his co-authors. Figure 3 depicts for each data paper cited at least once by the first author or the co-authors, the percentage of the uses that stem from the first author, the co-authors, and other unrelated teams. The data papers are sorted in ascending order based on the percentage of the sum of the strong citations made by the first author and the co-authors. For instance, 67% of the references to the collection of APIs usage information [77] were made by the first author.

The impact of published data papers can be deduced from Table V, which compares citations to data and non-data papers. (The three data papers missing from the table are those published in MSR '05, which are not tracked by Scopus.) The table shows that data papers are typically cited less often, compared to others of the MSR conference in terms of the median and average number of references. This occurs both in terms of yearly-weighted samples and as a whole. Also, MSR papers that cite data papers appear to be cited about the same as other MSR papers.

TABLE V
CITATION METRICS BY PAPER TYPE

Metric	Data Papers	Non-DP (Sample)	Non-DP (All)	Citing DP
N	78	78	429	49
Min	0	0	0	0
Max	158	107	306	147
Median	5	10	8	10
Avg	9.8	16.9	17.0	15.7
Stddev	19.9	21.7	27.7	25.3

TABLE VI
VENUES WITH RESEARCH BASED ON DATA PAPERS

Venue	Papers	Percentage
MSR	52	13.8
ICSE	24	6.4
CoRR	21	5.6
ICSME	16	4.2
SANER	14	3.7
EmpSE	13	3.4
ESEM	6	1.6
IEEE TSE	6	1.6
Other conference	176	46.7
Other journal	49	13.0

Table VI shows the venues where research that is based on data papers has been published. We see that more than a third of the corresponding papers are published in top-tier conferences and journals. This showcases the high quality of research that is conducted based on data papers. We examined by hand the papers published in the Computing Research Repository (CoRR), and found that almost all of them (19) are fairly recent (published in 2017 or 2018). This indicates that they are probably archival submissions of material that will eventually also end up in a conference or journal.

The timeline of the data paper uses is depicted in Figure 1. The strong citations of all data papers were summed up and illustrated as yearly records. We see that citations have risen since 2014, which was expected after the data showcase track's introduction in 2013. Only six studies were identified before the category's establishment.

In addition, we studied the growth of data paper use in a five-year window after the data papers' publication. The limit five was preferred because it provided us with sufficient insight, without excluding too many papers that were less than five years old. Consequently, we included data studies published in the years 2005–2014. The majority of them reveal a peak in the number of strong citations during the second year of their existence, but appear to have a significant decrease of uses the following year (see Figure 2). Research based on data papers seems to plateau after the third year of their life.

IV. DISCUSSION

As evidenced by the large increase in the published data papers since the MSR data showcase track was formalized, it is apparent that the track has catalyzed the publication of data papers. With data papers being more than 15% of the MSR

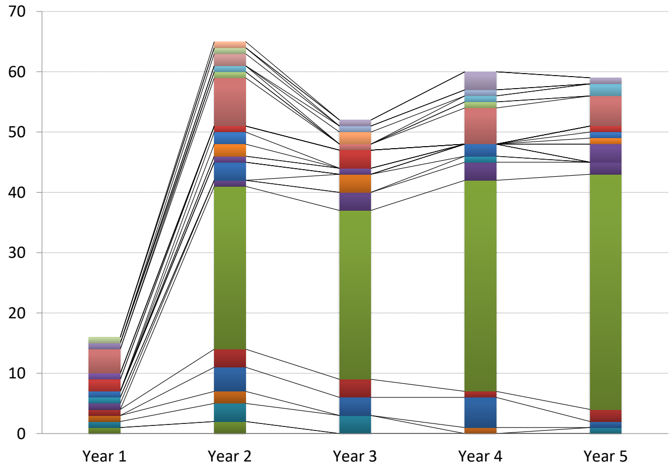

Fig. 2. Timeline of strong citations to data papers published from 2005–2014 over a five-year window. Each cited data paper is represented with the same color along the years.

publications, it is clear that the MSR data track has spurred a new type of publication yielding each year a notable number of studies. More generally, the data track's success in driving the publication of data papers indicates that a suitably themed conference track can in some cases drive research toward a given direction.

The categories of data papers (Table II) span equally product and process, but product-oriented papers outnumber the process ones. This can be explained by the preponderance of publicly available product data, which is associated with open source software projects, over process data, which is more difficult to come by. To overcome this bias it might be worth to focus the MSR call for data papers on specific topics each year, although past experience with calling for the publication of particular data types has not been encouraging [92].

The studies that cite data papers span the SWEBOK knowledge areas fairly unequally. It seems that software quality and maintenance can be profitably studied using materials from MSR data papers, but software design, requirements, and economics less so. Given the, by definition, primary importance of all SWEBOK areas, it would seem that the MSR data showcase track chairs could promote studies associated with the less covered areas by adjusting the track's call for papers to specifically invite data sets targeting them. We acknowledge, however, that for certain SWEBOK areas, such as software economics, the release of data sets is hard due to the often proprietary nature of the corresponding data. Nevertheless, data sets for underrepresented SWEBOK areas might really have lasting impact in their subfield despite being less popular.

With each data paper cited on average 5.4 times, it appears that data papers are in general useful for conducting other empirical studies. Many of these studies are published in top-notch venues (see Table VI), indicating the high quality of studies that can be performed through data papers. On the other hand, at least for MSR papers that cite data papers, their basis on published empirical data does not seem to increase their impact in terms of citations to them (see last column of Table V).

Regarding impact, the number of strong citations to data papers is constantly rising (Figure 1), indicating that the concept of data papers has long-term value. The enduring usefulness of specific data papers is also apparent by looking at the timeline of strong citations to specific MSR data showcase papers over a five-year period (Figure 2). The trend of the most cited papers retaining their citation number or obtaining ever more citations is yet another manifestation of the Matthew effect in science [54]. A survey or interview study of authors of data papers or research based on them might provide insights on what motivates authors to conduct data research and the reasons why particular data sets are more frequently preferred.

Yet, surprisingly for an artifact whose main purpose is for others to build on, data papers are cited significantly less than other MSR papers. One might think that this is due to the 28 out of 81 (35%) of the data papers that are never strongly used. The citation's distribution long tail—just 9% of the data papers are cited by 67% of all citing studies—could be another reason. However, by comparing the distribution of citations to data papers (according to Scopus) with that of citations to non-data papers (Figure 4), we see that the two distributions are similar in shape. It is apparent that the reason for the lower citation count of MSR data papers is the overall lower number of citations to each data paper compared to the citations to each non-data paper.

There are two reasons that could explain this phenomenon. First, data papers may not publish data that is actually useful for conducting other studies. To address this problem the MSR program committee could adopt more stringent criteria for accepting data papers, though this will certainly lead to a decline in the number of accepted papers, and there is no guarantee that a more selective track will still select the papers that will be most frequently cited. The track's toughening of data sharing can be counterbalanced by promoting open science initiatives, such as the ACM Artifact Review and Badging policy [10]. Second, software engineering researchers may be reluctant to use data stemming from MSR data papers in their research. Reasons behind this could be mistrust in the data's quality [49], difficulty to use the data, the researchers' reluctance to work with data coming outside their organization—also known as the not invented here syndrome [67], or a fear that working with publicly available data is less likely to yield original results. The high number of papers used by their authors (Figure 3) corroborates this second reason.

Although one might also expect that a data paper is typically only cited mainly when it is actually used, our findings do not support this assertion. We manually identified 440 strong citations; far fewer than half of the 1169 total citations that were made to data papers according to our results. This demonstrates that citations to any kind of published studies (including data research) can be made for a variety of reasons: to set the context, to replicate a method, to describe related work, or even to set a given study apart from unrelated work.

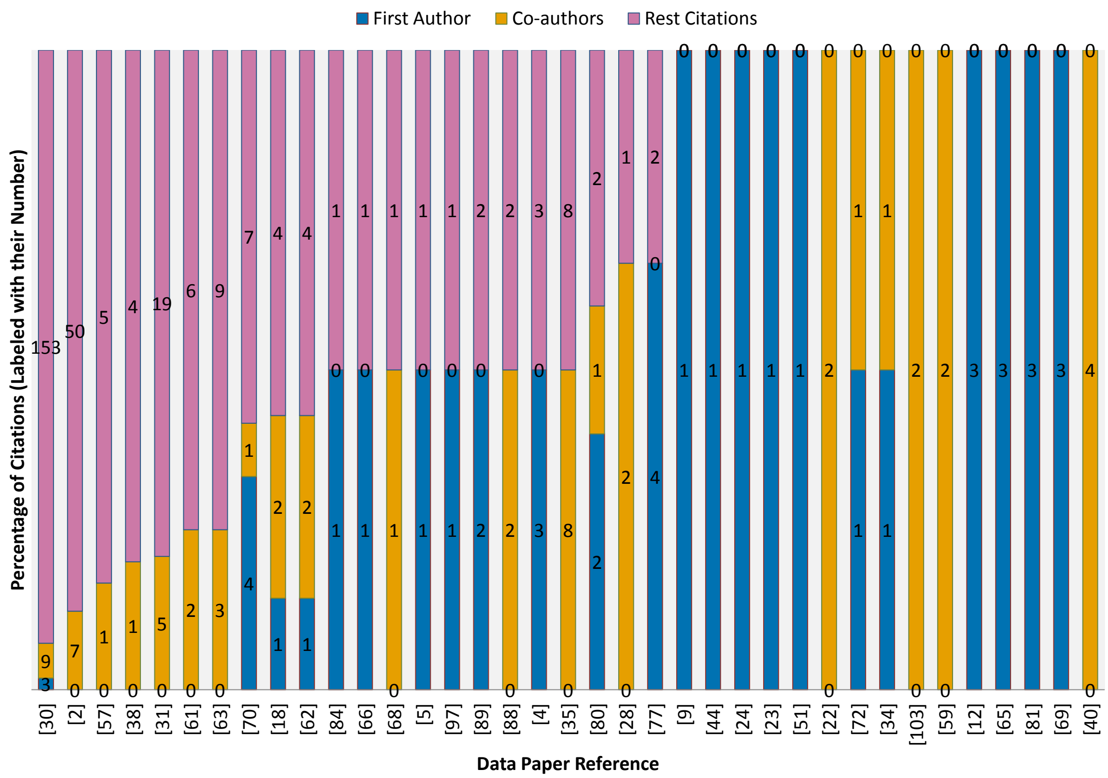

Fig. 3. Use of data papers by their authors (%). Data papers used at least once by the same first author or any of his co-authors are represented by the number of strong citations made by the first author, the co-authors, and other unrelated teams.

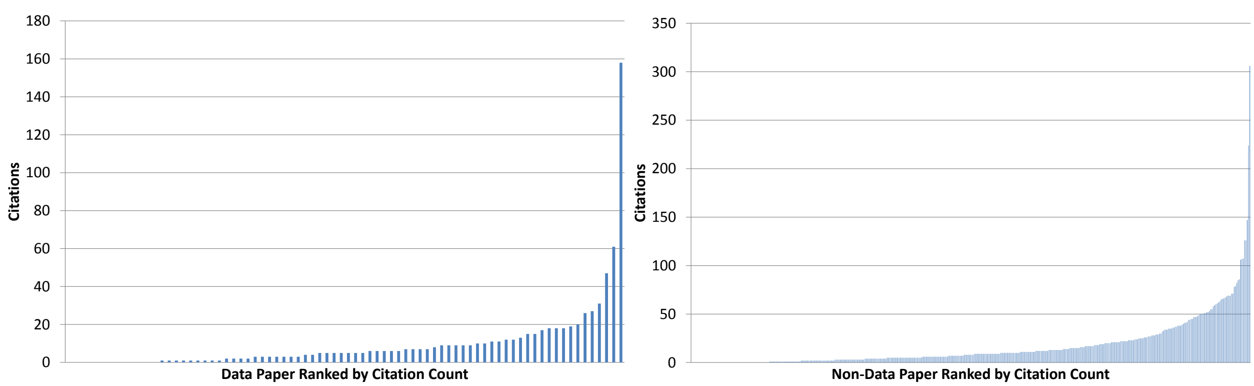

Fig. 4. Distribution in the number of citations to MSR data papers (left) and non-MSR data papers (right)

V. Threats to Validity

The study's external validity in terms of generalizability, obviously suffers by studying only data papers that have been published within the framework of the MSR conference and ignoring venues such as the PROMISE conference (consider e.g. reference [21]) or the *Empirical Software Engineering* (e.g. reference [86]). However, studying the MSR conference in isolation allowed us to analyze the effect of establishing the MSR data showcase track, and to compare citation counts among different groups of papers (Section II-C), without the bias associated with a paper's publication venue.

The major threats to the study's internal validity stem from the steps where we followed manual processes involving subjective judgment: the selection of data papers before the showcase track was introduced, the filtering of studies that actually use data papers, the clustering of data papers, and the categorization of studies using data papers. The trustworthiness of all these could be improved by having them performed by multiple raters and calculating statistics regarding interrater reliability.

Apart from subjective judgment regarding assignment to specific categories, the clustering of data papers introduced in Table II holds another serious threat associated with the establishment of the categories themselves. As elaborated in Section II-A, categories resulted from a conceptual analysis of the corresponding data studies. The validity risk associated with the particular process could be handled through the use of multiple raters, as stated above, and also through the implementation of topic analysis, followed by clustering using, for example, machine learning methods. Particularly, topic analysis could be used to infer the subject of study of each data paper, while machine learning clustering would provide insight regarding the accuracy of the categories that were formed by hand, by comparing them to the ones produced automatically.

VI. Related Work

A variety of evaluations have been conducted through research analysis. We recognize two major fields of evaluations: surveys and bibliometrics. Surveys review and summarize previously published studies of a particular topic through qualitative analysis. Webster and Watson [93] have authored a thorough guide on writing high quality literature reviews. On the other hand, bibliometrics are statistical analyses of written publications. We consider our work part of the bibliometric research, and to the best of our knowledge, we are the first to conduct a quantitative review of data paper usage.

A first step regarding bibliometric research in the field of software engineering models was made in 2004 [16] by the organisers of the PROMISE workshop, in their attempt "to strengthen the community's faith in software engineering models". Authors of such models were asked to submit, along with their work, a related data set to the PROMISE repository.

Many individuals have also carried out interesting quantitative research on various topics. Robles [71] conducted bibliometric research on papers that contained experimental analyses of software projects and were published in the MSR proceedings from 2004–2009. His objective was to review their potential replicability. The outcome proves that MSR authors prefer publicly available data sources from free software repositories. However, the amount of publicly available processed data collections then was very low, a fact we also stated in our results. Concerning replicability, Robles found that only a limited number of publications are replication friendly.

Liebchen and Shepperd [49] performed a different quantitative analysis on data sets. Their aim was to assess quality management techniques used by authors when producing data collections. They found that a surprisingly small percentage of studies take data quality into consideration. The authors of this work stress the need for more quality data rather than quantity data. To achieve this, they advise researchers to provide clear description of the procedures they follow prior to their analysis and data archiving. They also encourage the use of automated tools for assessing quality and the use of sensitivity analysis.

Another related publication is Cheikhi and Abran's [14] survey on data repositories. They noticed that the lack of structured documentation of PROMISE and ISBSG repositories impeded researchers' attempts to find specific types of data collection. To address this problem, they supplemented these data collections with additional information such as the subject of the study, the availability of data files and of further descriptions, and also their usefulness for benchmarking studies. Information on the subject of study was retrieved after the classification of the data studies based on the subject, reflecting our data paper classification.

In the field of Systems and Software Engineering, the five-year study of Glass and Chen [27] assesses scholars and institutions based on the number of papers they have published in related journals. Their results indicate that the high-ranked institutions are mainly academes, most of which are located in the United States. The rest are from the Asia-Pacific region and lastly, Europe. The leading institution of this list is the Carnegie Mellon University, and the top scholar is Khaled El Emam of the Canadian National Research Council.

A second evaluation of the ISBSG software project repository was carried out by Almakadmeh and Abran [3]. Their purpose was to assess the repository from Six Sigma measurement perspective and correlate this assessment with software defect estimation. They found that the ISBSG MS-Excel data extract contains a high ratio of missing data within the fields related to the total number of defects. They consider this outcome a serious challenge, especially for studies that use the particular data set for software defect estimation purposes.

The analysis on the Search Based Software Engineering (SBSE) publications [17] is the first bibliometric research of this community, covering a ten-year list of studies, from 2001–2010. The evaluation is concentrated on the categories of Publication, Sources, Authorship, and Collaboration. Estimations of various publication metrics are included for the following years. Along with the metric forecasting, the authors also studied the applicability of bibliometric laws in SBSE, such as Bradfords and Lotka.

In the same context, Harman et al. [37] assessed research trends, techniques and their applications in SBSE. They classified literature of SBSE, in order to extract specific knowledge on distinct areas of study. Then they performed a trend analysis, which supplied them with information on activity in these areas. Finally, for each area of study, they recognize and present opportunities for further improvement, and avenues for supplementary research.

The work of Gu in [33] is another interesting bibliometric analysis. The main point of evaluation in this study is the productivity of authors in the field of knowledge management (KM). To conduct the analysis, Gu collected articles published in the (former) ISI Web of Science from 1975–2004. He then recorded all unique productive authors, along with their contribution and authorship position, in order to examine their productivity and degree of involvement in their research publications. The results indicate that 86% of authors have only written one publication. As far as citation frequency is concerned, Gu proves its significant correlation with the reputation of the journal it has been published to. On the other hand, his findings reveal no correlation between R&D expenditures and research productivity or citation counts.

VII. Conclusions

The MSR data showcase track has been successful in encouraging the publication of data papers. Data papers are generally used by other empirical studies, though not as much as one might expect or hope for. The gatekeepers of science, such as journal editors and program committees, can address this by setting a higher bar for the publication of data papers and by encouraging their use. An additional policy to improve the use and impact of data papers might be to provide incentives for researchers to enrich existing collections of data instead of reproducing similar data sets from scratch. Such incentives could involve awarding a most influential data paper award or inviting papers where researchers describe how they expanded upon a data track study.

Acknowledgments

Panos Louridas provided insightful comments on this manuscript. This work has received funding from: the European Union's Horizon 2020 research and innovation programme under grant agreement No 825328; the GSRT 2016–2017 Research Support (EP-2844-01); and the Research Centre of the Athens University of Economics and Business, under the Original Scientific Publications framework 2019.

References

[1] Efthimia Aivaloglou, Felienne Hermans, Jesús Moreno-León, and Gregorio Robles. A dataset of Scratch programs: Scraped, shaped and scored. In *Proceedings of the 14th International Conference on Mining Software Repositories*, MSR '17, pages 511–514, Piscataway, NJ, USA, 2017. IEEE Press.

[2] Kevin Allix, Tegawendé F. Bissyandé, Jacques Klein, and Yves Le Traon. Androzoo: Collecting millions of Android apps for the research community. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 468–471, New York, NY, USA, 2016. ACM.

[3] Mhammed Almakadmeh and Alain Abran. The ISBSG software project repository: An analysis from Six Sigma measurement perspective for software defect estimation. *Journal of Software Engineering and Applications*, 10(8):693–720, July 2017.

[4] Harald Altinger, Sebastian Siegl, Yanja Dajsuren, and Franz Wotawa. A novel industry grade dataset for fault prediction based on model-driven developed automotive embedded software. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 494–497, Piscataway, NJ, USA, 2015. IEEE Press.

[5] Sven Amann, Sarah Nadi, Hoan A. Nguyen, Tien N. Nguyen, and Mira Mezini. MUBench: A benchmark for API-misuse detectors. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 464–467, New York, NY, USA, 2016. ACM.

[6] Boris Baldassari and Philippe Preux. Understanding software evolution: The Maisqual Ant data set. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 424–427, New York, NY, USA, 2014. ACM.

[7] Titus Barik, Kevin Lubick, Justin Smith, John Slankas, and Emerson Murphy-Hill. FUSE: A reproducible, extendable, internet-scale corpus of spreadsheets. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 486–489, Piscataway, NJ, USA, 2015. IEEE Press.

[8] David Binkley, Dawn Lawrie, Lori Pollock, Emily Hill, and K. Vijay-Shanker. A dataset for evaluating identifier splitters. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 401–404, Piscataway, NJ, USA, 2013. IEEE Press.

[9] Remco Bloemen, Chintan Amrit, Stefan Kuhlmann, and Gonzalo Ordóñez Matamoros. Gentoo package dependencies over time. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 404–407, New York, NY, USA, 2014. ACM.

[10] Ronald F. Boisvert. Incentivizing reproducibility. *Commun. ACM*, 59(10):5–5, September 2016.

[11] Pierre Bourque and Richard E. Fair, editors. *Guide to the Software Engineering Body of Knowledge*. IEEE Computer Society, New York, version 3.0 edition, 2014. Available online http://www.swebok.org.

[12] Simon Butler, Michel Wermelinger, Yijun Yu, and Helen Sharp. INVocD: Identifier name vocabulary dataset. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 405–408, Piscataway, NJ, USA, 2013. IEEE Press.

[13] Kyriakos C. Chatzidimitriou, Michail D. Papamichail, Themistoklis Diamantopoulos, Michail Tsapanos, and Andreas L. Symeonidis. npm-miner: An infrastructure for measuring the quality of the npm registry. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 42–45, New York, NY, USA, 2018. ACM.

[14] Laila Cheikhi and Alain Abran. PROMISE and ISBSG Software Engineering data repositories: A survey. In *Proceedings of the Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, IWSM-Mensura '13, pages 17–24, Piscataway, NJ, USA, October 2013. IEEE Press.

[15] Megan Conklin, James Howison, and Kevin Crowston. Collaboration using OSSmole: A repository of FLOSS data and analyses. In *Proceedings of the 2nd International Workshop on Mining Software Repositories*, MSR '05, pages 1–5, New York, NY, USA, 2005. ACM.

[16] B. Cukic. Guest editor's introduction: The promise of public software engineering data repositories. *IEEE Software*, 22(6):20–22, November 2005.

[17] Fabrício Gomes de Freitas and Jerffeson Teixeira de Souza. Ten years of Search Based Software Engineering: A bibliometric analysis. In Myra B. Cohen and Mel Ó Cinnéide, editors, *Proceedings of the 3rd International Symposium on Search Based Software Engineering*, SSBSE '11, pages 18–32, Berlin, Heidelberg, September 2011. Springer Berlin Heidelberg.

[18] Bogdan Dit, Andrew Holtzhauer, Denys Poshyvanyk, and Huzefa Kagdi. A dataset from change history to support evaluation of software maintenance tasks. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 131–134, Piscataway, NJ, USA, 2013. IEEE Press.

[19] Vasiliki Efstathiou, Christos Chatzilenas, and Diomidis Spinellis. Word embeddings for the software engineering domain. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 38–41, New York, NY, USA, 2018. ACM.

[20] Gabriel Farah, Juan Sebastian Tejada, and Dario Correal. OpenHub: A scalable architecture for the analysis of software quality attributes. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 420–423, New York, NY, USA, 2014. ACM.

[21] Rudolf Ferenc, Zoltán Tóth, Gergely Ladányi, István Siket, and Tibor Gyimóthy. A public unified bug dataset for Java. In *Proceedings of the 14th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE'18, pages 12–21, New York, NY, USA, 2018. ACM.

[22] Kenji Fujiwara, Hideaki Hata, Erina Makihara, Yusuke Fujihara, Naoki Nakayama, Hajimu Iida, and Kenichi Matsumoto. Kataribe: A hosting service of historage repositories. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 380–383, New York, NY, USA, 2014. ACM.

[23] Jian Gao, Xin Yang, Yu Jiang, Han Liu, Weiliang Ying, and Xian Zhang. JBench: A dataset of data races for concurrency testing. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 6–9, New York, NY, USA, 2018. ACM.

[24] Franz-Xaver Geiger, Ivano Malavolta, Luca Pascarella, Fabio Palomba, Dario Di Nucci, and Alberto Bacchelli. A graph-based dataset of commit history of real-world Android apps. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 30–33, New York, NY, USA, 2018. ACM.

[25] Daniel M. German, Bram Adams, and Ahmed E. Hassan. A dataset of the activity of the Git super-repository of Linux in 2012. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 470–473, Piscataway, NJ, USA, 2015. IEEE Press.

[26] Antonios Gkortzis, Dimitris Mitropoulos, and Diomidis Spinellis. VulinOSS: A dataset of security vulnerabilities in open-source systems. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 18–21, New York, NY, USA, 2018. ACM.

[27] R.L Glass and T.Y Chen. An assessment of systems and software engineering scholars and institutions (19982002). *Journal of Systems and Software*, 68(1):77–84, 2003.

[28] Mathieu Goeminne, Maëlick Claes, and Tom Mens. A historical dataset for the Gnome ecosystem. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 225–228, Piscataway, NJ, USA, 2013. IEEE Press.

[29] Jesus M. Gonzalez-Barahona, Gregorio Robles, and Daniel Izquierdo-Cortazar. The MetricsGrimoire database collection. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 478–481, Piscataway, NJ, USA, 2015. IEEE Press.

[30] Georgios Gousios. The GHTorent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 233–236, Piscataway, NJ, USA, 2013. IEEE Press.

[31] Georgios Gousios, Bogdan Vasilescu, Alexander Serebrenik, and Andy Zaidman. Lean GHTorrent: GitHub data on demand. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 384–387, New York, NY, USA, 2014. ACM.

[32] Georgios Gousios and Andy Zaidman. A dataset for pull-based development research. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 368–371, New York, NY, USA, 2014. ACM.

[33] Yinian Gu. Global knowledge management research: A bibliometric analysis. *Scientometrics*, 61(2):171–190, October 2004.

[34] Mayy Habayeb, Andriy Miranskyy, Syed Shariyar Murtaza, Leotis Buchanan, and Ayse Basar Bener. The Firefox temporal defect dataset. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 498–501, Piscataway, NJ, USA, 2015. IEEE Press.

[35] Kazuki Hamasaki, Raula Gaikovina Kula, Norihiro Yoshida, A. E. Camargo Cruz, Kenji Fujiwara, and Hajimu Iida. Who does what during a code review? datasets of OSS peer review repositories. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 49–52, Piscataway, NJ, USA, 2013. IEEE Press.

[36] R. W. Hamming. One man's view of computer science. *Journal of the ACM*, 16(1):3–12, January 1969.

[37] Mark Harman, S. Afshin Mansouri, and Yuanyuan Zhang. Search Based Software Engineering: A comprehensive analysis and review of trends techniques and applications. Technical Report TR-09-03, Department of Computer Science, King's College London, and Brunel Business School, Brunel University, London, UK, April 2009.

[38] Werner Janjic, Oliver Hummel, Marcus Schumacher, and Colin Atkinson. An unabridged source code dataset for research in software reuse. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 339–342, Piscataway, NJ, USA, 2013. IEEE Press.

[39] Vassilios Karakoidas, Dimitris Mitropoulos, Panos Louridas, Georgios Gousios, and Diomidis Spinellis. Generating the blueprints of the Java ecosystem. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 510–513, Piscataway, NJ, USA, 2015. IEEE Press.

[40] Iman Keivanloo, Christopher Forbes, Aseel Hmood, Mostafa Erfani, Christopher Neal, George Peristerakis, and Juergen Rilling. A Linked Data platform for mining software repositories. In *Proceedings of the 9th Working Conference on Mining Software Repositories*, MSR '12, pages 32–35, Piscataway, NJ, USA, 2012. IEEE Press.

[41] Sunghun Kim, Thomas Zimmermann, Miryung Kim, Ahmed Hassan, Audris Mockus, Tudor Girba, Martin Pinzger, E. James Whitehead, Jr., and Andreas Zeller. TA-RE: An exchange language for mining software repositories. In *Proceedings of the 3rd International Workshop on Mining Software Repositories*, MSR '06, pages 22–25, New York, NY, USA, 2006. ACM.

[42] Barbara Kitchenham. Procedures for performing systematic reviews. Technical Report TR/SE-0401, Department of Computer Science, Keele University, Keele, Staffs, UK, July 2004.

[43] Barbara A. Kitchenham, Shari Lawrence Pfleeger, Lesley M. Pickard, Peter W. Jones, David C. Hoaglin, Khaled El Emam, and Jarrett Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Trans. Softw. Eng.*, 28(8):721–734, August 2002.

[44] Daniel E. Krutz and Wei Le. A code clone oracle. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 388–391, New York, NY, USA, 2014. ACM.

[45] Daniel E. Krutz, Mehdi Mirakhorli, Samuel A. Malachowsky, Andres Ruiz, Jacob Peterson, Andrew Filipski, and Jared Smith. A dataset of open-source Android applications. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 522–525, Piscataway, NJ, USA, 2015. IEEE Press.

[46] Ahmed Lamkanfi, Javier Pérez, and Serge Demeyer. The Eclipse and Mozilla defect tracking dataset: A genuine dataset for mining bug information. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 203–206, Piscataway, NJ, USA, 2013. IEEE Press.

[47] L. Lavazza and L. Santillo. Historical data repositories in software engineering: Status and possible improvements. In *2012 Joint Conference of the 22nd International Workshop on Software Measurement and the 2012 Seventh International Conference on Software Process and Product Measurement*, pages 221–225, October 2012.

[48] Alina Lazar, Sarah Ritchey, and Bonita Sharif. Generating duplicate bug datasets. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 392–395, New York, NY, USA, 2014. ACM.

[49] Gernot A. Liebchen and Martin Shepperd. Data sets and data quality in Software Engineering. In *Proceedings of the 4th International Workshop on Predictor Models in Software Engineering*, PROMISE '08, pages 39–44, New York, NY, USA, 2008. ACM.

[50] Alexander C. MacLean and Charles D. Knutson. Apache commits: Social network dataset. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 135–138, Piscataway, NJ, USA, 2013. IEEE Press.

[51] Vadim Markovtsev and Waren Long. Public Git archive: A big code dataset for all. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 34–37, New York, NY, USA, 2018. ACM.

[52] Pedro Martins, Rohan Achar, and Cristina V. Lopes. 50K-C: A dataset of compilable, and compiled, Java projects. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 1–5, New York, NY, USA, 2018. ACM.

[53] Andreas Mauczka, Florian Brosch, Christian Schanes, and Thomas Grechenig. Dataset of developer-labeled commit messages. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 490–493, Piscataway, NJ, USA, 2015. IEEE Press.

[54] Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, January 1968.

[55] Keir Mierle, Kevin Laven, Sam Roweis, and Greg Wilson. Mining student CVS repositories for performance indicators. In *Proceedings of the 2nd International Workshop on Mining Software Repositories*, MSR '05, pages 1–5, New York, NY, USA, 2005. ACM.

[56] Dimitris Mitropoulos, Vassilios Karakoidas, Panos Louridas, Georgios Gousios, and Diomidis Spinellis. The bug catalog of the Maven ecosystem. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 372–375, New York, NY, USA, 2014. ACM.

[57] Murtuza Mukadam, Christian Bird, and Peter C. Rigby. Gerrit software code review data from Android. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 45–48, Piscataway, NJ, USA, 2013. IEEE Press.

[58] Hiroaki Murakami, Yoshiki Higo, and Shinji Kusumoto. A dataset of clone references with gaps. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 412–415, New York, NY, USA, 2014. ACM.

[59] Jeroen Noten, Josh G. M. Mengerink, and Alexander Serebrenik. A data set of OCL expressions on GitHub. In *Proceedings of the 14th International Conference on Mining Software Repositories*, MSR '17, pages 531–534, Piscataway, NJ, USA, 2017. IEEE Press.

[60] Nicole Novielli, Fabio Calefato, and Filippo Lanubile. A gold standard for emotion annotation in Stack Overflow. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 14–17, New York, NY, USA, 2018. ACM.

[61] Lucas Nussbaum and Stefano Zacchiroli. The Ultimate Debian Database: Consolidating bazaar metadata for Quality Assurance and data mining. In *Proceedings of the 7th Working Conference on Mining Software Repositories*, MSR '10, page 10, Piscataway, NJ, USA, 2010. IEEE Press.

[62] Masao Ohira, Yutaro Kashiwa, Yosuke Yamatani, Hayato Yoshiyuki, Yoshiya Maeda, Nachai Limsettho, Keisuke Fujino, Hideaki Hata, Akinori Ihara, and Kenichi Matsumoto. A dataset of high impact bugs: Manually-classified issue reports. In *Proceedings of the 12th Working*

Conference on Mining Software Repositories, MSR '15, pages 518–521, Piscataway, NJ, USA, 2015. IEEE Press.

[63] Marco Ortu, Alessandro Murgia, Giuseppe Destefanis, Parastou Tourani, Roberto Tonelli, Michele Marchesi, and Bram Adams. The emotional side of software developers in JIRA. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 480–483, New York, NY, USA, 2016. ACM.

[64] Matheus Paixao, Jens Krinke, Donggyun Han, and Mark Harman. CROP: Linking code reviews to source code changes. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 46–49, New York, NY, USA, 2018. ACM.

[65] Fabio Palomba, Dario Di Nucci, Michele Tufano, Gabriele Bavota, Rocco Oliveto, Denys Poshyvanyk, and Andrea De Lucia. Landfill: An open dataset of code smells with public evaluation. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 482–485, Piscataway, NJ, USA, 2015. IEEE Press.

[66] Leonardo Passos and Krzysztof Czarnecki. A dataset of feature additions and feature removals from the Linux kernel. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 376–379, New York, NY, USA, 2014. ACM.

[67] Henning Piezunka and Linus Dahlander. Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58(3):856–880, 2015.

[68] Luca Ponzanelli, Andrea Mocci, and Michele Lanza. StORMeD: Stack Overflow ready made data. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 474–477, Piscataway, NJ, USA, 2015. IEEE Press.

[69] Sebastian Proksch, Sven Amann, Sarah Nadi, and Mira Mezini. A dataset of simplified syntax trees for C#. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 476–479, New York, NY, USA, 2016. ACM.

[70] Steven Raemaekers, Arie Van Deursen, and Joost Visser. The Maven repository dataset of metrics, changes, and dependencies. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 221–224, Piscataway, NJ, USA, 2013. IEEE Press.

[71] Gregorio Robles. Replicating MSR: A study of the potential replicability of papers published in the Mining Software Repositories proceedings. In *Proceedings of the 7th Working Conference on Mining Software Repositories*, MSR '10, pages 171–180, Piscataway, NJ, USA, May 2010. IEEE Press.

[72] Gregorio Robles, Laura Arjona Reina, Alexander Serebrenik, Bogdan Vasilescu, and Jesús M. González-Barahona. FLOSS 2013: A survey dataset about free software contributors: Challenges for curating, sharing, and combining. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 396–399, New York, NY, USA, 2014. ACM.

[73] Gregorio Robles, Truong Ho-Quang, Regina Hebig, Michel R. V. Chaudron, and Miguel Angel Fernandez. An extensive dataset of UML models in GitHub. In *Proceedings of the 14th International Conference on Mining Software Repositories*, MSR '17, pages 519–522, Piscataway, NJ, USA, 2017. IEEE Press.

[74] Mefta Sadat, Ayse Basar Bener, and Andriy V. Miranskyy. Rediscovery datasets: Connecting duplicate reports. In *Proceedings of the 14th International Conference on Mining Software Repositories*, MSR '17, pages 527–530, Piscataway, NJ, USA, 2017. IEEE Press.

[75] Ripon K. Saha, Yingjun Lyu, Wing Lam, Hiroaki Yoshida, and Mukul R. Prasad. Bugs.jar: A large-scale, diverse dataset of real-world Java bugs. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 10–13, New York, NY, USA, 2018. ACM.

[76] Vaibhav Saini, Hitesh Sajnani, Joel Ossher, and Cristina V. Lopes. A dataset for Maven artifacts and bug patterns found in them. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 416–419, New York, NY, USA, 2014. ACM.

[77] Anand Ashok Sawant and Alberto Bacchelli. A dataset for API usage. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 506–509, Piscataway, NJ, USA, 2015. IEEE Press.

[78] Gerald Schermann, Sali Zumberi, and Jürgen Cito. Structured information on state and evolution of Dockerfiles on GitHub. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 26–29, New York, NY, USA, 2018. ACM.

[79] Forrest J. Shull, Jeffrey C. Carver, Sira Vegas, and Natalia Juristo. The role of replications in empirical software engineering. *Empirical Software Engineering*, 13(2):211–218, April 2008.

[80] Jaime Spacco, Jaymie Strecker, David Hovemeyer, and William Pugh. Software repository mining with Marmoset: An automated programming project snapshot and testing system. In *Proceedings of the 2nd International Workshop on Mining Software Repositories*, MSR '05, pages 1–5, New York, NY, USA, 2005. ACM.

[81] Diomidis Spinellis. A repository with 44 years of Unix evolution. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 462–465, Piscataway, NJ, USA, 2015. IEEE Press.

[82] Diomidis Spinellis. Documented Unix facilities over 48 years. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 58–61, New York, NY, USA, 2018. ACM.

[83] Megan Squire. Apache-affiliated Twitter screen names: A dataset. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 305–308, Piscataway, NJ, USA, 2013. IEEE Press.

[84] Megan Squire. Project roles in the Apache software foundation: A dataset. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 301–304, Piscataway, NJ, USA, 2013. IEEE Press.

[85] Megan Squire. Data sets: The circle of life in Ruby hosting, 2003-2015. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 452–459, New York, NY, USA, 2016. ACM.

[86] Megan Squire. Data sets describing the circle of life in Ruby hosting, 2003–2016. *Empirical Software Engineering*, 23(2):1123–1152, April 2018.

[87] Asher Trockman, Shurui Zhou, Christian Kästner, and Bogdan Vasilescu. Adding sparkle to social coding: An empirical study of repository badges in the npm ecosystem. In *Proceedings of the 40th International Conference on Software Engineering*, ICSE '18, pages 511–522, New York, NY, USA, 2018. ACM.

[88] Bogdan Vasilescu, Alexander Serebrenik, and Vladimir Filkov. A data set for social diversity studies of GitHub teams. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 514–517, Piscataway, NJ, USA, 2015. IEEE Press.

[89] Bogdan Vasilescu, Alexander Serebrenik, and Tom Mens. A historical dataset of software engineering conferences. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 373–376, Piscataway, NJ, USA, 2013. IEEE Press.

[90] Georg von Krogh and Eric von Hippel. The promise of research on open source software. *Management Science*, 52(7):975–983, July 2006.

[91] Patrick Wagstrom, Corey Jergensen, and Anita Sarma. A network of Rails: A graph dataset of Ruby on Rails and associated projects. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 229–232, Piscataway, NJ, USA, 2013. IEEE Press.

[92] D. R. Wallace. Enhancing competitiveness via a public fault and failure data repository. In *Proceedings Third IEEE International High-Assurance Systems Engineering Symposium*, pages 178–185, November 1998.

[93] Jane Webster and Richard T. Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii, June 2002.

[94] Michel Wermelinger and Yijun Yu. An architectural evolution dataset. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 502–505, Piscataway, NJ, USA, 2015. IEEE Press.

[95] James R. Williams, Davide Di Ruscio, Nicholas Matragkas, Juri Di Rocco, and Dimitris S. Kolovos. Models of OSS project meta-information: A dataset of three forges. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 408–411, New York, NY, USA, 2014. ACM.

[96] Yulin Xu and Minghui Zhou. A multi-level dataset of Linux kernel patchwork. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 54–57, New York, NY, USA, 2018. ACM.

[97] Aiko Yamashita, S. Amirhossein Abtahizadeh, Foutse Khomh, and Yann-Gaël Guéhéneuc. Software evolution and quality data from controlled, multiple, industrial case studies. In *Proceedings of the 14th*

International Conference on Mining Software Repositories, MSR '17, pages 507–510, Piscataway, NJ, USA, 2017. IEEE Press.

[98] Aiko Yamashita, Fabio Petrillo, Foutse Khomh, and Yann-Gaël Guéhéneuc. Developer interaction traces backed by IDE screen recordings from Think aloud sessions. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 50–53, New York, NY, USA, 2018. ACM.

[99] Xin Yang, Raula Gaikovina Kula, Norihiro Yoshida, and Hajimu Iida. Mining the modern code review repositories: A dataset of people, process and product. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 460–463, New York, NY, USA, 2016. ACM.

[100] Yue Yu, Zhixing Li, Gang Yin, Tao Wang, and Huaimin Wang. A dataset of duplicate pull-requests in GitHub. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, pages 22–25, New York, NY, USA, 2018. ACM.

[101] Stefano Zacchiroli. The Debsources dataset: Two decades of Debian source code metadata. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR '15, pages 466–469, Piscataway, NJ, USA, 2015. IEEE Press.

[102] Chenlei Zhang and Abram Hindle. A green miner's dataset: Mining the impact of software change on energy consumption. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR '14, pages 400–403, New York, NY, USA, 2014. ACM.

[103] Chenguang Zhu, Yi Li, Julia Rubin, and Marsha Chechik. A dataset for dynamic discovery of semantic changes in version controlled software histories. In *Proceedings of the 14th International Conference on Mining Software Repositories*, MSR '17, pages 523–526, Piscataway, NJ, USA, 2017. IEEE Press.

[104] Jiaxin Zhu, Minghui Zhou, and Hong Mei. Multi-extract and multi-level dataset of Mozilla issue tracking history. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 472–475, New York, NY, USA, 2016. ACM.

[105] Thomas Zimmermann, Massimiliano Di Penta, Sunghun Kim, Daniel M. German, and Alberto Bacchelli. Welcome from the chairs. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages iii–viii, May 2013.