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

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

Abstract—Introduction: The establishment of the Mining Software Repositories (MSR) Data Showcase conference track has encouraged researchers to provide more data sets as a basis for further empirical studies.

Objectives: Examine the usage of the data papers published in the MSR proceedings in terms of use frequency, users, and use purpose.

Methods: Data track papers were collected from the MSR Data Showcase and through the manual inspection of older MSR proceedings. The use of data papers was established through citation searching followed by reading the studies that have cited them. Data papers were then clustered based on their content, whereas their citations were classified according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge.

Results: We found that 65% of the data papers have been used in other studies, with a long-tail distribution in the number of citations. MSR data papers are cited less than other MSR papers. A considerable number of the citations stem from the teams that authored the data papers. Publications providing repository data and metadata are the most frequent data papers and the most often cited ones. Mobile application data papers are the least common ones, but the second most frequently cited.

Conclusion: Data papers have provided the foundation for a significant number of studies, but there is room for improvement in their utilization. This can be done by setting a higher bar for their publication, by encouraging their use, and by providing incentives for the enrichment of existing data collections.

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

“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

11968 ACM Turing Award Lecture [36]

2https://github.com/dspinellis/awesome-msr

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

Copyright ©2019 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
The MSR data showcase track, established in 2013, aims at encouraging the research community to develop, share, and document software engineering research data sets. In the words of the 2013 MSR conference chairs [105],

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

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

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

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

### II. Methods

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

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

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

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

### A. Data Paper Collection and Clustering

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

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

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

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

### B. Data Paper Use Identification and Classification

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

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

---

\(^3\)http://doi.org/10.5281/zenodo.2544957

\(^4\)https://scholar.google.com/

\(^5\)https://www.scopus.com/

\(^6\)https://dl.acm.org/

<table>
<thead>
<tr>
<th>Year</th>
<th>Data Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>[80] [55] [15]</td>
</tr>
<tr>
<td>2006</td>
<td>[41]</td>
</tr>
<tr>
<td>2010</td>
<td>[61]</td>
</tr>
<tr>
<td>2012</td>
<td>[40]</td>
</tr>
<tr>
<td>2013</td>
<td>[8] [18] [28] [89] [91] [38] [50] [83] [57] [12] [84] [46] [30] [70] [35]</td>
</tr>
<tr>
<td>2014</td>
<td>[44] [76] [32] [58] [66] [102] [72] [48] [9] [22] [31] [95] [20] [56] [6]</td>
</tr>
<tr>
<td>2015</td>
<td>[88] [77] [62] [45] [25] [4] [81] [94] [53] [7] [39] [65] [68] [101] [34] [29]</td>
</tr>
<tr>
<td>2016</td>
<td>[69] [2] [85] [99] [5] [104] [63]</td>
</tr>
<tr>
<td>2017</td>
<td>[59] [1] [103] [73] [73] [74] [97]</td>
</tr>
<tr>
<td>2018</td>
<td>[52] [100] [60] [24] [96] [75] [64] [98] [82] [23] [13] [51] [78] [26] [19]</td>
</tr>
</tbody>
</table>
that were not publicly available were provided to us through personal communication with the authors.

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

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

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

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

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

C. Citation Analysis

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

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

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

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

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

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

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

\footnote{https://dblp.org/}
was published in 2016 and 2017. Nevertheless, 2018 indicates a double increase in data publications—15 (see Table I).

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

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

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

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>Title</th>
<th>Data Paper</th>
<th>Year</th>
<th>Category</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lean GitTorrent: GitHub Data on Demand</td>
<td>[31]</td>
<td>2014</td>
<td>Repository &amp; Data Metadata</td>
<td>24</td>
</tr>
<tr>
<td>The Maven Repository Dataset of Metrics, Changes, and Dependencies</td>
<td>[70]</td>
<td>2013</td>
<td>Repository Data &amp; Metadata</td>
<td>12</td>
</tr>
<tr>
<td>The Emotional Side of Software Developers in JIRA</td>
<td>[63]</td>
<td>2016</td>
<td>Human-centered Data</td>
<td>12</td>
</tr>
</tbody>
</table>
requirements [2] with 57 citations. GHTorrent’s complementing work, Lean GHTorrent [31], and the peer review data set [35] have also attracted the attention of the research community. Finally, the repository of Maven meta-information [70], along with the collections of Eclipse and Mozilla defects [46] and developers’ sentiments [63] share the same number of uses—twelve.

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

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

Furthermore, concerning the use of data papers by their respective authors, our findings show that 37 papers have been referenced by the teams that authored them. Specifically, 15 studies have been solely deployed either by their first author or his co-authors. Figure 3 depicts for each data paper cited at least once by the first author or the co-authors, the percentage of the uses that stem from the first author, the co-authors, and other unrelated teams. The data papers are sorted in ascending order based on the percentage of the sum of the strong citations made by the first author and the co-authors. Figure 3 shows that research is conducted based on data papers. We examined by hand the papers published in the Computing Research Repository (CoRR), and found that almost all of them (19) are fairly recent (published in 2017 or 2018). This indicates that they are probably archival submissions of material that will eventually also end up in a conference or journal.

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

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

### IV. Discussion

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

Table VI shows the venues where research that is based on data papers has been published. We see that more than a third of the corresponding papers are published in top-tier conferences and journals. This showcases the high quality of research that is conducted based on data papers. We examined by hand the papers published in the Computing Research Repository (CoRR), and found that almost all of them (19) are fairly recent (published in 2017 or 2018). This indicates that they are probably archival submissions of material that will eventually also end up in a conference or journal.
Nevertheless, data sets for underrepresented areas, such as software economics, the release of data sets is hard due to reluctance to work with data coming outside their organization—also known as the not invented here syndrome [67], or a fear that working with publicly available data is less likely to yield original results. The high number of papers used by their authors (Figure 3) corroborates this second reason.

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

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

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

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

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

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

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

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

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

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

![Figure 2: Timeline of strong citations to data papers published from 2005–2014](image-url)
V. Threats to Validity

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

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

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

VI. RELATED WORK

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

A first step regarding bibliometric research in the field of software engineering models was made in 2004 [16] by the organisers of the PROMISE workshop, in their attempt “to strengthen the community’s faith in software engineering models”. Authors of such models were asked to submit, along with their work, a related data set to the PROMISE repository. Many individuals have also carried out interesting quantitative research on various topics. Robles [71] conducted bibliometric research on papers that contained experimental analyses of software projects and were published in the MSR proceedings from 2004–2009. His objective was to review their potential replicability. The outcome proves that MSR authors prefer publicly available data sources from free software repositories. However, the amount of publicly available processed data collections then was very low, a fact we also stated in our results. Concerning replicability, Robles found that only a limited number of publications are replication friendly.

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

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

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

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

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

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

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

VII. CONCLUSIONS

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

ACKNOWLEDGMENTS

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

REFERENCES


