ReproDB: An Open Platform for Discovering Research Artifacts and Analyzing their Evaluation in Security and Systems

Anjo Vahldiek-Oberwagner
Intel
Berlin, Germany
anjovahldiek@gmail.com

Marton Bognar
DistriNet, KU Leuven
Leuven, Belgium
marton.bognar@kuleuven.be

Salvatore Signorello
NOVA University Lisbon
Lisbon, Portugal
s.signorello@fct.unl.pt

Abstract

Artifact evaluation (AE) has become a cornerstone of reproducible research in computer science, with dozens of security and systems conferences running formal AE processes. Unfortunately, the resulting data remains fragmented, as outcomes are scattered across venue-specific, inconsistent websites, lacking machine-readable metadata and cross-venue aggregation. This prevents the community from answering fundamental questions about AE adoption trends, institutional participation, evaluator workload and retention, and AE sustainability at scale. It also breaks the creation–evaluation–reuse loop: no search by, e.g., topic, exists across security and systems artifacts, limiting artifact discovery and reuse.

We present ReproDB, an open-source automated pipeline that scrapes AE results, committees, author metadata, and repository statistics from 13 security and systems conferences and homogenizes them into a unified dataset. Building on this foundation, the platform enables three capabilities: (1) the first cross-venue analysis of AE (2017-2026), revealing AE health challenges and open-science policies as the strongest lever for participation; (2) a combined metric that captures both artifact creation and AE committee service, visualizing reproducibility labor; and (3) the first cross-venue artifact search engine, closing the creation–evaluation–reuse loop by enabling discovery by topic, author, institution, or venue.

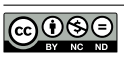
CCS Concepts

• **General and reference** → **Empirical studies**; **Measurement**.

Keywords

Artifact Evaluation, Reproducibility, Open Science, Systems Research, Security Research

ACM Reference Format:
Anjo Vahldiek-Oberwagner, Marton Bognar, and Salvatore Signorello. 2026. ReproDB: An Open Platform for Discovering Research Artifacts and Analyzing their Evaluation in Security and Systems. In *ACM Conference on Reproducibility and Replicability (ACM REP '26), July 20–22, 2026, Delft, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3820002.3828588



ACM REP '26, Delft, Netherlands

ACM ISBN 979-8-4007-2778-8/2026/07
https://doi.org/10.1145/3820002.3828588

1 Introduction

Artifact evaluation (AE) is a core mechanism for improving reproducibility in computer science [2, 19]. Since its adoption by security and systems conferences, the practice has expanded to encompass numerous conferences releasing thousands of artifacts reviewed by with hundreds of dedicated evaluators each year. Experience reports show that AE improves artifact quality and availability [10, 16, 17, 20, 25], but these studies only cover one community. Moreover, AE data remains siloed, making it difficult to compare AE practices across venues or discover artifacts beyond a single conference's portal. Community portals such as *sysartifacts/secartifacts.github.io* [24, 27] collect per-venue results but provide no unified schema or programmatic access. This state of the art leaves three open challenges:

Homogenization. AE outcomes are siloed in per-venue websites with heterogeneous formats and metadata, preventing aggregation.

Analysis. It is currently difficult to study AE adoption trends, geographical and institutional participation, evaluator workload and retention, or the sustainability of AE infrastructure as it scales.

Discoverability. Neither conference portals nor digital libraries track artifact repository locations, and searching Zenodo or GitHub by paper title yields incomplete results. Olszewski et al. [17] and Vansteenhuyse et al. [29] report that many security papers do not provide an associated artifact link, even when one exists. Locating the relevant artifact is often difficult, breaking the creation–evaluation–reuse loop.

We present ReproDB, an open-source platform that addresses these three challenges. Its automated pipeline scrapes, enriches, and homogenizes 2,831 evaluated artifacts from 13 conferences into a unified dataset covering 8,139 authors, 2,458 evaluators, and 902 institutions (§5). Unlike one-shot measurement studies, the pipeline is designed for recurring execution as new conferences are added. The platform enables three complementary capabilities (§3):

Metrics. We introduce a combined scoring methodology for authors and institutions, capturing artifact creation and AE committee service, making the full reproducibility labor spectrum visible (§4).

Cross-Venue Analysis. Using our dataset, we analyze 2,831 artifacts across 13 conferences, providing the first cross-community analysis of AE. Our analysis uncovers historical differences between communities, ecosystem roles as artifact creators and evaluators, repository engagement patterns, AE committee health (retention and workload), the impact of open-science policy on AE participation, and a negative result on citing artifacts directly (§6).

Artifact Search. We build the first cross-venue artifact search engine, allowing researchers to query the entire corpus of 2,831 artifacts by title, keywords, authors, institutions, venue, year, and research area. This closes the creation–evaluation–reuse loop by

enabling discovery across community boundaries, surfacing otherwise hidden connections between research areas (§7).

To the best of our knowledge, no prior effort has combined cross-venue AE aggregation and ecosystem analysis at this scale, with artifact discovery in a single open platform. This paper contributes the ReproDB platform and its automated pipeline, cross-community empirical findings, and cross-community artifact search engine.

2 Background and Motivation

ACM introduced an artifact review and badging policy in 2016 [1], defining several levels of recognition. The most common levels in the security and systems community are Artifacts Available (publicly accessible), Artifacts Evaluated—Functional (documented and exercisable), and Results Reproduced (key results independently verified). The adoption of these badges by top systems and security venues has transformed expectations around research artifacts, making source code availability and functional verification increasingly standard. Most recently, USENIX Security'25 introduced a mandatory open-science policy requiring authors to share research artifacts by default [28], signaling a shift from voluntary participation toward policy-driven artifact sharing requirements.

These badge tiers partially overlap with the FAIR principles for research software [3], adapted from the original data-stewardship framework [30]: Available corresponds to Findable and Accessible, while Functional touches on aspects of Interoperable. Finally, the Reproduced badge relates to the reusability principles.

Community-driven efforts such as sysartifacts and secartifacts have played a pivotal role by serving as persistent, public records of artifact evaluation outcomes. Their structured per-conference, per-year archives have enabled retrospective studies such as D'Elia et al.'s five-year analysis of EuroSys artifact evaluations [10], served as baseline data for Vansteenhuyse et al.'s security artifact availability study [29], and given AE chairs shared infrastructure for publishing results consistently across all tracked venues.

Despite these advances, most studies consider only one or two venues and examine only artifact health or artifact evaluation [10, 12, 17, 18]. The field lacks infrastructure to track whether proposed interventions [31] or citation effects [23] actually change community behavior over time, also across community boundaries. ReproDB complements one-time replication studies by providing a persistent observability infrastructure that reveals how AE behavior changes over years, venues, and communities. More broadly, we aid future AE experience reports by supplying ready-made cross-venue data for comparison, and provide a systematic artifact search that, beyond helping individual authors discover artifacts, enables reuse studies that today require ad-hoc corpus assembly.

3 Platform Design and Implementation

ReproDB follows an Extract–Enrich–Rank–Publish architecture, implemented as a modular Python pipeline. The pipeline proceeds in four stages (Figure 11 in Section A). First, conference-specific extractors collect artifact evaluation results and AE committee rosters from the sources (sysartifacts, secartifacts, and conference portals), normalizing heterogeneous inputs into a uniform schema of papers, badges, and committee members. Second, the normalized records are enriched with authors, bibliographic metadata (publication counts and co-author graphs from DBLP), institutional affiliations (resolved via CrossRef, OpenAlex, DBLP, and CSRankings), and repository engagement signals (GitHub stars/forks, Zenodo/Figshare downloads). Third, records are aggregated into per-author and per-institution scores using the combined ranking formula defined in §4.2, producing ranked tables, rate statistics, and temporal trends. Fourth, the resulting JSON/YAML outputs are published to the website, where visitors can browse rankings, search artifacts, and explore author profiles.

Data Extraction. Conference-specific scrapers extract artifact evaluation results from sysartifacts and secartifacts. Both portals are Jekyll sites whose conference data is stored in YAML frontmatter of Markdown files; the pipeline uses the GitHub API to enumerate conference directories and retrieve these files, then parses the YAML entries to extract paper titles, awarded badges, and repository URLs. When the data from sys-/secartifacts is incomplete, the scraper falls back to HTML parsing conference websites. AE committee rosters are similarly extracted from conference-specific pages when they are not hosted on the artifact portals. Badge extraction includes a normalization step: early security conferences (such as ACSAC and USENIX Security) did not tag outcomes with badge labels. The pipeline assigns these artifacts a separate Evaluated category.

Data Enrichment. The artifact portals record only paper titles, badges, and repository URLs. Author and affiliation information is typically incomplete. To bridge this gap, the pipeline employs a multi-source enrichment strategy that connects each artifact to external bibliographic databases. The primary source is DBLP [9]. Extracted paper titles are matched against the DBLP XML dump to resolve author identities, retrieve affiliation data, and compute total publication counts at tracked venues. Because DBLP data does not cover all author affiliations, the pipeline additionally queries CrossRef (by DOI and by title), OpenAlex (by title), and CSRankings (by faculty name) in this order, trying each author's newest paper first so the returned affiliation reflects their most recent institution. For authors still unresolved, a co-author bridge strategy identifies the author's OpenAlex profile through a co-author's publication list, then retrieves the affiliation from the author's most recent work. Afterwards, affiliations undergo a manually-constructed regex-based normalization to merge institutional variants (e.g., "ETH Zürich" / "ETHZ"). A canonical author index assigns each researcher a stable integer identifier and records affiliation provenance (source and date). This way, we always associate authors to their most recent affiliation, while also preserving the affiliation history for future analysis. The tradeoffs and limitations of this approach are discussed further in §8.1. We attribute 92.0% of artifacts to authors (excluding artifacts from 2026 not yet indexed by DBLP) and resolve affiliations for 57.3% of the 8,139 unique authors.

In addition to the affiliations, repository URLs are probed via the GitHub API for star and fork counts and via Zenodo/Figshare APIs for download statistics, providing engagement signals per artifact.

Outcome: Heterogeneous artifact results are homogenized and enriched into a uniform dataset covering 2,831 artifacts, 8,139 authors, and 902 institutions.

Analysis and Ranking. Once enriched, the pipeline computes per-author badge scores (Available, Functional, Reproduced), organizes authors by research area (systems vs. security), and merges artifact contributions with AE committee service data to produce the composite ranking defined in §4.2. Author scores are then aggregated to institution-level rankings and statistics. Beyond rankings, the pipeline derives ecosystem-wide statistics: badge adoption rates over time, geographical distribution of contributors based on resolved affiliations, AE committee composition trends, and repository health indicators (e.g., storage distribution across GitHub, Zenodo, and Figshare). Repository engagement metrics (GitHub stars/forks, Zenodo/Figshare downloads) are collected as independent signals alongside rankings. Per-author profiles consolidate each individual's artifact history, badge progression, and AE service timeline. All outputs are written as JSON/YAML files that feed the publication stage.

Publication. The analysis produces rich datasets, including ranked author and institution tables, temporal adoption trends, badge distributions, and repository engagement statistics that benefit from interactive exploration. A Jekyll-based website presents these results through ranking tables, visualizations, and per-author profile pages. The entire pipeline is automated via GitHub Actions with monthly runs and version-controlled intermediate results.

Outcome: All rankings, profiles, and trends are publicly accessible via the website https://ReproDB.github.io.

Artifact Discovery. The platform also provides a unified search interface for all 2,831 artifacts by keyword, author, institution, venue, and badge status—filling a gap left by venue-siloed portals and general-purpose repositories that lack AE context (§7).

Implementation Effort. The analysis pipeline comprises shell scripts and 103 Python modules organized into four categories—scrapers, enrichers, generators, and utilities—totaling 28,638 lines of code. The website adds 4,819 lines of Markdown pages, HTML, and JavaScript that render the interactive rankings and artifact search. The generated dataset (YAML/JSON) currently totals 12 MB. Section A provides additional implementation details, including module-level data-flow diagrams.

4 Methodology: Metrics for Reproducibility Labor and Impact

Reproducible research is not a single-axis achievement; rather, it comprises two complementary dimensions of labor. On the one hand, when researchers publish artifacts and get them evaluated, they bear direct costs: code documentation, environment setup, long-term maintenance, and support. On the other hand, evaluators and committee chairs invest substantial time into reading papers, running code, investigating failures, and providing feedback. While resulting badges are used to reward the former labor, the latter contributions have traditionally been confined to personal websites and omitted from institutional metrics.

Using the homogenized dataset from ReproDB, we define metrics that quantify both dimensions: (1) artifact and reproducibility rates that measure the breadth and depth of artifact production (§4.1), (2) a combined score that aggregates artifact creation and AE service into a single comparable measure (§4.2), and (3) an artifact-to-evaluation ratio that characterizes an entity's balance between the two (§4.2). We view these metrics as a deliberate first proposal and welcome community contributions to refine them over time.

4.1 Artifact and Reproducibility Rates

We compute and consider two key rates to measure the breadth and depth of artifact production, which can be applied at the level of authors, institutions, or venues.

Artifact Rate (AR). The share of papers at tracked conferences that received artifact badges:

$$\text{AR\%} = \frac{\text{\# papers with badged artifacts}}{\text{\# total papers at AE-active conferences}} \times 100$$

The AR is only calculated for conference editions that conducted artifact evaluation. For example, if an author published at ACSAC in 2010–2024 but ACSAC's AE program began in 2017, only papers from 2017–2024 are included in the denominator. This avoids long-standing career totals diluting recent reproducibility efforts.

Reproducibility Rate (RR). Among papers with artifacts, the share achieving the highest-tier badge (Reproduced/Reusable):

$$\text{RR\%} = \frac{\text{\# Reproduced/Reusable badges}}{\text{\# papers with any artifact badge}} \times 100$$

While AR measures the frequency of artifact submission, the RR metric captures the depth of the reproducibility effort beyond mere artifact availability.

4.2 Combined Score: Formula and Rationale

Our decomposition of reproducibility labor into artifact creation and evaluation service parallels the CRediT taxonomy [6], which distinguishes "Software" from "Validation" contributions—both recognized but traditionally invisible in standard bibliometrics. We define a composite Author Score (S) for each author as the sum of an Artifact Score (AS) and of an AE Service Score (AES):

$$S(a) = \sum_{i \in \text{Artifacts}(a)} (A_i + F_i + R_i) + \sum_{j \in \text{AE service}(a)} (M_j \times 3 + C_j \times 2)$$

where A_i, F_i, and R_i each score one point for the respective Available, Functional, Reproduced (or Reusable) badge. The incremental scoring reflects that higher-tier badges demand greater effort and deliver greater community benefit: Olszewski et al. [17] show that artifacts earning Reproduced or Reusable badges see measurably higher reuse than Available-only ones. We similarly value AE service: terms score 3 points for committee membership, plus a 2-point bonus if the role is a chair. The chair bonus of 2 reflects the additional organizational overhead beyond per-artifact review effort.

We kept the weights of the two dimensions in rough balance: a single artifact earning all three badges scores 3 points, while a single AE membership scores 3 points—reflecting that both dimensions demand comparable effort. Following the CRediT taxonomy's modular design [6], the additive structure is easy to extend with new badge tiers or service roles. We combine creation and evaluation into a single score because the two activities are complementary facets of reproducibility labor—separating them would obscure the

synergy between researchers who both produce and evaluate artifacts. The component scores remain independently queryable on the website for users who prefer dimensional analysis.

Based on the author scores, the institution-level scores aggregate the contributions of all currently affiliated authors:

$$\text{Institution Score}(I) = \sum_{a \in \text{Authors}(I)} \text{Author Score}(a)$$

The combined metric exposes an important signal that pure publication or artifact counts miss: high-impact reproducibility contributors are not concentrated on a single axis. Some researchers score highly through artifact volume, while others score highly through sustained AE service (§6.1). With this metric, we counter the traditional invisibility of evaluation work in academic metrics and incentivize sustained participation in the reproducibility ecosystem.

Artifact-to-Evaluation Ratio. To characterize the balance between artifact creation and evaluation service, we compute the ratio of *AS* to *AES*:

$$\text{A:E (a)} = \frac{\text{AS(a)}}{\text{AES(a)}}$$

This ratio helps interpret an author's or institution's role in the reproducibility community. A score near 1.0 indicates equal contributions in both dimensions; a higher ratio suggests artifact-creation focus, a lower ratio evaluation-service focus. When AES(a) = 0, the ratio is treated as infinite, indicating pure artifact creators.

5 Dataset Overview

Having defined the methodology for scoring artifacts, AE service, and their combination, we now characterize our dataset to which they are applied. The pipeline currently ingests 2,831 evaluated artifacts from 13 conferences (7 security, 6 systems) spanning 2017-2026, enriched with 8,139 unique authors, 2,458 AE members, and engagement data for 2,709 artifact repositories.

5.1 Data Scope

We collect artifact evaluation results from security conferences (ACSAC since 2017, WOOT since 2019, USENIX Security and PETS since 2020, CHES since 2021, NDSS and SysTEX since 2024) and systems conferences (OSDI since 2020, SOSP since 2019, EuroSys and SC since 2021, USENIX ATC since 2022, FAST since 2024). The platform is designed to support additional venues as they publish machine-readable results; current coverage limitations are discussed in §8. We report partial data for 2025–2026, pending venue publications.

Inclusion Policy. A conference edition is included when its AE results are publicly scrapable from the sysartifacts or secartifacts community portals, or directly from conference websites (e.g., USENIX technical-session pages). Results hosted exclusively on the ACM Digital Library are currently excluded due to scraping restrictions. Editions not yet published or lacking machine-readable results are also omitted. All numbers in this paper are generated from the frozen input snapshot extracted on May 29, 2026.

5.2 Data Characterization

To establish the factual basis for the cross-venue analysis in §6, we characterize eight properties in the dataset, divided into two main categories. The first category focuses on artifact-specific metrics, such as growth over time, badge distribution, and repository hosting. The second category considers author- and evaluator-specific metrics, such as contribution patterns, geographical and institutional distribution, AE committee composition, and evaluation load. Appendix B provides per-venue raw data on AE metrics.

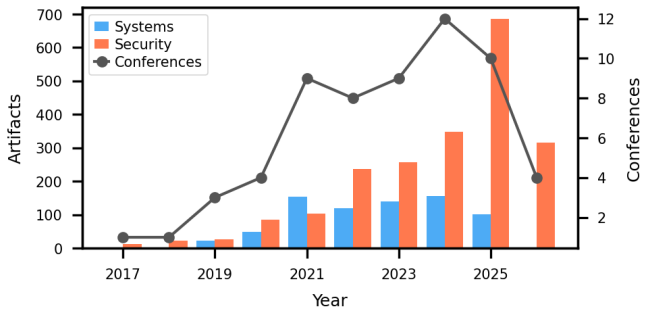


Figure 1: Volume of artifacts and conferences tracked by year.

Table 1: Badge distribution across areas.

	Systems	Security	Overall
Artifacts	740	2,091	2,831
Evaluated	0 (0.0%)	393 (18.8%)	393 (13.9%)
Available	722 (97.6%)	1,345 (64.3%)	2,067 (73.0%)
Functional	642 (86.8%)	1,189 (56.9%)	1,831 (64.7%)
Reproduced	461 (62.3%)	759 (36.3%)	1,220 (43.1%)

Growth Over Time. Figure 1 shows the number of evaluated artifacts per year by area. Among the editions in our dataset, security adopted AE earlier (2017) than systems (2019), and has grown faster: splitting each series at the midpoint (2017–2022 vs. 2023–2024), security's average artifacts per year increased 5.3× (from 80.7 to 430.3), driven by both venue expansion and rising per-venue counts—median artifacts per tracked venue rose from 12 in 2017 to 55 in 2025 across 7 venues. Systems grew more modestly (2.3×, from 57.2 to 132.3), reaching 157 artifacts across 5 venues in 2024 (2025 data is incomplete). The dataset overall roughly doubles every two years (133 in 2020, 355 in 2022, 505 in 2024). Venues with mandatory artifact evaluation [28] amplify these trends disproportionately: security alone jumped 97% year-over-year from 348 to 685 in 2024/2025, lifting the overall total.

Badge Distribution. Table 1 summarizes the distribution of badges across the dataset. The Evaluated column counts artifacts that passed AE at conferences without a badging system. Most notably, the 26.0-percentage-point gap in Reproduced rates (62.3% systems vs. 36.3% security) is substantial.

Badge Depth Over Time. Figure 2 tracks the average number of badges per artifact over time (0–3) by research area. Artifacts that passed AE without formal badges receive a depth of one. The mean across all areas rises from 1.00 in 2017 to 2.02 in 2026. Systems conferences immediately offered up to three badges, and artifacts consistently show high badge depth (2.36–2.75), possibly indicating the community's higher expectation to deliver reproduced research. Security artifacts show lower initial depths, mostly due to the lack of

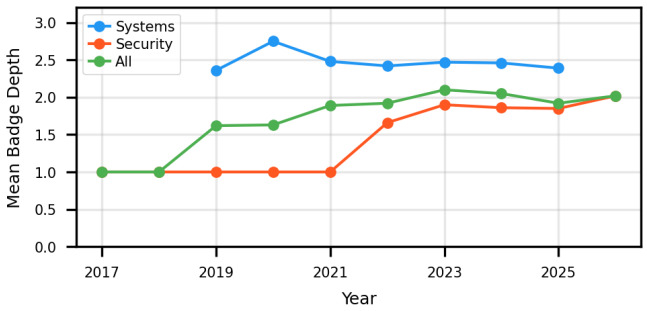


Figure 2: Average badge depth by area over time.

a badging system, but have risen steadily to 1.85 by 2025, suggesting that evaluation rigor is converging as the community matures.

Repository Hosting. The vast majority of artifacts are hosted on GitHub (51.0%) or Zenodo (11.2%). Security venues began moving toward Zenodo around 2022, as conferences such as USENIX Security and NDSS added DOI-based archival requirements, and systems venues are beginning to follow suit (Figure 6 in §6.2). One negative impact on our analysis is that these platforms typically collect less metadata than GitHub, resulting in limited information for recent years.

Author Geographical Diversity. Mapping institutions to continents reveals a striking geographical contrast. North America dominates systems more heavily (53.0% of artifacts vs. 40.7% in security). Asia contributes a comparable share of artifacts in both areas (34.6% systems, 26.2% security). European institutions account for only 11.2% of artifacts in systems, while they contribute 28.9% in security. Relative to all paper authors, artifact authors overrepresent North America (5.7 percentage point (pp) in systems, 6.3 pp in security), underrepresent Asia (-2.9 pp in systems, -9.4 pp in security), and slightly underrepresent Europe in systems (-2.9 pp) but overrepresent in security (4.9 pp).

Institutional and Area Coverage. The 8,139 authors span 902 institutions. With only 174 cross-community authors (2.1%), the communities are mostly separated. Affiliations cover 24 countries across 5 continents, with the largest author-affiliation counts in United States (620) and China (235).

AE Participation Rates. Cross-referencing AE results with total accepted papers from DBLP reveals stark historical differences in how many papers undergo artifact evaluation. In systems, annual AE participation ranges from 25.7–68.6% across 2020–2025. Similarly, in security, the average rises from 35.9% (2022–2024) to 63.4% in 2025, a shift we analyze more in §6.4.

AE Committee Diversity. The dataset records 4,393 AE committee memberships, spanning 58 countries and 475 institutions. The 2,458 unique AE members average 1.8 service terms; 987 (40.2%) have served 2+ times, while 1,471 (59.8%) serve one term. Only 91 members have served on committees in both communities. The geographical split mirrors artifact production: North America dominates both security and systems committees (49.2% and 58.4%), while Europe's share is larger in security (27.4% vs. 13.1%). Interestingly, we find that 71.4% of AE members have no artifacts published at any of our analyzed conferences. We report per-conference committee sizes and the full temporal coverage in Appendix Section B.

Insight: Both fields are dominated by North American authors, with Europeans showing more presence in security.

6 Analysis and Findings

The second capability enabled by the platform is cross-venue empirical analysis. The homogenized dataset allows us to examine dimensions that no single-venue study can address. We organize our analysis around four questions: (1) How does our suggested metric combine artifact creation and AE service into a single ranking, and do institutions specialize? (2) Do higher badge levels predict greater community uptake and scholarly impact? (3) Is the AE infrastructure sustainable? What do retention, workload, and citation patterns reveal about ecosystem health? (4) What effect do policy interventions like open-science mandates have?

6.1 Combined Metric for Authors and Institutes

The combined metric (§4.2) honors contributions from both artifact creation and evaluation service, recognizing that high-impact contributors and institutions span both axes rather than concentrating on a single dimension. We present the findings from applying this metric to our dataset separately for authors and institutions.

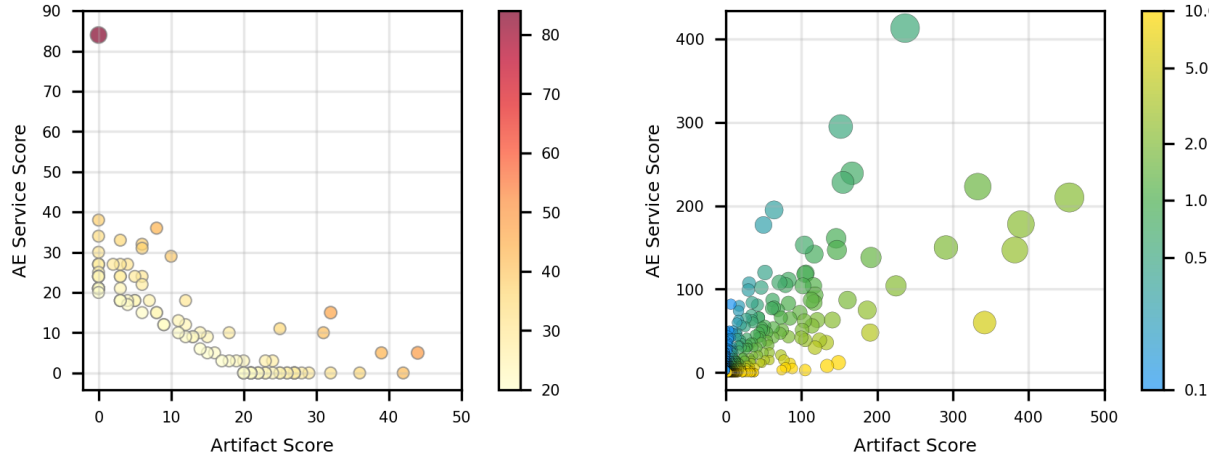
Author Metric. Figure 3a shows diversity in reproducibility contributions: while some researchers score highly through artifact volume and others through extensive AE service, the top cohorts display notable dual participation. Concretely, 50% of the top-50 and 52% of the top-100 contributors hold both artifact credits and AE memberships. This dual-role rate is striking, since only 859 (14.7%) of all contributors with ≥ 3 score in our dataset have both author and evaluator experience. The high rate for the top-scorers could reflect self-selection (high performers attracted to both), recognition effects (visibility from one role attracting opportunities), or genuine synergy between roles.

Outcome: A composite metric that rewards both artifact creation and AE service makes the full spectrum of reproducibility contributions visible and comparable.

Insight: Many top contributors are dual-role: 50% of the top-50 both create artifacts and serve on AE committees, far above the 14.7% population base rate.

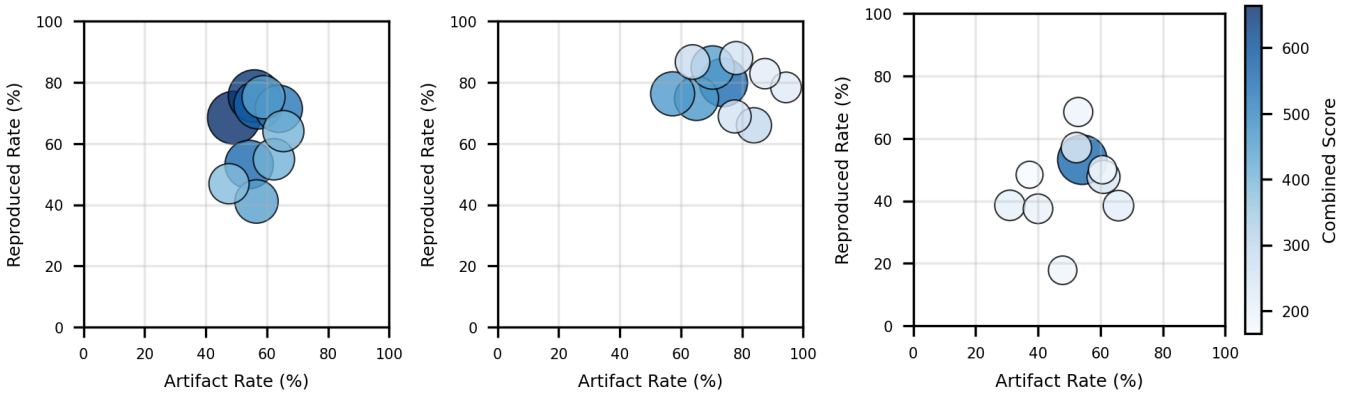
Institutional Ecosystem Roles: Creators vs. Evaluators. Examining the relationship between artifact creation and evaluation service reveals a diverse institutional ecosystem with specialized roles. Figure 3b visualizes all 902 institutions in artifact score versus AE service score space, with point size indicating combined score and color gradient representing the artifact:evaluation (A:E) ratio.

The ecosystem spans a continuum from evaluator-leaning institutions (A:E < 0.5) through balanced contributors (A:E 0.5–2.0) to creator-leaning institutions (A:E > 2.0). Among all ranked institutions, 34% are creator-oriented, 45% are evaluator-oriented, and 21% maintain balanced contributions. However, the top echelon shows a different profile: among the top-10 institutions by combined score, 40% are creator-leaning (e.g., ETH Zürich: artifact score 342, AE service score 60, A:E=5.7; Tsinghua University: artifact score 454,



(a) Top-50 individuals (the gradient shows the total score). **(b) All institutions (the gradient shows the A:E ratio).**

Figure 3: Artifact creation and AE service for individual authors and institutions.



(a) All conferences. **(b) Systems conferences.** **(c) Security conferences.**

Figure 4: Top-10 institutions by combined score in AR–RR space. Bubble size indicates the combined score. Color scale is shared.

AE service score 210, A:E=2.16) and 60% maintain balanced profiles (e.g., University of Illinois Urbana-Champaign: artifact score 237, AE service score 413, A:E=0.57). Top-10 institutions exhibit a median A:E ratio of 1.71, indicating that top-scoring institutions tend to accumulate higher artifact creation scores—likely because prolific research groups naturally produce more artifacts rather than artifact creation driving institutional prestige.

Figure 4 plots the top-10 institutions by combined score in AR vs. RR space (see §4.1 for definitions) for (a) all conferences, (b) systems, and (c) security. Bubble size and color intensity represent the combined score. These plots reflect different community norms. The security cohort generally shows lower AR than systems: no institution in the security cohort achieves AR > 80%. Similarly, RR for systems is typically higher than security, with some institutions exceeding 80% RR. Only 6 institutions appear in both security and systems top-20 lists, underscoring distinct ecosystems.

Insight: Institutions specialize: 34% creators, 45% evaluators, 21% balanced. The top 10 are creators (median A:E 1.71).

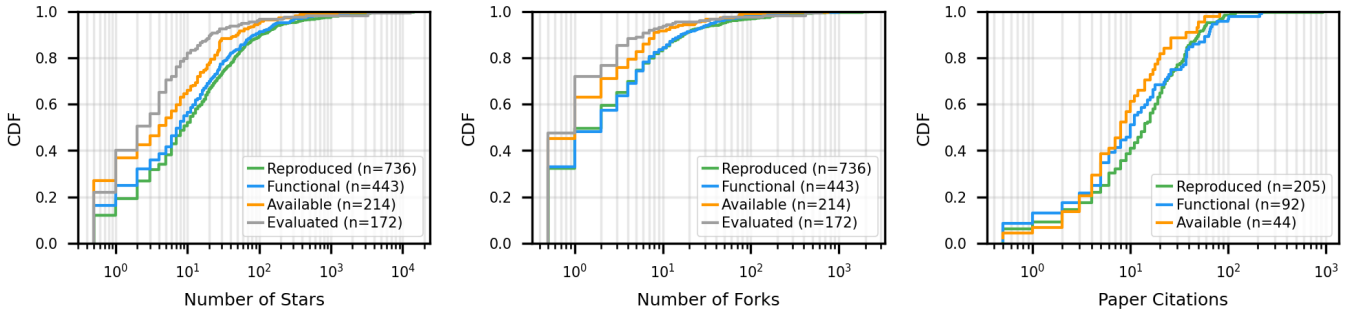
6.2 Community Uptake

A natural question about badge incentives is whether higher badge levels correlate with greater community uptake. We find that GitHub stars and forks provide early, observable signals of badge value, whereas the correlation with citations is less clear.

Storage Platforms. The two communities exhibit divergent storage preferences and engagement patterns (Figure 6). While GitHub is the primary platform for both (45.3% security, 69.1% systems), recently, security's Zenodo usage surged to 80% in 2026 following USENIX Security's open-science policy [28].

Artifact Availability. Long-term availability is high overall (96.8% of URLs accessible). Platform choice is the dominant driver: DOI-backed archives (Zenodo, Figshare) retain 98.9% availability, compared with 95.6% for non-DOI hosting, with self-managed URLs (personal and institutional pages) showing the highest decay.

GitHub Engagement. We consider 1,565 artifacts linking to GitHub repositories and compare stars and forks, considering the artifact's highest badge level: Reproduced (n=736), Functional (n=443), Available (n=214), and Evaluated (n=172). Figures 5a and 5b show



(a) Stars by badge level. **(b) Forks by badge level.** **(c) Systems-only citations by badge level.**

Figure 5: CDFs of GitHub stars, forks, and paper citations by highest badge level.

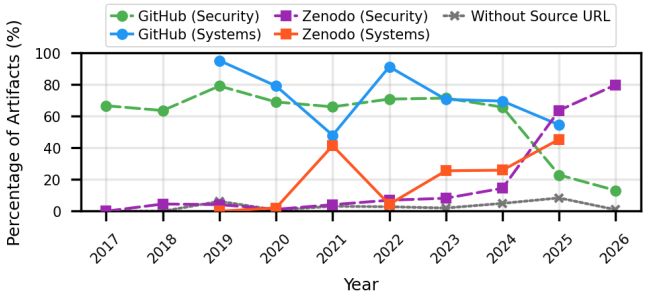


Figure 6: Evolution of artifact storage platforms over time.

that badge level is significantly associated with both stars and forks. Reproduced artifacts have a median of 9 stars and 2.0 forks, compared to 4 and 1.0 for Available-only artifacts. The effect is particularly pronounced in security, where Reproduced artifacts attract 6 median stars, 3.0× that of Available.

Paper Citations. We retrieve Semantic Scholar citation data for 759 AE papers (26.8% coverage of the full AE set). Available-only (n=74) and Evaluated (n=171) subgroups are comparatively small, limiting the statistical power of per-badge comparisons. In systems, where badge-based evaluation has been established longer, the value of the Reproduced badge is highly pronounced: Reproduced papers earn a median of 14 citations, 1.8× that of Available-only (8), consistent with Raff [23]. This systems-side signal is supported by deeper and older citation coverage in our data, which provides more stable per-badge estimates than the newer security cohorts (Figure 5c). For security, we refrain from a strong citation-level claim because the citation-observable subset is more recent, increasing variance.

Insight: Higher badge levels correlate with greater GitHub engagement overall. In systems, Reproduced papers earn 1.8× the median citations of Available-only papers.

Artifact Citation: A Negative Result A notable question for reproducibility incentives is whether artifacts are cited directly (instead of citing the research paper). We examined 782 artifact DOIs from our dataset. Of these, OpenAlex reported 14 artifacts as having citations, totaling 43 citing DOIs. Manual verification of each citation revealed zero genuine third-party artifact citations. All reported citations were either false positives (citations that referenced the paper DOI, not the artifact DOI due to identical titles) or self-citations (authors citing their own artifacts in the implementation section). This finding stands in tension with the FORCE11 Software Citation Principles [26], which advocate treating software as an independently citable scholarly object; our data shows that despite growing artifact production, the community has not yet adopted this in practice. Until artifacts are cited in their own right, the primary metric for evaluating artifact quality remains badge counts. This underscores the need for communities to encourage artifact citations as a first-class metric.

Insight: Despite growing artifact production, we found zero artifact citations across 782 DOIs.

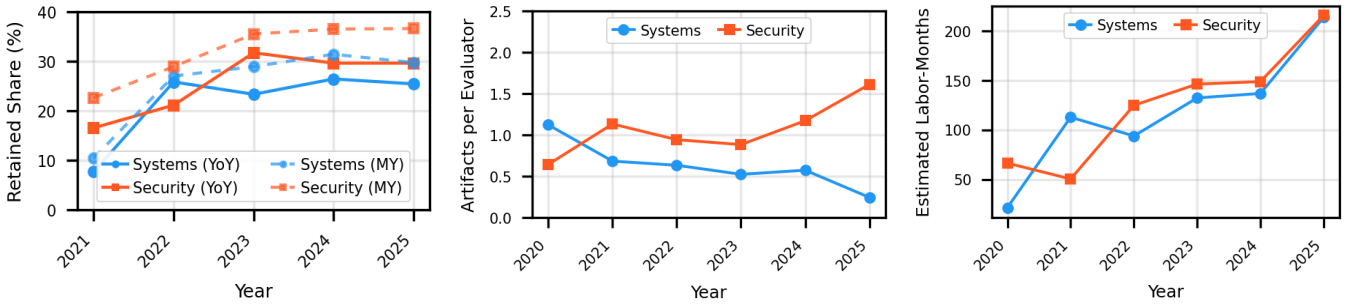
6.3 AE Committee Health and Sustainability

As artifact evaluation scales, understanding committee sustainability is essential for long-term AE infrastructure. We examine three dimensions of AE ecosystem health: committee stability, evaluator workload, and the aggregate volunteer labor investment.

Committee Stability and Retention. Figure 7a shows the retained share—the fraction of a given year's committee that served in the same area the previous year, i.e., year-over-year (YoY) retention. The retention rate in systems ranges from 7.6% to 26.4% (2021–2025), while security ranges from 16.5% to 31.7%. While the majority of each year's committees consists of members who did not serve in the immediately preceding year, the upward retention trend suggests steadily accumulating systematic knowledge in both areas. We also report a multi-year (MY) retention for committee members who served in any prior year, which is 3.6–5.7 % higher than YoY retention. The retention across research areas is negligible.

AE Chair Leadership Pipeline. Beyond evaluator retention, leadership continuity is equally important. Of the 82 unique AE chairs in our dataset, 23 (28.0%) have chaired more than once and 13 have led AE processes at multiple conference series. Examining the member-to-chair promotion pathway, 31 chairs (37.8%) previously served as regular AE members before being appointed as chairs, with a median transition time of 2 years.

Insight: The AE chair leadership pipeline is healthy with recurring and cross-venue chairs and a member-to-chair path.



(a) Committee retention. **(b) Evaluator workload.** **(c) Estimated volunteer labor.**

Figure 7: AE committee health: retention, workload, and volunteer labor by community.

Evaluator Workload Imbalance. Aggregating across all tracked years, security conferences handled 2,091 artifacts with 1,348 unique evaluators, while systems conferences handled 740 artifacts with 1,201 unique evaluators. Workload distribution reveals a critical disparity (Figure 7b). In systems, the per-year ratio of artifacts to committee size has remained low and stable, while security workload is consistently higher and in recent years increased to 1.6. As the corresponding data is not available, we assume that each artifact is reviewed by 2–4 evaluators, making the actual per-evaluator workload higher than this ratio suggests. This structural imbalance reflects security's larger submission volume, and it raises sustainability concerns: overloaded committees risk inconsistent evaluations and burnout.

Volunteer Labor. To quantify the aggregate effort behind artifact evaluation, we estimate volunteer labor-months in Figure 7c by assuming two weeks of effort per evaluator in a given year, following the evaluator survey of Malik et al. [16]. Across all tracked years, AE committees have volunteered an estimated 2,196 person-months. The effort in security and systems is similar due to comparable committee sizes. Further data and exploration are needed to understand whether systems artifacts are reviewed by substantially more evaluators, and whether this leads to positive effects compared to artifacts in security. This largely unrecognized volunteer investment underscores the need for institutional support and sustainable committee sizing.

Insight: Security evaluators potentially face 2.5 × the per-capita workload of systems evaluators; rising retention helps, but sustainable scaling requires proactive committee sizing.

6.4 Open-Science Policy and AE Participation

Across the many changes to artifact evaluation processes over the past decade, such as expanded badge tiers, larger committees, or streamlined review workflows, no single intervention has had as large an effect on participation as mandating artifact availability. Figure 8 shows Available, Functional, and Reproduced badge rates as a percentage of all accepted papers. Systems venues see a slow decline (especially in papers with Reproduced badges), while security surged to 59.6% Available rate in 2025.

The USENIX Security Effect. USENIX Security's 2025 open-science policy [28] requires all accepted papers to include an open-science appendix describing the artifact and to participate in AE

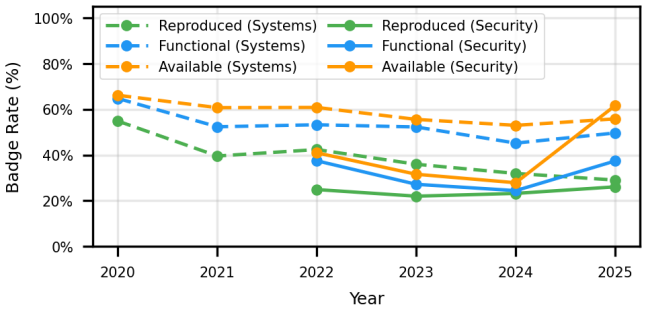


Figure 8: Badge rates as a fraction of all accepted papers.

at least to the Available badge. This effectively made AE opt-out rather than opt-in: participation jumped from 29.5% in 2024 to 89.7% in 2025. The mandate's impact extends beyond the Available badge: Reproduced rates also rose from 20.6% to 34.2%, indicating that funneling more artifacts into the pipeline lifts attainment across all tiers, though gains are smaller at higher levels.

Insight: Open-science mandates are the strongest lever for AE adoption, tripling participation at USENIX Security.

7 Benefits of Cross-Community Artifact Search

The third capability enabled by ReproDB is cross-community artifact discovery, which is not supported by today's venue-siloed portals. Without a cross-venue index, a researcher would need to know all relevant papers and manually browse each conference's artifact page or appendices to find related implementations. Today, researchers typically discover artifacts through their associated papers, while topic-based search enables complementary workflows: locating reusable research artifacts for comparative studies, finding baseline implementations across venues, or surveying the landscape of available artifacts in a research area. By surfacing these cross-community connections, ReproDB closes the creation–evaluation–reuse loop: artifacts are not only produced and assessed but can also be discovered and built upon across venue boundaries. To quantify the overlap, we extract keywords from 2,831 indexed artifacts and measure the span of venues and research areas.

Keyword Extraction. After removing standard stop words and academic boilerplate (e.g., "novel", "efficient", "approach"), artifact

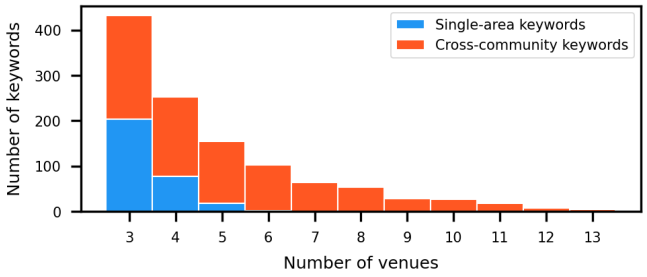


Figure 9: Distribution of venue coverage for title keywords.

titles yield 7,125 unique keywords. Of these, 1,133 (15.9%) cross-community keywords appear in artifacts from both security and systems. The remaining 5,992 keywords are confined to a single area. Cross-community keywords have a median venue span of 4 and a median year span of 4, with the broadest reaching 13 venues and 10 years of continuous presence.

Cross-Venue and Cross-Year Reach. Figure 9 shows the distribution of venue coverage among keywords spanning three or more venues. Among the 1,133 cross-community keywords, 441 (38.9%) span five or more venues, 140 (12.4%) span eight or more, and 58 (5.1%) span ten or more. The temporal dimension is equally broad: 415 cross-community keywords (36.6%) appear in five or more distinct years, and 173 (15.3%) persist for seven or more years, showing that the overlap is not a recent artifact of growing venue coverage, but reflects long-standing shared research themes.

Concrete examples across both areas illustrate the practical value: "federated learning" yields 26 artifacts spanning 7 venues and 6 years, "learning" yields 161 artifacts across 13 venues and 8 years, and "differential privacy" appears in 56 artifacts across 7 venues and 7 years. None of these cross-venue, cross-year connections are visible when browsing a single conference's artifact page.

Outcome: By providing a longitudinal index, ReproDB facilitates cross-community artifact discovery and reuse across diverse venues and extended temporal scales.

8 Limitations and Future Work

The ReproDB platform shows that a unified, automated platform for aggregating and analyzing artifact evaluation and making artifact data discoverable is both feasible and valuable. Yet, several limitations remain, each representing an opportunity for the community to strengthen the reproducibility infrastructure. We organize these into three themes: data coverage and quality, metric biases and interpretability, and future capabilities.

8.1 Data Coverage and Quality

AE Coverage. We capture only venues and years that publish machine-readable AE outcomes. Conferences with private or inconsistently published results are not represented. USENIX ATC and OSDI share a single joint AE committee in years when both run (2022–2024). Our pipeline counts joint-committee members once per shared cycle. These caveats affect committee diversity analyses but do not bias artifact-level findings. ACM CCS publishes artifact evaluation results in the ACM Digital Library (DL) rather than on the secartifacts portal. The ACM DL does not expose these results in a machine-readable format and prevents automated scraping, so CCS artifacts are currently excluded from our dataset despite being publicly visible. Similarly, other ACM-hosted venues with active AE programs (ASPLOS, PLDI, OOPSLA, SIGMOD) publish results only on ACM DL, placing them beyond our current reach. Our cross-venue analysis thus covers systems and security as defined by the sysartifacts and secartifacts portals, not computer science broadly. We call on conference organizers and digital libraries to publish machine-readable AE results, enabling platforms like ReproDB.

Artifact URLs. Vansteenhuyse et al. [29] showed that many papers include a different artifact URL in the final paper than the one submitted for artifact evaluation (e.g., GitHub repository in the paper, Zenodo for AE). This behavior can influence some of the metrics and behaviors we studied, such as popularity indicators and citations. In the future, these papers and the impact of this practice should be studied further.

Data Quality and Affiliation Coverage. Resolving affiliations is complex (cf. §3). Affiliations in ReproDB reflect the most recent record found across our sources (OpenAlex, CrossRef, CSRankings) rather than the affiliation held at evaluation time. This approach mirrors CSRankings' methodology and, while imperfect, reveals which institutions currently value and invest in reproducibility. Repeated pipeline runs may reassign institutions as researchers move, which we view as a feature (tracking current engagement) rather than a bug. Of the 42.7% authors with unresolved affiliations, 99% appear on fewer than three tracked papers, limiting the co-author evidence available for disambiguation; and single-paper authors lack the cross-references needed by any automated enrichment heuristic. We aim to improve coverage by adding an interface for AE chairs to upload data directly. DBLP's multi-month indexing lag further impedes timely enrichment for recent papers.

Badge Semantics. Badge names and criteria vary by venue and community, and evolve over time. We assume that identically named badges are equivalent. Early security conferences evaluated artifacts without awarding separate badges. For the combined score, we include them with a score of 1 (recognizing that the artifact passed evaluation). We include them because excluding early results would systematically undercount AE adoption. A common badge ontology would eliminate this ambiguity for the future.

8.2 Metrics: Biases and Responsible Use

While ReproDB makes reproducibility contributions visible and allows for the benchmarking of institutional reproducibility practices, these rankings are observational. Users should interpret component scores carefully and account for potential data biases before using this data for high-stakes decisions or sensitive conclusions. Specifically, we caution against the use of raw ReproDB data in hiring, promotion, or funding decisions.

AE Selection Bias. AE members typically self-nominate and are selected by chairs; chairs are invited by program/steering committees. Not all community members have equal opportunity to accumulate AE service score, so this axis reflects opportunity as much as willingness to serve. Transparent, open AE recruitment could broaden the evaluator pool, making metrics more equitable.

Publication Volume Bias. Prolific authors have more opportunities to submit artifacts, inflating absolute scores. The rate metrics (AR%, RR%) normalize for this: low rates despite high volume indicate that reproducibility is not a priority, while high rates from less prolific authors reveal consistent effort. However, the combined score tries to counteract this bias by awarding more points per additional badge, rewarding depth of reproducibility over volume.

Combined vs. Separate Scores. A single combined score may obscure dimension-specific contributions: a prolific artifact creator and a dedicated evaluator can receive similar combined scores for very different profiles. The dataset and website mitigate this by allowing filtering and sorting by each score type independently.

Metric Integrity. To mitigate the risk of researchers or institutions "gaming" the system by submitting minimal-effort artifacts (receiving available badges) to inflate the artifact score, we utilize rate-based metrics (AR%, RR%). Furthermore, ReproDB's absolute artifact counts run contrary to CSRankings' author-count-normalized publication scores [4]: inflating artifact volume by adding authors to many papers would improve ReproDB rankings for institutions but dilute CSRankings scores, making it difficult to exploit both systems simultaneously.

8.3 Community-Driven Roadmap

Standardized AE Reporting. Transitioning ReproDB into a fully automated, real-time, comprehensive index requires efforts that extend beyond the platform itself; it necessitates a shift in how communities publish AE results. If every venue published machine-readable AE results, if bibliographic databases exposed up-to-date affiliation data, and if the community adopted a shared badge ontology aligned with FAIR4RS principles [3], platforms like ReproDB could provide a comprehensive, always-current view of reproducibility across all of computer science. We call on conference organizers, digital libraries, and bibliographic databases to enable this vision.

Sustainability. The platform is designed for zero-cost long-term operation: GitHub Actions provides free CI/CD for public repositories, and GitHub Pages hosts the website at no cost. New conferences self-register by submitting a scraper module via pull request; no central authority or funding is required. We are committed to supporting the platform long-term and actively seek to grow a contributing community of conference organizers, AE chairs, and researchers to extend coverage and maintain data quality.

Additional Metadata Enrichment. GitHub and Zenodo repositories offer additional auxiliary metadata (e.g., licenses, hardware requirements, programming languages) that could be used to further enrich the dataset, enabling researchers to quickly assess reuse eligibility or filter artifacts by technology stack. Beyond structured APIs, semi-automated scraping of paper appendices could recover artifact repository links even when conferences migrate results to non-machine-readable formats such as the ACM DL, as demonstrated by Vansteenhuyse et al. [29]. Their tool, ArtiFinder, could also allow extending our search database with artifacts that did not participate in formal artifact evaluation and whose URL is only listed in the corresponding paper.

Robust Combined Score. This paper suggests a combined scoring for artifact creation and AE service. Developing a robust combined score that stands the test of time requires careful consideration of the evolving landscape of reproducibility metrics. The score should balance artifact creation and evaluation service, adapt to new types of contributions, and remain resistant to gaming. By continuously refining the scoring methodology and incorporating community feedback, ReproDB can ensure that the combined score accurately reflects meaningful contributions to reproducibility.

9 Related Work

AE Experience Reports. D'Elia et al. [10] report on five years of EuroSys artifact evaluations, and Olszewski et al. [17] provide an 11-year longitudinal review of AE in security venues and a large-scale study on machine learning artifacts in security [18], offering rich qualitative lessons, while ReproDB focuses on quantitative cross-community comparisons. Malik et al. [16] describe expanding AE scope at SC21, while Pallez et al. [20] and Sirvent et al. [25] describe HPC reproducibility initiatives. These efforts focus on process design, while ReproDB complements them with quantitative outcomes informing future process improvements.

Artifact Longevity and Quality. Guilloteau et al. [12] studied artifact longevity in parallel and distributed systems, and their follow-up work [13] measures Docker-based environment decay. Costa et al. [8] introduced CompRep, a dataset for computational reproducibility assessment. Olszewski et al. [18] find that artifacts that went through AE work more often than those that did not. These efforts assess whether individual artifacts still work, while ReproDB addresses the complementary question of how the AE ecosystem evolves. Our artifact index could serve as a structured input corpus for future longevity studies across communities.

Incentives and Community. Raff [23] investigated whether citations reward reproducibility, while our negative artifact-citation analysis (§6.2) expands upon the usage of citation counts. Fund [11] advocated for integrating reproducibility into CS curricula, and Lieggi et al. [15] described the Summer of Reproducibility mentorship program. These are both interventions that ReproDB's evaluator-retention and newcomer data (§6.3) could help evaluate longitudinally. CyCoAnalysis [5] is a database that collects conference policies in a unified format, including data on artifact evaluation. However, unlike ReproDB, this data collection is not automated and does not extend to the artifacts themselves. In the future, CyCoAnalysis could be extended with our automatically scraped data on, e.g., AEC composition. Wonsil et al. [31] proposed mechanisms to raise the reproducibility bar. Our combined metric offers a concrete way to measure whether such mechanisms change community behavior.

Broader Meta-Science Context. Within computer systems, Collberg and Proebsting [7] measured repeatability outcomes in major systems conferences and documented practical barriers. In machine learning, Hutson [14] highlighted similar concerns around incomplete artifact disclosure and experimental opacity. Vansteenhuyse et al. [29] designed a tool to automatically find artifact links in papers and used this to build a dataset of security artifacts from the past 25 years. These studies motivate ReproDB as infrastructure for

continuous, cross-venue measurement. ReproDB operationalizes longitudinal analysis of artifact practices at the ecosystem scale.

Publication Rankings and Benchmarking Infrastructure. Our platform draws inspiration from publication-statistics and benchmarking platforms such as Systems Circus [22] and CSRankings [4], as well as large-scale bibliographic infrastructure such as DBLP [9]. These systems focus on publication output, venue-level visibility, and author/institution discoverability rather than artifact-evaluation activity. In machine learning, Papers With Code [21] links papers to code repositories and organizes them by task and benchmark, substantially improving artifact discoverability. However, its taxonomy is ML-specific (datasets, tasks, benchmarks) without adoption beyond ML. ReproDB brings a similar search to the security and systems communities.

Artifact Reuse. Prior work measures artifact creation and evaluation, but largely overlooks whether artifacts are actually reused. Olszewski et al. [17] provide initial evidence that higher-badge artifacts see more downstream reuse. ReproDB's cross-venue search engine (§7) is designed to facilitate exactly this: by surfacing related artifacts across community boundaries, it lowers the barrier to discovering and building on existing evaluated code. Closing this creation–evaluation–reuse loop is a key motivation for the search capability, and measuring reuse at scale is a natural next step.

Positioning ReproDB. Prior work examines individual venues, tests artifact health, or tracks publication output, but none combines cross-community AE process measurement, evaluator labor analysis, and artifact discoverability into one automated platform. ReproDB complements one-time replication studies by providing a persistent observability infrastructure that reveals how reproducibility behavior changes over years, venues, and communities. More broadly, we aid future AE experience reports by supplying ready-made cross-venue data for comparison, and a systematic artifact search enabling longevity and reuse studies that today require ad-hoc corpus assembly.

10 Conclusion

We present ReproDB, an open platform that addresses three challenges. The platform's automated pipeline homogenizes 2,831 artifacts from 13 conferences spanning 2017-2026 into a unified, continuously updated dataset with a single schema. The homogenized dataset enables cross-venue analysis and findings on adoption trajectories, ecosystem roles, and evaluator sustainability, while revealing the continued invisibility of artifacts in formal citation practices; insights that no single-venue study can provide. ReproDB also closes the creation–evaluation–reuse loop with the first cross-venue artifact search engine for security and systems. Unlike one-shot AE reports, ReproDB is designed as a continuously operating measurement infrastructure, enabling the community to monitor whether interventions and policy changes improve reproducibility efforts over time and assess key indicators of AE ecosystem health.

Our findings already demonstrate this utility across multiple dimensions. We show how policy-driven requirements are the strongest lever for broadening artifact evaluation adoption and that higher badge levels correlate strongly with community engagement, both in terms of repository statistics and paper citations. Additionally, we characterize the AE ecosystem—authors and institutions—in terms of artifact creation and evaluation service. We encourage the community to adopt, extend, and contribute to the platform, and to treat AE health as an ongoing concern, requiring regular measurement and coordinated intervention.

Reproducibility and Open Science Statement. All figures, tables, and inline statistics are generated by self-contained scripts in the supplementary material. The data presented in this paper was scraped and analyzed on May 29, 2026. All code, data, and the live website are available at https://github.com/ReproDB and https://ReproDB.github.io.

AI Usage. GitHub Copilot assisted by generating text, tables, graphs, and code. All AI-generated content was reviewed and edited by the authors for accuracy and clarity.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and constructive suggestions. We are grateful to Alexios Voulimeneas, Carlos Maltzahn, Kerstin Oberwagner, Mathias Payer, Marcela Melara, Miguel Matos, Pedro Fonseca, Tanu Malik, and Thaleia Doudali for reviewing the website and early drafts, and providing insightful comments that improved the website and paper's clarity and rigor. This research is partially funded by the Internal Funds KU Leuven, the Cybersecurity Research Program Flanders, and FCT – Fundação para a Ciência e a Tecnologia, I.P., under project 2022.09325.PTDC.

References

[1] ACM. 2020. Artifact Review and Badging — Version 1.1. https://www.acm.org/publications/policies/artifact-review-and-badging-current. Accessed: 2026-03-01.

[2] Monya Baker. 2016. 1,500 Scientists Lift the Lid on Reproducibility. ***Nature*** 533, 7604 (2016), 452–454.

[3] Michelle Barker, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, and Tom Honeyman. 2022. Introducing the FAIR Principles for Research Software. ***Scientific Data*** 9 (2022), 622. doi:10.1038/s41597-022-01710-x

[4] Emery D. Berger. 2024. CSRankings: Computer Science Rankings. https://csrankings.org/. Accessed: 2026-03-01.

[5] Marton Bognar, Lieven Desmet, and Frank Piessens. 2026. CyCoAnalysis: A Dataset of Cybersecurity Conference Metadata for Meta-Science and Policy Decisions. In ***IEEE Security and Privacy Workshops (SPW)***.

[6] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond Authorship: Attribution, Contribution, Collaboration, and Credit. ***Learned Publishing*** 28, 2 (2015), 151–155. doi:10.1087/20150211

[7] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in Computer Systems Research. ***Commun. ACM*** 59, 3 (2016).

[8] Lázaro Costa, Susana Barbosa, and Jácome Cunha. 2025. CompRep: A Dataset For Computational Reproducibility. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746160

[9] DBLP Team. 2024. DBLP Computer Science Bibliography. https://dblp.org/. Accessed: 2026-03-01.

[10] Daniele Cono D'Elia, Thaleia Dimitra Doudali, Cristiano Giuffrida, Miguel Matos, Mathias Payer, Solal Pirelli, Georgios Portokalidis, Valerio Schiavoni, Salvatore Signorello, and Anjo Vahldiek-Oberwagner. 2025. Lessons Learned from Five Years of Artifact Evaluations at EuroSys. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746152

[11] Fraida Fund. 2023. We Need More Reproducibility Content Across the Computer Science Curriculum. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3589806.3600033

[12] Quentin Guilloteau, Florina M. Ciorba, Millian Poquet, Dorian Goepp, and Olivier Richard. 2024. Longevity of Artifacts in Leading Parallel and Distributed Systems Conferences: a Review of the State of the Practice in 2023. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3641525.3663631

[13] Quentin Guilloteau, Antoine Waehren, and Florina M. Ciorba. 2025. Longitudinal Study of the Software Environments Produced by Dockerfiles from Research Artifacts: Initial Design. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746146
[14] Matthew Hutson. 2018. Artificial Intelligence Faces Reproducibility Crisis. ***Science*** 359, 6377 (2018).
[15] Stephanie Lieggi, Fraida Fund, Kate Keahey, and Marc Richardson. 2025. Summer of Reproducibility: Building Global Capacity for Practical Reproducibility through Hands-On Mentorship. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746149
[16] Tanu Malik, Anjo Vahldiek-Oberwagner, Ivo Jimenez, and Carlos Maltzahn. 2022. Expanding the Scope of Artifact Evaluation at HPC Conferences: Experience of SC21. In ***Proceedings of the International Workshop on Practical Reproducible Evaluation of Computer Systems***.
[17] Daniel Olszewski, Allison Lu, Anna Crowder, Nathaniel Bennett, Seth Layton, Sri Hrushikesh Varma Bhupathiraju, Tyler Tucker, Siddhant Kalgutkar, Hunter Ver Helst, Carson Stillman, Kevin R. B. Butler, Sara Rampazzi, and Patrick Traynor. 2025. Reproducibility in Applied Security Conferences: An 11-Year Review on Artifacts and Evaluation Committees. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746151
[18] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. 2023. "Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In ***Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)***.
[19] Open Science Collaboration. 2015. Estimating the Reproducibility of Psychological Science. ***Science*** 349, 6251 (2015), aac4716.
[20] Guillaume Pallez, Judith C. Hill, and Sascha Hunold. 2025. Implementing a Reproducibility Initiative in HPC: Experiences from SC'24. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746148
[21] Papers with Code. 2024. Papers with Code — Machine Learning Research Discovery. https://paperswithcode.com/. Accessed: 2026-03-01.
[22] Mathias Payer. 2024. Systems Circus: Publication Statistics for Systems Conferences. https://nebelwelt.net/pubstats/. Accessed: 2026-03-01.
[23] Edward Raff. 2023. Does the Market of Citations Reward Reproducible Work?. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3589806.3600041
[24] secartifacts. 2024. Security Research Artifacts. https://secartifacts.github.io/. Accessed: 2026-03-01.
[25] Raül Sirvent, Rocío Carratalá-Sáez, Amal Gueroudji, Tanzima Z. Islam, Line Pouchard, and Michela Taufer. 2025. Reproducibility for HPC and Distributed Environments: Committees, Nondeterminism, Performance and Workflows. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746141
[26] Arfon M. Smith, Daniel S. Katz, and Kyle E. Niemeyer. 2016. Software Citation Principles. ***PeerJ Computer Science*** 2 (2016), e86. doi:10.7717/peerj-cs.86
[27] sysartifacts. 2024. Systems Research Artifacts. https://sysartifacts.github.io/. Accessed: 2026-03-01.
[28] USENIX Association. 2025. USENIX Security '25 Call for Papers. https://www.usenix.org/conference/usenixsecurity25/call-for-papers. Accessed: 2026-03-06.
[29] Daan Vansteenhuyse, Arthur Bols, Lieven Desmet, Victor Le Pochat, Jo Van Bulck, and Marton Bognar. 2026. Not All Those Who Share Are Lost: Analyzing 25 Years of Cybersecurity Artifact Sharing Practices Through Automated Discovery. In ***Proceedings of the USENIX Security Symposium***.
[30] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. ***Scientific Data*** 3 (2016), 160018. doi:10.1038/sdata.2016.18
[31] Joseph Wonsil, Rúbia Reis Guerra, Adam Craig Pocock, Jack Sullivan, and Margo I. Seltzer. 2025. Raising the Reproducibility Bar. In ***Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)***. doi:10.1145/3736731.3746157

A Pipeline Details

The implementation of our contributions is split across multiple open-source repositories under the ReproDB GitHub organization[1]:

(1) `reprodb-pipeline`[2] — the Python data pipeline (scraping, enrichment, ranking, and output generation);
(2) `reprodb.github.io`[3] — the Jekyll-based public website rendering rankings and visualizations;
(3) `data-schemas`[4] — versioned JSON Schema definitions (auto-exported from the pipeline's Pydantic models);
(4) `reprodb-pipeline-results`[5] — auto-generated output archive for reproducibility.

Pipeline Architecture. Figure 11 illustrates the four-stage Extract–Enrich–Rank–Publish pipeline. External sources (AE portals, DBLP, CrossRef, OpenAlex, CSRankings, and repository hosts) feed into conference-specific scrapers that produce normalized JSON. Enrichment and ranking stages resolve author identities and affiliations, collect repository engagement signals, and compute per-author and per-institution scores. The final stage publishes ranked JSON and YAML files consumed by the public website.

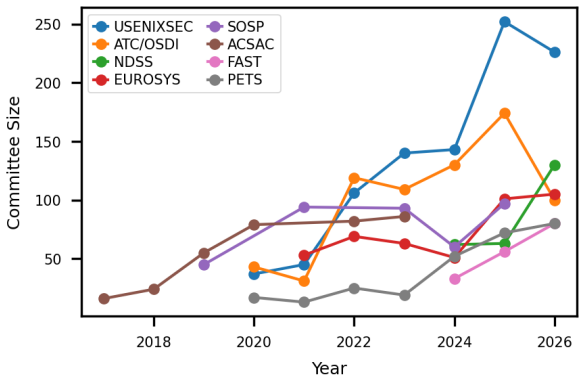


Figure 10: AE Committee size growth over time (top 8 by size conferences by tracked history).

B Artifact Evaluation Details

Scaling AE Infrastructure. Figure 10 illustrates rapid growth in AE committee sizes: USENIX Security, ATC/OSDI, and EuroSys have expanded 1–7 × between 2020 and 2025. USENIX ATC and OSDI share a joint AE committee; SOSP appears with gaps due to biennial scheduling and limited data availability. In 2026, we see an uptick so far, although the numbers are preliminary.

GitHub Engagement. Table 2 breaks down repository engagement by conference. OSDI leads in median stars (30.5) and total stars (38,955 across 135 repositories), consistent with the engagement gap observed in the CDF analysis above (§6.2). Security venues show a wider spread: USENIX Security has the most repositories (527) but a lower median (12 stars), while WOOT has only 31 repositories but a high median (10 stars), driven by popular offensive-security tools. Conferences without GitHub repository data in our dataset (CHES, FAST, SC) are omitted from this table.

Committee Sizes. Table 3 shows per-conference committee sizes over time. USENIX Security's committee grew from 37 in 2020 to 252 in 2025, reflecting its open-science mandate. Newer venues (SYSTEX, FAST) begin with smaller committees.

[1] https://github.com/ReproDB
[2] https://github.com/ReproDB/reprodb-pipeline
[3] https://github.com/ReproDB/reprodb.github.io
[4] https://github.com/ReproDB/data-schemas
[5] https://github.com/ReproDB/reprodb-pipeline-results

Conference Timeline Coverage. Table 4 shows the number of evaluated artifacts per conference per year. For the editions in our dataset, security conferences began AE earlier (ACSAC in 2017) and contribute more artifacts in aggregate, while systems venues show consistently high per-venue counts. Gaps reflect biennial scheduling (SOSP) or editions with no published results.

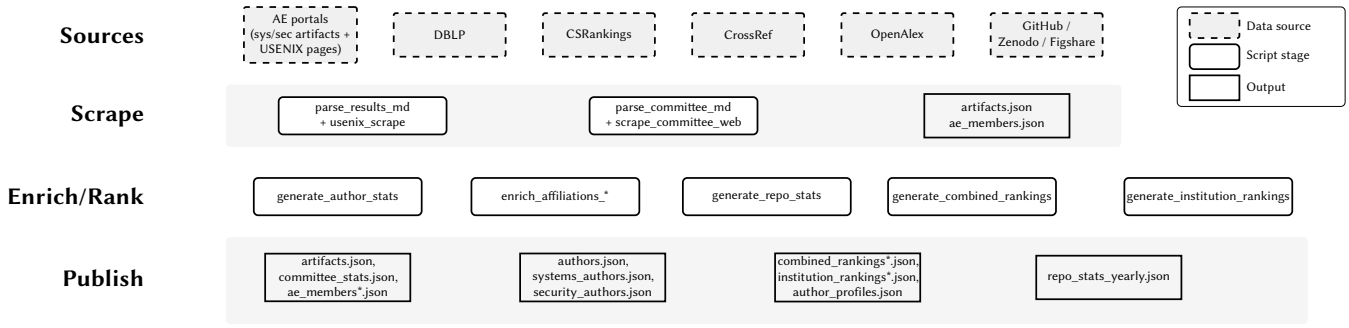


Figure 11: High-level steps, inputs, scripts and outputs of the ReproDB pipeline.

Table 2: Repository engagement metrics by conference.

Conference	Area	Repos	Stars	Med. Stars	Forks	Med. Forks	Max Stars
ACSAC	Security	205	3,092	2	747	0	566
NDSS	Security	74	4,073	5.0	1,348	0.5	1,606
PETS	Security	294	18,677	1	1,667	0	12,895
SYSTEX	Security	17	60	1	14	0	27
USENIXSEC	Security	527	64,026	12	10,277	2	12,867
WOOT	Security	31	16,219	10	1,303	1	12,381
ATC	Systems	110	4,827	9.5	837	1.0	1,652
EUROSYS	Systems	127	3,948	7	935	2	596
OSDI	Systems	135	38,955	30.5	4,807	5.0	14,349
SOSP	Systems	106	6,681	19.5	1,298	5.0	1,219

Table 3: AE committee sizes by conference and year. "–" indicates not available.

Conference	Area	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
ACSAC	Security	16	24	55	79	3	82	86	–	–	–
CHES	Security	–	–	–	–	33	26	26	17	23	–
NDSS	Security	–	–	–	–	–	–	–	62	63	130
PETS	Security	–	–	–	17	13	25	19	52	72	80
SYSTEX	Security	–	–	–	–	–	–	–	–	6	9
USENIXSEC	Security	–	–	–	37	45	106	140	143	252	226
WOOT	Security	–	–	15	–	7	11	22	24	16	16
ATC	Systems	–	–	–	–	–	119	109	130	174	–
EUROSYS	Systems	–	–	–	–	53	69	63	51	101	105
FAST	Systems	–	–	–	–	–	–	–	33	56	80
OSDI	Systems	–	–	–	43	31	119	109	130	174	100
SC	Systems	–	–	–	–	48	–	–	–	–	–
SOSP	Systems	–	–	45	–	94	–	93	60	97	2

Table 4: Number of artifacts in AE by conference and year.

Conference	Area	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
ACSAC	Security	12	22	20	26	20	43	38	53	55	–
CHES	Security	–	–	–	–	20	25	18	30	52	–
NDSS	Security	–	–	–	–	–	–	–	38	63	114
PETS	Security	–	–	–	22	25	48	53	68	99	45
SYSTEX	Security	–	–	–	–	–	–	–	7	5	8
USENIXSEC	Security	–	–	–	37	34	114	140	142	394	149
WOOT	Security	–	–	6	–	4	6	9	10	17	–
ATC	Systems	–	–	–	–	–	51	41	49	–	–
EUROSYS	Systems	–	–	–	–	21	33	32	32	45	–
FAST	Systems	–	–	–	–	–	–	–	17	18	–
OSDI	Systems	–	–	–	48	26	35	32	34	38	–
SC	Systems	–	–	–	–	67	–	–	–	–	–
SOSP	Systems	–	–	22	–	40	–	34	25	–	–