ARTICLE

SIGIR Initiative to Implement ACM Artifact Review and Badging

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Diane Kelly
University of Tennessee, USA
dianek@utk.edu

Abstract

Recently, the ACM created a policy on Artifact Review and Badging, which presents a framework to help SIGs recognize repeatability, replicability and reproducibility in published research. While the ACM policy established a vocabulary and definitions, it did not prescribe procedures for implementation. Rather, the ACM has left this to each SIG to define given the variety of research traditions and approaches that exist with the ACM community. SIGs are not required to implement badging, but given the growing interest in this topic in the SIGIR community, a task force has been assembled to determine how badging might be implemented. This report describes the ACM policy on Artifact Review and Badging, introduces the task force, and presents survey data describing the SIGIR community's opinions about this initiative.

1 Replicability and Reproducibility of Experimental Results

Replicability and *reproducibility* are becoming a primary concern in many areas of science [4, 6] and, in particular, in computer science as also witnessed by the recent ACM policy on *Artifact Review and Badging*[1].

IR is a discipline strongly rooted in experimentation and replicability and reproducibility of the experimental results are becoming a more central discussion item in the research community [1–3, 5, 7]. We now commonly ask questions about the extent of reproducibility of the reported experiments in the review forms of all the major IR conferences, such as SIGIR, CHIIR, ICTIR and ECIR, as well as journals, such as ACM TOIS. We have also witnessed the rise of new activities aimed at verifying the reproducibility of results: for example, the "Reproducibility Track" at ECIR since 2015 hosts papers which replicate, reproduce and/or generalize previous research results while CLEF/NTCIR/TREC

[1] https://www.acm.org/publications/policies/artifact-review-badging

REproducibility[2] (CENTRE) is a new joint evaluation activity, started in 2018, to assess and quantify the extent of replicability and reproducibility of our experimental results.

For all these reasons, SIGIR has decided to explore the adoption and implementation of the ACM's policy on Artifact Review and Badging. The ACM does not specify procedures to assign badges to published research, but rather leaves this up to each individual SIG. To this end, SIGIR has setup a task force dedicated to define how the ACM policies should be implemented in the SIGIR community and has conducted a survey among its members to hear their opinion and suggestions about artifact review and badging.

1.1 The ACM Policy on Artifact Review and Badging

We summarize here the main aspects of the ACM policy which are relevant to the SIGIR community. Our summary draws heavily on the ACM policy, which can be found at: https://www.acm.org/publications/policies/artifact-review-badging. We present this summary here as a convenience for the reader and do not claim any originality of the text in the section.

1.1.1 Terminology

Repeatability (Same team, same experimental setup). The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

Replicability (Different team, same experimental setup). The measurement 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 original authors own artifacts.

Reproducibility (Different team, different experimental setup). The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently of the original work.

1.1.2 Badges

There are three main types of badges – evaluated, available, and validated – which are described in the following.

Artifacts Evaluated. This badge is applied to papers whose associated artifacts have successfully completed an independent audit. Artifacts need not be made publicly available to be considered for this badge. However, they do need to be made available to reviewers.

[2] http://www.centre-eval.org/

Artifacts Evaluated – Functional. The artifacts associated with the research are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.

Artifacts Evaluated – Reusable. The artifacts associated with the paper are of a quality that significantly exceeds minimal functionality. That is, they have all the qualities of the *Artifacts Evaluated – Functional level*, but, in addition, they are very carefully documented and well-structured to the extent that reuse and repurposing is facilitated. In particular, norms and standards of the research community for artifacts of this type are strictly adhered to.

Artifacts Available. This badge is applied to papers in which associated artifacts have been made permanently available for retrieval. Author-created artifacts relevant to the paper have been placed on a publicly accessible archival repository. A DOI or link to an artifact repository along with a unique identifier for the object is provided. Repositories used to archive data should have a declared plan to enable permanent accessibility. Personal web pages are not acceptable for this purpose. Artifacts do not need to have been formally evaluated in order for an article to receive this badge.

Results Validated. This badge is applied to papers in which the main results of the paper have been successfully obtained by a person or team other than the author. Exact replication or reproduction of results is not required, or even expected. In particular, differences in the results should not change the main claims made in the paper.

Results Replicated. The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.

Results Reproduced. The main results of the paper have been independently obtained in a subsequent study by a person or team other than the authors, without the use of author supplied artifacts.

1.2 The SIGIR Badging Task Force

The task force has been put together considering a mix of expertise and background which cover both the academic and industrial side of IR, as well as competencies in interactive IR, user-oriented evaluation, IR systems development and algorithmic approaches, infrastructures, and system-oriented evaluation.

The members of the task force are as follow:

- Nicola Ferro (chair), University of Padua, Italy
- Diane Kelly (ex-officio, SIGIR chair), University of Tennessee, USA
- Maarten de Rijke (ex-officio, ACM TOIS EiC), University of Amsterdam, The Netherlands
- Leif Azzopardi, University of Strathclyde, UK
- Peter Bailey, Microsoft, USA
- Hannah Bast, University of Freiburg, Germany
- Rob Capra, University of North Carolina at Chapel Hill, USA
- Norbert Fuhr, University of Duisburg-Essen, Germany
- Yiqun Liu, Tsinghua University, China
- Martin Potthast, Leipzig University, Germany
- Filip Radlinski, Google London, UK
- Tetsuya Sakai, Waseda University, Japan

- Ian Soboroff, National Institute of Standards and Technology (NIST), USA
- Arjen de Vries, Radboud University, The Netherlands

2 The Survey: SIGIR Community Opinion

The survey described in the preceding SIGIR Forum article by Kelly included three questions asking for opinions on different aspects of artifact badging. For some of these questions, there was an additional open-ended question asking why that score was assigned.

In the following sections, we report the distribution of the answers for each of the questions plus a summary of the comments provided in response to the open-ended questions, helpful to give a feeling about the motivations behind the main trends we observe.

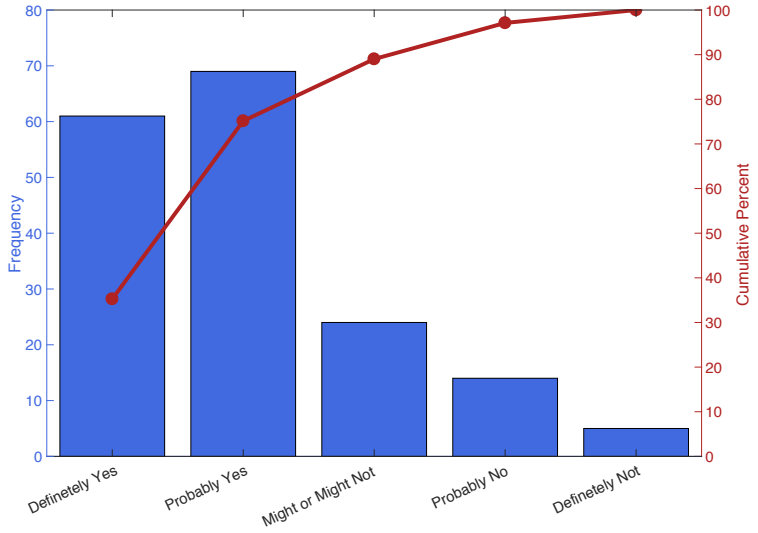

Figure 1: Do you support the implementation of procedures to assign ACM Artifact Badges to papers published at SIGIR, CHIIR and ICTIR Conferences, as well as the journal, ACM Transactions on Information Systems?

2.1.1 Q1: What about Introducing Badges?

Figure 1 shows the distribution of the answers for the first question, asking the opinion about the introduction of a badging policy for SIGIR conferences and journals. The introduction of a badging policy is seen in a positive or very positive way by about 75% of the respondents and this grows up to roughly 90% if we consider people with a neutral opinion; only about 10% of respondents have a negative or very negative position with respect to introducing artifact badging.

The main positive motivations for introducing artifact badging, among the neutral, positive and very positive respondents, were:

- badging and reproducibility lead to better science;
- production of better baselines and less effort to compare against them, since their reproducibility is guaranteed;
- overall improvement of the IR community and its experimental practices.

However, the neutral and positive respondents also raised some concerns about artifact badging, which were also the main negative motivations for the respondents with negative and very negative opinions:

- it could represent an issue for industry and for proprietary/confidential/sensitive data;
- it is a not negligible additional effort for authors and reviewers;
- partially linked to the above item, the pressure and the need for publish-or-perish is a potential barrier to reproducibility;
- the logistics needed to implement it is not trivial;
- possible issues of trust in the actual ability/willingness of reviewers in reproducing the results;
- risk of introducing class A and class B papers, leading of a decreased impact of not badged papers;
- a possible slow-down in the publication and diffusion of new ideas, due to the additional burden needed to ensure reproducibility;
- the problem that not all the kinds of results are replicable and/or reproducible, e.g. those based on ephemeral data;
- it may represent an high-barrier for smaller or junior groups which may not have all the expertise and/or resources to implement it.

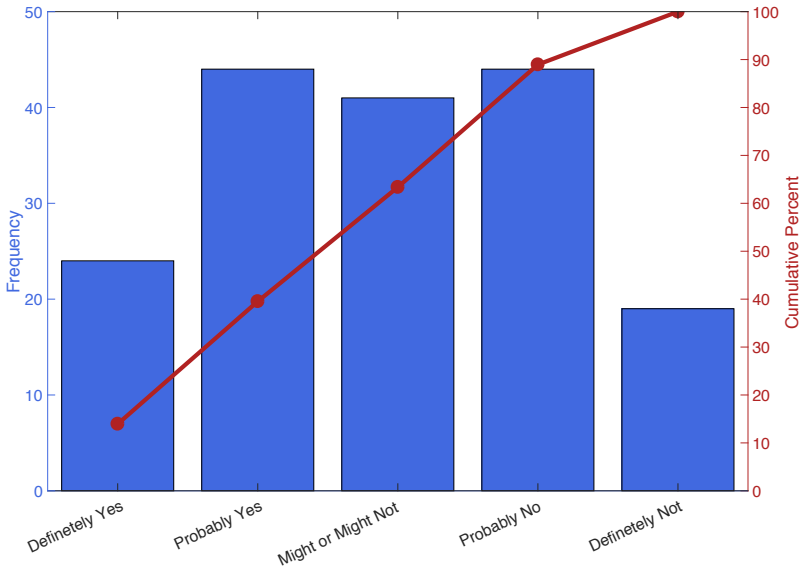

Figure 2: Would the potential to earn an ACM Artifact Badge change how you conduct and report research?

2.1.2 Q2: Would Badges Change Your Research?

Figure 2 shows the distribution of the answers for the second question, asking if badging would change how respondents conducted and reported research. About 40% of respondents indicated that badging would definitely or probably change their way of doing and reporting research. The main motivations for this opinion were that it would lead them to do better science, putting more care and effort in their work; also the possibility of producing more open data sets was considered an added value.

Around 20% of respondents thought that badging might or might not change their work practices while about 40% indicated that it would probably or definitely not change their way of doing research. One

interesting aspect is in the majority of these cases, respondents felt that they already conduct and report research in a way that makes it reproducible. This optimistic view might be a bit biased and contrasts with anecdotal experiences on the difficulty of reproducing previous research results. The other main reasons given for badging not changing practice were that badging does not represent a real incentive, e.g., in terms of career advancement, that reproducibility and badging could be an issue for industry, and possible lack of trust in the badging process.

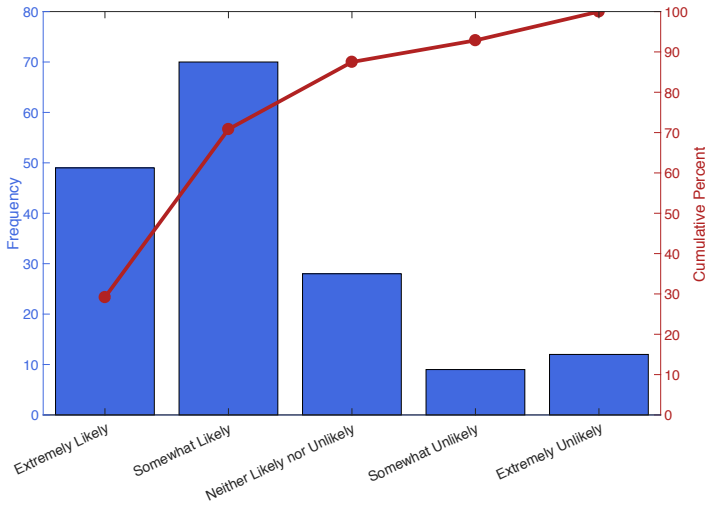

Figure 3: How likely would you be to request that your paper be considered for an ACM Artifact Badge?

2.1.3 Q3: Would Badges Change Your Research?

Figure 3 shows the distribution of responses to the third question, which asked how likely respondents would be to request their work be considered for a badge. More than 70% of respondents say that they would likely or very likely ask for their research to be badged and this ratio grows up to almost 90% if you also consider those who have a neutral attitude; only roughly 10% of the respondents said that they would be unlikely or very unlikely to request badges.

3 Conclusion

For many years, discussions regarding the repeatability, replicability and reproducibility of IR research have been common in the community. These include both informal discussions amongst community members and research teams, as well as formal discussions held at workshops, symposia and conferences. These discussion have led to the creation of special tracks at conferences (e.g., ECIR), adding questions to peer-review forms, and recent initiatives such as CENTRE.

The ACM policy on Artifact Review and Badging provides a framework for SIGs to develop procedures to recognize research that makes available functional and reusable artifacts, as well as research results that can be validated, replicated or reproduced. This article described the policy, introduced a SIGIR task force who will determine how this policy might be implemented, and presented results describing the community's opinions about the initiative.

One of the major goals of pursuing an implementation plan for ACM badges is to recognize efforts that make research more repeatable, replicable and reproducible. The goal is not to create different classes of papers. It is understood that there are many valid reasons why it is not possible to produce or share the artifacts needed to lead to a badge. It is also understood that it is not always possible to reproduce results even if they are valid, or were valid at the time they were initially produced. The hope is that badges will provide a way to formally recognize an important aspect of scientific research, which is currently not recognized, and improve research practice and the quality of our results.

4 References

[1] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2):107–116, December 2015.

[2] N. Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2):8:1–8:4, February 2017.

[3] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June 2016.

[4] J. Freire, N. Fuhr, and A. Rauber, editors. Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science, *Dagstuhl Reports*, Volume 6, Number 1, 2016. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.

[5] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, editors, *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, pages 357–368. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany, 2016.

[6] M. R. Munafò, BB. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021:1–0021:9, January 2017.

[7] J. Zobel, W. Webber, M. Sanderson, and A. Moffat. Principles for Robust Evaluation Infrastructure. In M. Agosti, N. Ferro, and C. Thanos, editors, *Proc. Workshop on Data Infrastructures for Supporting Information Retrieval Evaluation (DESIRE 2011)*, pages 3–6. ACM Press, New York, USA, 2011.