A Literature Review on Reproducibility Studies in Computer Science

Tobias Hummel^{1,*}, Johannes Manner¹

¹Distributed Systems Group, University of Bamberg, Germany

Abstract

Researchers expect a clear and well-documented experiment from industry experience reports and experimental research papers. All necessary configuration parameters, the source code, the experiment's machine configuration etc. should be documented in such a way that readers can interpret the results of these publication in detail. If this is the case, an interested reader is capable of redoing the experiments stated and verifying the results of others. This ability to properly interpret experiments and even reproduce them is a cornerstone of good scientific practice.

The experience from reading papers and consulting secondary studies reveals a different picture. A lot of papers are only partly interpretable since some information is missing. To understand the state of reproducibility in computer science, we conducted a Structured Literature Review (SLR) about reproducibility studies to list their motivations and challenges. These studies already tried to reproduce other research.

Two outcomes are of particular interest. First, the Information Retrieval (IR) domain is the role model w.r.t. reproducibility efforts. Most of the papers included in the SLR are from this domain. Second, publishers and conference formats start to create incentives by awarding badges for papers. Before the badges are awarded, the papers are checked for compliance with the rules of data submission and experiment reproducibility set by the conference respectively the publisher.

Keywords

Reproducibility, Replicability, Repeatability, Experiment Documentation

1. Introduction

"Reproducible Research in all sciences is critical to the advancement of knowledge. It is what enables a researcher to build upon, or refute, previous research allowing the field to act as a collective of knowledge rather than as tiny uncommunicated clusters" - [1, Cacho and Taghva p. 519].

This quote illustrates the importance of reproducibility of scientific work and describes an ideal state to gain knowledge. Nevertheless, the reality does not reflect this ideal state. In 2016, a survey with 1,576 respondents was published in Nature [2]. 90% of them stated that the scientific community in general is facing a reproducibility crisis. To understand this crisis, this paper tries to shed some light into this complex topic and states the current situation for

D 0000-0002-7298-3574 (J. Manner)

S. Böhm and D. Lübke (Eds.): 16th ZEUS Workshop, ZEUS 2024, Ulm, Germany, 29 February–1 March 2024, published at http://ceur-ws.org

Vorkshop ISSN 1613-0073

ZEUS2024: 16th Central European Workshop on Services and their Composition, Ulm, Germany, 2024. *Corresponding author.

tobias.hummel@stud.uni-bamberg.de (T. Hummel); johannes.manner@uni-bamberg.de (J. Manner)

https://www.uni-bamberg.de/pi/team/manner-johannes/ (J. Manner)

^{© 024} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

reproducibility and related terms within the computer science domain. It furthermore highlights the efforts made towards publishing raw data and source code and shares some ideas on how to improve the current situation.

According to the Association for Computing Machinery (ACM) a measurement is reproducible if it "can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts"¹. Artifacts can be, for example, program code or datasets. In addition to reproducibility, the ACM lists two further often used terms at the mentioned website, namely repeatability and replicability. The following list briefly distinguishes the terms from each other:

- reproducibility different team, same experimental setup
- repeatability same team, same experimental setup
- replicability different team, different experimental setup

However, in research these terms are sometimes used interchangeably [3]. Reproducibility and repeatability are even used the other way around [4]. There is also a standard called Reproducibility Badging and Definitions published by the National Information Standards Organization (NISO) [5]. There, they define further nuances of the three introduced terms above. Despite their depth, this standard being around for already three years is only referenced by 27 Google Scholar hits² when searching for the title of this publication. For this paper, we will stick to the ACM definitions introduced above due to their clarity and adaptation in the community. COLLBERG and others expected computer science to be in a special role compared to other disciplines: "reproducing the work published in a systems conference or journal should be as simple as going to the authors' website, downloading their code and data, typing 'make,' and seeing if the results correspond to the published ones" [4, p. 1]. However, in a 30 minutes time frame they were able to retrieve and build the source code of only 32.3% of the papers analyzed. Other secondary studies, e.g. [6, 7, 8], confirm this issue. Only 3 out of 26 experiments are reproducible based on an assessment within the early Function as a Service (FaaS) research domain [6]. Another study concluded that a *majority* of 122 is not reproducible [7]. COUTURE and others [8] revealed that only 26% of 315 data projects published raw data.

All these studies should be a wake-up call for the computer science domain to publish raw data, source code and scripts. For without this background information a correct interpretation of results is not always possible for others which prevents the correct evaluation of the merit of a paper. This can be illustrated with an example from FaaS research: Two SLRs [9, 10] revealed papers reporting about unusual performance results when running a function on a cloud provider. Properly interpreting the original results based on the provider documentation showed that the measurements where due to a memory setting. It unintentionally assigned more than one CPU to the function. Without enough details about the experimental setup, it would not have been possible to falsify the misinterpretations of the original authors. As a consequence, conference chairs and publishers should incentivise researchers to make their experiments open to the public and enable others to reproduce experimental results.

¹https://www.acm.org/publications/policies/artifact-review-and-badging-current

²Google Scholar search was performed on 19th of February 2024.

Since we are interested in the current state of reproducibility studies, the objective of this paper is not to conduct a further reproducibility study but to analyze the already published studies by answering the following three research questions:

- RQ1 What are publication trends for reproducibility studies in computer science?
- RQ2 Why do authors try to reproduce the work of others?
- **RQ3** How successful are the reproducibility studies and which challenges do the authors encounter?

RQ1 focuses on the number of studies and visible trends. The reasons for performing such studies are questioned in **RQ2**. And the last question, **RQ3**, reveals success numbers on how many studies were able to reproduce the original work. One caveat here could be a publication bias, where reproducibility problems could be more likely to be published than success [11]. For upcoming reproducibility studies, we also look at challenges the included publications faced to share aggregated learnings.

The agenda of our paper is as follows: In Section 2, we shortly introduce the SLR methodology and present some numbers and an overview of included papers. Results for our three research questions are presented as subsections in Section 3. Section 4 concludes the paper with a short summary and an outlook to future work.

2. Methodology

We conducted a SLR to understand the current state of reproducibility studies in the computer science domain. A SLR is well suited for summarizing and synthesizing the current status on a topic in a fair way [11]. For the search phase, ACM Digital Library, IEEE Digital Library, DBLP and Google Scholar were used. While the first three libraries have a computer science focus, Google Scholar was chosen as a more general search engine with a large corpus. Therefore, we only used the first 50 entries at Google Scholar, sorted by relevance to get a manageable set of literature as already done by other SLRs [12, 13, 14]. Figure 1 summarizes the search process.



Figure 1: The initial search of the SLR was conducted on 23rd of May 2023 to identify reproducibility studies.

The search term required the phrase "reproducibility study" to be part of the title or abstract. As the search engines differ in their search options, the exact search strings differ [15]. In total,

198 entries were found in the initial search phase. The inclusion and exclusion criteria are based on the research questions and were applied to title, abstract and conclusions [15]. In cases of doubt, the full text was skimmed. The selection criteria were developed upfront and piloted on some studies as recommended by KITCHENHAM and CHARTERS [11]. Nevertheless, they evolved over the process. The inclusion criteria were as follows:

- Reproducibility studies in the area of computer science, other disciplines were excluded immediately.
- Publications in which an attempt is made to reproduce the findings of one or several prior publications.
- Publications by universities, public research institutes or industry.
- In addition to peer-reviewed publications, also preprints.

The following exclusion criteria were used:

- Publications in which the authors do not use or do not mention a specific software implementation.
- Publications that belong to the field of computational science rather than computer science, i.e. that are more concerned with the application of computer science in fields such as medicine.
- Publications in languages other than English (language is often used as an exclusion criterion [16, 17, 11]).

All raw data for the SLR are archived at Zenodo [18]. Inclusion and exclusion decisions as well as all other data related to the SLR can be found there. In case of exclusion, a reason is stated. Although required by KITCHENHAM and CHARTERS, the quality of the studies was not assessed for every paper [11]. It is assumed that in particular search engines from ACM, IEEE and DBLP have results with high quality. Otherwise, studies of questionable quality would have been excluded during the data extraction phase.

After the merge and deduplication step, 20 results remained. During data extraction, one paper turned out to be a replicability study and was thus excluded [19]. Within the remaining, 14 belonged to the information retrieval domain [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. So as not to over-represent this field, out of these 14 papers the first four papers based on the first author names [20, 21, 22, 23] were chosen as examples from this computer science domain to discuss RQ2 and RQ3. Table 1 shows the remaining nine publications and the computer science sub-field to which they belong.

3. Results

3.1. RQ1: Publication Facts and Trends

The most obvious fact in our SLR investigation is that the information retrieval domain is a role model for reproducibility studies. 74% of our identified papers (14/19) are from this domain. The implementation of different ranking and retrieval algorithms is inherent to this computer science sub-field. For those an interpretation of results is only feasible by checking the raw data

Table 1

Publications included in the SLR by computer science sub-field.

Title	Ref	Sub-Field
A Comparison between Term-Independence Retrieval Models for	[23]	
A Reproducibility Study of Question Retrieval for Clarifying Ques- tions	[21]	Information Retrieval
Cross-Domain Retrieval in the Legal and Patent Domains: A Repro- ducibility Study	[20]	
Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study	[22]	
A Thorough Reproducibility Study on Sentiment Classification: Methodology, Experimental Setting, Results	[34]	Natural Language Processing
Reproducibility in Computational Linguistics: Is Source Code Enough?	[35]	8
Examining the Reproducibility of Using Dynamic Loop Scheduling Techniques in Scientific Applications	[36]	Distributed Systems
IPAL: Breaking up Silos of Protocol-Dependent and Domain-Specific Industrial Intrusion Detection Systems	[37]	Security and Privacy
Machine Learning Based Invariant Generation: A Framework and Reproducibility Study	[38]	Software Verifica- tion

and experiment details. The same holds true for another little cluster, namely Natural Language Processing (NLP). Here, the research objectives are also highly dependent on input data and its processing.



Figure 2: Number of reproducibility papers by publication year.

When we searched for literature, we did not limit the publication year. Nevertheless, the oldest paper in our set of filtered literature was published in 2016. Figure 2 shows the number of publications by their publication year. The bar for 2023 is colored orange since the search phase was conducted in May 2023, so further reproducibility studies might have been published in the rest of 2023. The distribution of publications over time shows that the number of papers which target reproducibility concerns has been rising over the last years.

Another facet of publication trends are the venues where papers get submitted and presented.

The *European Conference on Information Retrieval* is the top venue for reproducibility studies. A majority of the information retrieval papers were published there (9/14). One reason for this high number of papers is a special reproducibility track with a dedicated call for papers³. This shows that dedicated conference formats can support the reproducibility effort to bring experts together in one place.

To summarize the insights for **RQ1**: Information retrieval is the top domain for reproducibility research due to the domain specific challenges and supported by dedicated conference tracks. Reproducibility studies are a trending topic when looking at the rising publication numbers.

For answering the remaining research questions, we focus on the nine papers of Table 1.

3.2. RQ2: Reasons for Reproducibility

Five of the nine papers from Table 1, focus upon the reproducibility assessment [35, 21, 22, 36, 34]. In the other cases, the main contribution lies on other aspects but all publications included implicit or explicit statements why the reproducibility study was conducted. The motivations are summarized in the following list which answers **RQ2**, *Why do authors try to reproduce the work of others*?:

- Creating awareness for reproducibility [35, 21, 22, 36, 34].
- The analyzed paper is important for the research domain but the artifacts of the original study were not available [21].
- Starting point for own research [20, 22].
- Follow-up research after introducing a tool or framework to show that it works correctly [23, 38, 36, 37].

As an additional finding it could be shown that some studies mix terms and approaches. As introduced, we distinguish based on the ACM terms reproducibility, repeatability and replicability. Three of the publications analyzed in the SLR had a replicability study as an additional part [20, 22, 37]. That is, the authors first conducted a reproducibility study and then reused the setup for another dataset. There is a trade-off between an experimental setup as close as possible to the original setup, a reasonable use of resources, and the re-usability for other research questions. Nevertheless, the distinction is important since the motivation and results are influenced by the combination of team members and experimental setup.

3.3. RQ3: Success Ratios and Challenges

The success ratios for the nine included papers reveal a mixed picture. Four studies were successful, three were partly successful and two were not successful at all. The two biggest challenges were missing artifacts and lack of documentation.

Studies were considered successful when the measured values were close to the originally reported ones [22, 23, 34, 37]. For instance, DADVAR and others stated that "the majority of the reproduced results were within the standard deviation of the reference results" [22, p. 6]. NUNZIO and others report that they "have quite comparable results in terms of processing time"

³https://ecir2023.org/calls/reproducibility.html?v=3.8

and the "final scores differ from the original values by almost 2 percentage points" [34, p. 34]. For three publications, partly successful in this context means that the authors investigated several primary studies where some reproductions failed [35, 38, 36]. The two remaining studies failed in their efforts [20, 21].

One challenge of the studies was the availability of artifacts like source code or raw data. For three of the publications considered in this SLR the necessary artifacts were available [22, 23, 34] and the reproducibility for these studies succeeded. Regarding the other six, only incomplete artifacts were available.

Another big challenge was insufficient documentation of experimental setups. NUNZIO and others, for example, identified unclarities regarding the environment of the primary study. The hardware was not sufficiently described in the original publication, especially concerning the CPU and GPU as both could influence the execution time. In the original paper, there were two contradictory statements about the used GPU model. Additionally, the original authors trained one of their models in the cloud because they needed a more powerful environment. In comparison, the authors of the reproducibility study needed the cloud environment for two models which prevented the comparison of the training time for the additional model.

4. Conclusion

Reproducibility is a key requirement for science and has to be considered when starting a new research project. In reproducibility studies, researchers other than the original authors use a setup as close as possible and reasonable to reproduce and validate the work of others. Starting from this, the objective of this paper was to give an overview about reproducibility studies in computer science.

The publication trends showed that the information retrieval domain is a role model for other domains, supporting its researchers with dedicated conference tracks. Based on the number of publications per year, we saw that the reproducibility topic gains traction. Additionally, we could confirm other meta studies that showed that a lack of raw data and artifacts as well as an incomplete or missing documentation are the most serious challenges for good research.

For future work, we propose a follow-up study on reproducibility which should include additional search engines and incorporate further SLR tasks like snowballing⁴. We also plan to incorporate the feedback from reviewers to include the term "executable paper" which was a hyped term around 2011 as an additional search term. In addition, we want to broaden our scope by incorporating similar terms like "analysis" and "survey" paired with "reproducibility", "replicability" and "repeatability" to also include publications which use another term but target "reproducibility studies" in the sense of this paper.

⁴Snowballing was already performed for this study and the raw results are already available at Zenodo. Due to time and space constraints, these snowballed publications were not assessed but published for follow-up research.

References

- [1] J. R. F. Cacho, K. Taghva, The state of reproducible research in computer science, in: Proc. of ITNG, 2020.
- [2] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature 533 (2016) 452-454.
- [3] H. E. Plesser, Reproducibility vs. replicability: A brief history of a confused terminology, Frontiers in Neuroinformatics 11 (2018).
- [4] C. Collberg, T. Proebsting, A. M. Warren, Repeatability and benefaction in computer systems research - a study and a modest proposal, University of Arizona Technical Report 14 (2015).
- [5] NISO, Reproducibility Badging and Definitions: A Recommended Practice of the National Information Standards Organization, Technical Report niso-rp-31-2021, National Information Standards Organization (NISO), 2021.
- [6] J. Kuhlenkamp, S. Werner, Benchmarking FaaS Platforms: Call for Community Participation, in: Proc. of WoSC, 2018.
- [7] T. Kalibera, R. Jones, Rigorous benchmarking in reasonable time, in: Proc. of ISMM, 2013.
- [8] J. L. Couture, et al., A funder-imposed data publication requirement seldom inspired data sharing, PLOS ONE 13 (2018).
- [9] J. Manner, SeMoDe simulation and benchmarking pipeline for function as a service, in: Bamberger Beiträge zur Wirtschaftsinformatik und Angewandten Informatik, 105, Otto-Friedrich-University, 2021.
- [10] J. Manner, A Simulation Framework for Function as a Service, University of Bamberg Press, 2024.
- [11] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Keele University and University of Durham Technical Report (2007).
- [12] S. Kolb, On the Portability of Applications in Platform as a Service, Ph.D. thesis, Bamberg, 2019.
- [13] J. Manner, A structured literature review approach to define serverless computing and function as a service, in: Proc. of CLOUD, 2023.
- [14] D. Taibi, et al., Patterns for serverless functions (function-as-a-service): A multivocal literature review, in: Proc. of CLOSER, 2020.
- [15] P. Brereton, et al., Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (2007) 571–583.
- [16] J. Scheuner, P. Leitner, Function-as-a-service performance evaluation: A multivocal literature review, Journal of Systems and Software 170 (2020) 110708.
- [17] V. Yussupov, et al., A systematic mapping study on engineering function-as-a-service platforms and tools, in: Proc. of UCC, 2019.
- [18] T. Hummel, A Literature Review on Reproducibility Studies in Computer Science: Supporting Material, 2024.
- [19] R. F. G. Silva, K. Paixao, M. de Almeida Maia, Duplicate question detection in stack overflow: A reproducibility study, in: Proc. of SANER, 2018.
- [20] S. Althammer, S. Hofstätter, A. Hanbury, Cross-domain retrieval in the legal and patent domains: A reproducibility study, in: Proc. of ECIR, 2021.
- [21] S. Cross, G. Zuccon, A. Mourad, A reproducibility study of question retrieval for clarifying

questions, in: Proc. of ECIR, 2023.

- [22] M. Dadvar, K. Eckert, Cyberbullying detection in social networks using deep learning based models; A reproducibility study, CoRR abs/1812.08046 (2018).
- [23] E. K. F. Dang, R. W. P. Luk, J. Allan, A comparison between term-independence retrieval models for ad hoc retrieval, ACM Transactions on Information Systems 40 (2021) 1–37.
- [24] M. Färber, T. Klein, J. Sigloch, Neural citation recommendation: A reproducibility study, in: Proc. of BIR, 2020.
- [25] M. Hendriksen, et al., Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study, in: Proc. of ECIR, 2023.
- [26] J. Huang, et al., State encoders in reinforcement learning for recommendation: A reproducibility study, in: Proc. of SIGIR, 2022.
- [27] W. Lajewska, K. Balog, From baseline to top performer: A reproducibility study of approaches at the TREC 2021 conversational assistance track, in: Proc. of ECIR, 2023.
- [28] H. Li, et al., Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study, in: Proc. of ECIR, 2022.
- [29] J. Mackenzie, et al., Compressing inverted indexes with recursive graph bisection: A reproducibility study, in: Proc. of ECIR, 2019.
- [30] R. Sequiera, L. Tan, Y. Zhang, J. Lin, Update delivery mechanisms for prospective information needs: A reproducibility study, in: Proc. of CHIIR, 2020.
- [31] G. Silvello, et al., Statistical stemmers: A reproducibility study, in: Proc. of ECIR, 2018.
- [32] S. Wang, S. Zhuang, G. Zuccon, Federated online learning to rank with evolution strategies: A reproducibility study, in: Proc. of ECIR, 2021.
- [33] P. Yang, H. Fang, A reproducibility study of information retrieval models, in: Proc. of ICTIR, 2016.
- [34] G. M. D. Nunzio, R. Minzoni, A thorough reproducibility study on sentiment classification: Methodology, experimental setting, results, Information 14 (2023) 76.
- [35] M. Arvan, L. Pina, N. Parde, Reproducibility in computational linguistics: Is source code enough?, in: Proc. of EMNLP, 2022.
- [36] F. Hoffeins, F. M. Ciorba, I. Banicescu, Examining the reproducibility of using dynamic loop scheduling techniques in scientific applications, in: Proc. of IPDPS, 2017.
- [37] K. Wolsing, et al., Ipal: Breaking up silos of protocol-dependent and domain-specific industrial intrusion detection systems, in: Proc. of RAID, 2022.
- [38] J. Haltermann, H. Wehrheim, Machine learning based invariant generation: A framework and reproducibility study, in: Proc. of ICST, 2022.