# The Information Retrieval Anthology

Martin Potthast[1], Sebastian Günther[2], Janek Bevendorff[3], Jan Philipp Bittner[2], Alexander Bondarenko[2], Maik Fröbe[2], Christian Kahmann[1], Andreas Niekler[1], Michael Völske[3], Benno Stein[3] and Matthias Hagen[2]

[1]*Leipzig University, Leipzig, Germany*
[2]*Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany*
[3]*Bauhaus-Universität Weimar, Weimar, Germany*

### Abstract

In this extended abstract,[4] we present the IR Anthology, a corpus of information retrieval publications accessible at IR.webis.de via a metadata browser and a full-text search engine. Following the example of the well-known ACL Anthology, the IR Anthology serves as a hub for scholars interested in information retrieval. Our search engine ChatNoir indexes the publications' full texts, enabling a focused search and linking users to the respective publisher's site for personal access.

### Keywords

Scientific literature, bibliography, scholarly search

## 1. Introduction

The Information Retrieval Anthology, or IR Anthology for short, compiles scientific publications on the subject of information retrieval. Published online as a metadata search and browsing tool, it provides the information retrieval community with a comprehensive overview of its own body of publications and eases scholarly search within a *closed-world* environment.

The search results of generic academic search engines contain a mixture of publications from various fields. For instance, the query `"query processing"` may yield publications from the perspectives of both databases and information retrieval. A user particularly interested in the topic of query processing in IR can improve the precision of the results by adding terms that frequently co-occur with the term `"query processing"` in IR-related publications but not with others. Yet, even when using more specialized queries, generic search engines may still rank off-topic publications higher than on-topic ones, since, for instance, a paper's global "importance" in terms of citations or recency may exceed the importance of a term-based ranking signal. Furthermore, generic academic search engines generally do not allow pagination of their search results beyond the initial top-1000. Altogether, this reduces the retrievability of

[4] Condensed version of a demo paper at SIGIR 2021 [1]; further ideas in a follow-up SIGIR Forum article [2].

*LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany*

martin.potthast@uni-leipzig.de (M. Potthast); sebastian.guenther@informatik.uni-halle.de (S. Günther); janek.bevendorff@uni-weimar.de (J. Bevendorff); jan.bittner@student.uni-halle.de (J. P. Bittner); alexander.bondarenko@informatik.uni-halle.de (A. Bondarenko); maik.froebe@informatik.uni-halle.de (M. Fröbe); christian.kahmann@uni-leipzig.de (C. Kahmann); andreas.niekler@uni-leipzig.de (A. Niekler); michael.voelske@uni-weimar.de (M. Völske); benno.stein@uni-weimar.de (B. Stein); matthias.hagen@informatik.uni-halle.de (M. Hagen)

CEUR Workshop Proceedings (CEUR-WS.org)

contributions without a sufficient number of citations which would render them “important” enough to outrank more recent or more frequently-cited (and potentially off-topic) work.

A dedicated IR Anthology and an accompanying retrieval system tailored to the IR community has the potential to become particularly useful, helping to mitigate some of the biases introduced by generic academic search engines. Although the individual IR scholar cannot be relieved from reviewing the relevant publications on their subfields of interest (even from beyond IR), a search engine that exclusively indexes the IR Anthology yields results with a higher precision, constituting a valuable addition to the scholarly tool set.

## 2. Corpus, Interface & Search Engine

The IR Anthology is based on a BibTeX database of metadata on IR publications, the respective full text documents, and a website on which the metadata can be browsed and the document collection be searched.

**Corpus Construction.** Compiling a complete corpus for the IR Anthology is not trivial. We started by exploiting an existing classification to bootstrap our corpus. All metadata for publications at 19 conferences and 11 journals that primarily specialize in information retrieval or that are very closely related are collected from a recent DBLP XML dump. To acquire the full texts of these publications, we use the Webis-CSP-15 corpus [3] but also searched and crawled more recent or otherwise missing publications. This process is ongoing, since especially “older” publications can be difficult to be obtained. Going forward, members of the IR community may later supply the IR Anthology with copies from their own collections.

**Browsing the Anthology.** To bootstrap the web-based metadata browser, we follow and build upon the example of the ACL Anthology.[5] But simply reusing their website’s source with minimal changes to enable the exchange of bug fixes both ways was not possible due to significant changes that were necessary for our deployment.

Our revised web interface has four basic views: (1) landing page with an overview of all conferences, journals, workshops, summer schools, and societies; (2) volume pages listing all publications belonging to a selected venue and year; (3) publication pages showing metadata about a given publication; and (4) author profile pages with their respective publications. In listings of publications, the IR Anthology shows basic information like title and authors, and also directly links to the full text PDF on the publisher’s site, if a DOI is available, and to the BibTeX entry at DBLP. On the individual publication pages, further links allow for searching the respective publication’s title at Google Scholar, Microsoft Academic, or Semantic Scholar.

**Searching the IR Anthology.** To allow users to easily search the IR Anthology, we provide a dedicated search index accessible via our search engine ChatNoir [4]. We extracted the contents of all available papers using GROBID [5] and indexed titles, abstracts, full-text bodies, as well as metadata like authors, venue, year, and DOI in a BM25F setup [6]. Users are able to perform full-text search across all fields and can filter by individual metadata using operators.

Integrating other search functionalities like starting a related work search from a set of known publications [3] or simply offering different retrieval models are next steps on our roadmap. In particular for the comparison of different retrieval models, the organization of a shared task “IR 4 IR: Information Retrieval for Information Retrieval” may be promising.

[5] Its website is available open source at https://github.com/acl-org/acl-anthology

## References

[1] M. Potthast, S. Günther, J. Bevendorff, J. P. Bittner, A. Bondarenko, M. Fröbe, C. Kahmann, A. Niekler, M. Völske, B. Stein, M. Hagen, The Information Retrieval Anthology, in: 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021), ACM, 2021. URL: https://dl.acm.org/doi/10.1145/3404835.3462798. doi:10.1145/3404835.3462798.

[2] M. Potthast, B. Stein, M. Hagen, The information retrieval anthology 2021, SIGIR Forum 55 (2021).

[3] M. Hagen, A. Beyer, T. Gollub, K. Komlossy, B. Stein, Supporting Scholarly Search with Keyqueries, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. Di Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016), volume 9626 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2016, pp. 507–520. doi:10.1007/978-3-319-30671-1_37.

[4] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl, in: L. Azzopardi, A. Hanbury, G. Pasi, B. Piwowarski (Eds.), Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2018.

[5] GROBID, https://github.com/kermitt2/grobid, 2008-2021. arXiv:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

[6] S. Robertson, H. Zaragoza, M. Taylor, Simple BM25 extension to multiple weighted fields, in: D. Grossman, L. Gravano, C. Zhai, O. Herzog, D. Evans (Eds.), Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004, ACM, 2004, pp. 42–49. URL: https://doi.org/10.1145/1031171.1031181. doi:10.1145/1031171.1031181.