A Test Collection for Dataset Retrieval

Nikolay Kolyada,[1] Martin Potthast,[2] Benno Stein[1]

[1] Bauhaus-Universität Weimar
[2] University of Kassel, hessian.AI, and ScaDS.AI

Abstract Dataset search has become a relevant retrieval task which, with the current state of the art, remains difficult to be evaluated. The reason for this is probably not a lack of retrieval models, but a lack of suitable test collections that combine data-related information needs, such as descriptions of inductive learning tasks, with answers, namely suitable data sets for estimating model function parameters for the learning task. We have identified Reddit as a promising source for the creation of a dataset retrieval benchmark and describe its construction in detail here. The benchmark consists of a collection of datasets in the form of 9935 metadata descriptions from Papers with Code[3], a set of 814 topics in the form of queries or descriptions of the required dataset properties, and gold standard relevance assessments derived from Reddit discussions about which dataset is best suited. In a pilot study, we evaluate first baselines on answering test requests against the collection.

Keywords: Dataset Search · Test Collection · Benchmark · Dataset Retrieval

1 Introduction

As data science grows, so does the need for suitable data sets for machine learning tasks. The FAIR principles (Findable, Accessible, Interoperable, and Reusable) aim to improve the management of research data [30]. However, as Jacobsen et al. point out, their implementation requires effective retrieval technology, i.e. first and foremost improving findability [11], to precisely match a researcher's data needs with the datasets. Effective retrieval in turn requires evaluation, for which retrieval benchmarks play a key role.

Traditionally, retrieval benchmarks include a document collection, queries, and relevance judgments [6]. Dataset retrieval is more complex and may require a different approach. Published datasets are tagged with metadata supporting classical retrieval, such as names, descriptions, identifiers, or URLs. Existing search engines use this metadata (e.g., schema.org [10], DCAT [23]), approaching the task similarly to classic information retrieval.

This approach has weaknesses: metadata is often incomplete, noisy, and scattered, leading to linkage and disambiguation issues [1]. Users have different expectations on dataset ranking; sometimes the most recent is most relevant, other

[3] https://paperswithcode.com/

Table 1. Statistics of the original Reddit r/datasets snapshot (left), and the subset suitable for our dataset retrieval benchmark (right). Eligible posts had to be a 'Request' for a 'Known-item' or an 'Advice', and include a response mentioning a specific dataset, this way forming query relevance judgments (QRels).

Reddit r/datasets	**Count**
Original Posts (OP)	28,057
Total Replies	84,179
OP of 'Request' type	10,543
– Replies (total)	23,062
– Reply threads	5,163

QRels	**Post type (query)**		
Replies	Known-item	Advice	Sum
Relevant	186	515	701
Non-relevant	21	92	113
Benchmark	**207**	**607**	**814**

times the most comprehensive or highest quality [19]. Retrieval systems must weigh user requirements with query context to provide relevant results. Moreover, queries are often in natural language, requiring systems to understand context and intent [14], necessitating robust mechanisms to interpret and precisely answer such queries.

This paper presents a test collection for dataset retrieval and a novel method for bootstrapping it using curated Reddit "Datasets" threads.[4][5] The benchmark provides a way to evaluate dataset retrieval systems using users' natural language information needs. It includes 9,935 dataset metadata from Papers with Code, 1,054 topics (dataset requests), and gold-standard relevance judgments for 1013 datasets. As first baselines, we evaluate Okapi BM25 and ColBERT in comparison to state of the art LLM-based retrieval system on the collection.

New search applications are emerging, and machine learning techniques are redefining best practices. The importance of effective data search will continue to grow, along with the significance of large dataset corpora, retrieval benchmarks, and dataset search technologies. Developing this ecosystem will enable evaluation of different retrieval strategies and may form the basis for a new shared task (e.g., at TREC), raising awareness of dataset retrieval challenges and promoting progress.

2 Background and Related Work

Dataset retrieval originates from information retrieval, databases, and table searching [4], paralleling bibliographic search systems like Google Scholar, Microsoft Academic Graph, CrossRef [28], CiteSeerx [20], and Semantic Scholar [8] that use metadata, citation graphs, and complex ranking algorithms.

However, dataset retrieval presents unique challenges. Brickley et al. [1] found that enumerating existing dataset repositories, even within a single discipline, is significant. Maier et al. [25] noted scientists struggle to find relevant datasets due to increasing numbers making them less discoverable.

[4]We provide the dataset at https://doi.org/10.5281/zenodo.14679719 and publish the evaluation code at https://github.com/webis-de/ECIR-25.
[5]https://www.reddit.com/r/datasets/

Systems like Google Dataset Search, Figshare, and Zenodo have eased dataset discovery. Google Dataset Search relies on publisher-provided metadata, emphasizing the importance of well-curated metadata, while Zenodo and Figshare facilitate dataset publishing. Attempts to improve dataset retrieval include [13]'s ontology-based semantic search for open government data and [3]'s federated search using APIs like Zenodo's. However, these systems have limitations; effectiveness varies with metadata quality and structure [1] or index bias. Dataset metadata may focus on different attributes than traditional documents, and the "ecosystem" required to exploit a dataset includes programming issues, hardware, experience, and runtime. This illustrates difficulties due to dataset diversity, dispersion, dynamic nature, and metadata quality [27].

The diversity of user intents and their query expressions adds complexity. Dataset retrieval can be a known-item search when the user knows attributes like name or acronym [2]. Alternatively, users may seek recommendations for suitable datasets without specific ones in mind. Recently, [29] addressed dataset recommendation given research needs from paper abstracts.

Recent attempts to create test collections for ad-hoc dataset retrieval include the NTCIR-15 "Data Search" shared task,[6] where Kato et al. [16] introduced a collection based on Japanese and US governmental open data portals and 192 queries from a Japanese Yahoo! Q&A platform. Lin et al. [21] focus on content-based retrieval for RDF datasets from open data portals. In biomedical research, [7] proposed a reference standard for the 2016 bioCADDIE dataset retrieval challenge. Löffler et al. [22] constructed a test collection for biodiversity research with 14 questions, 372 datasets, and binary relevance judgments. [5] proposed enhancing dataset search with data snippets.

Our benchmark differs by expanding the search space and using relevance judgments derived from explicitly approved user recommendations, reflecting real information needs instead of synthetic or crowd-sourced data. It exceeds existing collections in size, with more queries and relevance assessments, and allows incorporating new information needs to ensure continued relevance.

3 Benchmark Construction

The construction of our dataset retrieval benchmark involves (1) collecting dataset requests as queries, (2) extracting gold-standard relevance judgments for them within written dataset recommendations together with the ensuing discussion threads about them as context, and (3) linking datasets found relevant against a large collection of dataset metadata.

3.1 Dataset Requests

Dataset queries are declarative descriptions of user information needs–essentially "long queries". We identified Reddit's `r/datasets`[7] as a rich source for such

[6] https://ntcir.datasearch.jp/data_search_1/
[7] https://www.reddit.com/r/datasets/

queries, using a recent Reddit crawl[8] to extract data. Since its launch in 2009, `r/datasets` has accumulated 28,057 original queries (see Table 1, left), covering a wide range of datasets over the past fifteen years, making the benchmark representative of real-world scenarios. Other platforms like Stackoverflow and Quora were considered but yielded fewer suitable queries and posed extraction challenges due to less homogeneous data.

The community guidelines of r/datasets require posts to be labeled, allowing filtering by discussion type. Most posts are labeled "Request", forming declarative queries about datasets in various scientific fields, making them diverse and representative. Other labels like "Question", "Discussion", or "Announcement" indicate posts not seeking datasets.

We distinguish between two types of dataset requests: known-item searches and open-ended, ad-hoc requests. Known-item searches assume the user knows certain identifying information but cannot locate the dataset:

> **OP.** *Does anyone have the "pirate-bay-torrent-dumps-2004-2016" dataset or a similar piratebay dataset?*
>
> **Reply.** *Here's the web page if directories scare you:* (URL omitted)

Ad-hoc requests are general queries where the user seeks recommendations suitable for their task without a specific dataset in mind:

> **OP.** *I am looking for a dataset which has music files available in mp3 / wav / ogg or any other audio format. It would be plus if that dataset has only the music sounds and not the lyrics stuff.*
>
> **Reply.***: Maybe MusicBrainz can be of help. There's also audio files in the Million Song Dataset.*

3.2 Query Relevance Judgments

Central to our benchmark is a collection of query relevance judgments for the dataset requests (see Table 1, right), which is non-trivial to extract. We use the original poster's approval or rejection of recommendations on r/datasets, assuming their response indicates the dataset's relevance to the query.

On r/datasets, users (OPs) post detailed dataset requests, and others reply with recommendations. We group replies into threads starting from the root post, extracting pairs of OP and direct replies, and treat further replies as independent threads. These pairs are classified as accepted" (relevant) or not accepted" (non-relevant) based on the OP's explicit approval. We omit overly broad requests (e.g., *Looking for some datasets with multiple tables to practice sql with ...*"), threads without answers or deleted content, recommendations to scrape data, and references to data portals without explicit dataset recommendations. The OP's responses serve as a gold standard" for relevance in our benchmark.

While determining relevance is not binary [14], our analysis shows most OPs clearly confirm their assessments with strong affirmative answers. Relevance is

[8] https://files.pushshift.io/reddit/

subjective [9]; what one user finds relevant may differ from another [18], and true suitability may only become apparent during experiments. Therefore, our benchmark provides a relative measure to compare retrieval systems, not an absolute one. We aim for a robust and flexible relevance and ranking framework that considers different users' needs and perspectives.

3.3 Dataset Collection

The collection of dataset metadata, serving as retrieval units or "documents" in our case, represents the corpus to be indexed in the test retrieval systems. The metadata collection for our benchmark, formatted as Schema.org Dataset type,[9] is manually constructed from data sources such as Papers with Code, verified for completeness and augmented with data from Google's "Dataset Search: Metadata for Datasets" collection [1], multiple open data portals, and data sharing repositories such as Zenodo, Kaggle.

4 Pilot Study

To demonstrate the efficacy of the proposed dataset retrieval benchmark, we conduct a pilot study evaluating baseline models BM25, ColBERT [17] implemented by PyTerrier [24], as well as the specialized deep passage retrieval (DPR) [15] with BERT, and in addition RAG system using Mistral7B [12] model.

Each topic in our evaluation includes the original Reddit request, keywords extracted from the request [26], and relevant responses containing a recommended dataset in our metadata collection. The original posts are preprocessed by removing special characters and punctuation, while the keyword queries are the unmodified list of words.

Table 2 summarizes our evaluation results using the effectiveness measures Recall@5 (R@5), Average Rank, and Mean Reciprocal Rank (MRR) as we typically have only one relevant dataset per query. Since the original requests are natural language questions, the BM25 baseline model, which is not optimized for question answering, shows limited performance. However, we observed a significant improvement with keyword queries only with BM25 model. ColBERT outperforms BM25 in terms of R@5., whereas DPR with BERT shows marginal improvement over it. RAG-Mistral7B with DPR retrieval improves significantly while ranking the relevant datasets higher.

5 Limitations

While the implemented benchmark offers a novel approach for evaluating dataset retrieval systems, it is not without limitations. These are largely derived from the inherent complexities of dataset retrieval and the constraints imposed by the source of our dataset request and relevance judgments.

Dataset collection limitations The dataset metadata corpus has limitations compared to web search. Multiple sources publish metadata on the same dataset,

[9] https://schema.org/Dataset

Table 2. Retrieval performance of various dataset search systems for the benchmark presented.

Retrieval Model	R@5	Avg. Rank	MRR
BM25 - OP	0.307	4.7	0.308
BM25 - Keywords	0.336	3.6	0.321
ColBERT - OP	0.365	4.1	0.298
DPR with BERT - OP	0.372	3.7	0.335
RAG - Mistral 7B - OP	**0.385**	**3.4**	**0.348**

making duplicates hard to recognize, and several versions often exist. Therefore, we represent each dataset with a single record in our collection, restricting it to datasets from Papers with Code.

Dataset request limitations Our benchmark's dataset requests come from Reddit's r/datasets. Though diverse, they are few and may be biased, not representing all dataset users. Ambiguous requests can hinder assessing relevance beyond what the original poster found relevant.

Relevance judgment limitations Our benchmark's relevance judgments rely on the original poster's assessment of replies in the r/datasets community, assuming they accurately indicate the relevance of suggested datasets. However, this may not always hold true; the poster might lack expertise, and participants may not provide all relevant datasets. Thus, only the original poster's explicit confirmation of a recommendation is considered a relevance judgment.

Benchmark limitations Our benchmark assesses retrieval systems based on their ability to meet specific dataset needs from requests. It doesn't annotate other aspects like the quality of Google's metadata collection, ambiguous or poorly specified requests, or the completeness of recommendations. However, some discussion threads were answered exhaustively, making our benchmark realistic in those cases.

Despite limitations, our benchmark offers a valuable resource for evaluating dataset retrieval systems. It addresses unique challenges and points to future work: expanding datasets, diversifying requests, and refining relevance judgments.

6 Conclusion and Future Work

We present a novel dataset retrieval benchmark using natural language requests from real users, addressing the need for standardized evaluation in dataset search. Comprising a dataset collection, user requests, and gold-standard relevance judgments, it offers a straightforward way to assess retrieval systems' effectiveness. We demonstrated its utility by providing baselines. Future research could address limitations like the number of requests and potential biases in their distribution and relevance judgments. Refining the generation of requests and judgments, possibly by leveraging additional open data communities, would be beneficial. In conclusion, our benchmark is an important step toward rigorous evaluation of dataset retrieval systems, and we anticipate it will serve as a foundation for future research in this evolving field.

Bibliography

[1] Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference, pp. 1365–1375, WWW '19, Association for Computing Machinery, New York, NY, USA (May 2019)

[2] Broder, A.: A taxonomy of web search. SIGIR Forum **36**(2), 3–10 (Sep 2002)

[3] Castelo, S., Rampin, R., Santos, A., Bessa, A., Chirigati, F., Freire, J.: Auctus: a dataset search engine for data discovery and augmentation. Proceedings VLDB Endowment **14**(12), 2791–2794 (Jul 2021)

[4] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. VLDB J. **29**(1), 251–272 (Jan 2020)

[5] Chen, Q., Chen, J., Zhou, X., Cheng, G.: Enhancing dataset search with compact data snippets. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA (Jul 2024)

[6] Cleverdon, C.: The cranfield tests on index language devices. In: Aslib proceedings, vol. 19, pp. 173–194, MCB UP Ltd (1967)

[7] Cohen, T., Roberts, K., Gururaj, A.E., Chen, X., Pournejati, S., Alter, G., Hersh, W.R., Demner-Fushman, D., Ohno-Machado, L., Xu, H.: A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 biocaddie dataset retrieval challenge. Database J. Biol. Databases Curation **2017**, bax061 (2017), https://doi.org/10.1093/DATABASE/BAX061, URL https://doi.org/10.1093/database/bax061

[8] Fricke, S.: Semantic scholar. J. Med. Libr. Assoc. **106**(1), 145 (Jan 2018)

[9] Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S.: Understanding data search as a socio-technical practice. J. Inf. Sci. Eng. **46**(4), 459–475 (Aug 2020)

[10] Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Commun. ACM **59**(2), 44?51 (jan 2016), ISSN 0001-0782, https://doi.org/10.1145/2844544, URL https://doi.org/10.1145/2844544

[11] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C.T., et al.: Fair principles: interpretations and implementation considerations (2020)

[12] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B. arXiv [cs.CL] (Oct 2023)

[13] Jiang, S., Hagelien, T.F., Natvig, M.K., Li, J.: Ontology-based semantic search for open government data. In: 13th IEEE International Conference on Semantic Computing, ICSC 2019, Newport Beach, CA, USA, January 30 - February 1, 2019, pp. 7–15, IEEE, Newport Beach, CA, USA (2019), https://doi.org/10.1109/ICOSC.2019.8665522, URL https://doi.org/10.1109/ICOSC.2019.8665522

[14] Kacprzak, E., Koesten, L., Tennison, J., Simperl, E.: Characterising dataset search queries. In: Companion Proceedings of the The Web Conference 2018, pp. 1485–1488, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (Apr 2018)

[15] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.T.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Stroudsburg, PA, USA (2020)

[16] Kato, M.P., Ohshima, H., Liu, Y., Chen, H.O.: A test collection for ad-hoc dataset retrieval. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pp. 2450–2456, ACM, Virtual Event, Canada (2021), https://doi.org/10.1145/3404835.3463261, URL https://doi.org/10.1145/3404835.3463261

[17] Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pp. 39–48, ACM, Virtual Event, China (2020), https://doi.org/10.1145/3397271.3401075, URL https://doi.org/10.1145/3397271.3401075

[18] Koesten, L., Simperl, E., Blount, T., Kacprzak, E., Tennison, J.: Everything you always wanted to know about a dataset: Studies in data summarisation. International journal of human-computer studies **135**, 102367 (2020)

[19] Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E.: The trials and tribulations of working with structured data: -a study on information seeking behaviour. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1277–1289, CHI '17, Association for Computing Machinery, New York, NY, USA (May 2017)

[20] Li, H., Councill, I., Lee, W.C., Giles, C.L.: CiteSeerx: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th international conference on World Wide Web, pp. 883–884, WWW '06, Association for Computing Machinery, New York, NY, USA (May 2006)

[21] Lin, T., Chen, Q., Cheng, G., Soylu, A., Ell, B., Zhao, R., Shi, Q., Wang, X., Gu, Y., Kharlamov, E.: ACORDAR: A test collection for ad hoc content-based (RDF) dataset retrieval. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, pp. 2981–2991, ACM, Madrid, Spain (2022), https://doi.org/10.1145/3477495.3531729, URL https://doi.org/10.1145/3477495.3531729

[22] Löffler, F., Schuldt, A., König-Ries, B., Bruelheide, H., Klan, F.: A test collection for dataset retrieval in biodiversity research. Res. Ideas Outcomes **7** (May 2021)

[23] Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (dcat). w3c recommendation. World Wide Web Consortium pp. 29–126 (2014)

[24] Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using pyterrier. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, pp. 161–168, ACM, Virtual Event, Norway (2020),

https://doi.org/10.1145/3409256.3409829, URL https://doi.org/10.1145/3409256.3409829

[25] Maier, D., Megler, V.M., Tufte, K.: Challenges for dataset search (2014)

[26] Martinc, M., Škrlj, B., Pollak, S.: KID: Transformer-based neural tagger for keyword identification. Natural Language Engineering **28**(4), 409–448 (Jul 2022)

[27] Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. J. Data and Information Quality **8**(1), 1–29 (Oct 2016)

[28] Pentz, E.: Crossref: The missing link. College & research libraries news **62**(2), 206–228 (2001)

[29] Viswanathan, V., Gao, L., Wu, T., Liu, P., Neubig, G.: Datafinder: Scientific dataset recommendation from natural language descriptions. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 10288–10303, Association for Computational Linguistics, Toronto, Canada (2023), https://doi.org/10.18653/V1/2023.ACL-LONG.573, URL https://doi.org/10.18653/v1/2023.acl-long.573

[30] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)