Review Data Collection at *ACL

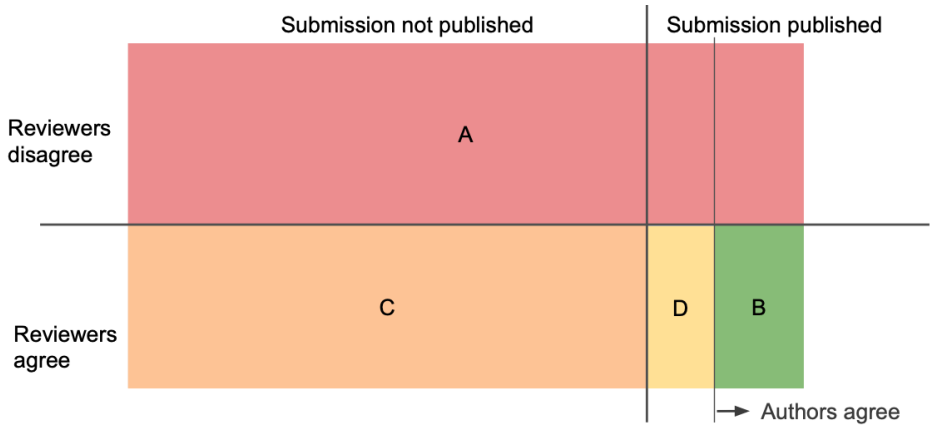
March 2022 Update [last edited 16.03.2022]

Nils Dycke, Ilia Kuznetsov, Iryna Gurevych

Recap: Computational study of peer review is an emerging area of NLP research that suffers from the lack of clearly licensed, open datasets. In March 2021 ACL Exec approved the initiative by UKP Lab (Iryna Gurevych, Ilia Kuznetsov) to collect peer reviewing data from *ACL conferences and create open, clearly licensed and ethically sound datasets of peer reviews in the NLP domain. The initiative was prepared in collaboration with the ACL Reviewing Committee, community stakeholders, as well as legal counseling, and is detailed here. The initiative grants UKP Lab permission to design workflows, collect, preprocess and clean peer reviewing data from *ACL conferences in coordination with their respective program chairs. The collected data is returned to the community in form of public datasets. The following requirements are agreed upon: the data is anonymized (no names, no unique identifiers, etc.), both authors and reviewers need to give explicit informed consent for data collection, and transfer license for their data to ACL (similar to ACL Anthology); only data for accepted publications is made public; yet, data for rejected publications can be archived for internal research and optionally published after a substantial time period. The initiative covered two separate data collection workflows: metadata (scores, etc.), and full peer review collection (incl. texts).

Status: As ACL moved to the Rolling Review (ARR) system, the original proposal required adjustments. In collaboration with ARR editors-in-chief, we have adapted our workflow to the ARR publishing lifecycle. With the OpenReview technical team, we have automated many data-related routines, minimizing the effort for consent collection and data extraction while ensuring safety of the confidential data. The code is publicly available and open for community inspection, the operational workflow at ARR is detailed in a preprint (subject to updates). The license and consent collection workflow is now constantly running at ARR, with ~50% consent rate among reviewers over the past five months. After the publication decisions for ACL-2022 are known, the first batch of public data (authors and reviewers agree, submission accepted for publication) will be ready for release. Substantial protected data has been accumulated as well, however the conditions of its use are underspecified in the original proposal. Metadata consent collection has not brought desired results, and alternative tools were developed independently by the NLP community.

Open questions: Multiple positions in the original proposal need to be revisited and refined based on our experience so far and on the community feedback. While the original proposal also involved consent-based *metadata* collection from peer reviews, due to the recent developments in the community we propose to focus the future efforts on collecting full peer review data only (which includes metadata and peer review texts). While the original proposal suggests making protected data available for internal research, it leaves the conditions undefined; practical aspects of handling protected data require careful consideration. Finally, it is necessary to formalise future work with the data by outlining and agreeing upon the responsibilities and permissions of both ACL and UKP Lab (or another future data controller) with regard to data management and data publication.



Data summary: below we use "metadata" to refer to numerical scores, track, etc., and "full data" to refer to metadata plus peer review texts, submission drafts, etc. No unique identifiers or reviewer names are ever stored in either case. We focus on collecting the full data. Current process implemented at ARR separates the data into four "buckets". If reviewers disagree, the data is discarded from collection (A). If reviewers agree, and submission is accepted for publication, and authors agree, the data is added to the public dataset (B). The remaining data is protected due to maintain confidentiality (C) and give the authors agency over the publication of their peer reviews (D).

Q1: Protected data. As the infographic shows, consent-driven collection of peer reviewing data can substantially reduce the dataset size and introduce bias. While dataset size is not an issue in the long term as more peer review data is collected, the bias of the public data (B) compared to overall data (A+B+C+D) is likely. It is yet unclear whether this bias is substantial or relevant for NLP applications. To study this bias, the original proposal introduced a protected dataset (C+D) which contains confidential data and could be released after a substantial time period or made available for research purposes through a protected environment (TIRA) or limited, strictly controlled access authorized by ACL Exec on a case by case basis. This, however, introduces high administrative overhead and might lead to legal challenges. Besides, as the infographic demonstrates, even protected data is not exempt from bias. We thereby propose to 1) nor collect neither use protected *full* data (C+D) for research at this stage; 2) instead, extract and report aggregated *numerical and metadata statistics* from B, C, D *as well as A* to highlight the differences between the public dataset (B) and the overall distribution. If systematic bias is encountered, measures can be taken to enable studies on protected data at a later point.

Q2: Public data. As per the current license agreement, the data to be released publicly (B) includes "text, review form scores and metadata, charts, graphics, spreadsheets, and any other materials prepared by Peer Reviewer in connection with the peer review process", as well as the blind submission versions of the publications (see preprint, Appendix A). To facilitate development of derivative datasets, authors transfer a CC-BY-NC-SA license for their data to the ACL; unlike consent, license can not be easily revoked, ensuring stability of the data and thereby replicability. To reduce overhead, we propose releasing data in batches, e.g. once a year, and to only publish data for the work that has been published **and** presented by the time of the data release. Two open questions with respect to public data release are license management and data archival. **License management** entails keeping

the record of license agreements digitally signed by data contributors and dealing with the unlikely cases of license violation or data withdrawal requests. Importantly, license management includes license versioning, i.e. keeping track of the changes to the license over time. **Data archival** entails storing public copy of the dataset and providing either anonymous or personalized access to it. While anonymous access (anyone on the Internet) is easier to implement, personalized access allows tracking dataset use and license compliance, as well notifying researchers in cases the dataset composition has changed.

Two available options for license management and data archival are **UKP on-site** and via **ACL Anthology**. The collected peer reviewing data is licensed to ACL, and ACL Anthology would be a natural choice for hosting this data; yet, this results in a communication overhead, and ACL Anthology currently does not allow tracking access to data. UKP Lab has capacity for data archival and controlled access via a specialized Europe-based TU Darmstadt university library server, e.g. ACL-2018 numerical data repository. Yet, UKP Lab is *not the licensee* of the collected data and should *not* store the license agreements, as they are personalized and can deanonymize the data. We propose to publish the first batch of public data via TU Darmstadt university library and devise the long-term data publishing and licensing strategy at a later point, with the end goal of transferring this responsibility to ACL. Yet, since the license is transferred to ACL, we deem it necessary to store the associated licenses on the ACL side from the beginning. The signed licenses constitute personal and sensitive data as they allow de-anonymizing the public dataset entries, which might be necessary in the *exceptional* case of data withdrawal request. The confidentiality of this data must be ensured correspondingly at the ACL side.

In summary, the next step for peer reviewing data collection at *ACL is the data release and subsequent maintenance. To enable it, we request ACL's approval for the following:

1. UKP Lab stops active efforts on metadata-only consent collection from ARR;
2. Instead, UKP Lab gets access to aggregated meta-statistics on different subsets of collected data, including the previously ignored non-consented metadata (A), which is already used in a range of community based efforts, see example. The *aggregated statistics* for A, B, C and D are published alongside the full review data to reflect bias of the published data compared to the overall distribution.
3. Publication and release plans for protected data (C, D) are halted until further notice. No research is performed with protected data apart from extracting numerical aggregate statistics. Current license agreement allows the use of protected data for internal research by ACL. We propose omitting this clause for the time being, as what constitutes internal research and the rules of data access is not defined at this moment.
4. UKP Lab takes responsibility for releasing the first batches of public data via its University server, with a long-term plan to transfer this responsibility to ACL Anthology. Yet, since license transfer implies dealing with reviewer identities, the licenses are transferred to ACL Anthology management from the beginning and stored securely with strictly regulated access only in exceptional cases (legal, data withdrawal).