The Information Retrieval Anthology 2021
Inaugural Status Report and Challenges Ahead

Martin Potthast
Leipzig University
martin.potthast@uni-leipzig.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Matthias Hagen
Martin-Luther-Universität Halle-Wittenberg
matthias.hagen@informatik.uni-halle.de

Abstract

The Information Retrieval Anthology, IR Anthology for short, is an endeavor to create a comprehensive collection of metadata and full texts of IR-related publications. We report on its first release, the use cases it can serve, as well as the challenges lying ahead to develop it towards a resource that serves the IR community for years to come. The IR Anthology's metadata browser and full text search engine are available at IR.webis.de.

1 Introduction

The development of technology for information retrieval was one of the first goals, right at the outset, of the digital revolution. As World War II came to an end, Vannevar Bush (1945) proposed in his landmark article "As We May Think" that scientists, who had succeeded in dramatically extending the capabilities of the human senses through dedicated instruments and our abilities to manipulate the physical world, should now turn their attention to accomplish the same for the human mind. Describing what later came to be known as "information overload," Bush observes:[1]

> "There is a growing mountain of research. [...] The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. [...]
>
> The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present-day interests, but rather that publication has been extended far beyond our present ability to make real use of the record."

He envisions in remarkable detail how the aforementioned scientific achievements could be implemented to tackle this problem:

[1] Quotes taken from a reprint of the original article [Bush, 1996].

> "Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, 'memex' will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

The memex became real in the form of personal computing devices coupled with the World Wide Web, and its envisioned key functionality as an extension of the human mind is largely driven by IR technology. However, despite this success, we are still fighting information overload, which, in IR terms, often boils down to finding new ways to distinguish what is relevant from what is not for a given "information need." Compiling the Information Retrieval Anthology could be seen as a "self-referred" contribution in this regard. The IR Anthology is supposed to collect all kinds of information retrieval research to make IR scholars' daily lives a bit easier. The first version of this anthology has been accepted as a demo at SIGIR 2021 [Potthast et al., 2021].

After reviewing related work in Section 2, Section 3 outlines the current building blocks of the IR Anthology, Section 4 makes the case for an initiative to experiment with and develop new IR technology specifically for the anthology, Section 5 reviews potential goals for future development, and Section 6 discusses organizational matters.

2 Related Work[2]

In a recent SIGIR Forum opinion article, Hiemstra et al. [2021] make the case for "Transitioning the Information Retrieval Literature to a Fully Open Access Model", observing that various research communities thrive on 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 exemplary in this regard and serves as the main inspiration and basis for this initiative. After reviewing related endeavors from the ACL Anthology and its offspring projects, we present a wider context of scholarly information utilities, both generic and specific to other fields. Table 1 compares a selection of popular services.

The ACL Anthology is an online platform that provides a curated collection of publications from the areas of computational linguistics and natural language processing [Gildea et al., 2018]. Organized as a table-based overview, it provides easy access to publication lists by venue, year, or both. The ACL Anthology's open archives have inspired a prospering ecosystem of research projects on academic literature search and exploration, among them the ACL Anthology Searchbench [Schäfer et al., 2011], LT Expert Finder [Fischer et al., 2019], NLP Scholar [Mohammad, 2020b,a], NLPExplorer [Parmar et al., 2020], and Talk to Papers [Zhao and Lee, 2020]. Due to its commitment to open-access, the ACL Anthology can implement powerful search features using nothing more than a general-purpose web search engine's `site`-operator. Beyond targeting literature exploration itself, projects built on the ACL Anthology have investigated scientometric research questions analyzing large-scale and long-term trends in NLP research [Mohammad, 2019] or temporal bias in citation patterns [Bollmann and Elliott, 2020]. Such studies are of interest to the IR community as well [Hiemstra et al., 2007], and we hope that the IR Anthology corpus

[2]The related work section is borrowed from Potthast et al. [2021].

[3]https://www.aclweb.org/anthology/

Service	Launch	Scope	Bib.	OA	Search	Social	Link and References
MEDLINE	1964	S	C		M, C		Arms [2000]
Web of Science	1964	G	C		M		Garfield [1964]; Birkle et al. [2020]
arXiv	1991	S	C	✓	M, T		Taubes [1993]
ACM Digital Library	1993	S	C		M, T, C	P	Arms [2000]
DBLP	1993	S	C	✓	M		Ley [2009]
The Coll. of CS Bib.	1995	S	C	✓	M		
PubMed	1996	S	C		M, T, C		Lindberg [2000 Sep-Oct]
Math. Genealogy Pr.	1996	S	C	✓	M		Coonce [2004]; Jackson [2007]; Mulcahy [2017]
CiteSeer^x	1997	S	C, F	✓	M, T		Giles et al. [1998]; Wu et al. [2019]
CoRR	1998	S	C	✓	M, T		Halpern [2000]
Crossref	1999	G	C		M		Hendricks et al. [2020]
ACL Anthology	2002	S	C	✓	M, T		Gildea et al. [2018]
Google Scholar	2004	G	C		M, T	P	Giles [2005]
Bibsonomy	2006	G	F		M, T	P, D	Benz et al. [2010]
Microsoft Academic	2006	G	C		M, T	P	Sinha et al. [2015]
Zotero	2006	G	P		M, T	P, D	Morrison [2019]
Academia.edu	2008	G	F		M, (T)	P, D	Jordan [2019]
Mendeley	2008	G	C, P		M	P, D	Henning and Reichelt [2008]; Zaugg et al. [2010]
ResearchGate	2008	G	F		M	P, D	O'Brien [2019]; Jordan [2019]
ORCID	2012	G			M	P	Butler [2012]
Semantic Scholar	2015	G	C		M, T		Bohannon [2016]; Lo et al. [2020]
IA Scholar	2021	G	C	✓	M, T		Newbold [2021]
IR Anthology	2021	S	C		M, T		Potthast et al. [2021]

Table 1: Popular scholarly information utilities by launch year (Column 2). The table informs also about the scope (Column 3) as either field-specific (S) or generic (G), the bibliography management (Column 4) as central database (C), folksonomy (F), or personal database (P), the open-access commitment (Column 5) as fulfilled (✓) if it covers 100 % of the content, the search facilities (Column 6) distinguishing metadata (M), full-texts (T), and controlled vocabularies (C), and the social networking support (Column 7) distinguishing author profiles (P) and discussions (D).

will facilitate them going forward. However, since most IR publications are currently not openly accessible, a custom search engine for the IR Anthology is probably the only way of matching the ACL Anthology's basic functionality.

A number of services implement search in scientific publications, of which Google Scholar is the longest established, with the most comprehensive index [Gusenbauer, 2019; Harzing, 2019]. Other contenders include Microsoft Academic[4] based on a large-scale entity graph [Sinha et al., 2015], Arnetminer [Wan et al., 2019], Semantic Scholar, and the associated Semantic Scholar Open Research Corpus of 80 million publications [Lo et al., 2020]. More specialized search engines focus on dataset retrieval [Akujuobi and Zhang, 2017; Brickley et al., 2019] for instance, or, as of recently, search in publications hosted at the Internet Archive [Newbold, 2021].

Notable among a great variety of other academic information utilities are bibliographic databases such as DBLP [Ley, 2009] (whose open metadata supports our efforts) as well as crowdsourced "folksonomy-style" [Benz et al., 2010], and personal bibliography alternatives such as Mendeley [Henning and Reichelt, 2008; Zaugg et al., 2010] and Zotero. ResearchGate and Academia.edu address, at least in part, a similar purpose but are academic social networks in first place [O'Brien, 2019; Jordan, 2019]. Preprint servers have long been an important part of the open-access ecosystem: the originally physics-focused arXiv [Taubes, 1993] has been active for three decades, its offshoot Computing Research Repository (CoRR) [Halpern, 2000] for two. Much of other fields'

[4]Microsoft Academic will be discontinued soon: https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

bibliographic information resides in large centralized databases (e.g., MEDLINE for the life sciences [Arms, 2000]) while an endeavor like The Mathematics Genealogy Project sets itself apart with a unique focus on student–advisor relations, tracing who taught whom throughout math history [Coonce, 2004; Jackson, 2007; Mulcahy, 2017].

3 Bootstrapping the IR Anthology

Four key elements form the basis of the ACL Anthology's success: (1) its unified collection of bibliographic metadata, and (2) its complete collection of full texts, which both (3) cover the vast majority of relevant publications, and which (4) can be searched with the most recent state-of-the-art retrieval technology. Their scale of operations demands for a centralized web service. To achieve a comparable success with the IR Anthology within the IR community, each of the above elements is essential. In order to bootstrap a "minimum viable product" within a reasonable time frame, the first edition of the IR Anthology harnesses many existing services.

Bibliographic Metadata Compiling reliable bibliographic metadata is a laborious task. In this regard, we are indebted to the DBLP computer science bibliography, which serves as the primary source of the IR Anthology's collection. Seeking to connect to their service, the IR Anthology links its users to it, so that it could be maintained metadata-wise as a DBLP subset of venues with an IR focus. The DBLP provides their complete collection not only via their website, but also as XML data dumps, which is parsed for relevant venues and their associated publications' metadata. An automatic import of new venues from DBLP as well as an upgrade of existing metadata is facilitated by tailored scripts. Nevertheless, it is likely that the DBLP does not cover every last venue of interest to the IR community, so that future editions of the IR Anthology will have to combine external and own metadata. Perhaps such venues can eventually be fed back to DBLP so as to improve their record for posterity. At the time of writing, the IR Anthology covers about 41,500 metadata records across 22 venues.

Full Text Corpus Equipped with the bibliographic metadata of a large number of venues with IR focus, we set out to collecting their full texts. Starting with Webis-CSP-15, a corpus compiled as part of previous work on scholarly search by Hagen et al. [2016], and by searching for and crawling publications available online, a full text coverage of about 88% could be reached. Completing the corpus, however, may take considerable more time than it took to reach this level of coverage, since many papers are not easily accessible via our universities' libraries or in digital form. In this regard, data donations from the IR community can help closing the gap.

This corpus cannot be openly shared with the IR community since the copyrights of most publications within prevent that. This is a significant limitation compared to the ACL Anthology, where the full texts of all papers are openly accessible. If Hiemstra et al.'s recently proposed move to adopting an open access publication model at IR venues takes hold, this will rectify the problem for future publications. Whether past publications would be covered as well remains questionable.

Coverage of IR Publications The question of which publications should be included in the IR Anthology is currently answered as follows: all publications which have been published at one

of the 22 venues (16 conferences and six journals) which are dedicated to IR or at which IR-related contributions are frequently published, e.g., in dedicated tracks.

Deciding which venues to include or exclude when compiling a field-specific literature collection is called "field delineation," an important task within the science studies [Zitt et al., 2019]. To obtain an initial set of publications, the fastest route presented itself in the form of exploiting an existing classification system covering leading IR venues. The list is not yet exhaustive, and we have already been made aware of a number of missing venues, as well as some (national) conferences and workshops organized by local IR societies.

The world is not black and white, and there are many more publications and venues related to IR. Likewise, not all publications published at venues currently included are related to IR since these venues sometimes organize tracks from different fields. Eventually, a decision must be made for every candidate venue/publication; before, however, all candidate venues/publications have to be discovered by applying further heuristics from science studies: next to exploiting existing classification systems, extracting results from generic scholarly search engines, and analyzing the bibliometric network of the publications already included. Altogether, a publication qualifies for inclusion if it contributes to IR, be it explicitly, or implicitly.

Full Text Search Though sharing the full text corpus is currently out of the question due to copyright restrictions, this fact is no roadblock for the IR Anthology to be just as useful to the IR community as the fully open access ACL Anthology is to the ACL community. Note that, in most cases, links alongside each publication point to its publisher's site as well as to generic scholarly search engines. A publication may be rather close at hand to an IR Anthology's user, either through their university's subscription or via author-supplied versions on personal web pages. The latter type of source can be systematically crawled, and corresponding links can be added directly to the IR Anthology. In addition, a full-text search engine can be built to enable focused retrieval within the IR Anthology's corpus.

To provide for a first basic search engine for the IR Anthology, we employ ChatNoir [Potthast et al., 2012; Bevendorff et al., 2018]. Originally developed as a reproducible baseline search engine for the ClueWeb09, the ClueWeb12, and the Common Crawl, ChatNoir's distributed Elasticsearch backend also enables the inclusion of other kinds of indexes. The full text corpus of the IR Anthology has been indexed, exposed via an API endpoint as well as a dedicated version of ChatNoir's web interface found at IR.chatnoir.eu. Before indexing, the PDFs of the publications were converted to plain text using GROBID [2008-2021]; as a retrieval model, ChatNoir employs BM25F [Robertson et al., 2004] including the fields title, abstract, and full text body, where matches in titles and abstracts are weighted higher than in bodies. For publications where full texts could not be obtained as of yet, only their titles and, if available, their abstracts were indexed.

In its current form, neither ChatNoir's index of the IR Anthology nor its corresponding interface are competitive with a hypothetical `site`-search via a commercial search engine as exemplified by the ACL Anthology, nor with a more tailored, custom-built scholarly search engine. This limitation may delay the adoption of the IR Anthology by the IR community as a go-to resource for scholarly search. However, we plan to set up an open research environment for the development and incorporation of new search engines, not just of our own design, but also engines contributed by interested community members.

Metadata Browser To enable browsing the IR Anthology, we are indebted to the ACL Anthology, both for providing an excellent example to emulate and for sharing their entire tried-and-tested code base open source. The IR Anthology's website is based on that of the ACL Anthology. Though the initial plan was to keep changes to an absolute minimum in order to easily pull upgrades from the ACL Anthology's developer team as they appear, and to feed back bug fixes, unfortunately, this goal could not be achieved due to limitations of extensibility and modularity of the original code base. The current processing pipeline thus constructs the IR Anthology's website from scratch by obtaining the latest metadata for the venues covered by DBLP and by converting it into the format required by the IR Anthology's variant of the ACL Anthology's website. To streamline this process, a corresponding command line utility has been developed.

4 IR4IR: Information Retrieval for Information Retrieval

Due to the non-open-access reasons outlined before, one cannot rely on existing commercial search engines to query the IR Anthology. While developing and hosting an own (both effective and efficient) full text search engine should not be underestimated, this effort is less of a problem for researchers working in IR. Taking a shortcut, the search infrastructure underlying the ChatNoir research search engine has been extended to also index the IR Anthology. But ChatNoir should not be considered a serious contender for searching, for instance, the ACL Anthology since a commercial search engine usually provides a shorter turnaround time and hence better timeliness—along with more sophisticated retrieval models. This also means that the limitations imposed on the IR Anthology by being barred from openly sharing its underlying full texts is only partially overcome by ChatNoir in its current form: it has not been designed with scholarly search in mind.[5]

Perhaps the idea of building a *single* scholarly search engine for the IR Anthology is misguiding. Playing to the IR community's strengths, we should build *many* to choose from right away. To facilitate this goal, a next step on the roadmap to maturing the IR Anthology may be the organization of the shared task "IR4IR: Information Retrieval for Information Retrieval." The copyright protections of the full texts may not be an issue for such a shared task, after all. For instance, the MIREX shared task (Music Information Retrieval Evaluation eXchange)[6] has been using copyright-protected collections of music from the start. Instead of distributing the data, the organizers asked the participants to submit working software.

Software submissions—albeit long-established also at non-IR shared task, such as the International SAT Competition on satisfiability of formulae in propositional logics—take a toll on task organizers' work overhead. To reduce this overhead, we develop the TIRA Integrated Research Architecture [Potthast et al., 2019], a cloud-based shared task platform which implements the evaluation-as-a-service paradigm [Hopfgartner et al., 2018]. It may serve as a basis to give participants of the IR4IR shared task self-service processing access to the full text corpus of the IR Anthology while preventing leakage. Using the ACL Anthology's open access corpus of full

[5]Technically, a `site`-search by commercial search engines may be possible via an exclusive exchange of full texts for indexing. The full texts of the IR Anthology can be provided to everyone who already owns a copy of them, which may well be the case for commercial (scholarly) search engines. As many active members of the IR community work at the respective companies, this may facilitate such an agreement. Finally, the possibility of enlisting a commercial search engine as a paid contractor has also not been explored yet.

[6]https://www.music-ir.org/mirex/wiki/MIREX_HOME

texts, training data can be provided which is on the order of magnitude of the IR Anthology's full text corpus, and which is pretty similar to what can be expected in the latter. The integration of the resulting search engines into the IR Anthology can be easily done if their authors are willing to provide a unified API and documentation. We hope that members of the IR community will pick up this challenge: there are many academic and commercial frameworks, libraries, and tools out there that may be used to establish a state-of-the-art scholarly search with an IR focus.

5 Use Cases and Future Challenges

The basic use case of the IR Anthology in its current form is the same as that of the ACL Anthology: a unified metadata collection serves scholars as a source for reliable and well-maintained bibliographic data, and a full text search engine enables focused and high-precision retrieval. Both improve current best practices in the sense that bibliographic metadata of poorer quality may get used less, and that more in-depth related work research can be done by authors as well as by reviewers. In our opinion these two benefits alone justify the expense.

Nevertheless, the IR Anthology has the potential to serve more purposes. In what follows, we collect extensions and use cases related to searching the IR Anthology, to covering more IR-related resources, and to harnessing it as a resource for teaching and learning IR.

5.1 Advanced Search and Indexing

Boolean Retrieval Scholarly search in the IR Anthology can be supported by a number of advanced search operators beyond basic keyword search. These include basic and advanced Boolean retrieval operators, some of which are already supported by ChatNoir's Elasticsearch backend, whereas others may require specialized retrieval models and indexes. The importance of Boolean retrieval for scholarly search should not be underestimated: many scholarly search tasks are high-recall tasks, if not total recall tasks. Developing and maintaining Boolean queries that capture publications related to a certain sub-field of IR thus appears to be a worthwhile goal, especially for scholars who have a long-lasting interest in a given sub-field. Using more sophisticated algorithms, the process of formulating such queries may be supported or even automated, e.g., by deriving keyqueries for a given set of relevant publications [Hagen et al., 2016].

Question Answering and Conversational Search At the other end of the "recency spectrum" (compared to Boolean retrieval) are neural retrieval models. Among others, they are employed for tackling passage retrieval with applications in question answering, and, building on top of that, conversational search assistants. Equipped with a sophisticated question answering system, the IR Anthology would be accessible through new ways of interaction. Moreover, the application of conversational AI in professional contexts appears to present a number of unique challenges compared to consumer uses, such as answering complex questions, mining and retrieving scientific arguments [Wachsmuth et al., 2017; Stede and Schneider, 2018; Balog et al., 2020], and leading a conversation about a given IR topic.

Expert Search and Reviewer Assignment Besides scholarly search in IR with relation to the search for related work, also the search for IR experts on a given subject can be supported by the IR Anthology. This pertains to journal editors searching for an expert as a candidate reviewer for a submitted article, as well as to program committee chairs who need to recruit reviewers for specialized conference tracks or workshops. In this regard, the assignment of reviewers to submitted papers at a conference might be supported as well. The ACL community has recently accomplished this task in collaboration with Semantic Scholar.[7]

Query by Example Reviewer assignment is related to querying by example, where the example is a to-be-reviewed paper, and the retrieval results correspond to people whose publication record features the most closely related work, disregarding the submitted paper's authors. Processing example queries in the form of publications, or even sets of publications, has other use cases when it comes to updating a previously undertaken related work search after a couple of years with the newest publications. Here, recommender systems or the aforementioned approach to compute keyqueries can be employed. If personal accounts were implemented at the IR Anthology, scholars may be automatically notified of newly arriving publications closely related to their own previous work and/or interests. Finally, another query-by-example application is the analysis and search for text reuse [Hagen and Stein, 2011; Hagen et al., 2017].

Modeling IR Research Despite being unable to share the full texts themselves, sharing derived data and models is a distinct concept. In this regard, training (and/or fine-tuning) a (transformer) model based on the IR Anthology enables participants in an IR4IR shared task to build advanced retrieval systems that are tailored to IR—instead of having to rely only on generic embedding models. Besides, such models may also serve various further purposes outlined below, especially with respect to teaching and learning IR, as well as with respect to literature analyses.

5.2 Coverage of IR Publications and Resources

Field Delineation through Citation Network Analysis The task of field delineation (to decide whether a given publication should be included in the IR Anthology or not) does not need to be a binary decision. Including only the publications of a selected set of venues yields a subset of all publications with a "sufficient" relation to IR, whereas, as a heuristic to construct a super set of all IR-related publications, one might consider a collection of all publications that have been cited by or that cite at least a given number publications already included. Such a "citation hull" of the IR Anthology very likely contains many rather unrelated publications, though. If indexed as part of its search engine, they would reduce retrieval precision compared to a more focused collection. In the anthology search engine, however, publications found outside the "core set" of publications covered by the IR Anthology may be indicated or offered as a search facet, enabling scholars to filter and contextualize their search as needed.

Monographs and Textbooks Two important types of publications presently not covered by the IR Anthology include monographs in the form of textbooks but also dissertation and habil-

[7] https://acl2020.org/blog/conflict-of-interest/

itation theses. The latter are typically published at the author's university and often available online; still, collecting the theses of all authors whose work addresses information retrieval aspects may be a Herculean task in itself. Nevertheless, it is an interesting endeavor, not only for the sake of completing the IR Anthology's coverage, but also to trace the genealogy of student–advisor relations throughout IR's history, or to envisage the development of certain research strands. As a shining example in this respect, the Mathematics Genealogy Project of Coonce [2004] stands out, which traces back the genealogy of math scholars throughout its long history.

Textbooks, although they cannot be openly shared as well, may still be easier to collect, especially since some of the IR-related book series have previously been made publicly available for free for a limited amount of time.

Blog Posts, Society Reports, and Web Pages Quite a number of IR scholars have been and continue to engage in blogging about their subjects of interest. Like in letters to the editor at journals, blog posts contain personal opinions and comments on current developments, as well as additional context and background both from a publication's authors and its readers. When lost, they may leave gaps in the record of the development of a particular sub-field of IR. Similarly, the business communications compiled as newsletters and reports of the various IR societies all over the world may shed light on the development of (parts of) the IR community.

Besides blog posts, all kinds of web pages are cited as sources in publications, whereas link rot ensures that most URLs eventually become unavailable, some already shortly after the paper has been published. Given the full texts of the IR Anthology, the URLs might be retrieved and then linked to their archived version at the Internet Archive's Web Archive. This way, if one finds a defunct URL within a publication, there is a chance that the original page can still be retrieved. Likewise, also the websites and web pages of IR events such as conferences and workshops often go missing after the event has passed. Following up on a past conference, though not a routine task, would be easier if the respective web pages were also archived.

Publication Artifacts, Software, Datasets, and Presentation Material Empirical research, which makes up for the vast majority of research in IR, hardly ever results in self-contained publications. To reproduce them, one requires certain number artifacts, including datasets and software. These artifacts are still not necessarily provided alongside a given publication. Recently, a trend throughout computer science has emerged both to recognize extra efforts made by authors, and to make reproducibility a part of peer review. The ACM meanwhile adopted an additional way of recognizing well-crafted reproducibility through its Artifact Review and Badging program,[8] which has just now also been introduced as an optional additional review round for publications published at major IR venues.[9] By way of awarding badges to publications that successfully pass an additional, post-acceptance evaluation for reproducibility, these publications serve as examples to emulate. The IR Anthology can support this initiative by tagging its publications and by providing links to, or even archiving the artifacts associated with a given publication.

Another artifact that goes along conference publications includes the material that authors prepare for on-site/online presentation, including talk slides, posters, and, in times of COVID-19,

[8] https://www.acm.org/publications/policies/artifact-review-badging
[9] https://sigir.org/general-information/acm-sigir-artifact-badging/

a significant amount of pre-recorded videos. Linking or archiving this presentation material will give readers an easier time to follow up on others' research.

Shared Tasks and Leaderboards Not least, shared tasks as one of IR's primary evaluation events are of special interest to the IR Anthology as well as to the IR community. Ever since the TREC conference has been organized for the first time, shared tasks have been key to IR progress. However, just as with other conferences and workshops, the web pages on which shared tasks are announced and where results are reported disappear frequently after the event has passed. A collection of all shared tasks that have been published throughout IR's history would enable scholars to follow up on progress on certain tasks of interest. Moreover, if all of a shared task's underlying artifacts were archived alongside the web page itself, scholars would have an easier time following up as well as "participating" or "replaying" afterwards. In this regard, again, the ACL community can serve as role model, where Sebastian Ruder has started the initiative "NLP Progress", compiling the community resources for some of the most important NLP tasks, and how well authors claim to have solved them.[10]

5.3 Teaching and Learning IR

Lecture Notes, Tutorials, and Videos Apart from serving scholarly research, the IR Anthology should also be harnessed for teaching and learning IR. Including teaching material, such as lecture slides, slides of tutorials given at conferences, and potentially associated video lectures would provide both teachers and students of IR with a diverse and rich resource. Shared teaching material may as well be a way of harmonizing the curriculum of foundational and advanced lectures. Conceivably, exchanging also exercises used in basic tutorials held as part of lectures, as well as examination material in the form of question catalogs and exercises is possible.

Terminology and Writing Support Scholarly newcomers to the research field of IR face the burden of learning what basically amounts to "a new language." The terminology that emerges as part of a specialized research field is often opaque to outsiders; in order to write and converse at a professionally recognized level with established scholars, newcomers need to catch up. In this regard, foundational textbooks serve as the initial guide, yet, a glossary of IR terminology, derived from the collected publications of the IR Anthology, would help to speed up and possibly provide for an even more diverse set of views and definitions. With the support of text mining technology, an initial glossary may be (semi-)automatically compiled.

Furthermore, writing down IR research in a way conforming to expectations of reviewers, even if one looks up the terminology, is still a challenge for newcomers that requires training. Here, based on the full texts compiled for the IR Anthology, tailored writing support technology may be devised. Services like the Netspeak search engine can be used to index common phrases found throughout the IR Anthology, enabling IR-specific wordsmithing [Stein et al., 2010]. For instance, which of the two phrases would an IR scholar consider as being more commonly used: "recall and precision" or "precision and recall"?[11]

[10] http://nlpprogress.com/

[11] Take a guess, then see here: https://netspeak.org/#q={precision+and+recall}.

Full Text Translations To further lower the barrier of entry into IR, and to broaden accessibility in general, the strong advances in machine translation of recent years may be exploited by automatically translating the full text corpus of the IR Anthology into different languages, as well as non-English IR publications to English (e.g., for the francophone CORIA conference).

5.4 Literature Analysis and Introspection

The IR Anthology enables literature analysis of information retrieval at scale. The SIGIR demo paper introducing the IR Anthology also showcases a corresponding analysis, modeling the most widespread topics found throughout the IR literature and their trends over the years Potthast et al. [2021] Many more such analyses are conceivable.

6 Organization

The opportunities and goals outlined in the previous section must be taken with a grain of salt. Developing the IR Anthology and its search engine to maturity already poses quite a challenge, let alone inventing and developing more advanced technologies on top of it. The IR Anthology is currently hosted at GitHub within a dedicated organization.[12] Its underlying code is open source, so that future developments may involve volunteer contributions. We are committed to the IR Anthology, but that is not to say that we aim at monopolizing it. It is perfectly alright, for example, for community members to host mirrors. The governance model of the IR Anthology may follow that of the ACL Anthology or that of journals, and be sponsored by an IR society and/or by a committee of senior community members.

Recruiting volunteers to help out building and maintaining the IR Anthology is a key goal, especially when considering the aforementioned *additional* use cases that it might support, or the future challenges. To involve the community, social media channels as well as IR-related mailing lists are of help; a key step towards raising attention to the IR Anthology is the planned IR4IR shared task. We foresee continued commitment on our part for the time being and will try to combine the development of the IR Anthology with ongoing and future (funded) projects. This pertains especially to the development of an advanced search infrastructure as well as science studies. In this respect, collaborations with other community members, e.g., within joint proposals, may become an interesting option.

A key long-term maintenance task is the addition of new conference proceedings to the IR Anthology as they are published. For conferences covered by the DBLP, this boils down to running a number of scripts to update the anthology's metadata. For conferences not covered by the DBLP, the metadata must be created manually. Obtaining the associated full texts may require more effort, unless conference attendees can be convinced to share access to their copy of the proceedings. Should the IR Anthology gain sufficient importance, conference organizers themselves may want to ensure a timely addition of their respective event's proceedings. Streamlining this process and its robustness, and minimizing its overhead is key to sustainability.

[12] https://github.com/ir-anthology

7 Conclusion

The IR Anthology may become an important addition to the already rich set of IR resources, in case the outlined challenges can be met. Especially improving accessibility to its proprietary full text corpus through tailored search engines is an important next development step. We hope the IR community will pick up the challenge of developing their own dedicated systems to solve information retrieval for information retrieval. Maybe there's even one among them who manages to get past the last IR conundrum: "*Das, was man sucht, findet man immer erst zum Schluss.*"

References

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

William Y. Arms. *Digital Libraries*. MIT Press, 2000. ISBN 0-262-01180-8. URL http://www.cs.cornell.edu/wya/DigLib/.

Krisztian Balog, Lucie Flekova, Matthias Hagen, Rosie Jones, Martin Potthast, Filip Radlinski, Mark Sanderson, Svitlana Vakulenko, and Hamed Zamani. Common Conversational Community Prototype: Scholarly Conversational Assistant. *CoRR*, abs/2001.06910, 2020. URL https://arxiv.org/abs/2001.06910.

Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. The Social Bookmark and Publication Management System BibSonomy - A Platform for Evaluating and Demonstrating Web 2.0 Research. *VLDB J.*, 19(6):849–875, 2010. doi: 10.1007/s00778-010-0208-4. URL https://doi.org/10.1007/s00778-010-0208-4.

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

Caroline Birkle, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. Web of Science as a Data Source for Research on Scientific and Scholarly Activity. *Quantitative Science Studies*, 1(1): 363–376, 2020. doi: 10.1162/qss_a_00018. URL https://doi.org/10.1162/qss_a_00018.

John Bohannon. A Computer Program Just Ranked the Most Influential Brain Scientists of the Modern Era. *Science*, November 2016. ISSN 0036-8075, 1095-9203. doi: 10/gh77gw.

Marcel Bollmann and Desmond Elliott. On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,*

ACL 2020, Online, July 5-10, 2020, pages 7819–7827. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.699. URL https://doi.org/10.18653/v1/2020.acl-main.699.

Dan Brickley, Matthew Burgess, and Natasha F. Noy. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1365–1375. ACM, 2019. doi: 10.1145/3308558.3313685. URL https://doi.org/10.1145/3308558.3313685.

Vannevar Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, 1945. URL http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm.

Vannevar Bush. As We May Think (Reprint). *Interactions*, 3(2):35–46, 1996. doi: 10.1145/227181.227186. URL https://doi.org/10.1145/227181.227186.

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

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

Tim Fischer, Steffen Remus, and Chris Biemann. LT Expertfinder: An Evaluation Framework for Expert Finding Methods. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, *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*, pages 98–104. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-4017. URL https://doi.org/10.18653/v1/n19-4017.

Eugene Garfield. "Science Citation Index"-A New Dimension in Indexing. *Science*, 144(3619):649–654, May 1964. ISSN 0036-8075, 1095-9203. doi: 10/d9qt5m.

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

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

Jim Giles. Science in the Web Age: Start Your Engines. *Nature*, 438(7068):554–555, December 2005. ISSN 1476-4687. doi: 10/dcz432.

GROBID. GROBID. https://github.com/kermitt2/grobid, 2008-2021.

Michael Gusenbauer. Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases. *Scientometrics*, 118(1):177–214, 2019. doi: 10.1007/s11192-018-2958-5. URL https://doi.org/10.1007/s11192-018-2958-5.

Matthias Hagen and Benno Stein. Candidate Document Retrieval for Web-Scale Text Reuse Detection. In Roberto Grossi, Fabrizio Sebastiani, and Fabrizio Silvestri, editors, *String Processing and Information Retrieval, 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011. Proceedings*, volume 7024 of *Lecture Notes in Computer Science*, pages 356–367. Springer, 2011. doi: 10.1007/978-3-642-24583-1_35. URL https://doi.org/10.1007/978-3-642-24583-1_35.

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

Matthias Hagen, Martin Potthast, Payam Adineh, Ehsan Fatehifar, and Benno Stein. Source Retrieval for Web-Scale Text Reuse Detection. In Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li, editors, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2091–2094. ACM, 2017. doi: 10.1145/3132847.3133097. URL https://doi.org/10.1145/3132847.3133097.

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

Anne-Wil Harzing. 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):341–349, 2019. doi: 10.1007/s11192-019-03114-y. URL https://doi.org/10.1007/s11192-019-03114-y.

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. Crossref: The Sustainable Source of Community-owned Scholarly Metadata. *Quantitative Science Studies*, 1(1):414–427, 2020. doi: 10.1162/qss_a_00022. URL https://doi.org/10.1162/qss_a_00022.

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

Djoerd Hiemstra, Claudia Hauff, Franciska de Jong, and Wessel Kraaij. SIGIR's 30th Anniversary: An Analysis of Trends in IR Research and the Topology of its Community. *SIGIR Forum*, 41 (2):18–24, 2007. doi: 10.1145/1328964.1328966. URL https://doi.org/10.1145/1328964.1328966.

Djoerd Hiemstra, Marie-Francine Moens, Raffaele Perego, and Fabrizio Sebastiani. Transitioning the Information Retrieval Literature to a Fully Open Access Model. *SIGIR Forum*, 54(1), February 2021. ISSN 0163-5840. doi: 10.1145/3451964.3451977.

Frank Hopfgartner, Allan Hanbury, Henning Müller, Ivan Eggel, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Jimmy Lin, Jayashree Kalpathy-Cramer, Noriko Kando, Makoto P. Kato, Anastasia Krithara, Tim Gollub, Martin Potthast, Evelyne Viegas, and Simon Mercer. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. *Journal of Data and Information Quality (JDIQ)*, 10(4):15:1–15:32, October 2018. doi: 10.1145/3239570.

Allyn Jackson. A Labor of Love: The Mathematics Genealogy Project. *Notices Of The American Mathematical Society*, 54(8):1002–1003, 2007.

Katy Jordan. From Social Networks to Publishing Platforms: A Review of the History and Scholarship of Academic Social Network Sites. *Frontiers in Digital Humanities*, 6:5, 2019. doi: 10.3389/fdigh.2019.00005. URL https://doi.org/10.3389/fdigh.2019.00005.

Michael Ley. DBLP - Some Lessons Learned. *Proceedings of the VLDB Endowment*, 2(2): 1493–1500, 2009. doi: 10.14778/1687553.1687577. URL http://www.vldb.org/pvldb/vol2/vldb09-98.pdf.

D. A. Lindberg. Internet Access to the National Library of Medicine. *Effective clinical practice: ECP*, 3(5):256–260, 2000 Sep-Oct. ISSN 1099-8128.

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

Saif M. Mohammad. The State of NLP Literature: A Diachronic Analysis of the ACL Anthology. *CoRR*, abs/1911.03562, 2019. URL http://arxiv.org/abs/1911.03562.

Saif M. Mohammad. NLP Scholar: A Dataset for Examining the State of NLP Research. In 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, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 868–877. European Language Resources Association, 2020a. URL https://www.aclweb.org/anthology/2020.lrec-1.109/.

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

Greg Morrison. Explorations in Bibliography: Zotero Goes Public. *Atla Summary of Proceedings*, pages 218–221, 2019. ISSN 0066-0868. doi: 10/gh77x8.

Colm Mulcahy. The Mathematics Genealogy Project Comes of Age at Twenty-one. *Notices Of The American Mathematical Society*, 64(5):466–470, 2017.

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

Kevin O'Brien. Resource Review: ResearchGate. *Journal of the Medical Library Association*, 107 (2):284–285, April 2019. ISSN 1558-9439. doi: 10/gh7rp4.

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

Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2012)*, page 1004. ACM, August 2012. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348429.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series. Springer, Berlin Heidelberg New York, September 2019. ISBN 978-3-030-22948-1. doi: 10.1007/978-3-030-22948-1_5.

Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. The Information Retrieval Anthology. In *44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021)*. ACM, July 2021. doi: 10.1145/3404835.3462798. URL https://dl.acm.org/doi/10.1145/3404835.3462798.

Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. Simple BM25 Extension to Multiple Weighted Fields. In David A. Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans, editors, *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, pages 42–49. ACM, 2004. doi: 10.1145/1031171.1031181. URL https://doi.org/10.1145/1031171.1031181.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 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, pages 7–13. The Association for Computer Linguistics, 2011. URL https://www.aclweb.org/anthology/P11-4002/.

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

Manfred Stede and Jodi Schneider. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool, 2018.

Benno Stein, Martin Potthast, and Martin Trenkmann. Retrieving Customary Web Language to Assist Writers. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan M. Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval. 32nd European Conference on Information Retrieval (ECIR 2010)*, volume 5993 of *Lecture Notes in Computer Science*, pages 631–635, Berlin Heidelberg New York, March 2010. Springer. ISBN 978-3-642-12274-3. doi: 10.1007/978-3-642-12275-0_64.

Gary Taubes. Publication by Electronic Mail Takes Physics by Storm. *Science*, 259(5099):1246–1248, February 1993. ISSN 0036-8075, 1095-9203. doi: 10/bwqfwv.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web. In Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Ivan Habernal, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker, editors, *4th Workshop on Argument Mining (ArgMining 2017) at EMNLP*, pages 49–59. Association for Computational Linguistics, September 2017. URL https://www.aclweb.org/anthology/W17-5106.

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

Jian Wu, Kunho Kim, and C. Lee Giles. CiteSeerX: 20 Years of Service to Scholarly Big Data. In Huajin Wang and Keith Webster, editors, *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse, AIDR 2019, Pittsburgh, PA, USA, May 13-15, 2019*, pages 1:1–1:4. ACM, 2019. doi: 10.1145/3359115.3359119. URL https://doi.org/10.1145/3359115.3359119.

Holt Zaugg, Richard E. West, Isaku Tateishi, and Daniel L. Randall. Mendeley: Creating Communities of Scholarly Inquiry through Research Collaboration. *TechTrends: Linking Research and Practice to Improve Learning*, 55(1):32–36, July 2010. ISSN 8756-3894. doi: 10/d4vbh8.

Tiancheng Zhao and Kyusong Lee. Talk to Papers: Bringing Neural Question Answering to Academic Search. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 30–36. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-demos.5. URL https://doi.org/10.18653/v1/2020.acl-demos.5.

Michel Zitt, Alain Lelu, Martine Cadot, and Guillaume Cabanac. Bibliometric Delineation of Scientific Fields. In Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, editors, *Springer Handbook of Science and Technology Indicators*, Springer Handbooks, pages 25–68. Springer, 2019. doi: 10.1007/978-3-030-02511-3_2. URL https://doi.org/10.1007/978-3-030-02511-3_2.