SBSeg 2026 Armação dos Búzios - RJ
VEJA OS SELOS DE ARTEFATOS

# TopVenues: An Executable Corpus and Research Tool for Cybersecurity Literature Reviews

## Código Aberto

**Sidnei Barbieri[1], Ágney Lopes Roth Ferraz[1] and Lourenço Alves Pereira Júnior[1]**

[1]Divisão de Ciência da Computação – Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50 – 12228-900
São José dos Campos – SP – Brasil

sidneisb@ita.br, roth@ita.br and ljr@ita.br

***Abstract.*** *Cybersecurity literature reviews depend on a stable denominator: the papers eligible for screening before synthesis begins. In practice, researchers reconstruct this denominator from changing portals, APIs, and exports, making the resulting corpus difficult to audit or reuse. TopVenues makes corpus construction an executable research artifact. Given a declared venue and year scope, it normalizes DBLP metadata, enriches records with abstracts and BibTeX, and materializes a monotonic SQLite snapshot exposed through a command-line interface, a web application, ranked search, and review-oriented exports. The release declares its venue policy, distinguishes bibliographic records from abstract-enriched records, and records field-level provenance. A reviewer can verify a pinned snapshot offline, exercise ranked search and exports, and reproduce the snapshot-bound workflows included with the release. The tool therefore provides both a usable literature workbench and a fixed, inspectable population for cybersecurity measurement.*

## 1. Problem and Motivation

A literature review begins before screening: researchers must first define which papers could enter the review. For topics such as intrusion detection, Cyber Threat Intelligence (CTI), or Large Language Model (LLM) security, this step spans the ACM Digital Library, IEEE Xplore, USENIX, NDSS, and scholarly APIs. Each source has different query semantics and export formats. A recent CTI survey, for example, required manual curation across hundreds of papers from leading security venues [Barbieri et al. 2025]. That work precedes the scientific synthesis, yet it is rarely packaged as a reproducible object.

Existing tools address adjacent stages. Rayyan [Ouzzani et al. 2016] and ASReview [van de Schoot et al. 2021] prioritize or coordinate screening after a candidate set exists. Zotero [Center for History and New Media 2006] manages collected references. DBLP [Ley 2009] provides stable bibliographic metadata but not abstracts. Semantic Scholar [Ammar et al. 2018], OpenAlex [Priem et al. 2022], and S2ORC [Lo et al. 2020] provide broad scholarly data without declaring a security-venue denominator. None of these systems records the full decision boundary from venue scope to a local, versioned corpus.

*TopVenues* fills this gap. Its input is a declared set of venues and years. Its output is a versioned corpus whose records, enrichment state, provenance, and exports can be

inspected offline. Researchers can therefore share the same population rather than only the final list of included papers. This paper presents the tool in the **Open-Source** category and evaluates the released artifact and researcher-facing workflows.

## 2. Architecture and Research Workflows

*TopVenues* centers every operation on a stable corpus snapshot. Figure 1 contrasts this design with an ad hoc workflow. Portal queries and manual merges create an implicit, mutable denominator. *TopVenues* instead binds a declared scope, a DBLP metadata spine, enrichment provenance, and snapshot lineage before exposing the corpus to search, export, or measurement.

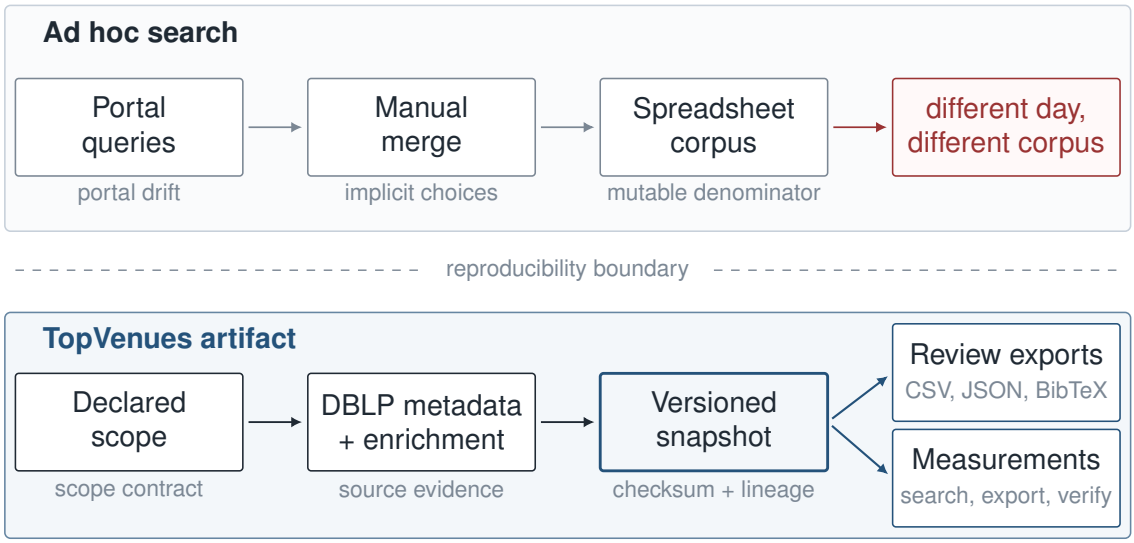


**Figure 1. Ad hoc search reconstructs a mutable corpus; *TopVenues* materializes a versioned denominator for review exports and measurements.**

**Acquisition and consolidation.** The pipeline has four incremental stages. `download` caches DBLP JSON records for the configured venues and years. For large expansions, `materialize-from-dump` derives the same records from the public DBLP XML dump without depending on the live API. `consolidate` normalizes venue labels, deduplicates records by DBLP key, and performs SQLite upserts. The update rule uses `COALESCE`: a missing value in a later source cannot erase an abstract already present. `extract` queries registered abstract sources in priority order, while circuit breakers suppress unstable sources. Finally, `bibtex` retrieves DBLP BibTeX and falls back to a local generator; the dump-based path supports bulk materialization.

The abstract registry currently integrates Semantic Scholar, OpenAlex, CrossRef, the ACM Digital Library, IEEE Xplore, USENIX, and NDSS. Every adapter implements the same contract and either returns a validated abstract or no result. Adding a source does not change pipeline orchestration. This separation is important operationally: live enrichment may fail, but a released snapshot remains usable and reproducible offline.

**Snapshot and lineage.** SQLite is the sole persistent database. The repository distributes `papers.db.gz`; first use materializes the working database locally. A sync identifier

records the hashes of the compressed and working copies. Read-only users receive snapshot updates automatically, while users with local changes receive an explicit warning before replacement. The resulting file can be versioned, cited, copied, and queried without a server.

**Search and export.** Substring filters answer which records *mention* a term. Ranked search (`search --rank`) answers which records are most likely *about* that term. An FTS5 index covers titles, abstracts, and authors; BM25 gives title matches greater weight and ranks results deterministically. The index builds in well under a second for the full snapshot and remains synchronized through database triggers. Results can be exported as BibTeX, CSV, or JSON, which connects corpus definition directly to screening and writing workflows.

**Transparent author visibility.** Researchers also ask which authors recur within a topic or area. The `authors` command reports a declared, venue-weighted visibility heuristic: papers in CCS, IEEE S&P, USENIX Security, or NDSS receive weight 5.0; other configured top-tier venues receive 3.0; survey journals receive 2.0; strong venues receive 1.5; and workshops receive 0.5. The registry is a tool configuration, not a universal quality judgment. Each row exposes the underlying paper count, top-four count, recorded awards, active years, and main venues. DBLP identity suffixes are preserved to avoid merging homonyms. The output should guide corpus exploration, not evaluate individual researchers.

**Topic trends.** The `trends` command traces a topic over time: papers per year, the topic's share of each year's corpus, and the venues that publish it most. Share normalizes by yearly corpus size, so overall corpus growth does not masquerade as topic growth; counts are a lower bound outside the abstract-enriched layers. Together with `authors`, this answers three recurring review questions from the same fixed denominator: who works on a topic, how it evolves, and where it is published.

The Streamlit application exposes the same operations through interactive search, coverage panels, topic trends, author visibility, and one-click exports. Table 1 summarizes the corresponding command-line workflows.

## 3. Interactive Use and Research Application

The interactive application is the reviewer-facing surface of the same local database used by the command-line interface. A researcher first verifies the named profile, then searches within that fixed population rather than relying on an unrecorded portal query. Figure 2 shows a live run against the released `security-20` profile: the sidebar records the abstract-text query and scope controls, the header exposes the corpus state, and the result panel keeps CSV, JSON, and BibTeX export adjacent to the candidate set.

For a literature review, this supports a concrete, bounded workflow. The researcher formulates a query such as *intrusion detection*, restricts venue, year, or paper class when justified by the review protocol, and inspects the resulting records and abstracts. The export carries the selected records into the review's screening and citation workflow. The

**Table 1. Primary *TopVenues* command-line workflows.**

| Command | Purpose |
|---|---|
| `topvenues download` | Fetch DBLP records by venue and year |
| `topvenues consolidate` | Normalize and upsert records into SQLite |
| `topvenues extract` | Enrich records with available abstracts |
| `topvenues bibtex` | Materialize canonical BibTeX entries |
| `topvenues search` | Filter or rank papers by relevance |
| `topvenues authors` | Inspect author visibility by topic or area |
| `topvenues trends` | Trace a topic's yearly volume, share, and venues |
| `topvenues export` | Export filtered BibTeX, CSV, or JSON |
| `topvenues export-hf` | Build Parquet data and a dataset card |
| `topvenues stats` | Report corpus and coverage statistics |
| `topvenues web` | Launch the local web application |

tool does not decide relevance, infer inclusion criteria, or write a synthesis: it makes the pre-screening population, query parameters, and exported inputs inspectable.

The same interface supports corpus audit and descriptive exploration. Its Insights view reports venue and year distributions, paper classes, and abstract coverage from the named snapshot; author visibility and topic trends expose their assumptions rather than presenting an authority ranking. These views help a researcher detect scope gaps, state what population a review considered, and generate reproducible inputs for subsequent scientific analysis. They do not turn a search result into a research conclusion.

## 4. The Corpus as an Artifact

The released snapshot contains 20,305 papers over 2017–2026, refreshed against DBLP in July 2026 so that the 2025 proceedings are nearly complete. The scope is deliberately cybersecurity-only: 17 security venues (including the four classified as CORE A* [CORE 2023]) and 3 survey venues that anchor security literature reviews. Keeping the scope focused, rather than mixing in off-topic AI, systems, or networking venues, is what keeps abstract coverage high: 17,491 of 20,305 records carry an abstract (86.1%), including 14,290 of the 16,806 security-core records (85.0%). Every record has a canonical BibTeX entry. Table 2 exposes the largest remaining gaps rather than hiding them through record deletion.

Missing abstracts remain searchable by title, venue, year, author, and DBLP key; they are excluded only from abstract-dependent queries. Every record has a canonical BibTeX entry. The deterministic `export-hf` command also builds a Parquet mirror and dataset card from the frozen SQLite state:

```
from datasets import load_dataset
corpus = load_dataset("sidneibarbieri/topvenues", split="train")
security = corpus.filter(
    lambda paper: paper["area"] == "security"
)
```

Bibliographic metadata follows DBLP's CC0 terms. Original abstract text remains subject to its source's rights and terms; the artifact documents provenance and does not claim ownership of that text.

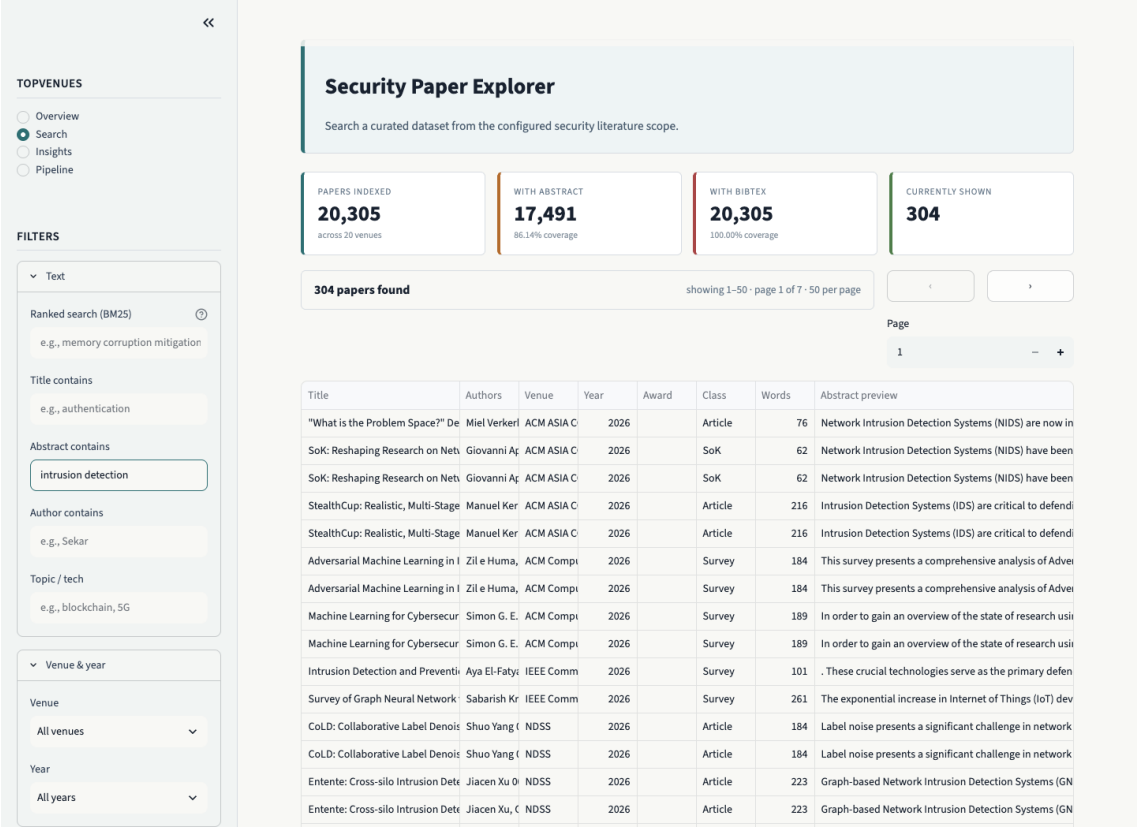


**Figure 2. Live abstract-text search for *intrusion detection* in the released `security-20` profile. Every displayed row has an abstract preview, and the application exposes the resulting candidate set for inspection and export; screening remains a researcher decision.**

## 5. Evaluation and Artifact Readiness

The evaluation tests the properties required by a working review corpus: metadata completeness, query behavior, offline reproducibility, and explicit scope and snapshot identity.

**Search behavior and profile isolation.** The reproduction script builds the FTS5 index and exercises both substring and BM25-ranked search on a freshly materialized copy of the declared snapshot. It reports timings for the reviewer's machine, but the artifact does not treat machine-dependent latency as a scientific result. The two search modes are tested for nonempty, deterministic results; their result counts can differ because substring matching accepts fragments whereas ranked search applies token semantics.

The release reports only evidence produced from its own 20,305-record snapshot: coverage, search, exports, and checksum verification. Any future analysis must name the snapshot hash that produced its denominator; values from another corpus are not transferable merely because the software is shared.

**Reviewer-path validation.** The public artifact packages the compressed snapshot, source code, Dockerfile, reviewer guide, citation metadata, and automated tests. From a fresh clone, `bash reproduce.sh -profile security-20` verifies the headline

**Table 2. Abstract coverage in the security core (July 2026 snapshot).**

| Venue | Papers | Abstracts | Coverage |
|---|---|---|---|
| ACM CCS | 4,329 | 4,090 | 94.5% |
| USENIX Security | 2,542 | 2,056 | 80.9% |
| TrustCom | 2,311 | 1,914 | 82.8% |
| NDSS | 1,535 | 1,534 | 99.9% |
| IEEE S&P | 1,420 | 1,165 | 82.0% |
| ACM ASIA CCS | 1,062 | 957 | 90.1% |
| ESORICS | 598 | 39 | 6.5% |
| ACSAC | 593 | 593 | 100.0% |
| Other nine venues | 2,416 | 1,942 | 80.4% |
| **Security core** | **16,806** | **14,290** | **85.0%** |

counts and snapshot hash, runs 238 tests offline, rebuilds the FTS5 index, exercises both search paths, and exports a BibTeX sample. This path tests the object a reviewer receives rather than a development checkout. Table 3 maps these checks to the artifact dimensions named by the SBSeg evaluation process; the table reports evidence, not self-awarded badges.

**Table 3. Evidence for the SBSeg artifact-evaluation dimensions.**

| Dimension | Verifiable evidence |
|---|---|
| Available | Public MIT-licensed repository, compressed SQLite snapshot, Dockerfile, reviewer guide, and citation metadata. |
| Functional | CLI and web workflows, ranked search, author visibility, exports, DBLP ingestion, abstract/BibTeX enrichment, and 238 automated tests. |
| Reproducible | One offline command checks the named snapshot, exercises search, and produces a sample export without external sources. |
| Sustainable | Typed Python/SQLite implementation, modular source adapters, configuration-driven venue registry, lineage hashes, and CI on Python 3.11 and 3.12. |

The main operational limitation is external-source drift. Publishers change page structure, APIs impose rate limits, and some papers have no openly available abstract. *TopVenues* isolates this uncertainty in live enrichment. Search, statistics and export reproduce from the released snapshot even when every external source is unavailable.

## 6. Planned Demonstration

The demonstration follows a researcher's path from verification to use on a laptop with Python 3.11+ and a web browser. After the initial clone, the reproduction path requires no network connection:

```
git clone https://github.com/sidneibarbieri/topvenues-tool
cd topvenues-tool
bash reproduce.sh --profile security-20
```

The script materializes the database, runs all tests, exercises both search paths, and exports sample BibTeX. We then use `topvenues stats` to inspect venue coverage; run ranked search for "intrusion detection"; inspect venue-stratified author visibility for "LLM"; and export the selected papers to BibTeX. The final step repeats the workflow in the Streamlit application, including interactive search, coverage, author evidence, and CSV/JSON/BibTeX export.

The demonstration requires one author laptop, a browser, and an HDMI display or projector. The artifact itself runs locally. A technical video covering installation, verification, and these workflows is available at: the immutable release.

## 7. Availability, Lineage, and Sustainability

The open-source tool and snapshot are available under the MIT license at release `sbseg2026-sf-submission-r1`. Its Hugging Face export is generated from the same named SQLite snapshot. The dataset card records the profile identifier, snapshot hash, and source release tag, so a reviewer can reject an export that does not match the artifact.

The pipeline supports monotonic updates. Running `topvenues run-all` after a DBLP release adds new records without erasing existing enrichment. The venue scope is declared in `config.yaml`; source adapters and venue mappings are additive extension points. The runtime uses Python, SQLite, and widely maintained scientific libraries rather than a managed service.

## 8. Related Tools

Rayyan [Ouzzani et al. 2016] and ASReview [van de Schoot et al. 2021] support collaborative screening and active-learning prioritization after researchers assemble a candidate corpus. *TopVenues* operates at the preceding boundary: it declares and materializes that corpus. DBLP [Ley 2009], Semantic Scholar [Ammar et al. 2018], OpenAlex [Priem et al. 2022], and S2ORC [Lo et al. 2020] are scholarly data sources, whereas *TopVenues* packages a venue-bounded local snapshot, its enrichment state, ranked search, and review exports. Zotero organizes references already collected; *TopVenues* produces the auditable population and BibTeX that such managers consume. Its reviewer guide also follows recent guidance that security artifacts should expose the problem, domain, method, data, and analysis needed for replication reasoning [Olszewski et al. 2025].

## 9. Conclusion

*TopVenues* turns cybersecurity corpus construction into an executable and citable research object. It binds a declared venue scope to normalized DBLP metadata, monotonic enrichment, a versioned SQLite snapshot, ranked search, transparent author evidence, and review-oriented exports. The current artifact contains 20,305 cybersecurity and security-relevant papers with 86.1% abstract coverage (85.0% on the 16,806-paper security core) and BibTeX for every record. More importantly, a fresh public clone can verify those profile-bound properties without borrowing results from the separately frozen companion-paper corpus. The contribution is therefore not another search portal, but a stable boundary between corpus construction, literature analysis, and auditable cybersecurity measurement.

## Acknowledgments

This work is funded by FAPESP #2020/09850-0, #2024/21006-1, #2022/00741-0; CNPq 312294/2025-5, 409743/2025-9; and CAPES.

## References

Ammar, W. et al. (2018). Construction of the literature graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 84–91, New Orleans, Louisiana. Association for Computational Linguistics.

Barbieri, S., De Souza, F. L. D. S., Teixeira, M. A., Marcondes, C. A. C., and Pereira, L. A. (2025). Searching for diamonds: Cross-domain opportunities in cyber threat intelligence. *IEEE Access*, 13:189554–189588.

Center for History and New Media (2006). Zotero: A free, easy-to-use research tool. Available: `https://www.zotero.org`.

CORE (2023). CORE – computing research and education: Conference rankings. Available: `https://core.edu.au/conference-portal`.

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Olszewski, D., Tucker, T., Butler, K. R. B., and Traynor, P. (2025). SoK: Towards a unified approach to applied replicability for computer security. In *34th USENIX Security Symposium*, pages 469–488. USENIX Association.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):210.

Priem, J., Piwowar, H., and Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

van de Schoot, R. et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3:125–133.