Empirical Software Engineering
https://doi.org/10.1007/s10664-020-09834-7

Standing on shoulders or feet? An extended study on the usage of the MSR data papers

Zoe Kotti[1] · **Konstantinos Kravvaritis**[1] · **Konstantina Dritsa**[1] · **Diomidis Spinellis**[1]

Published online: 18 July 2020

Abstract

The establishment of the Mining Software Repositories (MSR) data showcase conference track has encouraged researchers to provide data sets as a basis for further empirical studies. The objective of this study is to examine the usage of data papers published in the MSR proceedings in terms of use frequency, users, and use purpose. Data track papers were collected from the MSR data showcase track and through the manual inspection of older MSR proceedings. The use of data papers was established through manual citation searching followed by reading the citing studies and dividing them into strong and weak citations. Contrary to weak, strong citations truly use the data set of a data paper. Data papers were then manually clustered based on their content, whereas their strong citations were classified by hand according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge. A survey study on 108 authors and users of data papers provided further insights regarding motivation and effort in data paper production, encouraging and discouraging factors in data set use, and future desired direction regarding data papers. We found that 65% of the data papers have been used in other studies, with a long-tail distribution in the number of strong citations. Weak citations to data papers usually refer to them as an example. MSR data papers are cited in total less than other MSR papers. A considerable number of the strong citations stem from the teams that authored the data papers. Publications providing Version Control System (VCS) primary and derived data are the most frequent data papers and the most often strongly cited ones. Enhanced developer data papers are the least common ones, and the second least frequently strongly cited. Data paper authors tend to gather data in the context of other research. Users of data sets appreciate high data quality and are discouraged by lack of replicability of data set construction. Data related to machine learning or derived from the manufacturing sector are two suggestions of the respondents for future data papers. Overall, 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, by

Communicated by: Yasutaka Kamei and Andy Zaidman

✉ Zoe Kotti
zoekotti@aueb.gr

Extended author information available on the last page of the article.

promoting open science initiatives, and by providing incentives for the enrichment of existing data collections.

Keywords Software engineering data · Bibliometrics · Survey study · Mining software repositories · Data paper · Reproducibility

> 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]

1 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 (Cukic 2005). They can be used to empirically evaluate software product quality and development process attributes and also to create or verify estimation models (Lavazza and Santillo 2012). In addition, publicly available data sets can help researchers perform so-called *exact* replications of existing studies and thus address potential internal validity problems (Shull et al. 2008). 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 (Shull et al. 2008).

Yet, at least in the past, data sets for software engineering research were small in size and difficult to obtain (Kitchenham et al. 2002). The situation has improved over the past decades with the emergence of open source software (von Krogh and von Hippel 2006), and the growing interest in sharing artifacts, including data sets, along with research publications. Such efforts are encouraged by initiatives such as the ACM SIGSOFT Artifact Evaluation Working Group,[2] which aims to integrate the artifact evaluation in the publication process, or the *Recognizing and Rewarding Open Science in Software Engineering* (ROSE) festival, which salutes replication and reproducibility in software engineering. For these reasons researchers have collaborated (Cukic 2005) through various initiatives to develop data set repositories, such as the *International Conference on Predictive Models and Data Analytics for Software Engineering* (PROMISE) (Sayyad Shirabad and Menzies 2005), 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" (Wallace 1998), the Mining Software Repositories (MSR) conference data showcase track (Zimmermann et al. 2013a), or the *awesome-msr* GitHub project.[3]

[1] 1968 ACM Turing Award Lecture (Hamming 1969)
[2] https://github.com/acmsigsoft/artifact-evaluation
[3] 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 (Zimmermann et al. 2013b):

> 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 — to evaluate the data track's actual impact, users — to examine researchers' potential reluctance to work with data coming outside their organization, and use purpose — to identify the most used types of data papers, and the types of studies that mainly use them. 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,
- the quantitative analysis of the MSR data papers' impact, and
- the analysis of the community's opinion regarding data paper publication and use.

In the following Section 2 we present an overview of related work. We then describe our study's methods in Section 3, present our results in Section 4, discuss the findings, and identify our study's implications in Section 5. The study is complemented by the associated validity threats in Section 6, followed by our conclusions in Section 7. The data sets associated with our study (data papers, citing papers, categorizations, MSR papers, citations, survey questionnaire and responses) are made available online.[4]

A shorter version of this study appeared in the 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR '19) (Kotti and Spinellis 2019). This work extends the conference paper by using multiple raters and established research methods for the manual clustering of data papers and the classification of strong citations. Furthermore, this work introduces a questionnaire survey study on all identified primary authors and users of data papers, and an analysis of weak citations (defined in Section 3.2) to data papers.

2 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 (2002) have authored a thorough guide on writing high quality literature reviews. On the other hand, bibliometric studies are statistical analyses of publication data. We consider our work part of the bibliometric research strand, and to the best of our knowledge, we are the first to conduct a quantitative review of data paper usage.

A first step toward the assistance of bibliometric research in the field of software engineering models was made in 2004 by the organizers of the PROMISE workshop, in their

[4] http://doi.org/10.5281/zenodo.3709219

attempt "to strengthen the community's faith in software engineering models" (Cukic 2005). Authors of such models were asked to submit, along with their work, a related data set to the PROMISE repository.

Many individuals have carried out interesting quantitative bibliometric research on various topics. Robles (2010) 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 was very low at the time, a fact we also state in Section 4.1. Concerning replicability, Robles found that only a limited number of publications are replication friendly.

Liebchen and Shepperd (2008) 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 that work stress the need for more quality data rather than quantity data. To achieve this, they advise researchers to provide a 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 (2013) survey on data repositories. They noticed that the lack of structured documentation of PROMISE and ISBSG repositories hampered 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 each study's subject was analyzed following the corresponding classification of the data studies, reflecting the classification we subsequently perform on the MSR data papers (Section 4.1).

In the field of Systems and Software Engineering, the 13-part study series by Glass (1994) and Wong et al. (2011) assesses scholars and institutions based on the number of papers they have published in related journals. The majority of the studies span a five-year period covering overall the years 1993–2008. The progress of the study results indicates that the top three institutions up to 2003 were mainly from the USA and involved an equal number of industry research centers and academic institutions. Since 2004, the top three institutions are mainly academic from Korea, Taiwan, and, lastly, Norway. This change is also observed in the entire list of the 15 top-ranked institutions presented in the studies. USA was first in number of top-ranked institutions up to 2002, followed by Asia-Pacific, Europe, and, lastly, Middle East. Middle East has disappeared from the list since 2001. Additionally, the Asia-Pacific institutions have surpassed USA's since 2003, setting Europe in the last place. Regarding the type of the top-ranked institutions of the list, during the years 1993–2008, an average of 82% were academic institutions, as opposed to the remaining 18% which were industry research centers.

A second evaluation of the ISBSG software project repository was carried out by Almakadmeh and Abran (2017). Their purpose was to assess the repository from the Six Sigma measurement perspective and to correlate this assessment with software defect estimation. They found that the ISBSG Microsoft 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 (de Freitas and de Souza 2011) is the first bibliometric research of this community, covering a

ten-year list of studies, from 2001–2010. The evaluation focuses 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 those by Bradford (1985) and (Lotka 1926).

In the same context, Harman et al. (2009) assessed research trends, techniques and their applications in SBSE. They classified SBSE literature 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 recognized and presented opportunities for further improvement, and avenues for supplementary research.

The work of Gu (2004) 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[5] 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 the authors have only written one publication. As far as citation frequency is concerned, Gu demonstrates its significant correlation with the reputation of the journal it has been published in. On the other hand, his findings reveal no correlation between R&D expenditures and research productivity or citation counts.

In the field of Requirements Engineering, Zogaan et al. (2017) conducted a systematic literature review on 73 data sets used for software traceability over a fifteen-year period, between 2000–2016. Using both manual and automated methods they selected studies that have used data sets, case studies, or empirical data, to develop, validate, train, or test traceability techniques. Analyzing these studies they identified that healthcare and aerospace are the two most frequent domains represented by traceability data sets. The majority of the data sets are OSS, followed by academic and industrial. Concerning availability, almost 40% of the data sets examined are not available for reuse, originating mainly from industry and academia. On the contrary, almost all OSS data sets are available. To assess the quality of traceability data sets, the authors designed a framework consisting of quality characteristics such as availability, licensing, completeness, trustworthiness, and interpretability.

3 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 by hand, and further elaborate by manually clustering them based on the year of publication and the content of the data sets.

RQ2 *How are data papers used?* We answer this by collecting all citations to MSR data papers by hand and manually 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.

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

RQ4 *What is the community's opinion regarding data paper publication and use?* We answer the following subquestions through a web-based survey study on 108 authors and users of data papers.

- Q4.1 *What motivates people to produce a data paper?*
- Q4.2 *How much effort is required to produce a data paper?*
- Q4.3 *What are the characteristics of a useful data set?*
- Q4.4 *What characteristics prevent a data set from being used?*
- Q4.5 *What direction should data papers follow in the future?*

To answer the above research questions, we performed a mixed methods study. In order to ensure consistency of the manual processes that we performed for the analysis of the research usage of the data papers (Section 3.2), the clustering of them (Section 3.1), and the classification of their strong citations (Section 3.2), we employed certain guidelines for systematic literature reviews and systematic mapping studies. Furthermore, to answer RQ4, we performed a survey study employing survey research principles (Section 3.4).

3.1 RQ1: Data Paper Collection and Clustering

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 — consider e.g. Conklin et al.'s (2005) collection of FLOSS data and analyses.

To acquire the aforementioned papers, we searched through the programs of the MSR conferences on their respective websites. Programs that contained an explicit *Data Showcase* section immediately informed us about the data papers of the particular year. In programs that did not include the aforementioned section, we 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.

Following the collection, we sorted the data papers into distinct clusters according to their topic by combining methods of two prominent studies. From the systematic mapping studies in software engineering (Petersen et al. 2008), we applied the classification scheme of abstract keywording. In addition, we followed two data extraction approaches suggested in the work of Brereton et al. (2007) on systematic literature review within the software engineering domain. The first approach introduces the use of two reviewers performing individually the data extraction process and discussing their disagreements, while the second approach proposes the use of a data extractor and a data checker.

According to the above methods, the first and third authors of this paper individually labeled all data papers with keywords. For each data paper, the two authors read the abstract and extracted keywords related to the content of the data set provided by the particular data paper. For ambiguous abstracts that hampered the extraction of meaningful keywords, the two authors also studied the introduction or conclusion sections of the paper (Petersen et al. 2008). The keywords were either *in vivo* (Glaser and Strauss 1967) — when representative phrases could be extracted as is from the aforementioned sections, or otherwise constructed by the authors.

Following the individual keyword-labeling process, the two authors met, discussed their keywords and agreed on a final set of keywords by refining, merging, and renaming the initial ones. In this way, a structured keyword set was formed consisting of the general keywords *code* and *people* (after observing that all individual keywords could be divided into

these two groups), and more specialized keywords. *Code* and *people* were used to signify whether a data paper mostly targeted the software development process or the human factor respectively, while multiple specialized keywords could be assigned to it. Using the latter keyword set, the same authors repeated the labeling process for the data papers together. Supplementary keywords that appeared during the second round of keyword assignment were added to the final keyword set. Once the second round was completed, the two authors grouped together the conceptually related keywords. Through this process, the clusters of data papers were formed and then named accordingly.

Finally, the first and third authors assigned each data paper to the most conceptually relevant cluster (i.e., in case a paper could be assigned to more than one cluster, the authors selected the one they considered most descriptive of its content). To ensure the correct mapping of the papers to the clusters, the second author also assessed the cluster assignments, and then discussed and resolved his objections with the first and third authors. The agreement rate of the second author with the first and third authors was 91%. From the 9% of the disagreements, 57% were resolved in favor of the second author, while the remaining 43% were resolved in favor of the other two authors.

3.2 RQ2: Data Paper Use Identification and Classification

To conduct the analysis on the research usage of the data papers, we implemented the *Identification of Research* and *Study Selection* processes, as proposed in Kitchenham's (2004) work on procedures for performing systematic reviews.

The identification of research was made through widely used and established platforms that provide citation data: *Google Scholar*,[6] *Scopus* — Elsevier's abstract and citation database[7] and the *ACM Digital Library*.[8] Most research papers 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 of our analysis. First, we applied the whitelisting practice and kept studies of conference proceedings, journal articles, master's and doctoral theses, books, and technical reports. Studies published in multiple venues, such as conference publications that are later published in journals — e.g. Krueger et al.'s (2018; 2019) study on the usage of cryptographic APIs, were only listed once. Priority was given sequentially to books, journal articles, conference proceedings, reports, and, lastly, theses. We additionally decided to retain only studies written in the English language, due to its widespread adoption for scientific communication.

The main criterion for retaining citing studies was their actual 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 (*weak*) citation is the study of repository badges in the npm ecosystem (Trockman et al. 2018), which has cited the collection of social diversity attributes of programmers (Vasilescu et al. 2015), although it has not used its data. Weak citations were manually analyzed to determine the most common types of uses of data papers in these cases.

[6] https://scholar.google.com/
[7] https://www.scopus.com/
[8] https://dl.acm.org/

The process of citation collection and segregation into strong and weak was held from July to November 2018. For data papers that had been strongly cited by at least one of their authors, 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 strong 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.

Furthermore, we classified the collected strong citations according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge (Bourque and Fair 2014) (SWEBOK). Again, we followed a combination of the two data extraction approaches of Brereton et al. (2007) described in Section 3.1. The first and second authors of this paper individually assigned each strong citation to a particular SWEBOK knowledge area after reading their abstract. Similarly to Section 3.1, in cases of ambiguous abstracts, they also read the introduction or conclusion sections. Next, the two authors met, discussed, and partially resolved their disagreements. After that, 16% of the total citation categorizations remained conflicting. These disagreements were resolved by the paper's last author who again selected among all knowledge areas. The selections of the first two authors were not divulged to him to avoid bias. Through this process, 30% of the pending disagreements were resolved in favor of the first author, and 15% in favor of the second author. Hence, the overall agreement rate between the last and the first two authors was 45%. The last author's opinion prevailed over the first two authors' in the remaining 55% of disagreements, where all three assigned knowledge areas were conflicting, due to his long experience in software engineering and his extended familiarity with the SWEBOK knowledge areas.

3.3 RQ3: Citation Analysis

To assess in an objective manner the impact of MSR 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 3.2, 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,[9] 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 3.1) and the rest. We also split the MSR papers by year to simplify the selection of samples.

Furthermore, 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 3.2). Although this new set of papers citing data papers is not exhaustive (it only contains MSR papers), it allows us to compare

[9] https://dblp.org/

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

3.4 RQ4: Survey Planning, Execution, and Analysis

To conduct our survey study on authors and users of data papers in order to explore our community's view regarding data paper publication and use, we followed the set of ten activities introduced in Kitchenham and Pfleeger's (2001; 2002a, b, c, d; 2003) six-part series of principles of survey research.

Survey Design We adopted a cross sectional, case control observational study design (i.e., participants were surveyed about their past experiences at a particular fixed point in time), which is typical of surveys in software engineering (Kitchenham and Pfleeger 2002a). The goal of this study was *to obtain further insights on the production, use, and future desired direction of data papers*. Hence, we framed the objectives of our survey in terms of the following questions.

Q1 *What motivates people to produce a data paper?*
Q2 *How much effort is required to produce a data paper?*
Q3 *What are the characteristics of a useful data set?*
Q4 *What characteristics prevent a data set from being used?*
Q5 *What direction should data papers follow in the future?*

To prevent a low response rate due to the summer vacation period that co-occurred with the study preparation, the survey was scheduled to run in early September 2019.

Survey Sample The survey was conducted on two different samples in two different time periods — September 2019 and January 2020. The reason behind the second conduction was to include more strong citation authors. In the first conduction, all primary authors of the 81 data papers comprised our sample, along with an equal set of primary authors of strong citations. The unique primary authors of data papers were in total 71. For the selection of the 71 (out of 419) primary authors of strong citations, we implemented the probabilistic sampling method of simple random sample (Kitchenham and Pfleeger 2002d). For that purpose, we used the random number generator of the Python 3.7 *random* library with the default seed value and replacements. From the collected candidate respondents, 14 were primary authors of both data papers and strong citations, leading to a total of 128 unique candidates (instead of 142). In the second conduction, the remaining primary authors of strong citations comprised our sample excluding the ones already included in the first

sample. In this manner, our second sample was composed of 189 unique candidates, while the overall sample size was 317.

Survey Instrument The survey questionnaire was organized into eight sections. In the first section participants specified whether they were authors of data papers, along with the number of their data paper publications in MSR and in any other venue. The next five sections were organized according to the objective questions, and were composed of both mandatory and optional open-ended, multiple choice, and Likert scale questions. We intentionally used even Likert scales to force participants to make a choice (Gousios et al. 2016). To address Q1, data paper authors applied on a four-level Likert scale the extent to which a set of specific attributes motivate them to produce a data paper. There was also an open-ended question for additional motivational factors. For Q2, data paper authors specified through a multiple choice question the effort-months they need on average to produce a data paper. For Q3 and Q4, both authors and users of data papers evaluated on a four-level Likert scale the importance of a set of particular characteristics for the selection or avoidance of a data set for their research. The same characteristics were included in both questions. In our view, an attribute considered as useful to some extent is not necessarily considered as discouraging to the exact same extent, thus affecting the overall ranking of the characteristics. An open-ended question was included for additional useful or preventing attributes of data sets. To address Q5, respondents listed through open-ended questions what data papers they would like to see published in the future, and from what sources new data could be derived. The following section included demographic questions aiming to assess the diversity of the responses. Through the final section of the survey we retrieved feedback regarding the completeness of the questionnaire; respondents assessed whether the objective questions were sufficiently or weakly addressed, and left their comments. Finally, they could leave their e-mail address, to receive a report with the survey results. To ensure anonymity, the collected e-mail addresses are not publicly distributed within the online available data set of survey responses.

Survey Evaluation To evaluate and refine the questions of the survey, and to calculate the average time required to complete it, we initially performed a pilot study. The sample of the pilot study was composed of 23 members of our laboratory (two faculty members, three senior, six associate, and twelve junior researchers), four external faculty members, and one more senior researcher. In total 28 potential respondents constituted the sample of the pilot study. The pilot study ran from July 25th to August 1st, 2019, and nine responses were received (32% response rate). Respondents of the pilot study were asked at the beginning of the questionnaire to complete their current local time. Through subtraction from the reported completion timestamp (which was automatically recorded unlike the starting timestamp), the average time required was calculated at 18 minutes.

Survey Operation Both the pilot and the final survey were distributed as a *Google form*, which the candidate participants were invited to complete through an invitational mail. Although the mailing process was automated, it was personalized by addressing the candidates by name, and by including details on how they were selected and which of their research papers (data papers and/or strong citation papers) the survey involved. The candidates were informed on the average time required for the questionnaire completion (rounded up to 20 minutes), along with the goal and the objectives of the survey study.

From the total of 317 mails that were sent as part of the final survey, 39 failed to be delivered. These failures involved twelve e-mail addresses no longer in use, 26 rejected

recipients, and one wrong e-mail address. We consider our final sample a total of 278 potential participants.

The final survey ran from September 2nd to 24th, 2019, and from January 24th to February 16th, 2020. In both rounds we aimed for a three week duration, but we would briefly reopen the survey when candidate participants requested it.

A reminder mail was distributed to potential respondents ten days after the invitational mail in both rounds. Verified respondents who had either answered our initial mail or had left their e-mail address in their survey response were excluded from the recipient list of the reminder mail. The final survey received 108 responses (39% response rate calculated on the basis of the final sample size — 278).

Survey Analysis Q3–Q5 were analyzed from the perspective of authors and users combined, and of users independently. We applied manual pair coding (Salinger et al. 2008) to summarize the results of the six open-ended questions. For the first survey conduction, the first and second authors of this paper applied together codes to all open-ended responses following a mixed approach of line-by-line and sentence-by-sentence coding (Chametzky 2016). Next, they combined conceptually-related codes by generalizing or specializing them, and integrated them into distinct groups. For the second survey conduction, the same authors used the first group of codes to annotate the new responses (Gousios et al. 2016). In case a response was not connected to any group, or was connected but further ideas were also introduced, the authors would apply new codes to it. At the end, the generalization-specialization and grouping process was repeated for the new codes.

3.5 RQ4: Survey Participants

The questionnaire was completed by 108 respondents of various age groups. Concerning their current occupation, 37% (24) were academic staff, 24% (15) worked in industry, 19% (11) were post doctoral researchers, and 19% (5) were doctoral students.

From the 108 respondents of the survey, 30% (32) were primary authors only of strong citation papers, as opposed to the remaining 70% (76) who were primary authors of data papers, with a portion of them also being authors of strong citation papers. From the 76 data paper authors, one's data papers have not been published in any venue, 42% (32) have published only one data paper, 26% (20) have published two, 11% (8) of the authors have published three, 5% (4) have published four, and three authors own five, six, and seven publications respectively. Furthermore, 3% (2) of the authors have published eight data papers, followed by another 3% (2) with ten publications. Lastly, four authors own eleven, 20, 25, and 30 data paper publications respectively. Concerning data paper publications in the MSR conference, 22% (17) of the authors have no MSR data papers, 55% (42) have one, 15% (11) of the authors have two MSR publications, 3% (2) own three publications in MSR, 4% (3) own four, and one is the primary author of five MSR data papers.

4 Results

The findings of our study are framed in respect to the four research questions.

4.1 RQ1: What Data Papers Have Been Published?

We identified the 81 data papers presented in Table 1. These comprise about 15% of the 507 papers published in the MSR conference in the years when data papers appeared. The

Table 1 MSR data papers by year

Year	Data Papers
2005	(Spacco et al. 2005; Mierle et al. 2005; Conklin et al. 2005)
2006	(Kim et al. 2006)
2010	(Nussbaum and Zacchiroli 2010)
2012	(Keivanloo et al. 2012)
2013	(Binkley et al. 2013; Dit et al. 2013; Goeminne et al. 2013; Vasilescu et al. 2013; Wagstrom et al. 2013; Janjic et al. 2013; MacLean and Knutson 2013; Squire 2013a; Mukadam et al. 2013; Butler et al. 2013; Squire 2013b; Lamkanfi et al. 2013; Gousios 2013; Raemaekers et al. 2013; Hamasaki et al. 2013)
2014	(Krutz and Le 2014; Saini et al. 2014; Gousios and Zaidman 2014; Murakami et al. 2014; Passos and Czarnecki 2014; Zhang and Hindle 2014; Robles et al. 2014; Lazar et al. 2014; Bloemen et al. 2014; Fujiwara et al. 2014; Gousios et al. 2014; Williams et al. 2014; Farah et al. 2014; Mitropoulos et al. 2014; Baldassari and Preux 2014)
2015	(Vasilescu et al. 2015; Sawant and Bacchelli 2015; Ohira et al. 2015; Krutz et al. 2015; German et al. 2015; Altinger et al. 2015; Spinellis 2015; Wermelinger and Yu 2015; Mauczka et al. 2015; Barik et al. 2015; Karakoidas et al. 2015; Palomba et al. 2015; Ponzanelli et al. 2015; Zacchiroli 2015; Habayeb et al. 2015; Gonzalez-Barahona et al. 2015)
2016	(Proksch et al. 2016; Allix et al. 2016; Squire 2016; Yang et al. 2016; Amann et al. 2016; Zhu et al. 2016; Ortu et al. 2016)
2017	(Noten et al. 2017; Aivaloglou et al. 2017; Zhu et al. 2017; Robles et al. 2017; Madeyski and Kawalerowicz 2017; Sadat et al. 2017; Yamashita et al. 2017)
2018	(Martins et al. 2018; Yu et al. 2018; Novielli et al. 2018; Geiger et al. 2018; Xu and Zhou 2018; Saha et al. 2018; Paixao et al. 2018; Yamashita et al. 2018; Spinellis 2018; Gao et al. 2018; Chatzidimitriou et al. 2018; Markovtsev and Long 2018; Schermann et al. 2018; Gkortzis et al. 2018; Efstathiou et al. 2018)

There has been a significant rise in the number of data papers since 2013, the year that the MSR data showcase track was established

timeline of the data papers and the research based on them is depicted in Fig. 1. For each year, the number of published data papers is showcased, along with the number of studies published in the particular year, which have been based on any of these data papers. (It should be noted that this is not a cumulative graph; each year's outcome is independent of the previous.) 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 (Zimmermann et al. 2013b). Until then, 2005 was the year with the most data papers. 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 1).

From the clustering of the data papers, as described in Section 3.1, eight data clusters emerged. Table 2 shows for each cluster the number of data papers it comprises, the number of strongly cited and non-cited data papers, and the strong citations that have been made to them. We consider as *non-cited* the data papers with either weak citations or no citations at all. The clusters are sorted in descending order according to the number of data papers they contain.

VCS Primary and Derived Data preponderate. The particular cluster consists of 29 studies that provide Version Control System (VCS) raw or processed data, along with descriptive statistics and analyses. The collection of Java source code of the Merobase Component Finder project (Janjic et al. 2013) is part of this cluster.

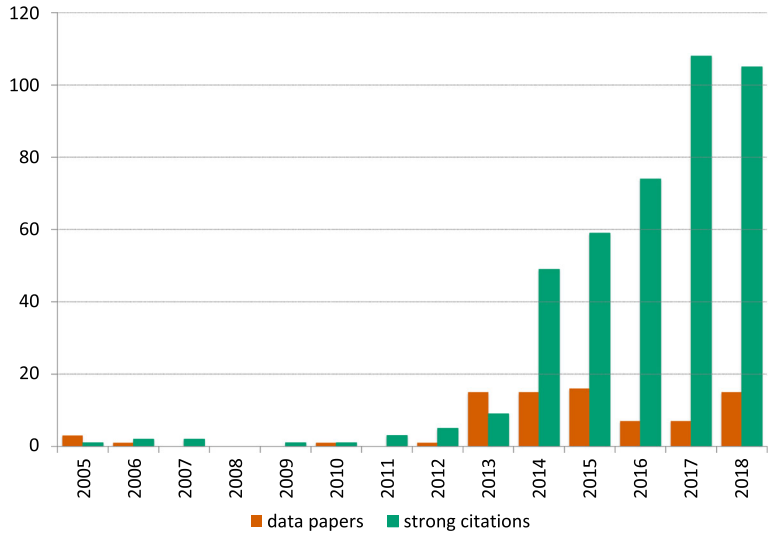

Fig. 1 Timeline of the data papers and the strong citations. Each year depicts the number of data papers published in the particular year and the number of studies published in the particular year that are based on any data paper. The number of strong citations to data papers is constantly rising, indicating that the concept of data papers has long-term value

Software Faults, Failures, Smells concern 17 data sets of security failures, software inconsistencies, and bad 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 (Gkortzis et al. 2018).

Software Evolution involves eleven collections with information on the evolution of artifacts, such as operating systems (Spinellis 2018), software architectures (Wermelinger and Yu 2015), and software packages (Bloemen et al. 2014).

Table 2 Data paper clusters and strong citations

Cluster	Data Papers	Str. cited DPs	Non-cited DPs	Str. Citation Ratio (%)	Str. Citations
VCS Primary and Derived Data	29	20	9	69	312
Software Faults, Failures, Smells	17	11	6	65	42
Software Evolution	11	6	5	55	17
Group Dynamics	9	7	2	78	41
Computational Linguistics	7	5	2	71	20
Software Models	3	2	1	67	5
Computing Education and Programming Practices	3	1	2	33	1
Enhanced Developer Data	2	1	1	50	2
Total	81	53	28	65	440

Publications providing VCS primary and derived data are the most frequent data papers and the most often strongly cited ones

The cluster of *Group Dynamics* is composed of nine data papers that focus on social networks (MacLean and Knutson 2013), code reviewing (Mukadam et al. 2013), project roles (Squire 2013b), and social diversity in programming teams (Vasilescu et al. 2015).

Seven data papers were grouped together due to their common theme of facilitating studies related to natural language processing and information extraction (Ponzanelli et al. 2015; Binkley et al. 2013). These papers constitute the *Computational Linguistics* cluster.

Software Models provide simplified visual representations of software processes, such as simplified syntax trees (Proksch et al. 2016) and UML models (Robles et al. 2017).

Papers that share records regarding novices' and experts' programming practices and abilities — e.g. the list of Scratch programs of students (Aivaloglou et al. 2017) — were classified in *Computing Education and Programming Practices*. The aim of this cluster is to facilitate studies on computing education.

The last defined cluster contains papers offering *Enhanced Developer Data*, such as screen and real names extracted from Twitter (Squire 2013a) and personal characteristics (e.g. gender, age, civil status, nationality), education and level of English, and professional status (Robles et al. 2014). Only two papers represent this cluster, however the uniqueness of their data sets segregates them from the other clusters.

4.2 RQ2: How are Data Papers Used?

The 81 MSR data papers are associated with 1169 citations to them, coming from 982 distinct studies (some studies cite multiple data papers). Out of the 1169 citations, 440 (419 distinct studies) use the data sets provided by the data papers (*strong* citations). The remaining 729 citations (610 distinct studies) refer to data papers without utilizing the particular data sets (*weak* citations). We were able to obtain most citations from digital libraries and the web. Six citations that were publicly unavailable were received from their respective authors through personal communication, as stated in Section 3.2, but no access was obtained for another three. (These three studies have been excluded from the total citations.) Table 3 depicts the most strongly cited data papers.

Through manual analysis we found that the most common uses of weak citations were mentioning the work as an example ($n = 524$, 77%), attributing a work's statement ($n = 59$, 9%), using the work's methods ($n = 47$, 7%), presenting the study as related work ($n = 12$, 2%), and reporting obtained statistics ($n = 13$, 2%).

Table 4 depicts the classification of the studies based on data papers according to the knowledge areas of the SWEBOK. This suggests that research on *Software Maintenance*, *Software Engineering Management*, and *Software Engineering Professional Practice* uses data papers to a considerable extent. On the other hand, only a slight portion of research on *Software Requirements*, *Software Engineering Models and Methods*, and *Software Engineering Economics* is facilitated by data showcase papers.

Furthermore, concerning the use of data papers by their respective authors, our findings show that 37 out of the 81 papers have been used by the teams that authored them. Specifically, 15 studies have been solely used either by their first author or his/her co-authors. Figure 2 depicts for each data paper strongly 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 downwards 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 strong citations to the collection of APIs usage information (Sawant and Bacchelli 2015) were made by the first author.

Table 3 Top five data papers in number of strong citations

Title	Data Paper	Year	Cluster	Str. Citations
The GHTorrent Dataset and Tool Suite	(Gousios 2013)	2013	VCS Primary and Derived Data	165
AndroZoo: Collecting Millions of Android Apps for the Research Community	(Allix et al. 2016)	2016	VCS Primary and Derived Data	57
Lean GHTorrent: GitHub Data on Demand	(Gousios et al. 2014)	2014	VCS Primary and Derived Data	24
Who Does What During a Code Review? Datasets of OSS Peer Review Repositories	(Hamasaki et al. 2013)	2013	Group Dynamics	16
The Maven Repository Dataset of Metrics, Changes, and Dependencies	(Raemaekers et al. 2013)	2013	VCS Primary and Derived Data	12
The Eclipse and Mozilla Defect Tracking Dataset: A Genuine Dataset for Mining Bug Information	(Lamkanfi et al. 2013)	2013	Software Faults, Failures, Smells	12
The Emotional Side of Software Developers in JIRA	(Ortu et al. 2016)	2016	Computational Linguistics	12

The most strongly cited data paper offers a collection of primary and derived data extracted from GitHub

Table 4 Areas of strong citing studies

SWEBOK Knowledge Area	Studies	Percentage
Software Maintenance	89	21.2
Software Engineering Management	63	15.0
Software Engineering Professional Practice	57	13.6
Software Quality	55	13.1
Software Configuration Management	46	11.0
Software Construction	43	10.3
Software Design	20	4.8
Software Engineering Process	19	4.5
Software Testing	15	3.6
Software Engineering Economics	6	1.4
Software Engineering Models and Methods	5	1.2
Software Requirements	1	0.2

The studies that strongly cite data papers span the SWEBOK knowledge areas fairly unequally

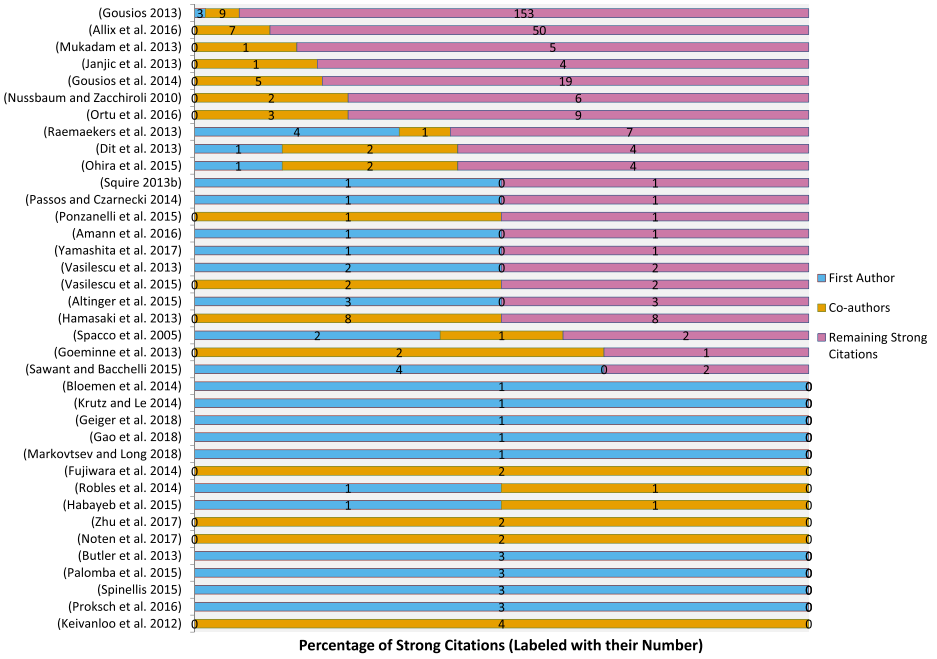

Fig. 2 Use of data papers by their authors (%). Data papers used at least once by the same first author or any of his/her co-authors are represented by the number of strong citations made by the first author, the co-authors, and other unrelated teams. From the total 81 data papers, 37 have been used by the teams that authored them

Table 5 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

Data papers are typically cited less often, compared to other papers of the MSR conference, in terms of the median and average number of references

4.3 RQ3: What is the Impact of Published Data Papers?

The relative impact of published data papers can be deduced from Table 5, which compares citations to data and non-data papers, collected in the way described in Section 3.3. (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 ($\mu = 10, \bar{x} = 15, 7$) as other MSR papers ($\mu = 8, \bar{x} = 17, 0$), meaning that citing an MSR data paper does not promise greater popularity concerning incoming citations.

Table 6 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

Table 6 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

The majority are top-notch venues, indicating the high quality of studies that can be performed through data papers

Repository (CoRR),[10] 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 Fig. 1. The strong citations of all data papers were summed up and illustrated as yearly records. We see that strong 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, and imprinted it on Fig. 3. The limit five was chosen because it provided us with sufficient insights, 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 in the following year. Research based on data papers seems to plateau after the third year of their life.

4.4 RQ4: What is the Community's Opinion Regarding Data Paper Publication and Use?

In this section we present the answers of the survey respondents to the questionnaire in respect to the objective questions, and their feedback on the survey study.

4.4.1 Q4.1: What Motivates People to Produce a Data Paper?

All 76 data paper authors were requested to assess on a four-level Likert scale the motivational impact of a set of eight predefined attributes for data paper production. The collected answers to this question are illustrated in Fig. 4, where the attributes are sorted downwards in descending order of the *Extremely* level. The majority of participants claim to a significant or extreme degree that publishing data papers is *a worthwhile way to contribute to our community*, and that their *contribution will be appreciated by the former*. In addition, they emphasize significantly that *certain data sets should be made available to the community*, while others *want to cover the lack of data in a particular research area*. To a similar extent of high significance, the responding authors of data papers mention that they usually *have gathered the data in the context of other research*, or *plan to use their published data in future research*.

Apart from the predefined set of attributes, some participants also stated in the related open-ended question regarding additional motivational aspects, the expectancy of obtaining a high number of citations, the challenging process of data paper production, validity, transparency, research reusability, and quality improvement as another set of aspiring characteristics. Others recognized that data papers provide a more thorough data set representation and description, promote open science and potential partnerships, and reveal new research trends. A few respondents identified their skillfulness in data paper production as another valuable factor.

4.4.2 Q4.2: How Much Effort is Required to Produce a Data Paper?

From the 76 data paper authors who responded to the survey, 8% (6) stated that they need less than an effort-month, 36% (27) need from one to three effort-months, 34% (26) require

[10] https://arxiv.org/corr

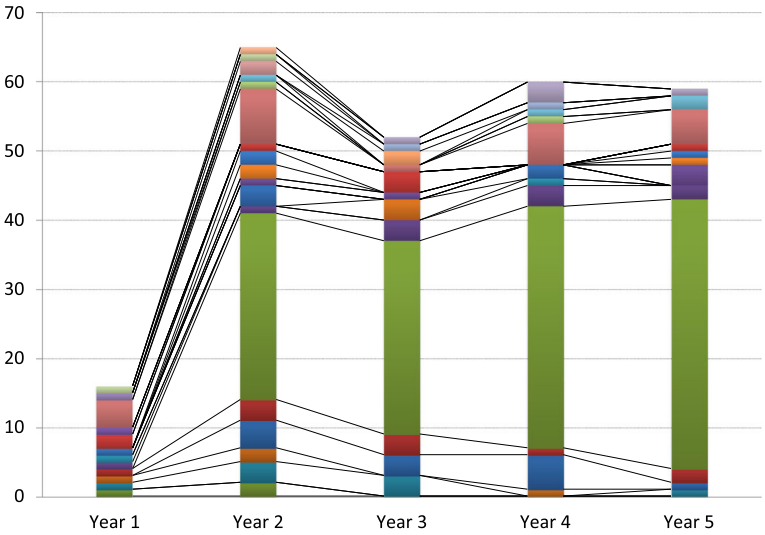

Fig. 3 Timeline of strong citations to data papers published from 2005–2014 over a five-year window. Each strongly cited data paper is represented with the same color along the years. The height of each color bar is relative to the number of strong citations. Data papers with the most strong citations during the second year of their existence seem to retain their citation number in the following years, or to obtain even more strong citations

a total of four to six effort-months, 15% (11) of the authors need from seven to nine effort-months, 7% (5) need from ten to twelve effort-months, and one needs more than a year of effort to produce a data paper.

4.4.3 Q4.3: What are the Characteristics of a Useful Data Set?

In Fig. 5, the assessment of a set of 15 characteristics regarding their importance in data set selection for research purposes is illustrated. The characteristics are sorted downwards in descending order of the *Very Important* level. All 108 respondents evaluated the particular attributes on a four-level Likert scale. The most important characteristics were found to be *ease of use* (i.e., the ability of users to effortlessly obtain the information they are after), *high data quality*, *data freshness*, *replicability of data set construction*, *data schema documentation*, and *documentation of the data collection methods*. Less important characteristics included *personal connection with the data set curators*, *data set having been published as a data paper*, and *data set having been highly cited*. Separating data paper users from authors in Fig. 5, we observe a similar ranking of the characteristics. The main difference is in the first place, which for users is *high data quality*, as opposed to *ease of use* selected by authors and users combined.

In addition to the above characteristics, through the complementary open-ended question, respondents stressed the importance of existing application examples, compatibility and extensibility with other data sets, data balance and integrity, data completeness, reproducibility, updatability, and ease of access and filtering. Participants also valued modification traceability, data novelty and diversity, data validation, continuous maintenance of data and support by the curators, and enhancement of existing data sets. Furthermore, documented threats and flaws, curation of duplicate data (e.g. repository forks) and code clones,

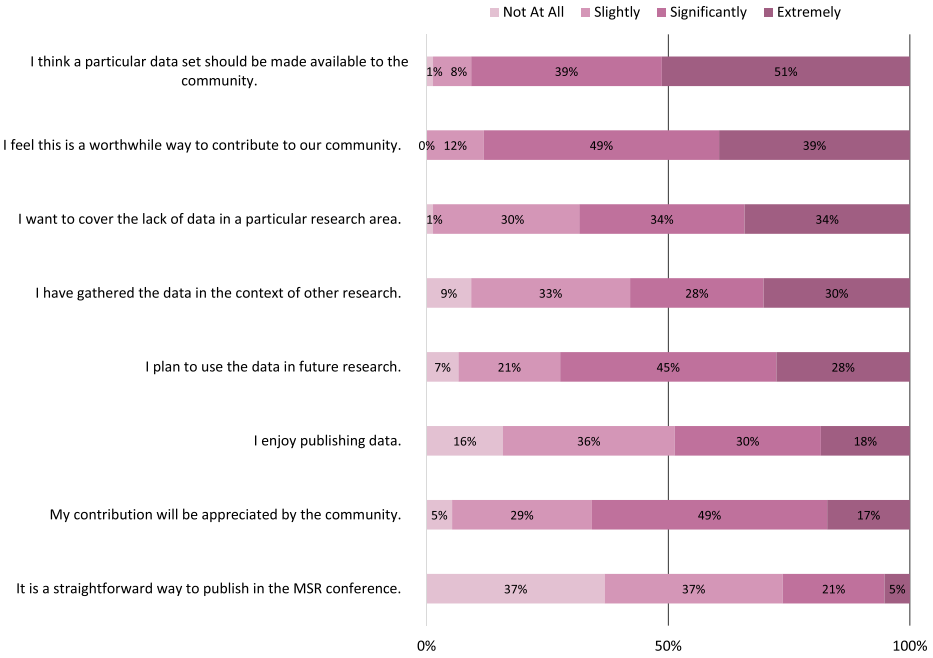

Fig. 4 Motivational impact of specific attributes for data paper production. According to the responses of the primary data paper authors, the motivational impact of a set of eight predefined characteristics for data paper production is depicted on a four-level Likert scale and on a percentile basis of the total data paper authors. The majority of the participants think that particular data sets should be made available, and feel that publishing data papers is a worthwhile way to contribute to our community

inclusion of timestamps, goal-oriented data, anonymity of subjects, contextuality (i.e., the data set captures its context), duality (i.e., the data set contains both positive and negative samples where applicable), and usage metadata of the data set are also appreciated. For software engineering fields with ongoing radical changes, such as malware detection, some remarked that the shelf life of a data set should be short.

4.4.4 Q4.4: What Characteristics Prevent a Data Set from Being Used?

For this question, all 108 respondents evaluated the degree to which a set comprising of the negations of the attributes presented in the previous question discourage them from using a data set. The evaluation was again done on a four-level Likert scale and the results are presented in Fig. 6. Again, the attributes are sorted downwards in descending order of the *Extremely* level. Among the most discouraging characteristics, we discerned *low data quality*, *difficulty of data set use*, *lack of documentation of the data collection methods*, *lack of replicability of data set construction*, *restrictive license*, and *lack of data schema documentation*. Less discouraging characteristics included *less known data set curators*, *lack of personal connection with the data set curators*, and *data set not having been published as a data paper*. Isolating data paper users from authors in Fig. 6, we notice the same characteristics in the last places and in the first position, with the intermediate attributes varying. Users placed *lack of replicability of data set construction*, *stale data*, and *no data verification methods employed in the construction* higher than authors and users combined.

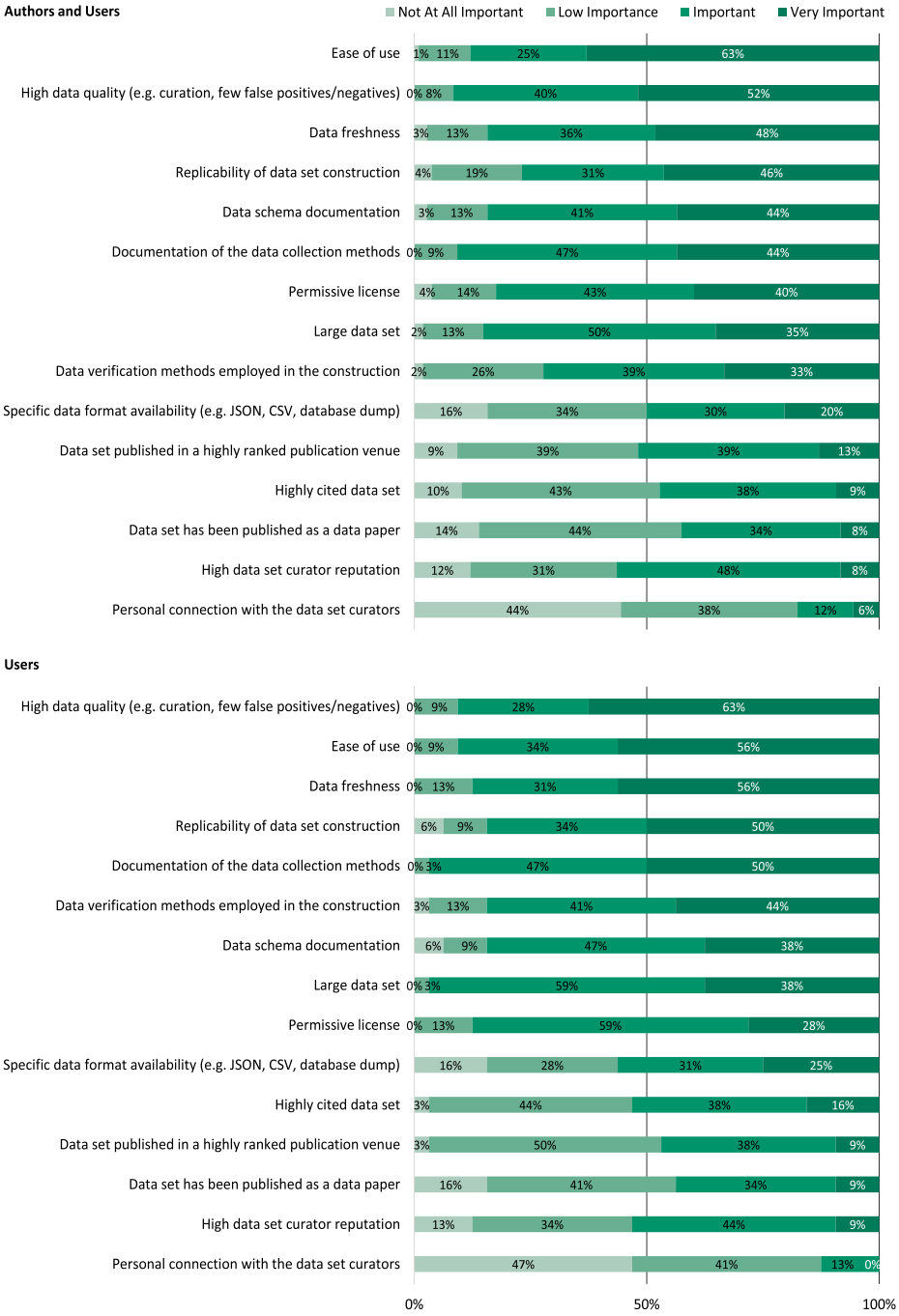

Fig. 5 Encouraging characteristics in data set selection for data paper authors and users (top), and users isolated (bottom). Respondents assessed a set of 15 predefined characteristics on a four-level Likert scale. Ease of use and high data quality appear to be the most important characteristic in data set selection

Fig. 6 Discouraging characteristics in data set selection for data paper authors and users (top), and users isolated (bottom). Respondents assessed a set of 15 predefined characteristics on a four-level Likert scale. Among the most discouraging characteristics, low data quality stands out

Except for the above characteristics, respondents further recognized through the related open-ended question, the discouragement derived from troubled access such as download issues, difficult access, and tool incompatibility/dependency, data isolation, lack of feature variety, novelty, support by the curators, or extensibility, and the limited scope of a data set.

4.4.5 Q4.5: What Direction Should Data Papers Follow in the Future?

Future Data Papers According to the 108 responses, future data papers could draw from artificial intelligence and machine learning, alternative and evolving software engineering, logs analysis (e.g. build logs, test failure logs), continuous integration and DevOps (particularly operations of DevOps), collaborative software development, code metrics and analyses, cross-disciplinary or domain specific processes. Furthermore, the respondents would appreciate shared data regarding health, fitness and performance, repository duplication and code cloning, human-centered and human-assessed data, concerning new technologies, and referring to security, social media, online computing courses, programming competitions, video material and video games. Data from software services, industrial ecosystems, grey literature, databases, Internet of things firmware, voluntarily collected data, and data on hyper-parameter optimization are also desired. Lastly, some participants suggested conducting systematic literature reviews on published data sets, and producing metadata after curating them.

Future Data Sources As far as data sources are concerned, many similar responses with the previous set of answers were observed. Authors and users of data papers would like to see data extracted from sources of entertainment, such as music and video streaming, smart devices, surveys, sources with documented and valid data, as well as data from the sectors of education, health, energy, defense and security, manufacturing and retail, blockchain, finance, and autonomous driving. Moreover, participants suggested exploiting Alexa Rank, execution logs, code review systems, integrated development environments, activity sensors, domain specific data sources, and sources complementary to software repositories. Industrial cloud systems, safety critical software systems, software industries, human resources, and industries with security breaches are also proposed as prospective sources of data.

Separating data paper users from authors, no variation was observed in their responses, since the majority of the above ideas regarding future data papers and data sources were also included in their answers.

4.4.6 Feedback

Overall 95% (103) of the respondents assessed the completeness of the questionnaire as *sufficient*, whereas the remaining 5% (5) reviewed it as *weak*. Overall, the survey was characterized as good and concise, with interesting aspects and sufficient maps to the objective questions. Still, a respondent was not sure whether the open-ended feedback question was enough for assessing completeness.

Objective Questions Q1 was characterized as answerable, as opposed to Q2 which was characterized as ambiguous. A few participants separated the process of data retrieval as part of another research paper requiring different effort from the data paper composition. This could be the reason we observe a significant wide spread in the responses. Q3 and Q4 were

considered as mirroring questions by some respondents. However, the results of both the pilot and the final study partially contradict this opinion due to the asymmetry observed in the answers on an individual level. For instance, although *ease of use* is ranked first in Fig. 5 with 63% of respondents considering it *very important*, it is placed second in Fig. 6 with 46% of respondents considering it *extremely* discouraging. Similarly, *high data set curator reputation* was evaluated as *important* by 48% of participants, but only 8% evaluated it as *significantly* discouraging. We also observe that *data freshness* is placed considerably higher than the equivalent *stale data*. Regarding Q5, a few respondents considered it enjoyable as opposed to some others who characterized it as ambiguous and unanswerable. Moreover, one participant remarked the lack of neutral option in the Likert scale questions.

General Comments Apart from comments on the questionnaire, some participants highlighted the significant effort required to produce a data paper, especially highly citable ones. According to them, these are a combination of good tooling architecture, documentation, novelty, generality, reproducibility, and clarity. Nevertheless, others expressed their concerns regarding data paper practical issues and troubled data sharing guidelines, such as the General Data Protection Regulation. Finally, an additional question was suggested concerning the number of existing data sets that researchers have used in their research.

5 Discussion and Implications

As evidenced by the large increase in the published data papers since the MSR data showcase track was formalized in 2013, it is apparent that the track has catalyzed the publication of data papers. With data papers being more than 15% of the MSR publications in 2019, it is clear that the MSR data showcase track has spurred a new type of publication, yielding each year a notable number of studies. More generally, the data showcase 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 2) 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 are associated with open source software projects, over process data, which are more difficult to come by. Although past experience with calling for the publication of particular data types has not been encouraging (Wallace 1998), many years have passed since then and it might be worth to try focusing the MSR call for data papers on specific topics each year, with an emphasis on software processes, in order to overcome the previous bias.

The studies that strongly cite data papers span the SWEBOK knowledge areas fairly unequally. It seems that software maintenance and engineering management can be profitably studied using materials from MSR data papers, but software requirements, economics, and engineering models and methods 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, while in others, such as software requirements, there is an established tradition to publish data sets together with research papers (Zogaan et al. 2017). Nevertheless, data sets for underrepresented SWEBOK areas might have lasting impact in their subfield despite being less popular.

Implication 1 *The* MSR *data showcase track chairs could target the call for data papers on process-oriented topics and less covered* SWEBOK *areas, to possibly improve their footprint.*

Although one might expect that a data paper is typically 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. According to the manual analysis of the weak citations, the most prominent reasons are mentioning the work as an example, attributing a work's statement, and using the work's methods. Be that as it may, based on the difference between the citations to data papers and to other studies, there seems to be room for improving the data papers' use.

Implication 2 *The actual use of data papers could potentially be increased through the promotion of open science initiatives by journal editors and conference program committees, such as the* ACM *Artifact Review and Badging policy* (Boisvert 2016).

With each data paper strongly 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 6), 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 5).

Regarding impact, the number of strong citations to data papers is constantly rising (Fig. 1), indicating that the concept of data papers has long-term value. The enduring usefulness of specific papers is also apparent by looking at the timeline of strong citations to specific MSR data showcase papers over a five-year period (Fig. 3). While the majority of papers indicate a short shelf life, the trend of the most strongly cited data papers retaining their citation number, or obtaining even more strong citations, is yet another manifestation of the Matthew effect in science (Merton 1968). However, this results in a constant need for new data papers, which was also expressed by some respondents of the survey stating that fields with ongoing changes result in a short shelf life for the corresponding data sets.

Implication 3 *The short shelf life of data sets implies a need for a constant stream of new data papers.*

Yet, surprisingly for an artifact whose main purpose is for others to build on, data papers are cited 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 used. The citation's distribution long tail — just 9% of the data papers are strongly 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 (Fig. 7), 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 three reasons that could explain this phenomenon. First, data papers may not publish data that are actually useful for conducting other studies. Our author survey suggests that high data quality, ease of use, and data freshness are the most valuable characteristics

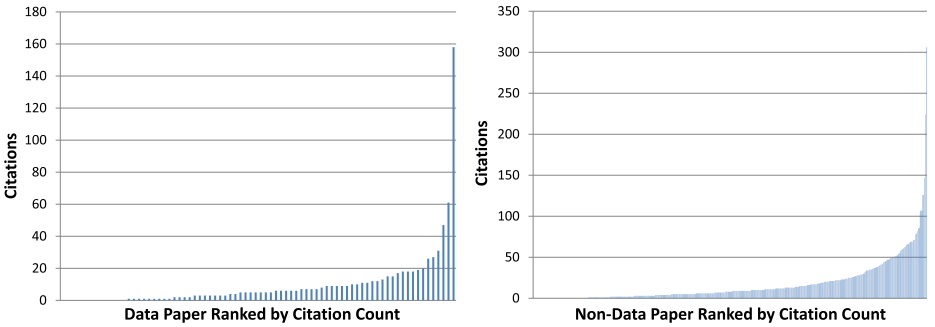

Fig. 7 Distribution in the number of citations to MSR data papers (left) and MSR non-data papers (right). The similar shape of the two distributions indicates that the reason for the lower citation count of MSR data papers is the overall lower number of citations to each data paper compared to the citations to each non-data paper

in data set selection. To address this problem, authors of new data papers can draw from these findings and ensure the satisfaction of the particular criteria in their work. In addition, 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 again by promoting open science initiatives.

Implication 4 *Program committees could consider adopting more stringent criteria for accepting data papers, to potentially improve their usage.*

Second, existing data papers may not satisfy the needs and interests of software researchers. The responses propose exploiting data related to topics such as artificial intelligence and machine learning, collaborative software development, health, fitness and performance, online computing courses, video material and video games. Suggested future data sources include the sectors of education, health, energy, manufacturing, and autonomous driving, entertainment, and smart devices.

Implication 5 *Prospective authors of data papers can exploit the survey's insights to produce quality work that will meet the community's expectations, needs, and interests.*

Third, researchers may be reluctant to work with data coming outside their organization — also known as the *not invented here syndrome* (Piezunka and Dahlander 2015), or fear that working with publicly available data is less likely to yield original results. Although respondents of the survey reported that personal connections with data paper authors play little role in selecting data papers, the high number of papers used by their authors (Fig. 2) contradicts this. This practice suggests the possibility of adopting a workflow similar to that of pre-registered studies (Hardwicke and Ioannidis 2018): publishing a data paper and then employing it for empirical software engineering research. Such a workflow may further strengthen the safeguards promoted by pre-registered studies against p-hacking and publication bias (Kupferschmidt 2018). In addition, encouraging the advance publication of a study's data would level the playing field between the scientists with access to rich empirical data and those without.

Implication 6 *Methodology researchers, conference program committees, and journal editorial boards could examine the opportunities and implications associated with a research paradigm where the data employed in empirical studies are published before the studies that analyze them.*

Looking at the most used data paper, GHTorrent (Gousios 2013), we observe that it is characterized by the majority of the attributes considered useful by the respondents (Fig. 5). Particularly, one highlighted that GHTorrent's updatability through its available source code "lends credence to the construct validity of the data set, since the instrument used to curate it is open for review". Adding to that the continuous human effort and attention by the curators for its regular maintenance, through daily and bi-monthly database dumps accessible from its web site,[11] addressing users' suggestions and bug reports submitted to the GitHub project,[12] this seems to create a self-reinforcing feedback loop between the curators' efforts and the continuing citations to it.

Implication 7 *Self-reinforcing feedback loops could affect positively a data set's citations over time, starting from the curators' regular data set updates, maintenance, and support.*

Overall, data paper authors tend to publish work with data they have gathered in the context of other research. Still, at the same time they seem motivated to benefit the community with new data, and they consider that a worthwhile way of doing so is by publishing data papers. Encouragingly, our author survey paints a picture of an open and meritocratic community, with authors failing to agree that drafting a data paper is an easy way to pad a CV with more publications. Moreover, they seem notable supporters of open science, through various open-ended responses where they expressed their desire for furthering open science goals. Hence, the suggestion regarding the promotion of open science initiatives is enhanced through the particular observation.

6 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. the work by Ferenc et al. (2018) — or the *Empirical Software Engineering* — e.g. the paper by Squire (2018). 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 3.3), without the bias associated with a paper's publication venue. Furthermore, external threats are also related to our ability to generalize the author survey results. Again, the sample selection only within the MSR boundaries prevents us from generalizing our conclusions to other venues. Still, it allowed meaningful insights to be derived, which could be enhanced and generalized through replication of the study to other venues.

The major threats to the study's internal validity stem from the steps during which 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,

[11] https://ghtorrent.org

[12] https://github.com/ghtorrent/ghtorrent.org

the analysis of the weak citations, the clustering of data papers, the classification of studies using data papers, and the pair coding of open-ended responses. Especially the clustering of data papers introduced in Table 2 holds another serious threat associated with the establishment of the clusters themselves. As elaborated in Section 3.1, clusters resulted from a conceptual analysis of the corresponding data studies. The risk stemming from the pair coding process is related to the loss of accuracy of the original response due to an increased level of categorization. The trustworthiness of the processes of clustering, classification, and pair coding were enhanced through the use of multiple raters and coders, and by grounding them on established research methods. However, we acknowledge that validity risks derived from manual processes requiring human judgment cannot be completely eliminated (Petersen et al. 2015). Another threat related to the survey responses is social desirability bias (Furnham 1986) (i.e., a respondent's possible tendency to appear in a positive light, such as by showing they are fair or rational). Particularly, the answers presented in Fig. 4 may lack some truthfulness. For instance, one should not over-interpret that few respondents consider the MSR data showcase track *a straightforward way to publish in the MSR conference*. To mitigate this bias, participants were informed that responses would be anonymous. Question-order effect (Sigelaman 1981) (e.g. one question may have provided context for the next one) may have led respondents to a specific answer, especially in the answers presented in Figs. 5 and 6. One approach to mitigate this bias could have been randomizing the order of questions. In our case, we decided to order the questions in a convenient manner for respondents to easily recall and understand the context of the questions asked.

7 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, by encouraging their constant stream and use, and by promoting open science initiatives. 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.

Acknowledgements Panos Louridas provided insightful comments on this manuscript. Furthermore, Georgios Gousios's suggestions regarding the refinement of the questionnaire were crucial for the survey attainment. 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.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

Aivaloglou E, Hermans F, Moreno-León J, Robles G (2017) A dataset of scratch programs: scraped, shaped and scored. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 511–514. https://doi.org/10.1109/MSR.2017.45

Allix K, Bissyandé TF, Klein J, Le Traon Y (2016) Androzoo: collecting millions of android apps for the research community. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 468–471. https://doi.org/10.1145/2901739.2903508

Almakadmeh M, Abran A (2017) 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. https://doi.org/10.4236/jsea.2017.108038

Altinger H, Siegl S, Dajsuren Y, Wotawa F (2015) 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. IEEE Press, Piscataway, MSR '15, pp 494–497. https://doi.org/10.1109/MSR.2015.72

Amann S, Nadi S, Nguyen HA, Nguyen TN, Mezini M (2016) MUBench: a benchmark for API-misuse detectors. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 464–467. https://doi.org/10.1145/2901739.2903506

Baldassari B, Preux P (2014) Understanding software evolution: the Maisqual Ant data set. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 424–427. https://doi.org/10.1145/2597073.2597136

Barik T, Lubick K, Smith J, Slankas J, Murphy-Hill E (2015) FUSE: a reproducible, extendable, internet-scale corpus of spreadsheets. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 486–489. https://doi.org/10.1109/MSR.2015.70

Binkley D, Lawrie D, Pollock L, Hill E, Vijay-Shanker K (2013) A dataset for evaluating identifier splitters. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 401–404. https://doi.org/10.1109/MSR.2013.6624055

Bloemen R, Amrit C, Kuhlmann S, Ordóñez Matamoros G (2014) Gentoo package dependencies over time. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 404–407. https://doi.org/10.1145/2597073.2597131

Boisvert RF (2016) Incentivizing reproducibility. Commun ACM 59(10):5–5. https://doi.org/10.1145/2994031

Bourque P, Fair RE (eds) (2014) Guide to the Software Engineering Body of Knowledge, version 3.0 edn. IEEE Computer Society, New York, http://www.swebok.org

Bradford SC (1985) Sources of information on specific subjects 1934. Journal of Information Science 10(4):176–180. https://doi.org/10.1177/016555158501000407

Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 80(4):571–583. https://doi.org/10.1016/j.jss.2006.07.009

Butler S, Wermelinger M, Yu Y, Sharp H (2013) INVocD: identifier name vocabulary dataset. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 405–408. https://doi.org/10.1109/MSR.2013.6624056

Chametzky B (2016) Coding in classic grounded theory: I've done an interview; now what? Sociology Mind 06:163–172. https://doi.org/10.4236/sm.2016.64014

Chatzidimitriou KC, Papamichail MD, Diamantopoulos T, Tsapanos M, Symeonidis AL (2018) npm-miner: an infrastructure for measuring the quality of the npm registry. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 42–45. https://doi.org/10.1145/3196398.3196465

Cheikhi L, Abran A (2013) 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. IEEE Press, Piscataway, IWSM-Mensura '13, pp 17–24. https://doi.org/10.1109/IWSM-Mensura.2013.13

Conklin M, Howison J, Crowston K (2005) Collaboration using OSSmole: a repository of FLOSS data and analyses. In: Proceedings of the 2nd international workshop on mining software repositories. ACM, New York, MSR '05, pp 1–5. https://doi.org/10.1145/1082983.1083164

Cukic B (2005) Guest editor's introduction: the promise of public software engineering data repositories. IEEE Software 22(6):20–22. https://doi.org/10.1109/MS.2005.153

Dit B, Holtzhauer A, Poshyvanyk D, Kagdi H (2013) A dataset from change history to support evaluation of software maintenance tasks. In: Proceedings of the 10th working conference on mining

software repositories. IEEE Press, Piscataway, MSR '13, pp 131–134. https://doi.org/10.1109/MSR.2013.6624019

Efstathiou V, Chatzilenas C, Spinellis D (2018) Word embeddings for the software engineering domain. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 38–41. https://doi.org/10.1145/3196398.3196448

Farah G, Tejada JS, Correal D (2014) OpenHub: a scalable architecture for the analysis of software quality attributes. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 420–423. https://doi.org/10.1145/2597073.2597135

Ferenc R, Tóth Z, Ladányi G, Siket I, Gyimóthy T (2018) A public unified bug dataset for Java. In: Proceedings of the 14th international conference on predictive models and data analytics in software engineering. ACM, New York, PROMISE '18, pp 12–21. https://doi.org/10.1145/3273934.3273936

de Freitas FG, de Souza JT (2011) Ten years of search based software engineering: a bibliometric analysis. In: Cohen MB, Ó Cinnéide M (eds) Proceedings of the 3rd international symposium on search based software engineering. Springer, Berlin, SSBSE '11, pp 18–32. https://link.springer.com/chapter/10.1007/978-3-642-23716-4_5

Fujiwara K, Hata H, Makihara E, Fujihara Y, Nakayama N, Iida H, Matsumoto K (2014) Kataribe: a hosting service of historage repositories. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 380–383. https://doi.org/10.1145/2597073.2597125

Furnham A (1986) Response bias, social desirability and dissimulation. Personality and Individual Differences 7(3):385–400. https://doi.org/10.1016/0191-8869(86)90014-0

Gao J, Yang X, Jiang Y, Liu H, Ying W, Zhang X (2018) JBench: a dataset of data races for concurrency testing. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 6–9. https://doi.org/10.1145/3196398.3196451

Geiger FX, Malavolta I, Pascarella L, Palomba F, Di Nucci D, Bacchelli A (2018) A graph-based dataset of commit history of real-world android apps. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 30–33. https://doi.org/10.1145/3196398.3196460

German DM, Adams B, Hassan AE (2015) 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. IEEE Press, Piscataway, MSR '15, pp 470–473. https://doi.org/10.1109/MSR.2015.66

Gkortzis A, Mitropoulos D, Spinellis D (2018) VulinOSS: a dataset of security vulnerabilities in open-source systems. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 18–21. https://doi.org/10.1145/3196398.3196454

Glaser B, Strauss A (1967) The Discovery of Grounded Theory: Strategies for Qualitative Research Observations (Chicago Ill.), Aldine Publishing

Glass RL (1994) An assessment of systems and software engineering scholars and institutions. Journal of Systems and Software 27(1):63–67. https://doi.org/10.1016/0164-1212(94)90115-5

Goeminne M, Claes M, Mens T (2013) A historical dataset for the Gnome ecosystem. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 225–228. https://doi.org/10.1109/MSR.2013.6624032

Gonzalez-Barahona JM, Robles G, Izquierdo-Cortazar D (2015) The MetricsGrimoire database collection. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 478–481. https://doi.org/10.1109/MSR.2015.68

Gousios G (2013) The GHTorrent dataset and tool suite. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 233–236. https://doi.org/10.1109/MSR.2013.6624034

Gousios G, Zaidman A (2014) A dataset for pull-based development research. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 368–371. https://doi.org/10.1145/2597073.2597122

Gousios G, Vasilescu B, Serebrenik A, Zaidman A (2014) Lean GHTorrent: GitHub data on demand. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 384–387. https://doi.org/10.1145/2597073.2597126

Gousios G, Storey MA, Bacchelli A (2016) Work practices and challenges in pull-based development: the contributor's perspective. In: Proceedings of the 38th international conference on software engineering. Association for Computing Machinery, New York, ICSE '16, pp 285–296. https://doi.org/10.1145/2884781.2884826

Gu Y (2004) Global knowledge management research: a bibliometric analysis. Scientometrics 61(2):171–190. https://doi.org/10.1023/B:SCIE.0000041647.01086.f4

Habayeb M, Miranskyy A, Murtaza SS, Buchanan L, Bener AB (2015) The Firefox temporal defect dataset. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 498–501. https://doi.org/10.1109/MSR.2015.73

Hamasaki K, Kula RG, Yoshida N, Cruz AEC, Fujiwara K, Iida H (2013) Who does what during a code review? Datasets of OSS peer review repositories. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 49–52. https://doi.org/10.1109/MSR.2013.6624003

Hamming R (1969) One man's view of computer science. Journal of the ACM 16(1):3–12. https://doi.org/10.1145/321495.321497

Hardwicke TE, Ioannidis JP (2018) Mapping the universe of registered reports. Nature Human Behaviour 2(11):793–796

Harman M, Mansouri SA, Zhang Y (2009) Search based software engineering: a comprehensive analysis and review of trends techniques and applications. Tech. Rep. TR-09-03, Department of Computer Science, King's College London, and Brunel Business School, Brunel University, London, UK, https://www.researchgate.net/profile/Yuanyuan_Zhang12/publication/228671024_Search_Based_Software_Engineering_A_Comprehensive_Analysis_and_Review_of_Trends_Techniques_and_Applications/links/00b4951811ba6a40eb000000.pdf

Janjic W, Hummel O, Schumacher M, Atkinson C (2013) An unabridged source code dataset for research in software reuse. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 339–342. https://doi.org/10.1109/MSR.2013.6624047

Karakoidas V, Mitropoulos D, Louridas P, Gousios G, Spinellis D (2015) Generating the blueprints of the Java ecosystem. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 510–513. https://doi.org/10.1109/MSR.2015.76

Keivanloo I, Forbes C, Hmood A, Erfani M, Neal C, Peristerakis G, Rilling J (2012) A linked data platform for mining software repositories. In: Proceedings of the 9th working conference on mining software repositories. IEEE Press, Piscataway, MSR '12, pp 32–35. https://doi.org/10.1109/MSR.2012.6224296

Kim S, Zimmermann T, Kim M, Hassan A, Mockus A, Girba T, Pinzger M, Whitehead EJ Jr, Zeller A (2006) TA-RE: an exchange language for mining software repositories. In: Proceedings of the 3rd international workshop on mining software repositories. ACM, New York, MSR '06, pp 22–25. https://doi.org/10.1145/1137983.1137990

Kitchenham B (2004) Procedures for performing systematic reviews. Tech. Rep. TR/SE-0401, Department of Computer Science, Keele University, Keele, Staffs, UK, http://www.it.hiof.no/~haraldh/misc/2016-08-22-smat/Kitchenham-Systematic-Review-2004.pdf

Kitchenham B, Pfleeger SL (2003) Principles of survey research: Part 6: data analysis. SIGSOFT Softw Eng Notes 28(2):24–27. https://doi.org/10.1145/638750.638758

Kitchenham BA, Pfleeger SL (2002a) Principles of survey research: Part 2: designing a survey. SIGSOFT Softw Eng Notes 27(1):18–20. https://doi.org/10.1145/566493.566495

Kitchenham BA, Pfleeger SL (2002b) Principles of survey research: Part 3: constructing a survey instrument, vol 27, pp 20–24. https://doi.org/10.1145/511152.511155

Kitchenham BA, Pfleeger SL (2002c) Principles of survey research: Part 4: questionnaire evaluation. SIGSOFT Softw Eng Notes 27(3):20–23. https://doi.org/10.1145/638574.638580

Kitchenham BA, Pfleeger SL (2002d) Principles of survey research: Part 5: populations and samples. SIGSOFT Softw Eng Notes 27(5):17–20. https://doi.org/10.1145/571681.571686

Kitchenham BA, Pfleeger SL, Pickard LM, Jones PW, Hoaglin DC, Emam KE, Rosenberg J (2002) Preliminary guidelines for empirical research in software engineering. IEEE Trans Softw Eng 28(8):721–734. https://doi.org/10.1109/TSE.2002.1027796

Kotti Z, Spinellis D (2019) Standing on shoulders or feet?: The usage of the MSR data papers. In: Proceedings of the 16th international conference on mining software repositories. IEEE Press, Piscataway, MSR '19, pp 565–576. https://doi.org/10.1109/MSR.2019.00085

von Krogh G, von Hippel E (2006) The promise of research on open source software. Management Science 52(7):975–983. https://doi.org/10.1287/mnsc.1060.0560

Krüger S, Späth J, Ali K, Bodden E, Mezini M (2018) CrySL: an extensible approach to validating the correct usage of cryptographic APIs. In: Millstein T (ed) Proceedings of the 32nd European conference on object-oriented programming, vol 109. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, ECOOP '18, pp 10:1–10:27. https://doi.org/10.4230/LIPIcs.ECOOP.2018.10

Krüger S, Späth J, Ali K, Bodden E, Mezini M (2019) CrySL: an extensible approach to validating the correct usage of cryptographic APIs. IEEE Transactions on Software Engineering. https://doi.org/10.1109/TSE.2019.2948910

Krutz DE, Le W (2014) A code clone oracle. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 388–391. https://doi.org/10.1145/2597073.2597127

Krutz DE, Mirakhorli M, Malachowsky SA, Ruiz A, Peterson J, Filipski A, Smith J (2015) A dataset of open-source android applications. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 522–525. https://doi.org/10.1109/MSR.2015.79

Kupferschmidt K (2018) More and more scientists are preregistering their studies. should you? Science. https://doi.org/10.1126/science.aav4786

Lamkanfi A, Pérez J, Demeyer S (2013) 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. IEEE Press, Piscataway, MSR '13, pp 203–206. https://doi.org/10.1109/MSR.2013.6624028

Lavazza L, Santillo L (2012) Historical data repositories in software engineering: status and possible improvements. In: Proceedings of the 2012 joint conference of the 22nd international workshop on software measurement and the 2012 seventh international conference on software process and product measurement, pp 221–225. https://doi.org/10.1109/IWSM-MENSURA.2012.39

Lazar A, Ritchey S, Sharif B (2014) Generating duplicate bug datasets. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 392–395. https://doi.org/10.1145/2597073.2597128

Liebchen GA, Shepperd M (2008) Data sets and data quality in software engineering. In: Proceedings of the 4th international workshop on predictor models in software engineering. ACM, New York, PROMISE '08, pp 39–44. https://doi.org/10.1145/1370788.1370799

Lotka AJ (1926) The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(12):317–323. http://www.jstor.org/stable/24529203

MacLean AC, Knutson CD (2013) Apache commits: social network dataset. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 135–138. https://doi.org/10.1109/MSR.2013.6624020

Madeyski L, Kawalerowicz M (2017) Continuous defect prediction: the idea and a related dataset. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 515–518. https://doi.org/10.1109/MSR.2017.46

Markovtsev V, Long W (2018) Public Git archive: a big code dataset for all. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 34–37. https://doi.org/10.1145/3196398.3196464

Martins P, Achar R, Lopes CV (2018) 50K-C: a dataset of compilable, and compiled, Java projects. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 1–5. https://doi.org/10.1145/3196398.3196450

Mauczka A, Brosch F, Schanes C, Grechenig T (2015) Dataset of developer-labeled commit messages. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 490–493. https://doi.org/10.1109/MSR.2015.71

Merton RK (1968) The Matthew effect in science. Science 159(3810):56–63

Mierle K, Laven K, Roweis S, Wilson G (2005) Mining student CVS repositories for performance indicators. In: Proceedings of the 2nd international workshop on mining software repositories. ACM, New York, MSR '05, pp 1–5. https://doi.org/10.1145/1082983.1083150

Mitropoulos D, Karakoidas V, Louridas P, Gousios G, Spinellis D (2014) The bug catalog of the Maven ecosystem. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 372–375. https://doi.org/10.1145/2597073.2597123

Mukadam M, Bird C, Rigby PC (2013) Gerrit software code review data from android. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 45–48. https://doi.org/10.1109/MSR.2013.6624002

Murakami H, Higo Y, Kusumoto S (2014) A dataset of clone references with gaps. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 412–415. https://doi.org/10.1145/2597073.2597133

Noten J, Mengerink JGM, Serebrenik A (2017) A data set of OCL expressions on GitHub. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 531–534. https://doi.org/10.1109/MSR.2017.52

Novielli N, Calefato F, Lanubile F (2018) A gold standard for emotion annotation in Stack Overflow. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 14–17. https://doi.org/10.1145/3196398.3196453

Nussbaum L, Zacchiroli S (2010) The ultimate debian database: consolidating bazaar metadata for quality assurance and data mining. In: Proceedings of the 7th working conference on mining software repositories. IEEE Press, Piscataway, MSR '10, p 10. https://doi.org/10.1109/MSR.2010.5463277

Ohira M, Kashiwa Y, Yamatani Y, Yoshiyuki H, Maeda Y, Limsettho N, Fujino K, Hata H, Ihara A, Matsumoto K (2015) A dataset of high impact bugs: manually-classified issue reports. In: Proceedings

of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 518–521. https://doi.org/10.1109/MSR.2015.78

Ortu M, Murgia A, Destefanis G, Tourani P, Tonelli R, Marchesi M, Adams B (2016) The emotional side of software developers in JIRA. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 480–483. https://doi.org/10.1145/2901739.2903505

Paixao M, Krinke J, Han D, Harman M (2018) CROP: linking code reviews to source code changes. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 46–49. https://doi.org/10.1145/3196398.3196466

Palomba F, Nucci DD, Tufano M, Bavota G, Oliveto R, Poshyvanyk D, De Lucia A (2015) Landfill: an open dataset of code smells with public evaluation. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 482–485. https://doi.org/10.1109/MSR.2015.69

Passos L, Czarnecki K (2014) A dataset of feature additions and feature removals from the Linux kernel. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 376–379. https://doi.org/10.1145/2597073.2597124

Petersen K, Feldt R, Mujtaba S, Mattsson M (2008) Systematic mapping studies in software engineering. In: Proceedings of the 12th international conference on evaluation and assessment in software engineering. BCS Learning & Development Ltd., Swindon, EASE '08, pp 68–77. http://dl.acm.org/citation.cfm?id=2227115.2227123

Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology 64:1–18. https://doi.org/10.1016/j.infsof.2015.03.007

Pfleeger SL, Kitchenham BA (2001) Principles of survey research: Part 1: turning lemons into lemonade. SIGSOFT Softw Eng Notes 26(6):16–18. https://doi.org/10.1145/505532.505535

Piezunka H, Dahlander L (2015) Distant search, narrow attention: how crowding alters organizations' filtering of suggestions in crowdsourcing. Academy of Management Journal 58(3):856–880. https://doi.org/10.5465/amj.2012.0458

Ponzanelli L, Mocci A, Lanza M (2015) StORMeD: stack overflow ready made data. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 474–477. https://doi.org/10.1109/MSR.2015.67

Proksch S, Amann S, Nadi S, Mezini M (2016) A dataset of simplified syntax trees for C#. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 476–479. https://doi.org/10.1145/2901739.2903507

Raemaekers S, Van Deursen A, Visser J (2013) The Maven repository dataset of metrics, changes, and dependencies. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 221–224. https://doi.org/10.1109/MSR.2013.6624031

Robles G (2010) 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. IEEE Press, Piscataway, MSR '10, pp 171–180. https://doi.org/10.1109/MSR.2010.5463348

Robles G, Arjona Reina L, Serebrenik A, Vasilescu B, González-Barahona JM (2014) 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. ACM, New York, MSR '14, pp 396–399. https://doi.org/10.1145/2597073.2597129

Robles G, Ho-Quang T, Hebig R, Chaudron MRV, Fernandez MA (2017) An extensive dataset of UML models in GitHub. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 519–522. https://doi.org/10.1109/MSR.2017.48

Sadat M, Bener AB, Miranskyy AV (2017) Rediscovery datasets: connecting duplicate reports. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 527–530. https://doi.org/10.1109/MSR.2017.50

Saha RK, Lyu Y, Lam W, Yoshida H, Prasad MR (2018) Bugs.jar: A large-scale, diverse dataset of real-world Java bugs. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 10–13. https://doi.org/10.1145/3196398.3196473

Saini V, Sajnani H, Ossher J, Lopes CV (2014) A dataset for Maven artifacts and bug patterns found in them. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 416–419. https://doi.org/10.1145/2597073.2597134

Salinger S, Plonka L, Prechelt L (2008) A coding scheme development methodology using Grounded Theory for qualitative analysis of pair programming. Human Technology: An Interdisciplinary Journal on Humans in ICT Environments 4. https://doi.org/10.17011/ht/urn.200804151350

Sawant AA, Bacchelli A (2015) A dataset for API usage. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 506–509. https://doi.org/10.1109/MSR.2015.75

Sayyad Shirabad J, Menzies T (2005) The PROMISE repository of software engineering databases, School of Information Technology and Engineering, University of Ottawa, Canada, http://promise.site.uottawa.ca/SERepository

Schermann G, Zumberi S, Cito J (2018) Structured information on state and evolution of Dockerfiles on GitHub. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 26–29. https://doi.org/10.1145/3196398.3196456

Shull FJ, Carver JC, Vegas S, Juristo N (2008) The role of replications in Empirical Software Engineering. Empirical Software Engineering 13(2):211–218. https://doi.org/10.1007/s10664-008-9060-1

Sigelaman L (1981) Question-order effects on presidential popularity. Public Opinion Quarterly 45(2):199–207. https://academic.oup.com/poq/article-pdf/45/2/199/5432386/45-2-199.pdf

Spacco J, Strecker J, Hovemeyer D, Pugh W (2005) Software repository mining with Marmoset: an automated programming project snapshot and testing system. In: Proceedings of the 2nd international workshop on mining software repositories. ACM, New York, MSR '05, pp 1–5. https://doi.org/10.1145/1082983.1083149

Spinellis D (2015) A repository with 44 years of Unix evolution. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 462–465. https://doi.org/10.1109/MSR.2015.64

Spinellis D (2018) Documented Unix facilities over 48 years. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 58–61. https://doi.org/10.1145/3196398.3196476

Squire M (2013a) Apache-affiliated twitter screen names: a dataset. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 305–308. https://doi.org/10.1109/MSR.2013.6624043

Squire M (2013b) Project roles in the Apache software foundation: a dataset. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 301–304. https://doi.org/10.1109/MSR.2013.6624042

Squire M (2016) Data sets: the circle of life in Ruby hosting, 2003-2015. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 452–459. https://doi.org/10.1145/2901739.2903509

Squire M (2018) Data sets describing the circle of life in Ruby hosting, 2003–2016. Empirical Software Engineering 23(2):1123–1152. https://doi.org/10.1007/s10664-017-9581-6

Trockman A, Zhou S, Kästner C, Vasilescu B (2018) 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. ACM, New York, ICSE '18, pp 511–522. https://doi.org/10.1145/3180155.3180209

Vasilescu B, Serebrenik A, Mens T (2013) A historical dataset of software engineering conferences. In: Proceedings of the 10th working conference on mining software repositories. IEEE Press, Piscataway, MSR '13, pp 373–376. https://doi.org/10.1109/MSR.2013.6624051

Vasilescu B, Serebrenik A, Filkov V (2015) A data set for social diversity studies of GitHub teams. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 514–517. https://doi.org/10.1109/MSR.2015.77

Wagstrom P, Jergensen C, Sarma A (2013) 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. IEEE Press, Piscataway, MSR '13, pp 229–232. https://doi.org/10.1109/MSR.2013.6624033

Wallace D (1998) Enhancing competitiveness via a public fault and failure data repository. In: Proceedings of the third ieee international high-assurance systems engineering symposium, pp 178–185. https://doi.org/10.1109/HASE.1998.731610

Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Quarterly 26(2):xiii–xxiii

Wermelinger M, Yu Y (2015) An architectural evolution dataset. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 502–505. https://doi.org/10.1109/MSR.2015.74

Williams JR, Di Ruscio D, Matragkas N, Di Rocco J, Kolovos DS (2014) Models of OSS project meta-information: a dataset of three forges. In: Proceedings of the 11th working conference on mining software repositories. ACM, New York, MSR '14, pp 408–411. https://doi.org/10.1145/2597073.2597132

Wong WE, Tse T, Glass RL, Basili VR, Chen TY (2011) An assessment of systems and software engineering scholars and institutions (2003–2007 and 2004–2008). Journal of Systems and Software 84(1):162–168. https://doi.org/10.1016/j.jss.2010.09.036

Xu Y, Zhou M (2018) A multi-level dataset of Linux kernel patchwork. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 54–57. https://doi.org/10.1145/3196398.3196475

Yamashita A, Abtahizadeh SA, Khomh F, Guéhéneuc YG (2017) Software evolution and quality data from controlled, multiple, industrial case studies. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 507–510. https://doi.org/10.1109/MSR.2017.44

Yamashita A, Petrillo F, Khomh F, Guéhéneuc YG (2018) Developer interaction traces backed by IDE screen recordings from Think aloud sessions. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 50–53. https://doi.org/10.1145/3196398.3196457

Yang X, Kula RG, Yoshida N, Iida H (2016) Mining the modern code review repositories: a dataset of people, process and product. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 460–463. https://doi.org/10.1145/2901739.2903504

Yu Y, Li Z, Yin G, Wang T, Wang H (2018) A dataset of duplicate pull-requests in GitHub. In: Proceedings of the 15th international conference on mining software repositories. ACM, New York, MSR '18, pp 22–25. https://doi.org/10.1145/3196398.3196455

Zacchiroli S (2015) The Debsources dataset: two decades of Debian source code metadata. In: Proceedings of the 12th working conference on mining software repositories. IEEE Press, Piscataway, MSR '15, pp 466–469. https://doi.org/10.1109/MSR.2015.65

Zhang C, Hindle A (2014) 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. ACM, New York, MSR '14, pp 400–403. https://doi.org/10.1145/2597073.2597130

Zhu C, Li Y, Rubin J, Chechik M (2017) A dataset for dynamic discovery of semantic changes in version controlled software histories. In: Proceedings of the 14th international conference on mining software repositories. IEEE Press, Piscataway, MSR '17, pp 523–526. https://doi.org/10.1109/MSR.2017.49

Zhu J, Zhou M, Mei H (2016) Multi-extract and multi-level dataset of Mozilla issue tracking history. In: Proceedings of the 13th international conference on mining software repositories. ACM, New York, MSR '16, pp 472–475. https://doi.org/10.1145/2901739.2903502

Zimmermann T, Di Penta M, Kim S(eds) (2013a) Proceedings of the 10th Working Conference on Mining Software Repositories, MSR '13, IEEE Computer Society, http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6597024

Zimmermann T, Di Penta M, Kim S, German DM, Bacchelli A (2013b) Welcome from the chairs. In: Proceedings of the 10th working conference on mining software repositories, MSR '13, pp iii–viii. https://doi.org/10.1109/MSR.2013.6623995

Zogaan W, Sharma P, Mirahkorli M, Arnaoudova V (2017) Datasets from fifteen years of automated requirements traceability research: current state, characteristics, and quality. In: Proceedings of the 25th international requirements engineering conference, IEEE, RE '17, pp 110–121. https://doi.org/10.1109/RE.2017.80

Affiliations

Zoe Kotti[1] · Konstantinos Kravvaritis[1] · Konstantina Dritsa[1] · Diomidis Spinellis[1]

Konstantinos Kravvaritis
kravvaritisk@aueb.gr

Konstantina Dritsa
dritsakon@aueb.gr

Diomidis Spinellis
dds@aueb.gr

[1] Department of Management Science and Technology, Athens University of Economics and Business, 76, Patission street, Athens, 104 34 Greece