2018 IEEE 34th International Conference on Data Engineering

Artifact Evaluation: *FAD* or *Real* News?

Bruce R. Childers
Department of Computer Science
School of Computing and Information
University of Pittsburgh, Pittsburgh PA 15260
Email: childers@cs.pitt.edu

Panos K. Chrysanthis
Department of Computer Science
School of Computing and Information
University of Pittsburgh, Pittsburgh PA 15260
Email: panos@cs.pitt.edu

***Abstract*—Data Management (DM), like many areas of computer science (CS), relies on empirical evaluation that uses software, data sets and benchmarks to evaluate new ideas and compare with past innovation. Despite the importance of these artifacts and associated information about experimental evaluations, few researchers make these available in a findable, accessible, interoperable and reusable (FAIR) manner, in this way hindering the scientific process by limiting open collaboration, credibility of published outcomes, and research progress. Fortunately, this problem is recognized and many CS communities, including the DM one, are advocating and providing incentives for software and analysis papers to follow FAIR principles and be treated equally to traditional publications. Some ACM/IEEE conferences adopted Artifact Evaluation (AE) to reward authors for doing a great job in conducting experiments with FAIR software and data. After half a decade since AE's inception, the question is whether the emerging emphasis on artifacts, is having a real impact in CS research.**

I. Introduction

It is an accepted fact that we learn hard lessons when implementing and re-evaluating systems, yet it is also acknowledged that science faces a *crisis in reproducibility* [3]. Data Management research (DM), like many areas of experimental computer science (CS) [5], is far from immune to this crisis, although it should be easier for DM than other sciences, given the emphasis on encapsulating experimental artifacts, such as source code, data sets, workflows, configuration parameters, etc. Despite the importance of these artifacts and associated information about experimental evaluations, few researchers make these available in a *findable, accessible, interoperable* and *reusable* (FAIR) manner, in this way hindering the scientific process by limiting open collaboration, credibility of published outcomes, and research progress.

Fortunately, there is growing recognition of this challenge in CS. DM is one of the first to consider the issue of reproducibility seriously. Early on at VLDB 2007, the panel on "Performance Evaluation and Experimental Assessment" debated the challenges in adopting reproducibility as a review criterion and promoted the idea that software, experiments and analysis papers should be valued equally to papers offering new solutions. The following year, ACM SIGMOD proposed for the first time to test the code of submitted papers in 2008. The initial effort was on *repeatability*, testing the code associated with conference submissions against the data sets used by the authors. In 2010, repeatability was expanded to include *workability*, running different/more experiments with different or more parameters than shown in the respective papers. The 2008–2012 repeatability and reproducibility efforts were renewed in 2016, creating the SIGMOD Reproducibility committee to advocate that software and analysis papers should follow FAIR principles and be treated equally to traditional publications. To incentivize authors, it established the *ACM SIGMOD Most Reproducible Paper Award*.

Many other ACM and IEEE conferences considered similar ideas, perhaps inspired by these early efforts of DM, and over the last half dozen years adopted *Artifact Evaluation* [6] (AE, artifact-eval.org). AE is an optional, post-acceptance process, which is conducted independently of paper review and by an artifact evaluation committee (AEC) separate from the program committee. Upon paper acceptance, authors are given the option of providing their artifacts to the AEC, which checks that results/conclusions of a paper are *consistent* with the artifacts. A paper that passes is rewarded with a badge which is put in the paper's PDF and the ACM Digital Library to distinguish it. ACM has adopted a set of badges for different award degrees [2]—ACM SIGMOD adopted two ACM badges, *Results Replicated* and *Artifacts Available*.

The question is whether the emerging emphasis on artifacts, in particular AE, provides a *real* incentive in computer systems research, or whether this is just another *fad*. This paper attempts to answer this question. More specifically, we aim at providing the scientific research community with our preliminary insights [4] and seek its assistance in furthering and broadening our study.

II. Our Study

To answer whether AE is an incentive for authors to provide their software artifacts, we used *level of participation* and *citation count* as proxy metrics. Levels of participation can be defined in several ways, e.g., to align with ACM badges. We adopted the simplest way to define participation, which is to count AE papers and authors without distinguishing award degrees. For a conference that adopts AE, the citation count metric sets up an experimental group of papers that successfully went through AE and a control group that either were not successful or did not participate. Citation count also permits us to answer our question quantitatively, albeit indirectly. These metrics are imprecise with many sources of bias, but they can show trends and indicate, at least in a preliminary way, whether AE can promote improved practices.

2375-026X/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDE.2018.00204

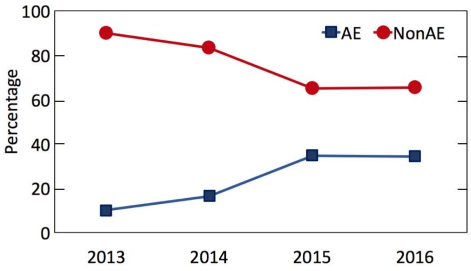

Fig. 1. Percentage of AE and non-AE papers.

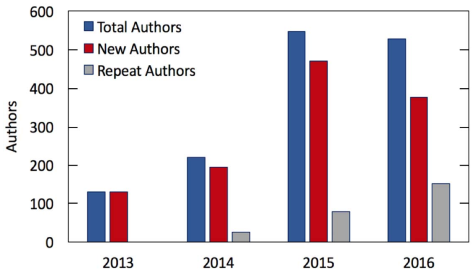

Fig. 2. Number of New and Repeated AE Authors

We did our best to identify all conferences that use the AE process, and we went through the proceedings of these conferences to find papers that successfully went through AE. We used Google Scholar to get citation counts for the papers[1].

III. The Analysis

In our analysis, we considered conferences in 2013-2016 and did not include any in 2017 since we are still collecting the data for 2017. Figure 1 shows an upward trend of papers that passed AE. Figure 2 illustrates that this increased number of AE papers is due to an increased number of authors who participated in AE for the first time. In fact, the number of first time authors is substantially higher compared to that of repeated ones. An interesting observation on repeated authors is that the Top-15 AE authors, corresponding to 5% of the total number of repeated authors, had 5-8 AE papers.

Figure 3 shows the summary data for the average number of citations per paper for AE and non-AE papers. It is clear that the citation counts for AE papers is higher on average than the control group. Table I shows citation counts on a per year basis for three conferences that regularly use AE. In these conferences, there is a trend that AE papers receive more citations. Interestingly, the table shows that older AE papers tend to collect even more citations than non-AE papers of equivalent age. For instance, in 2013, the AE papers for ECOOP had an average of 22.25 citations per paper, while the non-AE papers had 15.67. PLDI had a big spike for 2014 (60.83 vs. 26.54). However, this spike is due to one particularly influential paper in 2014 that successfully went through AE [1].

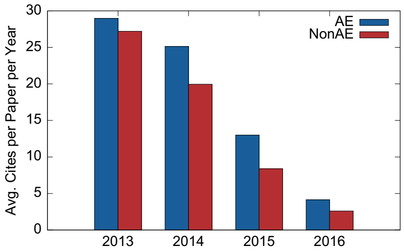

Fig. 3. Average citation counts of AE and non-AE papers.

TABLE I
Average citations per AE and non-AE per year.

Year	ECOOP		OOPSLA		PLDI	
	AE	Non-AE	AE	Non-AE	AE	Non-AE
2013	22.25	15.67	22.50	28.06	N/A	N/A
2014	11.67	11.44	13.35	12.86	60.83	26.54
2015	7.92	5.47	7.56	7.52	15.04	11.97
2016	2.50	1.00	1.00	1.34	4.55	4.33

IV. The Answer

Figures 1–3 illustrate that AE does seem to have an effect on reproducibility. The number of first time authors engaged in AE is increasing and the citation counts for AE papers is higher on average than for the non-AE ones. It is very important to note: There may not be a direct causation e.g., perhaps authors that participated in AE have a tendency to be more active and visible in the community, already have a history and a desire to release software, or have a history of producing high quality and innovative outcomes, etc. These biases may lead to the higher citation counts; deeper study will be necessary to understand and correct for possible bias.

Although we cannot draw a cause-and-effect conclusion yet, there is quantitative evidence that the AE badges influence participation in the AE process. Furthermore, there is quantitative evidence that AE is at least correlated with citation count. There are also qualitative indicators that AE and similar processes are creating incentives for the development of artifacts in FAIR manner. In conclusion, our current results suggest that AE is one potentially powerful incentive toward producing better software and other artifacts!

References

[1] S. Arzt, et al. D. Octeau, and P. McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. In *PLDI '14*, pp. 259–269, 2014.

[2] Result and artifact review and badging. http://www.acm.org/publications/policies/artifact-review-badging, 2016. [Online; accessed 18-Aug-2016].

[3] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, May 2016.

[4] B. R. Childers and P. K. Chrysanthis. Artifact evaluation: Is it a real incentive? In *eScience WSSSPE5.2*, pp. 488–489, 2017.

[5] C. Collberg and T. A. Proebsting. Repeatability in computer systems research. *Commun. ACM*, 59(3):62–69, Feb. 2016.

[6] S. Krishnamurthi and J. Vitek. The real software crisis: Repeatability as a core value. *Commun. ACM*, 58(3):34–36, Feb. 2015.

[1] This data was collected in Spring 2017 & 2018 by Anuradha Kulkarni, Seth M. Stayer and Christopher Corsi as part of their degree projects.

This material is based upon work supported by the National Science Foundation under grants ACI-1535232, CNS-1305220 and CBET-1609120.