

# Artifact Sharing for Information Retrieval Research

Sean MacAvaney

sean.macavaney@glasgow.ac.uk

University of Glasgow

Glasgow, United Kingdom

## Abstract

Sharing artifacts—such as trained models, pre-built indexes, and the code to use them—aids in reproducibility efforts by allowing researchers to validate intermediate steps and improves the sustainability of research by allowing multiple groups to build off one another’s prior computational work. Although there are *de facto* consensuses on how to share research code (through a git repository linked to from publications) and trained models (via HuggingFace Hub), there is no consensus for other types of artifacts, such as built indexes. Given the practical utility of using shared indexes, researchers have resorted to self-hosting these resources or performing *ad hoc* file transfers upon request, ultimately limiting the artifacts’ discoverability and reuse. This demonstration introduces a flexible and interoperable way to share artifacts for Information Retrieval research, improving both their accessibility and usability.

 <https://github.com/seanmacavaney/artifacts-demo>

## CCS Concepts

- Information systems → Information retrieval.

## Keywords

Reproducibility, Artifact Reuse, Green IR

### ACM Reference Format:

Sean MacAvaney. 2025. Artifact Sharing for Information Retrieval Research. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’25)*, July 13–18, 2025, Padua, Italy. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3726302.3730147>

## 1 Introduction

A variety of artifacts are produced throughout the course of research in Information Retrieval, including trained models, built indexes, caches, and the code to do it all. A few *de facto* standard approaches have emerged for distributing some of these kinds of artifacts. For instance, code is typically shared via git repositories linked in the research paper, and trained models are typically shared on the HuggingFace Hub. However, there is no such standard for other types of artifacts—especially built indexes. This demonstration presents a new system that unifies access to data-oriented artifacts (such as built indexes and caches), while maintaining the flexibility required to conduct research in this fast-paced field.

---

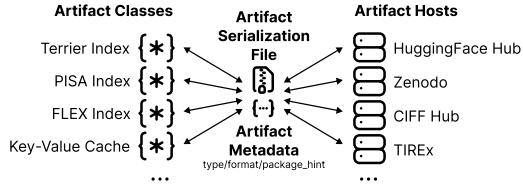
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*SIGIR ’25, July 13–18, 2025, Padua, Italy*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1592-1/2025/07

<https://doi.org/10.1145/3726302.3730147>



**Figure 1: An overview of the distributed design for sharing IR artifacts.** Artifacts (instances of *Artifact Classes*) can be serialized and shared to an *Artifact Host* via a *Artifact Serialization File*. Hosted artifacts can then be downloaded and loaded as the original *Artifact Class*.

There have been several attempts to improve the sharing and reuse of artifacts in IR. For example, PyTerrier [18] and Pyserini [10] platforms include repositories for pre-built indexes<sup>1</sup>. However, these repositories are not publicly writeable, limiting their contents to that of the platform maintainers rather than the IR community at large. TIREx [6] provides access to built indexes and is built around taking community contributions, but requires that the data generation and indexing code be containerized in Docker images. While this has major benefits in terms of reproducibility, it reduces flexibility by adding the overhead of containerization and makes it more appropriate for sharing final products rather than works-in-progress. BM25S [12] allows its built indexes to be shared on HuggingFace Hub, but is limited by only supporting its index format and the HuggingFace Hub platform for sharing. CIFF [11] provides a common exchange format across various inverted index formats and a centralized location for sharing them (CIFF Hub<sup>2</sup>), but it is limited to inverted indexes and uploads are limited to the platform maintainers.

The system described in this demonstration paper aims to address the limitations of existing approaches for sharing artifacts in IR by allowing all researchers to share (publicly or privately) any artifact with minimal overhead. The system’s design is distributed in nature, with support for virtually any class of artifact (Section 3.2) and hosting provider (Section 3.3), registered via Python packages. The exchange of artifacts is facilitated by a flexible artifact serialization file format (Section 3.1) that provides basic metadata about the type/format of the artifact alongside suggested software to process it. An overview of the design is given in Figure 1. In demonstration of the flexibility of the format, the system currently supports 13 classes of artifacts (with artifact implementations written in Python, Java, C++, and Rust), and 7 hosts (including HuggingFace Hub, Zenodo, and CIFF Hub). This nearly completely covers (and expands upon) the aforementioned existing artifact-sharing mechanisms.

<sup>1</sup><http://data.terrier.org/> and <https://github.com/castorini/pyserini/blob/master/docs/prebuilt-indexes.md>

<sup>2</sup><https://github.com/pisa-engine/ciff-hub>

Several factors have coincided to make this the "right time" to improve the ease of sharing artifacts in IR research. First, the computational demands of conducting IR research are increasing. With that, it's both more convenient and more environmentally responsible to build off computations that others have done when possible [29]. Second, there is an increasing interest within the community on reproducibility [2]. Sharing artifacts enables other researchers to validate intermediate results and assess the reproducibility of different individual components of a research procedure [33]. Finally, the field is arguably growing more collaborative [15, 30], which increases the need to share artifacts across institutions in the course of day-to-day research.

The system is integrated into PyTerrier [18]. PyTerrier is a natural platform for the system given its extension-oriented design that provides interoperable functionality across numerous established and emerging IR techniques (e.g., [3, 13, 17, 21]). However, the design of the system described in this paper is not dependent on PyTerrier, and we welcome contributions that apply the design in other toolkits since doing so would help further broaden artifact sharing and reuse in the field. We have already used this system to share artifacts for a published research paper [1] and to help facilitate cross-institutional collaborations [8, 26–28]. By providing this demonstration, we hope to help encourage more artifact sharing and ease more collaborations.

## 2 System at a Glance

The artifact-sharing system is geared towards IR researchers, especially those looking to conduct collaborative research (e.g., participate with one another for a shared task) and those looking to share the research artifacts upon publication. Core functionality of the system is demonstrated in a Google Colab notebook,<sup>3</sup> which will be made available for SIGIR attendees to interact with on a laptop during the demonstration period. The core functionality is also recorded in the remainder of this section.

The main functionality of the system bundled with the core PyTerrier Python package, which can be installed as follows:

```
pip install python-terrier
```

Artifacts can be loaded from a variety of hosts. For instance, the following code downloads and loads a Terrier [25] index from the HuggingFace Hub.<sup>4</sup> Once it's loaded, it is ready to use for retrieval:

```
import pyterrier as pt
# load a TerrierIndex object for the msmarco-passage corpus
index = pt.Artifact.from_hf('macavaney/msmarco-passage.terrier')
# the artifact is ready to use:
retriever = index.bm25()
retriever.search('my dear watson')
# qid      query      docno      score      rank
# 1 my dear watson 5341214 36.087756 0
# 1 my dear watson 2385137 30.109050 1
# ...
```

Note that the system automatically detects the type of artifact and returns an instance of an artifact class that can process it. For

instance, if we requested a PISA index [19], it will be loaded as that class (from the pyterrier-pisa [13] extension package<sup>5</sup>) instead:

```
# load a PisaIndex object
index = pt.Artifact.from_hf('macavaney/msmarco-passage.pisa')
retriever = index.bm25()
retriever.search('my dear watson')
# qid      query      docno      score      rank
# 1 my dear watson 5341214 21.715158 0
# 1 my dear watson 2385137 18.506784 1
# ...
```

(The results PISA returns for the query are different than Terrier due to differences in tokenization and BM25 score calculations.)

Next, we consider the case where a researcher is looking to share a built artifact (e.g., they are participating in a shared task and built a resource that might be useful for others.) The demonstration shows how you can load a dataset (in this case, a corpus of text from Sir Arthur Conan Doyle, hosted on HuggingFace datasets), index it, and share the artifact to the HuggingFace Hub:

```
from datasets import load_dataset
my_corpus = load_dataset('macavaney/arthur-conan-doyle')['corpus']
index = pt.terrier.TerrierIndex("my-index.terrier")
index.index(corpus)
index.to_hf('my-username/my-index.terrier')
```

After the upload is complete, the artifact is available for others to use using `Artifact.from_hf('my-username/my-index.terrier')`.

This demonstration only scratches the surface of the system's available functionality, which is covered in further detail in the following section. Nonetheless, it shows how the core functionality can be useful for those looking to build off somebody else's artifacts and for those looking to easily share artifacts with others.

## 3 Technical Details

There are three core elements to the design of this artifact-sharing system. At its core, the Artifact Serialization File and metadata (Section 3.1) allow artifacts to be represented as a file that includes basic information about the type of artifact stored and suggested software for how to load it. These files can be downloaded from (or, in some cases, uploaded to) Artifact Hosts (Section 3.3), which are responsible for the long-term file hosting of the artifact, plus auxiliary features (e.g., search/discoverability, access control, etc.) Finally, Artifact Classes (Section 3.2) define a Python interface for using the artifacts once downloaded.

Several principles guided the design of this system:

- **Flexibility.** The system should be able to support constraints from virtually any artifact class or host—including constraints like the maximum size of individual files (for hosts) or the implementation programming language (for classes).
- **Extendable and Decentralized.** The system should be able to be extended with new artifact classes and hosts without needing to change the core package. This is accomplished by defining extensions through Python packages that register their functionality through Python Entry Points.<sup>6</sup>
- **Usability.** Downloaded artifacts should be immediately ready for use through a convenient Python interface.

<sup>3</sup>Link on the first page of this paper.

<sup>4</sup>All available artifacts on the HuggingFace Hub can be found through a tag search: <https://huggingface.co/datasets?other=pyterrier-artifact>

<sup>5</sup>If the package isn't installed, an error will be raised that gives a hint about what package needs to be installed.

<sup>6</sup><https://packaging.python.org/en/latest/specifications/entry-points/>

### 3.1 Artifact Serialization File and Metadata

The Artifact Serialization File format enables any artifact to be represented as a single file for distribution. At its core, the format is a simple compressed TAR file containing the files that represent the artifact. Many artifacts are already in a binary format that does not benefit much from general compression algorithms, so LZ4 compression was chosen for its reasonably strong compression ratio and its fast compression/decompression rates. The TAR file may include attributes, such as file creation timestamps, though these attributes are removed by default. Similarly, the artifact files can be compressed in any order but are sorted lexically by default to help ensure that functionally-identical artifacts are encoded to the same serialization file.

**Metadata.** The serialization file should (but does not need to) include a file (`pt_meta.json`) that contains JSON-encoded metadata about the artifact. If present, two fields are required: “type” (which refers to the broad type of artifact, such as `sparse_index`), and “format” (which refers to the specific format of the data, such as `terrier`). Together, this metadata allows the system to identify Artifact Classes that support the type and format of the data. An optional field “package\_hint” provides a Python package name to show the user if no Artifact Class is found that supports the type/format. This can aid the user in installing any additional software that they need to use the artifact. Any other metadata may be stored in this file for use by the artifact; other fields are ignored by this system but made available to the loaded artifacts.

Sometimes, the software that creates an artifact may not include the metadata file. For instance, external software may create the artifact, or the artifact may have been created before the introduction of this system. In these case, `metadata-adapter` entry points allow the system to match an artifact to its corresponding metadata by inspecting its contents. For instance, the `anserini` adapter checks for the presence of files that suggest that the artifact is an `Anserini` index. Or, the `ciff` adapter checks whether the file has a `.ciff` file extension. Naturally, these checks are potentially error-prone and are more expensive than loading the metadata from a file, so these are only applied as a fallback when the metadata is unavailable.

**Segmented Serialization Files.** Some Artifact Hosts impose a maximum individual file size. For instance, HuggingFace Hub limits individual file sizes to 50GB<sup>7</sup>. In these cases, the artifact serialization file can be split into segments with numeric suffixes (`artifact.tar.lz4.0`, `artifact.tar.lz4.1`, ...). When the file is segmented, a `artifact.tar.lz4.json` file<sup>8</sup> must accompany the files, which provides information about how many segments to expect (among other file metadata, like a checksum). When the artifact is not segmented, the file may be present, but is not required.

### 3.2 Artifact Classes

To meet the usability design goal, Artifact Classes define a high-level interface to the functionality provided by the artifact. An Artifact Classes is a simple Python class:

```
class MyArtifact(pt.Artifact):
    def retriever(self):
        ... # return an object that retrieves over this index
```

<sup>7</sup><https://huggingface.co/docs/hub/en/storage-limits>

<sup>8</sup>example `artifact.tar.lz4.json` file

Table 1: Examples of Artifact Classes.

Class	Type/Format	Package
<b>Sparse Indexes</b>		
TerrierIndex	<code>sparse_index/terrier</code>	<code>python-terrier</code>
AnseriniIndex	<code>sparse_index/anserini</code>	<code>pyterrier-anserini</code>
PisaIndex	<code>sparse_index/pisa</code>	<code>pyterrier-pisa</code>
CiffIndex	<code>sparse_index/ciff</code>	<code>pyterrier-ciff</code>
BmpIndex	<code>sparse_index/bmp</code>	<code>bmp</code>
<b>Dense Indexes</b>		
FlexIndex	<code>dense_index/flex</code>	<code>pyterrier-dr</code>
<b>Auxiliary Index Structures</b>		
CorpusGraph	<code>corpus_graph/np_topk</code>	<code>pyterrier-adaptive</code>
<b>Caches</b>		
KeyValueCache	<code>key_value_cache/sqlite3</code>	<code>pyterrier-caching</code>
IndexerCache	<code>indexer_cache/lz4pickle</code>	<code>pyterrier-caching</code>
RetrieverCache	<code>retriever_cache/dbm.dumb</code>	<code>pyterrier-caching</code>
ScorerCache	<code>scorer_cache/sqlite3</code>	<code>pyterrier-caching</code>
DenseScorerCache	<code>scorer_cache/hdf5</code>	<code>pyterrier-caching</code>
CDECache	<code>cde_cache/np_pickle</code>	<code>pyterrier-dr</code>
QualCache	<code>quality_score_cache/numpy</code>	<code>pyterrier-quality</code>

To “register” the class, it should be included as `artifact` entry point, with a key containing the type and format. This design allows the system to import only the necessary packages.

The Artifact Class may also include fields named `ARTIFACT_TYPE` and `ARTIFACT_FORMAT`, which the system can use to help automatically generate the `pt_meta.json` file. The `pt.Artifact` base class provides a variety of functionality, including methods for downloading artifacts, uploading artifacts, and building the serialization file. Additional utilities for working with Artifact Classes are available in the `pyterrier-alpha` package<sup>9</sup>, including tools to aid in the construction of artifacts on disk and inspection of artifact objects. These utilities may be added to the core package in the future.

The system already supports several artifact classes, which are summarized in Table 1. There are five classes of sparse indexes, including general-purpose `Terrier` [25], `Anserini` [34], and `PISA` [19] classes, plus one for `CIFF` [11] (a sparse index interchange format) and `BMP` [20] (a specialized engine for Learned Sparse Retrieval [24]). The `FLEX` (FFlexible EXecution) Index supports and unifies a variety of dense retrieval engines, including `FAISS` [4], `SCANN` [7], `LADR` [9], and `FlatNav` [23]. We also plan to add dedicated Artifact Classes for these backends in the future. Finally, other classes are available for supporting corpus graph traversal [16], specialized caches for contextual document embeddings [22] and document quality scores [1], and general-purpose caches [14]. Naturally, the functionality provided by each Artifact Class will vary depending on the purpose of the artifact; documentation for each artifact is available in the PyTerrier documentation.<sup>10</sup>

In summary, Artifact Classes allow users to define the high-level functionality that an artifact provides, allowing users to immediately make use of artifacts after they are downloaded. A variety of Artifact Classes are already defined (with backends written in several languages, including Python, Java, C++, and Rust), and more can be added through extension packages.

<sup>9</sup><https://github.com/seanmacavaney/pyterrier-alpha>

<sup>10</sup><https://pyterrier.readthedocs.io/>

**Table 2: Examples of uploading and downloading from various Artifact Hosts.**

Platform	Upload Example	Download Example
HuggingFace Hub	<code>artifact.to_hf('user/repo')</code>	<code>Artifact.from_hf('user/repo')</code>
Zenodo	<code>artifact.to_zenodo()</code>	<code>Artifact.from_zenodo('zenodo-id')</code>
Magic-Wormhole	<code>artifact.to_p2p()</code>	<code>Artifact.from_p2p('code', '/path/to/index')</code>
PyTerrier Data Repository	N/A	<code>Artifact.from_dataset('dataset-id', 'index-id')</code>
CIFF Hub	N/A	<code>Artifact.from_url('ciff-hub:ciff-id') (via pyterrier-ciff)</code>
Pyserini Prebuilt Indexes	N/A	<code>Artifact.from_url('anserini:index-name') (via pyterrier-anserini)</code>
TIREX	N/A	<code>Artifact.from_url('tira:dataset-id/team/approach ') (via tira)</code>
Any URL	N/A	<code>Artifact.from_url('https://domain.com/artifact.tar.lz4')</code>

### 3.3 Artifact Hosts

To avoid relying on a single platform for the distribution of artifacts, the system integrates with several existing hosts. The system can also be extended to support new hosts through entry points. This section covers the current hosts available in the core PyTerrier package and extensions. An overview is given in Table 2.

**HuggingFace Hub.** The HuggingFace Hub<sup>11</sup> provides researchers with a platform, software, and free storage to share various artifacts. It is most well-known as a place to share trained machine learning models, though it supports other artifacts as well. There are a variety of benefits for choosing to share artifact on the HuggingFace hub. They are highly discoverable via its search and filtering features<sup>12</sup>. It also offers high uptime<sup>13</sup> and generous free storage limits<sup>14</sup>.

All Artifact Classes have the methods `.to_hf('user/repo')` and `.from_hf('user/repo')`, allowing users to both upload and download their built artifacts. The upload process automatically splits the archive into files below the maximum individual file size (using the approach described in Section 3.1), allowing for very large artifacts to be uploaded. Uploading an artifact generates a README file<sup>15</sup> that includes placeholders for information that will likely be useful to other researchers, such as an example usage script, benchmarks, and code to reproduce the artifact.

In line with other integrations, the HuggingFace Hub integration exposes a `hf:user/repo` “schema” that allows artifacts to be downloaded with the `.from_url` method. The integration also supports private and gated repositories for those who want to limit the access of their artifacts (e.g., for ongoing work-in-progress or artifacts only shareable under data use agreements).

**Zenodo.** Zenodo [5] is a free and dedicated platform for sharing research artifacts with a long-lasting service level agreement.<sup>16</sup> The functionality works similarly to that of HuggingFace Hub, with `.to_zenodo()` and `.from_zenodo()` methods.

**Magic-Wormhole (Peer-to-Peer).** In some cases, one-off peer-to-peer transfers are preferable to approaches that upload and host artifacts. For instance, two collaborators may want to share an artifact that is a work-in-progress. Magic-Wormhole<sup>17</sup> provides an

a way to transfer files between machines, and is built into the core PyTrrier package for performing peer-to-peer artifact transfers.

Due to the nature of peer-to-peer transfers, this integration works slightly differently than the others. The user looking to share an artifact first calls `artifact.to_p2p()`, which provides a one-time code for sharing the artifact (e.g., `66-antenna-transit`). The sharer gives this code to the one receiving it, who then calls `Artifact.from_p2p(code, path)`, specifying both the code and the desired path to store the artifact at.

**Other Hosts** The system also supports other artifact repositories, including the PyTerrier Data Repository [18], TIREX [6], CIFF Hub [11], and Pyserini Prebuilt Indexes [10]. These generally work through the registration of custom URL schemes (via entry points), e.g., `ciff-hub:ciff-id`. PyTerrier uses a `.from_dataset()` method to enable backward compatibility with its past functionality. The custom URL schemes can either translate the provided ID to a full HTTP(S) URL, or download the data through another means and return the path of the artifact on the local filesystem. Finally, the system supports providing a serialized artifact at any arbitrary URL. Collectively, these demonstrate a high degree of flexibility over different possible hosting methods.

## 4 Conclusion

Artifact reuse is an important part of IR research. It can help facilitate collaborations, enable reproduction studies, and reduce the environmental impact of doing IR research. This demonstration presented a new system for sharing artifacts, especially pre-built indexes, to support these research activities. Unlike prior efforts in this direction, this system is designed with flexibility at its core, allowing users to share many types of artifacts over many different hosts. This is facilitated by a simple artifact serialization file format and metadata scheme. Once downloaded, artifacts are ready for use in experiments.

The system currently provides over 100 artifacts on HuggingFace datasets, provides 14 Artifact Classes, and has been used to help facilitate a cross-institutional collaboration. We hope that this demonstration at SIGIR will help others share and re-use artifacts throughout the course of their research.

## Acknowledgments

I thank Craig Macdonald and Maik Fröbe for their helpful comments and suggestions on this system. I also thank Maik Fröbe and Patrick Stahl for contributing the TIREX integration.

<sup>11</sup><https://huggingface.co/>

<sup>12</sup>E.g., all artifacts can be found by filtering on the `pyterrier-artifact` tag: <https://huggingface.co/datasets?other=pyterrier-artifact>.

<sup>13</sup><https://status.huggingface.co/>

<sup>14</sup><https://huggingface.co/docs/hub/en/storage-limits>

<sup>15</sup>Example of README file for a Terrier index for CORD-19 [32] (the corpus for TREC-COVID [31]): <https://huggingface.co/datasets/pyterrier/trec-covid.terrier>

<sup>16</sup><https://help.zenodo.org/guides/nih/element4>

<sup>17</sup><https://magic-wormhole.readthedocs.io/>

## References

- [1] Xuejun Chang, Debabrata Mishra, Craig Macdonald, and Sean MacAvaney. 2024. Neural Passage Quality Estimation for Static Pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zucco, and Yi Zhang (Eds.). ACM, 174–185. <https://doi.org/10.1145/3626772.3657765>
- [2] Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu. 2019. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). 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, 1432–1434. <https://doi.org/10.1145/3331184.3331647>
- [3] Kaustubh D. Dhole. 2024. PyTerrier-GenRank: The PyTerrier Plugin for Reranking with Large Language Models. *CoRR* abs/2412.05339 (2024). <https://doi.org/10.48550/ARXIV.2412.05339> arXiv:2412.05339
- [4] Matthijs Douze, Alexandre Guizha, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *CoRR* abs/2401.08281 (2024). <https://doi.org/10.48550/ARXIV.2401.08281> arXiv:2401.08281
- [5] European Organization For Nuclear Research and OpenAIRE. 2013. Zenodo. <https://doi.org/10.25495/7GXK-RD71>
- [6] Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Information Retrieval Experiment Platform. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2826–2836. <https://doi.org/10.1145/3539618.3591888>
- [7] Ruqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3887–3896. <http://proceedings.mlr.press/v119/guo20h.html>
- [8] Krishkesh Kulkarni, Nazli Goharian, Ophir Frieder, and Sean MacAvaney. 2024. LexBoost: Improving Lexical Document Retrieval with Nearest Neighbors. In *Proceedings of the ACM Symposium on Document Engineering 2024, DocEng 2024, San Jose, CA, USA, August 20–23, 2024*. ACM, 16:1–16:10. <https://doi.org/10.1145/3685650.3685658>
- [9] Krishkesh Kulkarni, Sean MacAvaney, Nazli Goharian, and Ophir Frieder. 2023. Lexically-Accelerated Dense Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 152–162. <https://doi.org/10.1145/3539618.3591715>
- [10] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Frassetto Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2356–2362. <https://doi.org/10.1145/3404835.3463238>
- [11] Jimmy Lin, Joel M. Mackenzie, Chris Kamphuis, Craig Macdonald, Antonio Mallia, Michal Siedlaczek, Andrew Trotman, and Arjen P. de Vries. 2020. Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2149–2152. <https://doi.org/10.1145/3397271.3401404>
- [12] Xing Han Lü. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *CoRR* abs/2407.03618 (2024). <https://doi.org/10.48550/ARXIV.2407.03618> arXiv:2407.03618
- [13] Sean MacAvaney and Craig Macdonald. 2022. A Python Interface to PISA!. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 3339–3344. <https://doi.org/10.1145/3477495.3531656>
- [14] Sean MacAvaney and Craig Macdonald. 2025. On Precomputation and Caching in Information Retrieval Experiments with Pipeline Architectures. In *Proceedings of the 2nd International Workshop on Open Web Search (WOWS) co-located with 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025 (CEUR Workshop Proceedings)*. CEUR-WS.org.
- [15] Sean MacAvaney, Adam Roegiest, Aldo Lipani, Andrew Parry, Björn Engelmann, Christin Katharina Kreutz, Chuan Meng, Erlend Frayling, Eugene Yang, Ferdinand Schlatt, Guglielmo Faggioli, Harrisen Scells, Iana Atanassova, Jana Friese, Janek Bevendorff, Javier Sanz-Cruzado, Johanne Trippas, Kanaad Pathak, Kaustubh D. Dhole, Leif Azzopardi, Maik Fröbe, Marc Bertin, Nishchal Prasad, Saber Zerhoudi, Shuai Wang, Shubham Chatterjee, Thomas Jánich, Udo Kruschwitz, Xi Wang, and Zijun Long. 2024. Report on the Collab-a-Thon at ECIR 2024. *SIGIR Forum* 58, 1 (2024), 1–11. <https://doi.org/10.1145/3687273.3687287>
- [16] Sean MacAvaney, Nicola Tonello, and Craig Macdonald. 2022. Adaptive Ranking with a Corpus Graph. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17–21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 1491–1500. <https://doi.org/10.1145/3511808.3557231>
- [17] Craig Macdonald, Jinyuan Fang, Andrew Parry, Craig Macdonald, and Zaiqiao Meng. 2025. Constructing and Evaluating Declarative RAG Pipelines in PyTerrier. In *Proceedings of the 48th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2025*. ACM. <https://doi.org/10.1145/3726302.3730150>
- [18] Craig Macdonald, Nicola Tonello, Sean MacAvaney, and Iadh Ounis. 2021. PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 – 5, 2021*, Gianluca Demartini, Guido Zucco, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 4526–4533. <https://doi.org/10.1145/3459637.3482013>
- [19] Antoni Mallia, Michal Siedlaczek, Joel M. Mackenzie, and Torsten Suel. 2019. PISA: Performant Indexes and Search for Academia. In *Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019 (CEUR Workshop Proceedings, Vol. 2409)*, Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu (Eds.). CEUR-WS.org, 50–56. <https://ceur-ws.org/Vol-2409/docker08.pdf>
- [20] Antonio Mallia, Torsten Suel, and Nicola Tonello. 2024. Faster Learned Sparse Retrieval with Block-Max Pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zucco, and Yi Zhang (Eds.). ACM, 2411–2415. <https://doi.org/10.1145/3626772.3657906>
- [21] Jan Heinrich Merker, Janek Bevendorff, Maik Fröbe, Tim Hagen, Harrisen Scells, Matti Wiegmann, Benno Stein, Matthias Hagen, and Martin Potthast. 2025. Web-Scale Retrieval Experimentation with chatnoir-pyterrier. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 15576)*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonello (Eds.). Springer, 96–104. [https://doi.org/10.1007/978-3-031-88720-8\\_17](https://doi.org/10.1007/978-3-031-88720-8_17)
- [22] John X. Morris and Alexander M. Rush. 2024. Contextual Document Embeddings. *CoRR* abs/2410.02525 (2024). <https://doi.org/10.48550/ARXIV.2410.02525> arXiv:2410.02525
- [23] Blaise Munyampirwa, Vihan Lakshman, and Benjamin Coleman. 2024. Down with the Hierarchy: The ‘H’ in HNSW Stands for “Hubs”. *CoRR* abs/2412.01940 (2024). <https://doi.org/10.48550/ARXIV.2412.01940> arXiv:2412.01940
- [24] Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A Unified Framework for Learned Sparse Retrieval. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13982)*, Jaap Kamps, Lorraine Goeriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 101–116. [https://doi.org/10.1007/978-3-031-28241-6\\_7](https://doi.org/10.1007/978-3-031-28241-6_7)
- [25] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier Information Retrieval Platform. In *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21–23, 2005, Proceedings (Lecture Notes in Computer Science, Vol. 3408)*, David E. Losada and Juan M. Fernández-Luna (Eds.). Springer, 517–519. [https://doi.org/10.1007/978-3-540-31865-1\\_37](https://doi.org/10.1007/978-3-540-31865-1_37)
- [26] Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Guiding Retrieval Using LLM-Based Listwise Rankers. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15572)*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonello (Eds.). Springer, 230–246. [https://doi.org/10.1007/978-3-031-88708-6\\_15](https://doi.org/10.1007/978-3-031-88708-6_15)
- [27] Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Quam: Adaptive Retrieval through Query Affinity Modelling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10–14, 2025*, Wolfgang Nejdl, Sören Auer, Meeyoung Cha, Marie-Francine Moens, and Marc Najork (Eds.). ACM, 954–962. <https://doi.org/10.1145/3701551.3703584>
- [28] Mandeep Rathee, V Venkatesh, Sean MacAvaney, and Avishek Anand. 2025. Breaking the Lens of the Telescope: Online Relevance Estimation over Large Retrieval Sets. In *Proceedings of the 48th International ACM SIGIR conference on*

- research and development in Information Retrieval, SIGIR 2025.* ACM. <https://doi.org/10.1145/3726302.3729910>
- [29] Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2825–2837. <https://doi.org/10.1145/3477495.3531766>
- [30] Ellen M. Voorhees. 2020. Competition in IR research. *SIGIR Forum* 54, 2 (2020), 1:1–1:3. <https://doi.org/10.1145/3483382.3483384>
- [31] Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: constructing a pandemic information retrieval test collection. *SIGIR Forum* 54, 1 (2020), 1:1–1:12. <https://doi.org/10.1145/3451964.3451965>
- [32] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *CoRR* abs/2004.10706 (2020). arXiv:2004.10706. <https://arxiv.org/abs/2004.10706>
- [33] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. An Inspection of the Reproducibility and Replicability of TCT-ColBERT. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2790–2800. <https://doi.org/10.1145/3477495.3531721>
- [34] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *ACM J. Data Inf. Qual.* 10, 4 (2018), 16:1–16:20. <https://doi.org/10.1145/3239571>