The Information Retrieval Anthology

Martin Potthast,* Sebastian Günther,† Janek Bevendorff,‡ Jan Philipp Bittner,† Alexander Bondarenko,† Maik Fröbe,† Christian Kahmann,* Andreas Niekler,* Michael Völske,‡ Benno Stein,‡ Matthias Hagen†
*Leipzig University †Martin-Luther-Universität Halle-Wittenberg ‡Bauhaus-Universität Weimar

ABSTRACT

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.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Scientific literature analysis; bibliography; scholarly search

ACM Reference Format:
Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, Matthias Hagen. 2021. The Information Retrieval Anthology. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3404835.3462798

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 (1) provides the information retrieval community with a comprehensive overview of its own body of publications, (2) eases scholarly search within a *closed-world* environment, and (3) facilitates community introspection via quantitative publication analysis.

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, e.g., 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 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. Considering the growing body of publications from the IR community over the years in a wide array of subfields, staying on top of it in breadth and depth demands supporting information systems more than ever. Fortunately, the IR community is fittingly specialized and equipped with the expertise to support itself with the latest state-of-the-art search technology directly out of its labs.

In this paper, we describe the elements of the IR Anthology and its search engine ChatNoir,[1] present a basic evaluation and corpus analysis,[2] and briefly touch some ideas for future developments.

2 RELATED WORK

In a recent SIGIR Forum opinion article, Hiemstra et al. [28] make the case for "Transitioning the Information Retrieval Literature to a Fully Open Access Model", observing that various research communities thrive in such a setting. The ACL Anthology,[3] which for nearly two decades has maintained an open archive of the computational linguistics and natural language processing literature published at various venues, is a particularly salient example, and serves as the main inspiration and basis for our initiative. After reviewing related endeavors from among the ACL Anthology and its offspring projects, we present a wider context of scholarly information utilities, both generic and specific to other fields. The table in Figure 1a compares a selection of popular services.

The ACL Anthology is an online platform that provides a curated collection of publications from the computational linguistics and natural language processing [17]. From a table-based overview, it enables easy access to publication lists by venue, year, or both. The ACL Anthology's open archives have enabled a thriving ecosystem of research projects on academic literature search and exploration, among them the ACL Anthology Searchbench [51], NLP Scholar [39, 40], NLPExplorer [47], LT Expert Finder [14], or Talk to Papers [61].

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462798

[1] Code: https://github.com/ir-anthology and https://github.com/chatnoir-eu
[2] Code and Data: https://github.com/webis-de/SIGIR-21
[3] https://www.aclweb.org/anthology/

(a)

Service	Launch	Scope	Bib.	OA	Search	Social	References
MEDLINE	1964	S	C		M, C		[3]
Web of Science	1964	G	C		M		[7, 15]
arXiv	1991	S	C	✓	M, T		[54]
ACM Digital Library	1993	S	C		M, T, C	P	[3]
DBLP	1993	S	C	✓	M	P	[35]
The Coll. of CS Bib.	1995	S	C	✓	M		
PubMed	1996	S	C		M, T, C		[36]
Math. Genealogy Pr.	1996	S	C	✓	M		[13, 29, 42]
CiteSeer^x	1997	S	C, F	✓	M, T		[18, 58]
CoRR	1998	S	C	✓	M, T		[22]
Crossref	1999	G	C		M		[25]
ACL Anthology	2002	S	C	✓	M, T		[17]
Google Scholar	2004	G	C		M, T	P	[19]
Bibsonomy	2006	G	F		M, T	P, D	[5]
Microsoft Academic[4]	2006	G	C		M, T	P	[52]
Zotero	2006	G	P		M, T	P, D	[41]
Academia.edu	2008	G	F		M, (T)	P, D	[32]
Mendeley	2008	G	C, P		M	P, D	[26, 59]
ResearchGate	2008	G	F		M, T	P, D	[32, 46]
ORCID	2012	G	C		M	P	[12]
Semantic Scholar	2015	G	C		M, T	P	[9, 37]
IA Scholar	2021	G	C	✓	M, T		[43]
IR Anthology	2021	S	C		M, T		

(b)

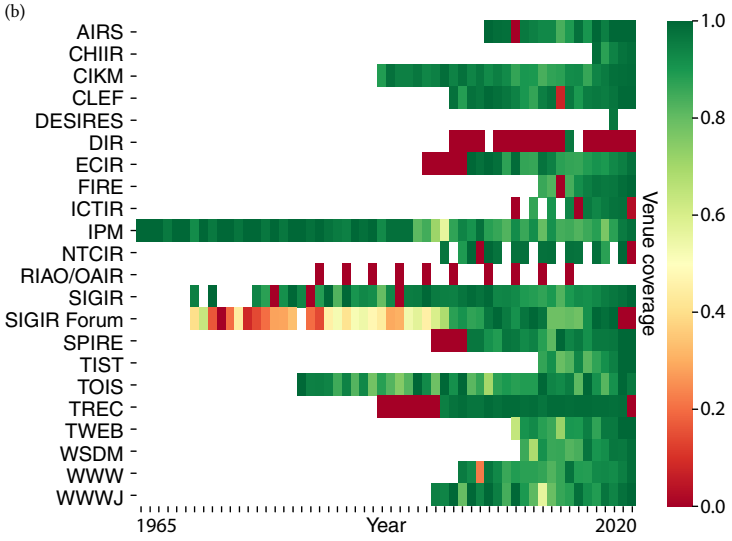

Figure 1: (a) Popular scholarly information utilities by launch year, depicting their scope as field-specific (S) or generic (G); bibliography management (Bib.) as central database (C), "folksonomy" (F), or personal database (P); 100 % open-access content (OA); search facilities for metadata (M), full-texts (T), or based primarily on controlled vocabularies (C); and social networking via author profiles (P) or discussions (D). (b) Current coverage of the IR Anthology in terms of URLs for PDFs and DOIs.

Beyond targeting literature exploration itself, projects built on the ACL Anthology have investigated scientometric research questions concerning large-scale and long-term trends in NLP research [38], or temporal bias in citation patterns [10]. Such studies are of interest to the IR community as well [27], and we hope that the IR Anthology corpus will facilitate them going forward. Due do being open-access, the ACL Anthology can very straightforwardly implement search using nothing more than a general-purpose web search engine's `site`-operator (via Google Custom Search). With most publications from IR currently not openly accessible, we provide a custom search index for the IR Anthology instead.

A number of services implement search in scientific publications, of which Google Scholar is the longest established, with the most comprehensive index [20, 23]. Other contenders include Microsoft Academic based on a large-scale entity graph [52],[4] Arnetminer [56], Semantic Scholar, and the associated Semantic Scholar Open Research Corpus of 80 million papers [37]. More specialized search engines focus, e.g., on dataset retrieval [2, 11], or, as of recently, search in publications hosted at the Internet Archive [43].

Notable among a great variety of other academic information utilities are bibliographic databases like DBLP [35]—whose open metadata supports our efforts—as well as crowdsourced "folksonomy-style" [5], and personal bibliography alternatives like Mendeley [26, 59] and Zotero. ResearchGate and Academia.edu address, in part, a similar purpose, but are primarily academic social networks [32, 46]. Preprint servers have long been an important part of the open-access ecosystem: the physics-focused arXiv [54] has been active for three decades, its offshoot Computing Research Repository (CoRR) [22] for two. Much of other fields' bibliographic information resides in large centralized databases (e.g., MEDLINE for the life sciences [3]), while an endeavor like The Mathematics Genealogy Project has a unique focus on thesis–advisor relations to trace who taught whom throughout math history [13, 29, 42].

[4]The shutdown of Microsoft Academic by the end of 2021 has been announced:

3 CORPUS, INTERFACE & SEARCH ENGINE

The IR Anthology is based on a number of components and design decisions: 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 searched. A key design goal of the IR Anthology is to fit in with and connect to relevant existing services, rather than starting from scratch.

Corpus Construction. The ultimate goal of the IR Anthology is to encompass *all* publications on information retrieval. Compiling a complete corpus, however, is not trivial. The typical first step in doing so is a bibliometric field delineation [62], combining manual or semi-automatic heuristics to determine a given field's "boundary" and whether or not a given publication belongs to the field. Three different kinds of heuristics are employed in practice, namely (1) exploiting existing classification systems, (2) searching scholarly search engines, and (3) analyzing bibliometric networks.

First, we exploited an existing classification to bootstrap our corpus. All metadata for publications at 16 conferences and 6 journals that primarily specialize in information retrieval or that are very closely related are collected from a recent DBLP XML dump (venues shown in Figure 1b). Besides bibliographic data, such as authors, venue, etc., various paper URLs and IDs/keys allow, for instance, to separate workshops from main conference tracks. Different authors with the same name are disambiguated via DBLP's author IDs.

Second, starting from the Webis-CSP-15 corpus comprising 35,000 publications from 20 top-tier conferences, plus their respective references, for a total of 200,000 publications [21],[5] we then searched and crawled copies of the missing ones, both from within our respective universities, and without. 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. As the

[5]https://webis.de/data.html#webis-csp-15

process of collecting all publications from what could be called the "core" venues of information retrieval will continue, further heuristics may be explored: for example, including venues listed by IR societies or identifying IR tracks at non-IR venues.

Browsing the Anthology. To bootstrap the web-based metadata browser, we follow and build upon the example of the ACL Anthology.[6] The goal to reuse their website's source with minimal changes to enable the exchange of bug fixes both ways could not be reached due to hard-coded variables and customized deployment procedures. Our revised web interface has four basic views: (1) landing page with an overview of all conferences and journals, giving direct access to their individual (proceedings) volumes; (2) volume page, which lists all papers belonging to the proceedings of a conference or the issues of a journal in a given year; (3) publication page, showing metadata about a given publication; (4) author page, showing all publications by a given author. The listings of publication entries on both the volume page and the author page display basic information about a publication like title and authors, and also directly link to its full text PDF on the publisher's site, if a DOI is available, and to its BibTeX entry at DBLP. On a publication page, 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 [6], a proven web search engine indexing around 5 billion web pages from the ClueWeb crawls and Common Crawl versions. We extracted the contents of all available papers using GROBID [1] and BM25F-indexed [50] titles, abstracts, and full-text bodies as fields, as well as additional metadata such as authors, venue, year, and DOI. Users are able to perform full-text search across all fields and can filter by individual metadata using keywords. Our experience with ChatNoir taught us the importance of titles in document retrieval, so that we also apply the highest weights to title and abstract matches and slightly lower weights to the body and other fields. The results contain snippets whenever available and link back into the IR Anthology. The IR Anthology can be searched via a conventional web interface at IR.chatnoir.eu, and via a simple yet powerful REST API.

Corpus Statistics, Discussion, and Limitations. At the time of writing, the IR Anthology covers 40,933 publications and has links to the publishers' full text for 35,763 of them (88 %, per venue shown in Figure 1b). ChatNoir indexes the full texts of these 88 % plus titles and sometimes the abstract for the remainder. Unlike for the ACL Anthology, most of the full texts in the IR Anthology cannot be publicly shared due to copyright restrictions. This means that Google's powerful `site`-operator cannot be used as a search engine against the IR Anthology, which is why we offer our own search engine. Eventually, we plan to organize a shared task via our TIRA platform [48], enabling the community to develop their own, improved search engines without the need to share public access to the corpus. Regarding metadata, we plan to further build on top of DBLP, which contains only few mistakes (e.g., coverage gaps for venues, incomplete titles, etc.) which we report back to them. We expect a future expansion to venues not covered by DBLP, which ideally could also be fed back into their database.

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

Table 1: (a) Effectiveness of ChatNoir (CN), Google Scholar (GS), and GS with "information retrieval" appended to queries (GS_{IR}). (b) Rank of a paper (Ref.) relevant to the aligned query in (a) when searching for its title.

(a)

Query	Precision@10		
	CN	GS	GS_{IR}
`semantic search`	1.0	1.0	1.0
`query understanding`	1.0	0.9	0.5
`link prediction`	1.0	0.9	1.0
`conversational systems`	1.0	0.7	1.0
`health search`	1.0	0.1	1.0
`content recommendation`	0.9	1.0	1.0
`question answering`	0.9	0.8	0.9
`social media search`	0.9	0.6	0.9
`neural network retrieval model`	0.9	0.5	0.9
`query processing`	0.6	0.1	0.4
Average	0.9	0.7	0.9

(b)

Known-item search		
Ref.	CN	GS
[16]	1	1
[4]	1	1
[33]	1	1
[60]	1	1
[31]	1	1
[30]	1	1
[24]	1	1
[49]	1	1
[53]	1	1
[57]	1	1
Avg.	1	1

4 EVALUATION AND CORPUS ANALYSIS

We evaluate ChatNoir's search for the IR Anthology in a Cranfield-style setup and present a first quantitative literature analysis.

4.1 ChatNoir Evaluation

We compare the search effectiveness of ChatNoir to searching on Google Scholar with and without adding the term `information retrieval` to the query as an indication of IR focus. As queries, we randomly selected ten topics related to the SIGIR 2021 call for papers. For each query, the top 10 results of each system were judged in random order by an experienced IR researcher as either relevant to a related work search in the respective sub-field of information retrieval (e.g., query understanding) or as irrelevant, belonging to another research area (e.g., linguistics or psychology).

Table 1a shows the systems' precision@10 scores. ChatNoir (CN) achieves perfect precision on half of the queries. Google Scholar (GS) often yields results unrelated to IR: For instance, for the query `health search` publications from psychology [34], biology [55], and from other unrelated fields are returned. Even the query `neural network retrieval model` returns publications from neuropsychology [45]. Adding the term IR (GS_{IR}) substantially improves the results across all queries with one exception: For the query `query understanding` GS_{IR} actually has a lower effectiveness since many text books with little to no content on query understanding are returned. All systems struggle a bit with the query `query processing`, where even ChatNoir returns some hits from conference tracks unrelated to IR (e.g., from CIKM). In fact, the non-IR tracks at some conferences included in the IR Anthology might need to be identified and omitted to improve precision in such cases. Finally, we also experimented with known-item search to demonstrate that specific publications can be found. We randomly selected one paper for each of the above topics from the IR Anthology and queried ChatNoir and Google Scholar with their titles (Table 1 right). Both returned each of the ten papers at the first rank.

The above study is not meant as a comprehensive comparison of ChatNoir to other academic search systems; it rather demonstrates that searching in a domain-specific focused collection provides more accurate results, even when using basic retrieval models.

(a)

Topic Keywords (top 5)	**Topic Category**	**Trend**
metric, function, score, rank, ranking	evaluation+users	evergreen
student, health, group, study, course	evaluation+users	evergreen
similarity, algorithm, cluster, distance, space	modeling	evergreen
search, user, session, task, behavior	evaluation+users	neogreen
question, answer, tag, expert, qa	evaluation+users	neogreen
image, video, music, multimedia, photo	multimodal/media	neogreen
feature, classify, learning, training, label	modeling	rising
recommend, user, item, preference, filtering	modeling	rising
review, product, opinion, sentiment, trust	multimodal/media	rising
privacy, security, email, app, attack	multimodal/media	rising
ad, advertising, game, worker, market	multimodal/media	rising
location, city, sensor, mobility, region	multimodal/media	rising
event, news, tweet, medium, blog	multimodal/media	rising
network, graph, node, community, edge	multimodal/media	rising
evaluation, collection, test, retrieval, judgment	evaluation+users	peaked
document, topic, sentence, text, summarization	modeling	peaked
word, language, translation, text, speech	modeling	peaked
term, query, document, retrieval, model	modeling	peaked
page, web, search, engine, query	web	peaked
service, web, ontology, application, integration	web	peaked
index, xml, query, tree, database	systems	peaked
peer, server, load, latency, memory	systems	settled
program, system, software, programming, design	systems	settled
entity, copy, knowledge, fee, permission	meta	
research, workshop, tutorial, researcher, book	meta	

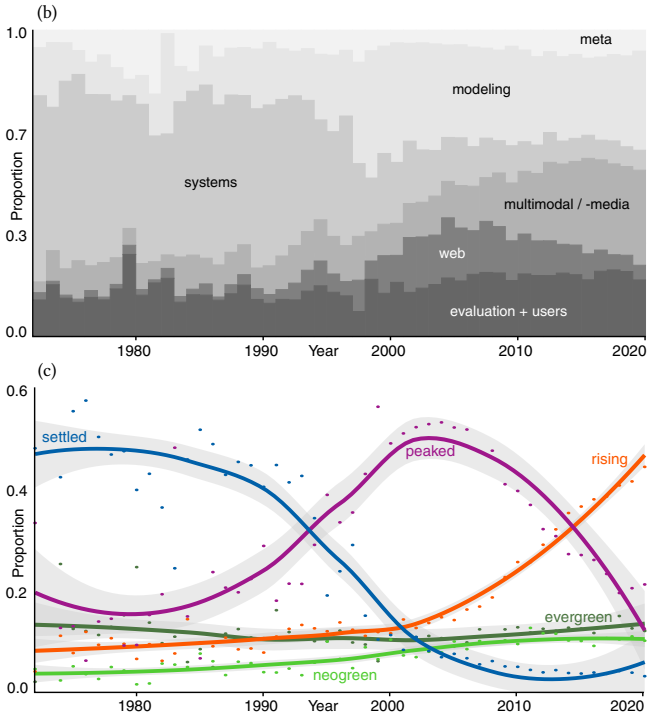

Figure 2: (a) Topics identified as sets of keywords along a manual categorization into groups of related topics, and into popularity trend groups. The categorizations were derived from a visual exploration of topic aggregation plots as shown in (b) and (c). (b) Final aggregation of topics by topic group over time. (c) Final aggregation of topics by trend over time. We identified five different trends, which were named evergreen, neogreen, rising, peaked, and settled as per their shape.

4.2 Mining the IR Anthology

Using the Leipzig Corpus Miner [44], our corpus analysis initialized an LDA topic model [8] with 25 noun-based topics on all abstracts in the IR Anthology. Figure 2a shows the obtained topics. Analyzing each topics' keywords, we manually grouped them into 6 categories relating to IR research. Figure 2b shows a stacked bar chart of the proportion of documents belonging to a given topic category over time. This shows the relative "importance" of a given topic category: IR systems (indexing, software, and efficiency) dominated before the turn of the millennium. The advent of the web caused a paradigm shift towards modeling and web search, whereas the latter was extended with a diverse mix of multimodal/-media topics. Evaluation and user-related research remained stable over time. Two topics relate more to meta-information in scientific literature.

Not all topics follow the same trend over time. Visualizing the relative proportions of each topic as individual line charts, common shapes could be distinguished, which enabled their categorization into five different groups. Figure 2c depicts these shapes as aggregated line charts of topics exerting a common shape. Two of the three systems-related topics have settled over time, whereas one relating to indexing and XML peaked alongside web search-, modeling-, and evaluation-related topics. Rising topics include the ones related to artificial intelligence and multimodal/-media research. Some evaluation and modeling topics continue to be evergreens, and some more evaluation and multimodal/-media topics increased in proportion to become new evergreens, i.e., neogreens.

5 CONCLUSION

To bootstrap an anthology of publications on information retrieval, we collect metadata and full texts from the field's primary venues. Since the majority of IR publications are not available open access, we cannot share them. Nevertheless, we can share access to a full-text search engine to enable visitors to search for papers of interest. The list of venues is not yet complete, and future work will require adding more venues including conferences and workshops organized by IR societies, IR tracks at related venues, preprints, and even papers referenced by any given IR paper—some of which might not be covered by DBLP. Further extensions may include artifacts other than publications, including code, data, videos, slides, posters, and even entire conference websites, IR blogs, and other relevant non-archival publications. Taking further inspiration from the ACL community and NLP-progress,[7] a dedicated listing of all IR-related shared tasks and their results appears worthwhile, too.

The IR Anthology welcomes and facilitates community contributions. In particular, we plan to organize a shared task on building better search engines for it. Through our evaluation-as-a-service platform TIRA we can grant access to full texts without sharing them publicly, so that interested groups are able develop their own search engines that may later be added to the anthology's website. Long-term, a governance model for the anthology's maintenance may also involve others besides us to ensure that our initiative sustainably supports and enhances everyone's daily work routines.

[7] http://nlpprogress.com

REFERENCES

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

[2] Uchenna Akujuobi and Xiangliang Zhang. 2017. Delve: A Dataset-Driven Scholarly Search and Analysis System. *SIGKDD Explor.* 19, 2 (2017), 36–46. https://doi.org/10.1145/3166054.3166059

[3] William Y. Arms. 2000. *Digital Libraries*. MIT Press. http://www.cs.cornell.edu/wya/DigLib/

[4] Ricardo Baeza-Yates. 2017. Semantic Query Understanding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 1357. https://doi.org/10.1145/3077136.3096472

[5] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. 2010. The social bookmark and publication management system bibsonomy - A platform for evaluating and demonstrating Web 2.0 research. *VLDB J.* 19, 6 (2010), 849–875. https://doi.org/10.1007/s00778-010-0208-4

[6] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018) (Lecture Notes in Computer Science)*, Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski (Eds.). Springer, Berlin Heidelberg New York.

[7] Caroline Birkle, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. 2020. Web of Science as a data source for research on scientific and scholarly activity. *Quant. Sci. Stud.* 1, 1 (2020), 363–376. https://doi.org/10.1162/qss_a_00018

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. http://jmlr.org/papers/v3/blei03a.html

[9] John Bohannon. 2016. A Computer Program Just Ranked the Most Influential Brain Scientists of the Modern Era. *Science* (Nov. 2016). https://doi.org/10/gh77gw

[10] Marcel Bollmann and Desmond Elliott. 2020. On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7819–7827. https://doi.org/10.18653/v1/2020.acl-main.699

[11] Dan Brickley, Matthew Burgess, and Natasha F. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 1365–1375. https://doi.org/10.1145/3308558.3313685

[12] Declan Butler. 2012. Scientists: your number is up. *Nat.* 485, 7400 (2012), 564. https://doi.org/10.1038/485564a

[13] Harry B. Coonce. 2004. Computer science and the mathematics genealogy project. *SIGACT News* 35, 4 (2004), 117. https://doi.org/10.1145/1054916.1054918

[14] Tim Fischer, Steffen Remus, and Chris Biemann. 2019. LT Expertfinder: An Evaluation Framework for Expert Finding Methods. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh (Eds.). Association for Computational Linguistics, 98–104. https://doi.org/10.18653/v1/n19-4017

[15] Eugene Garfield. 1964. "Science Citation Index"—A New Dimension in Indexing. *Science* 144, 3619 (May 1964), 649–654. https://doi.org/10/d9qt5m

[16] Darío Garigliotti. 2018. A Semantic Search Approach to Task-Completion Engines. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1457. https://doi.org/10.1145/3209978.3210224

[17] Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current State and Future Directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Melbourne, Australia, 23–28. https://doi.org/10.18653/v1/W18-2504

[18] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA*. ACM, 89–98. https://doi.org/10.1145/276675.276685

[19] Jim Giles. 2005. Science in the Web Age: Start Your Engines. *Nature* 438, 7068 (Dec. 2005), 554–555. https://doi.org/10/dcz432

[20] Michael Gusenbauer. 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118, 1 (2019), 177–214. https://doi.org/10.1007/s11192-018-2958-5

[21] Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Komlossy, and Benno Stein. 2016. Supporting Scholarly Search with Keyqueries. In *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016) (Lecture Notes in Computer Science, Vol. 9626)*, Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.). Springer, Berlin Heidelberg New York, 507–520. https://doi.org/10.1007/978-3-319-30671-1_37

[22] Joseph Y. Halpern. 2000. CoRR: a computing research repository. *ACM J. Comput. Documentation* 24, 2 (2000), 41–48. https://doi.org/10.1145/337271.337274

[23] Anne-Wil Harzing. 2019. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics* 120, 1 (2019), 341–349. https://doi.org/10.1007/s11192-019-03114-y

[24] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. ANTIQUE: A Non-factoid Question Answering Benchmark. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 166–173. https://doi.org/10.1007/978-3-030-45442-5_21

[25] Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quant. Sci. Stud.* 1, 1 (2020), 414–427. https://doi.org/10.1162/qss_a_00022

[26] Victor Henning and Jan Reichelt. 2008. Mendeley - A Last.Fm for Research?. In *Fourth International Conference on E-Science, e-Science 2008, 7-12 December 2008, Indianapolis, IN, USA*. IEEE Computer Society, 327–328. https://doi.org/10/cb9w22

[27] Djoerd Hiemstra, Claudia Hauff, Franciska de Jong, and Wessel Kraaij. 2007. SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community. *SIGIR Forum* 41, 2 (2007), 18–24. https://doi.org/10.1145/1328964.1328966

[28] Djoerd Hiemstra, Marie-Francine Moens, Raffaele Perego, and Fabrizio Sebastiani. 2021. Transitioning the Information Retrieval Literature to a Fully Open Access Model. *SIGIR Forum* 54, 1, Article 13 (Feb. 2021), 10 pages. https://doi.org/10.1145/3451964.3451977

[29] Allyn Jackson. 2007. A labor of love: The mathematics genealogy project. *Notices of the AMS* 54, 8 (2007), 1002–1003.

[30] Vidit Jain and Esther Galbrun. 2013. Topical organization of user comments and application to content recommendation. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, Leslie Carr, Alberto H. F. Laender, Bernadette Farias Lóscio, Irwin King, Marcus Fontoura, Denny Vrandecic, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, and Erik Wilde (Eds.). International World Wide Web Conferences Steering Committee / ACM, 61–62. https://doi.org/10.1145/2487788.2487812

[31] Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Cards for Consumer Health Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 35–44. https://doi.org/10.1145/3331184.3331194

[32] Katy Jordan. 2019. From Social Networks to Publishing Platforms: A Review of the History and Scholarship of Academic Social Network Sites. *Frontiers Digit. Humanit.* 6 (2019), 5. https://doi.org/10.3389/fdigh.2019.00005

[33] Jungeun Kim, Minsoo Choy, Daehoon Kim, and U Kang. 2014. Link prediction based on generalized cluster information. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel (Eds.). ACM, 317–318. https://doi.org/10.1145/2567948.2578807

[34] Howard Leventhal, John Weinman, Elaine A Leventhal, and L Alison Phillips. 2008. Health Psychology: The Search for Pathways Between Behavior and Health. *Annu. Rev. Psychol.* 59 (2008), 477–505.

[35] Michael Ley. 2009. DBLP - Some Lessons Learned. *Proc. VLDB Endow.* 2, 2 (2009), 1493–1500. https://doi.org/10.14778/1687553.1687577

[36] D. A. Lindberg. 2000 Sep-Oct. Internet Access to the National Library of Medicine. *Effective clinical practice: ECP* 3, 5 (2000 Sep-Oct), 256–260.

[37] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[38] Saif M. Mohammad. 2019. The State of NLP Literature: A Diachronic Analysis of the ACL Anthology. *CoRR* abs/1911.03562 (2019). arXiv:1911.03562 http://arxiv.org/abs/1911.03562

[39] Saif M. Mohammad. 2020. NLP Scholar: A Dataset for Examining the State of NLP Research. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, Nicoletta Calzolari, Frédéric Béchet,

Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 868–877. https://www.aclweb.org/anthology/2020.lrec-1.109/

[40] Saif M. Mohammad. 2020. NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 232–255. https://doi.org/10.18653/v1/2020.acl-demos.27

[41] Greg Morrison. 2019. Explorations in Bibliography: Zotero Goes Public. *Atla Summary of Proceedings* (2019), 218–221. https://doi.org/10/gh77x8

[42] Colm Mulcahy. 2017. The Mathematics Genealogy Project Comes of Age at Twenty-one. *Notices of the AMS* 64, 5 (2017), 466–470.

[43] Bryan Newbold. 2021. Search Scholarly Materials Preserved in the Internet Archive. https://blog.archive.org/2021/03/09/search-scholarly-materials-preserved-in-the-internet-archive/

[44] Andreas Niekler, Arnim Bleier, Christian Kahmann, Lisa Posch, Gregor Wiedemann, Kenan Erdogan, Gerhard Heyer, and Markus Strohmaier. 2018. ILCM - A Virtual Research Infrastructure for Large-Scale Qualitative Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/734.html

[45] Kenneth A Norman, Ehren L Newman, and Greg Detre. 2007. A Neural Network Model of Retrieval-Induced Forgetting. *Psychological Review* 114, 4 (2007), 887.

[46] Kevin O'Brien. 2019. Resource Review: ResearchGate. *Journal of the Medical Library Association* 107, 2 (April 2019), 284–285. https://doi.org/10/gh7rp4

[47] Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. NLPExplorer: Exploring the Universe of NLP Papers. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 476–480. https://doi.org/10.1007/978-3-030-45442-5_61

[48] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In *Information Retrieval Evaluation in a Changing World*, Nicola Ferro and Carol Peters (Eds.). Springer, Berlin Heidelberg New York. https://doi.org/10.1007/978-3-030-22948-1_5

[49] Jinfeng Rao, Ferhan Türe, Xing Niu, and Jimmy Lin. 2017. Mining the Temporal Statistics of Query Terms for Searching Social Media Posts. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz (Eds.). ACM, 133–140. https://doi.org/10.1145/3121050.3121052

[50] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, David A. Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans (Eds.). ACM, 42–49. https://doi.org/10.1145/1031171.1031181

[51] Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology Searchbench. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - System Demonstrations*. The Association for Computer Linguistics, 7–13. https://www.aclweb.org/anthology/P11-4002/

[52] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 243–246. https://doi.org/10.1145/2740908.2742839

[53] Inien Syu, Sheau-Dong Lang, and Narsingh Deo. 1996. Incorporating Latent Semantic Indexing into a Neural Network Model for Information Retrieval. In *CIKM '96, Proceedings of the Fifth International Conference on Information and Knowledge Management, November 12 - 16, 1996, Rockville, Maryland, USA*. ACM, 145–153. https://doi.org/10.1145/238355.238475

[54] Gary Taubes. 1993. Publication by Electronic Mail Takes Physics by Storm. *Science* 259, 5099 (Feb. 1993), 1246–1248. https://doi.org/10/bwqfwv

[55] Ariena HC van Bruggen and Alexander M. Semenov. 2000. In search of biological indicators for soil health and disease suppression. *Applied Soil Ecology* 15, 1 (2000), 13–24. https://doi.org/10.1016/S0929-1393(00)00068-8 Special issue: Managing the Biotic component of Soil Quality.

[56] Huaiyu Wan, Yutao Zhang, Jing Zhang, and Jie Tang. 2019. AMiner: Search and Mining of Academic Social Networks. *Data Intell.* 1, 1 (2019), 58–76. https://doi.org/10.1162/dint_a_00006

[57] Hao Wu and Hui Fang. 2014. Document Prioritization for Scalable Query Processing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang (Eds.). ACM, 1609–1618. https://doi.org/10.1145/2661829.2661914

[58] Jian Wu, Kunho Kim, and C. Lee Giles. 2019. CiteSeerX: 20 years of service to scholarly big data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse, AIDR 2019, Pittsburgh, PA, USA, May 13-15, 2019*, Huajin Wang and Keith Webster (Eds.). ACM, 1:1–1:4. https://doi.org/10.1145/3359115.3359119

[59] Holt Zaugg, Richard E. West, Isaku Tateishi, and Daniel L. Randall. 2010. Mendeley: Creating Communities of Scholarly Inquiry through Research Collaboration. *TechTrends: Linking Research and Practice to Improve Learning* 55, 1 (July 2010), 32–36. https://doi.org/10/d4vbh8

[60] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 177–186. https://doi.org/10.1145/3269206.3271776

[61] Tiancheng Zhao and Kyusong Lee. 2020. Talk to Papers: Bringing Neural Question Answering to Academic Search. (2020), 30–36. https://doi.org/10.18653/v1/2020.acl-demos.5

[62] Michel Zitt, Alain Lelu, Martine Cadot, and Guillaume Cabanac. 2019. *Bibliometric Delineation of Scientific Fields*. Springer, 25–68. https://doi.org/10.1007/978-3-030-02511-3_2