Content-Based Dataset Retrieval Methods: Reproducibility of the ACORDAR Test Collection

Laura Menotti[0000−0002−0676−682X], Manuel Barusco[0009−0000−0568−2432], Riccardo Forzan[0009−0001−6621−7208], and Gianmaria Silvello[0000−0003−4970−4554]

Department of Information Engineering, University of Padua, Padua, Italy
{name.surname}@unipd.it

Abstract. The FAIR principles constitute a cornerstone of contemporary scientific methodology, with the Digital Library (DL) community actively participating and providing significant advancements within this framework. By taking a reproducibility approach, this paper centers on findability, a pivotal aspect of scientific data management and stewardship. Specifically, we delve into the critical role of Data Search in enabling efficient retrieval across various contexts, including scholarly publications and scientific data management. Consequently, the convergence of Digital Library and Information Retrieval (IR) domains underscores the necessity to adapt document-level IR techniques to optimize dataset retrieval processes.

Dataset retrieval relies on dataset descriptions, hampered by incomplete and inconsistent metadata issues. Lately, there has been a growing emphasis on Content-Based Dataset Retrieval (CBDR), where metadata and dataset content are equally considered during indexing and retrieval. ACORDAR is the first open test collection to evaluate CBDR methods. It offered early insights into the benefits of integrating dataset content in retrieval.

Our study thoroughly assesses ACORDAR's quality and reusability while investigating the reproducibility of retrieval results. Concerns arise about accessibility to the collection's content due to broken links for 17.6 of datasets. Despite some errors and requiring non-trivial pre-processing steps, we replicated most but not all CBDR methods, thus raising some concerns about the suitability of ACORDAR as a reference test collection to further advance CBDR research and to employ these methods in the context of DL.

Keywords: Data Search and Discovery in Digital Libraries · FAIR Data · Data Management · Information Retrieval.

1 Introduction

The fundamental importance of data in scientific advancement has prompted institutions and the Digital Library (DL) community to collaborate on initiatives

to promote data accessibility by exploiting open and FAIR (Findable, Accessible, Interoperable, and Reusable) data policies. To effectively achieve these principles, it is imperative to represent datasets in a machine-readable and interoperable format. Thus, numerous scientific datasets are now shared online as Resource Description Framework (RDF) graphs of varying complexity and sizes. Dataset discovery is crucial for ensuring *Findability*, hence promoting the application of the FAIR principles in scientific data management. Such an approach will benefit many applications, ranging from Digital Libraries, Open Data Portals, to the web in general. Open Data Portals are web-based interfaces designed to ease the search for reusable information, as they provide a comprehensive ensemble of datasets. For instance, the data catalog registry [1] lists 597 open data portals worldwide. Examples are the US data portal [2], which comprises 250,723 datasets, and the European Commission's portal [3], which includes 1,610,143 datasets harvested from 176 national and European portals.

Dataset retrieval is emerging as a subfield of Information Retrieval (IR) [8] and involves the adaptation of document-level IR methods to facilitate the retrieval of datasets in Open Data Portals, Digital Libraries, and on the web [10]. Dataset retrieval also contributes to scholarly publication since enabling the discovery of datasets will facilitate integrating authors and publications in scholarly knowledge graphs.

Existing industry-scale retrieval systems rely on metadata for dataset search [2, 3]. Indeed, datasets are commonly accompanied by manually curated metadata, which includes crucial details such as the authors, title, and a concise description. Nevertheless, metadata quality is frequently marred by inconsistencies, incompleteness, and low reliability [1, 5, 7]. When metadata quality is subpar, dataset search systems struggle to provide useful results, leaving data practitioners to search among several portals in an exploratory manner. This process is time-consuming and often frustrating for users [1, 5]. To mitigate these issues, recent works have shifted from metadata-based methods to a mixed approach called Content-Based Dataset Retrieval (CBDR) that leverages the content of datasets to facilitate their discovery [9, 11]. However, several efforts are still needed to assess the impact of datasets' content in the retrieval task and to evaluate CBDR systems.

Reliable and shared test collections are required to evaluate and improve CBDR systems. However, in the context of dataset retrieval, few test collections have been released in the literature. The first is the NTCIR-15 (English) test collection [6], consisting of 46,615 datasets from the US Data portal [4] in RDF and CSV formats. It comprises 192 queries developed by mining real information needs from questions in a community question-answering device. The relevance judgments of the test collection were gathered by considering only the metadata of the dataset, making it not ideally suited for evaluating CBDR systems. Hence,

[1] http://datacatalogs.org/
[2] https://data.gov/
[3] https://data.europa.eu/en
[4] https://data.gov/

the Ad hoc COntent-based RDF DAtaset Retrieval (ACORDAR) collection [9] was introduced as the pioneering test collection for CBDR systems, incorporating metadata and content and focusing only on RDF datasets. The authors conducted an evaluation, leveraging standard retrieval models, to demonstrate that including dataset content alongside metadata results in a more effective search than using metadata alone.

In this work, we analyzed ACORDAR, focusing on its quality and reusability. Our analysis revealed the presence of duplicated datasets in ACORDAR. We also noticed that whenever a dataset imported a portion of an external resource, e.g., Friend Of A Friend (FOAF) or DBpedia, the whole content of such resource was imported into the dataset, introducing noise. Concerning the reusability of the collection, some concerns arise from using non-permanent URIs for the datasets; indeed, we could not download 17.6% datasets. Considering only the dataset with available content, we successfully reproduced the results presented in the reference paper for three out of four tested retrieval systems. To ensure the applicability of our study, we released in Zenodo the test collection with the content information used to run the experiments. [5] We published the code to build the test collection [6] and to reproduce the experiments. [7] In addition, we provide further experiments to investigate data's impact during retrieval. To this end, we discovered that content is marginal in the presented results as the retrieval models rely almost entirely on metadata to return relevant datasets.

The rest of the paper is organized as follows. Section 2 presents the ACORDAR test collection and all available resources to reproduce the work. Section 3 analyzes ACORDAR and describes our efforts to obtain the test collection. Section 4 presents and compares the experimental results with the reference paper. Section 5 investigates the impact of data during retrieval. Section 6 concludes the paper and draws final remarks.

2 Original Contribution

The ACORDAR test collection comprises 31,589 RDF datasets, 493 queries, and 10,671 relevance judgments. The corpus is based on the RDF datasets downloaded from 543 data portals. The queries are divided into "TREC queries" and "synthetic queries." The former comes from the Ad-Hoc Text REtrieval Conference (TREC) 1-8, [8] while the latter was created anew by the ACORDAR assessors. The ground truth has been built employing four standard retrieval models, namely TF-IDF, BM25F, Fielded Sequential Dependence Model (FSDM) [13], and Language Model using Dirichlet priors for smoothing (LMD) [12], which were used for pooling with depth 10. The authors of the reference paper also conducted some experimental analysis, evaluating the four retrieval models measuring nDCG@{5,10} and AP@{5,10}. Besides the release of the test collection,

[5] URL provided upon acceptance.
[6] https://github.com/mntlra/ACORDAR-Repro-py
[7] https://github.com/mntlra/ACORDAR-retrieval
[8] https://trec.nist.gov/data/topics_eng/index.html

the reference paper has shown that the joint use of content and metadata improved the retrieval performance for all the tested systems, compared with the same systems indexing only metadata. Overall, FSDM obtained the best performances across all evaluation measures.

ACORDAR has been released in GitHub to enhance reuse. [9] It provides the corpus, the `qrels` in the standard TREC format, all queries, and the runs produced by the experimental evaluation. The authors did not provide the source code nor the index used to produce the runs, and the reference paper lacks some details about the indexing configuration and the retrieval strategy. By exploring the authors' GitHub profile, we found a second GitHub repository (ACORDAR 2.0), [10] which was not referenced in the paper, where they released a second version of the test collection, with different queries, `qrels`, and retrieval results. This repository also comprises the code for the four retrieval models used in the reference paper. There is a third GitHub repository called CADDIE, [11] not directly related to ACORDAR, but providing some insights on the triples deduplication strategy employed to build the test collection.

3 Analysis of the ACORDAR Test Collection

ACORDAR is released as a JSON file with the metadata of each dataset and a list of URLs referring to external data portals to download their content. Concerning the usability of the test collection, only metadata can be straightforwardly used for indexing. To get the content, the RDF datasets files must be downloaded and parsed from the provided list of URLs. Analyzing the URLs in the test collection, we discovered that 98.4% of the datasets (31,094) provide a single URL pointing to a dump file. The remaining 495 datasets provide from 2 (203 datasets, 41% of the remaining) to 417 URLs. Thus, if we cannot download and parse all files, the datasets with multiple URLs would contain partial information. When a dataset imports an external ontology or dataset, such resource is listed as the content of the dataset. To give an idea, the same FOAF file is downloaded for 129 datasets, while DBpedia is downloaded in different versions for seven datasets. The presence of resources like DBpedia introduces general-purpose information unrelated to the content of the dataset, which may hinder retrieval performances. As an example, consider the dataset #14530, which contains information about the city of Madrid and comprises two URLs: `Madrid.nt`, the actual content of the dataset, and `dbpedia-2014.owl`, as a supporting vocabulary. Such a dataset is relevant to queries as "*English and Spanish Terminology*" and "*Spanish Terminology*." However, the system employing BM25F and indexing only data fields returns the dataset #14530 in the top 5 ranking list also for the query "*Non-commercial Satellite Launches*." This happens because the DBpedia dump associated with the dataset contains the keyword "satellite." This is

[9] https://github.com/nju-websoft/ACORDAR/
[10] https://github.com/nju-websoft/ACORDAR-2/
[11] https://github.com/nju-websoft/CADDIE

just an example of how DBpedia could be considered relevant for most queries given its encyclopedic nature.

Analyzing ACORDAR, we noticed the presence of duplicated datasets. We identified 13 dataset pairs differing from each other for the identifier while sharing the same content and metadata. If we restrict our analysis to the indexed metadata fields – i.e., *title*, *description*, *author*, *tags* – and the download URLs for the content, we found 196 non-unique datasets, with a notable case of a dataset with five duplicates. Duplicate datasets may arise because data portals were crawled at various times to construct the test collection. In certain instances, it proved challenging to ascertain whether a dataset was already included in the corpus. However, the influence of duplicated datasets is limited as most of these datasets were not assessed in the relevance judgments.

We performed two steps to build the test collection: (1) download of the collection to access the content of the datasets; and, (2) parsing of the datasets to build the index. We used a Python script to download the datasets, retrieving files via HTTP calls. If a download did not get through the first time, we retried up to twenty times at different moments to avoid losing files due to connection issues. The test collection comprises 34,484 URLs with only 14 invalid URLs. We kept all the URLs pointing to RDF datasets and discarded those pointing to HTML documents, images, and textual files. Overall, we managed to access 28,506 (82.7%) URLs, accounting for 25,930 (82.1%) complete dataset downloads and 93 (0.3%) partial datasets, for which only a part of the available files was correctly downloaded. We could not access 5,978 URLs, accounting for 5,566 datasets without content, 17.6% of the total, due to broken links (70.5% of the total errors) and forbidden resources (20%). Almost all the downloaded files are valid RDF serialization formats (98.9%), while the rest comprised textual files and compressed archives. We downloaded the collection several times, in different periods, and noticed that we obtained different results when downloading the collection after some weeks. Hence, sharing a test collection via non-permanent URLs to external data portals could cause significant information loss. The authors of the reference paper also struggled to retrieve datasets in the first place, stating that from the 111,017 RDF datasets available, they ended up with only 31,589 for the same issues we experienced.

We parsed all datasets and extracted the so-called *data fields* for indexing: classes, entities, properties, and literals. In the following, we refer to such information as "*indexable content*." In practice, for every graph, the four data fields are extracted and concatenated together to build a textual document – acting as a proxy for the dataset – that can be indexed and retrieved. The authors of the reference paper used Apache Jena for parsing the RDF and running the experiments. They also performed deduplication to remove redundant RDF triples from the datasets, but they did not provide any detail about this process.

To obtain results closer to the reference paper and maximize the extraction of the indexable content, we tested three parsing tools, namely Apache Jena,

LightRDF, [12] and RDFLib. [13] In all cases, we limited the number of parsed files per dataset to one hundred. This limitation only affected one dataset (# 14054), including an archive TAR file containing 73,204 small files. Following the reference paper, we developed a parser that exploits the Apache Jena library and resorts to LightRDF only for files that raise exceptions or are too large (larger than 300MB) to be handled with Apache Jena. For files bigger than 1GB, we considered only the first 500,000 triples. Subsequently, all triples referring to the same dataset are deduplicated using Minimum Spanning Graphs (MSGs) and a set-based approach implemented in the CADDIE repository. We also developed two parsing strategies entirely in Python, one leveraging only RDFLib and the other using an ensemble of RDFLib and LightRDF. RDFLib exploits a SPARQL-based parser, which loads the whole content of the RDF file in memory. For this reason, in the first parsing strategy, we limited the size of files to 150 MB, thus excluding 14 files. In the second strategy, we parsed files up to 200 MB with RDFLib, employing LightRDF for bigger files. Strategies employing Apache Jena and LightRDF treat files as a stream, allowing the parsing of partial files until the first exception. Nevertheless, we discarded partial files from the final results to limit the noise injection in the index.

Overall, all strategies correctly parsed almost the same amount of files and produced identical runs using the four retrieval models, showing no statistical difference. For this reason, we decided to keep the results from the RDFLib parser developed in Python, as it is best tailored for data processing pipelines and enhances reusability. We extracted the indexable content once we parsed all the files, limiting each data field to 100,000 items per dataset, as in the code provided in ACORDAR 2.0. We could correctly parse and extract indexable information for 28,207 files, 98.4% of the available ones. Concerning the test collection, we could completely retrieve content, i.e., download and parse all files, for 25,707 datasets (81.4%) and partially retrieve content-based information for 72 datasets (0.2%). We could not download and parse the content of 5,810 datasets, accounting for 18.4% of the total, due to exceptions in the download (5,566 datasets, 95.8%) and in the parsing phase (244 empty datasets, 4.2%).

4 Reproducibility Results

The test collection has been used in the experiments with three different configurations: the first indexing only metadata (*Metadata Configuration*), the second indexing only content (*Content Configuration*), and the third indexing metadata and content (*Full Configuration*). We analyzed the impact of empty datasets in the ground truth, with a particular interest in queries for which such datasets are considered relevant. Overall, we found 1,652 query-dataset pairs (15.5%) that refer to an empty dataset, of which 626 are judged as partially relevant (393) or highly relevant (233). Concerning the runs provided in the reference paper, we analyze the Content and the Full Configuration. The empty datasets will

[12] https://github.com/ozekik/lightrdf
[13] https://github.com/RDFLib/rdflib

not affect the Metadata Configuration since their fields are also available for the empty datasets. In the Content Configuration, 452 queries return at least one empty dataset across the four considered retrieval models; in particular, BM25F returns at least one empty dataset for 331 (out of 493) queries, TF-IDF for 343, LMD for 337, and FSDM for 346 queries. Thus, the retrieval models mark at least one empty dataset as relevant for more than half of the queries. In the Full Configuration instead, BM25F returns at least one empty dataset for 302 (out of 493) queries, TF-IDF for 319, LMD for 295, and FSDM for 330 queries. From this analysis, we expect an impact of empty datasets on retrieval results over the Complete Collection when comparing our results to the original ones.

All retrieval models rely on field weights tuned using grid search optimizing for nDCG@10. No other information about the experimental setting is provided in the reference paper. In the ACORDAR 2.0 GitHub repository, the field boost weights were provided, and we found that they used the `StandardAnalyzer` by Lucene, with the addition of the NLTK stopwords list for English. About the retrieval models, for TF-IDF, BM25F, and LMD we used the similarity function available in Lucene, i.e., `ClassicSimilarity`, `BM25Similarity`, and `LMDirichletSimilarity` respectively. FSDM is used as a reranking function, and it has been developed by the authors of the reference paper following [13]. Thus, we run our experiments using the code provided by the authors in the ACORDAR 2.0 repository. We noticed that FSDM field weights are required to sum to one, but the original work does not meet this constraint.

In the following, we compare the original results with the one we reproduced. We run the experiments on the Complete Collection, including empty datasets, and the Restricted Collection, excluding the empty datasets. Tables 1, 3, and 5 refers to the Complete Collection. Whereas Tables 4 and 6 report the results on a Restricted Collection. We removed the empty datasets from the original runs for the restricted version.

Metadata Configuration. Table 1 reports the experimental results for the four retrieval models indexing only the metadata fields and considering the Complete Collection. We report in bold the absolute differences larger than 0.01. Concerning the runs employing TF-IDF and BM25F, our results are consistent with those reported in the reference paper. We conducted a paired t-test ($p \leq 0.05$) for each run on each evaluation measure and observed no statistical difference. In particular, there is no statistical difference using BM25F for all evaluation measures. Regarding TF-IDF, runs are statistically different for the nDCG@10 and MAP@10 measures. In contrast, our LMD run reports a statistically significant ($p \leq 0.01$) improvement over the reference run for all the measures. Concerning FSDM, although we used the implementation provided by the authors, we could not achieve similar results to those reported in the reference paper, with a statistically significant ($p \leq 0.01$) performance gain. The only exception is the gain with MAP@5, which is statistically significant only with $p \leq 0.05$.

We can see that for TF-IDF and BM25F, we reproduced the original results; on the contrary, for FSDM and LMD, we obtained better results for all measures,

Table 1. Original and reproduced results for the *Metadata* Configuration on the *Complete Collection*. Differences greater than 0.01 are reported in bold. † ($p \leq 0.05$) and ‡($p \leq 0.01$) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.4743	0.5019	0.2676	0.3685
	Reproduced	0.4659	0.4855‡	0.2633	0.3519‡
	Difference	-0.0084	**-0.0164**	-0.0043	**-0.0166**
BM25F	Original	0.5045	0.5250	0.2859	0.3838
	Reproduced	0.5059	0.5163	0.2886	0.3778
	Difference	+0.0014	-0.0087	+0.0027	-0.0060
FSDM	Original	0.4853	0.4958	0.2770	0.3516
	Reproduced	0.5221‡	0.5401‡	0.2981†	0.3979‡
	Difference	**+0.0368**	**+0.0443**	**+0.0219**	**+0.0463**
LMD	Original	0.4363	0.4573	0.2543	0.3325
	Reproduced	0.4532‡	0.4702‡	0.2682‡	0.3456‡
	Difference	**+0.0169**	**+0.0129**	**+0.0139**	**+0.0131**

showing that the retrieval models were slightly underperforming in the reference paper. Excluding the empty datasets does not affect the Metadata Configuration.

Table 2 reports the evaluation measures for the same configuration presented in Table 1, but considering the Restricted Collection. Since this configuration is not influenced by the content of each dataset and Table 1 already showed comparable results, we expect the same behavior. Indeed, we obtain almost identical performances to the original results for the TF-IDF and BM25F runs. The nDCG@10 and MAP@10 of the TF-IDF run differ from the original result by only 0.0001 and 0.0004, respectively. In this case, we perform better on all measures using LMD and on nDCG@10 and MAP@10 using BM25F. However, we only have a statistically significant improvement in the nDCG@10 and MAP@10 using the LMD retrieval model. One peculiarity of these results is that using LMD as a retrieval model, we achieve better performances, even if there is no statistical difference on most of the measures, apart from nDCG@10 and MAP@10. Regarding the FSDM, there is a notable contrast in performance compared to the reference paper, as corroborated by the results of both statistical tests.

Content Configuration. Table 3 reports the evaluation measures for the four retrieval models on the Complete Collection indexing the four data fields (i.e., classes, entities, properties, and literals) and not the metadata fields. In this case, our results have a significant performance gap compared to the reference paper. In all cases, the difference with the original is larger than 0.01. As we saw in the Metadata Configuration, FSDM is the model experiencing the most significant shift in performance, with a peak gain of 0.0939 on nDCG@10. We conducted paired t-tests ($p \leq 0.01$) on all runs and measures, always showing statistical differences.

Table 2. Original and reproduced results for the *Metadata* Configuration. Results refer to the *Restricted Collection*, i.e. without the empty datasets. Differences greater than 0.01 are reported in bold. † (p-value$\leq$ 0.05) and ‡(p-value$\leq$ 0.01) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.4663	0.4735	0.2874	0.3621
	Reproduced	0.4594	0.4734	0.2833	0.3625
	Difference	-0.0069	-0.0001	-0.0041	+0.0004
BM25F	Original	0.4809	0.4855	0.2921	0.3679
	Reproduced	0.4761	0.4912	0.2913	0.3737
	Difference	-0.0048	+0.0057	-0.0008	+0.0058
FSDM	Original	0.4524	0.4482	0.2742	0.3312
	Reproduced	0.4934‡	0.5112‡	0.3045‡	0.3932‡
	Difference	**+0.0410**	**+0.0630**	**+0.0303**	**+0.0620**
LMD	Original	0.4164	0.4281	0.2601	0.3242
	Reproduced	0.4261	0.4454‡	0.2670	0.3381‡
	Difference	+0.0097	**+0.0173**	+0.0069	**+0.0139**

We recall that 5,810 datasets in our test collection do not have any content available due to download errors or parsing failures. Since, in this configuration, we only index fields related to the content of each dataset, such datasets are treated as empty documents. To check if the empty datasets cause the performance gap, Table 4 reports the results for the same configuration presented in Table 3 but considering the Restricted Collection. When we consider only datasets with content available, our results are consistent with those reported in the reference paper. The nDCG@10 is identical to the reference paper for the TF-IDF run, while there is a difference of 0.0006 for the LMD run. We conducted a paired t-test ($p \leq 0.05$) for each run on each measure and saw only a statistical difference between all measures of the FSDM run and three measures of BM25F for the reference paper. If we restrict the p-value to 0.01, only the FSDM run presents a statistical difference from the original results. From these outcomes, we can state that the experimental configuration for TF-IDF, BM25F, and LMD is the same as in the reference paper, and the empty datasets caused the gap between the original and reproduced runs. Concerning FSDM, we could not achieve similar results to those reported in the reference paper.

Full Configuration. Table 5 reports the results of the Complete Collection with the Full Configuration. Also, in this case, our results have a sizable performance gap compared to the reference paper. FSDM is the model experiencing the most significant shift in performance, where differences are an order of magnitude bigger than the other retrieval models, with a peak gap of -0.1775 on nDCG@10. This behavior confirms the trend of previous configurations. We conducted a paired t-test ($p \leq 0.05$) on all runs and evaluation measures and saw that all the runs were statistically different from those reported in the reference paper. If

Table 3. Original and reproduced results for the *Content* Configuration on the *Complete Collection*. Differences greater than 0.01 are in bold. † ($p \leq 0.05$) and ‡($p \leq 0.01$) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.1910	0.1963	0.0998	0.1199
	Reproduced	0.1561‡	0.1614‡	0.0753‡	0.0917‡
	Difference	**-0.0349**	**-0.0349**	**-0.0245**	**-0.0282**
BM25F	Original	0.2163	0.2196	0.1385	0.1550
	Reproduced	0.1730‡	0.1743‡	0.1019‡	0.1148‡
	Difference	**-0.0433**	**-0.0453**	**-0.0366**	**-0.0402**
FSDM	Original	0.2497	0.2606	0.14787	0.1758
	Reproduced	0.1614‡	0.1527‡	0.0973‡	0.1046‡
	Difference	**-0.0749**	**-0.0939**	**-0.0370**	**-0.0566**
LMD	Original	0.2398	0.2523	0.1415	0.1672
	Reproduced	0.2077‡	0.2119‡	0.1127‡	0.1331‡
	Difference	**-0.0321**	**-0.0404**	**-0.0288**	**-0.0341**

we restrict the p-value to 0.01, the only results that do not present a statistical difference with the reference paper are nDCG@5 and MAP@5 for the run that uses BM25F and MAP@5 for the TF-IDF run.

To verify if the empty datasets cause the performance gap, Table 6 reports the results for the Restricted Collection. In this case, the original results are reproduced. In particular, most of the measures for the TF-IDF and BM25F runs have a difference smaller than 0.01, and as low as 0.0001 for MAP@10 of the TF-IDF run. The LMD run presents differences smaller than 0.01 only for the nDCG@10. Results on the FSDM run do not show much improvement compared to the run on the Complete Collection, confirming the FSDM behavior is not entirely caused by the absence of content for some datasets. We conducted a paired t-test ($p \leq 0.05$) for each run on each measure and saw a statistical difference in half of the tested cases. If we restrict the p-value to 0.01, for FSDM there is a statistical difference on all measures. TF-IDF and BM25F do not present a statistical difference from the reference paper.

5 Further Analyses on the Impact of Data

One of the primary objectives of the ACORDAR initiative is to emphasize the importance of searching datasets based on their content rather than solely relying on metadata. The findings presented in Section 4 support this assertion. Specifically, when comparing the performance of TF-IDF, BM25F, and LMD retrieval models, it becomes evident that the Full Configuration yields superior results to the Metadata Configuration. This improvement is evident when the results presented in Tables 4 and 6 demonstrate overall enhancements across various measures and retrieval models.

Table 4. Original and reproduced results for the *Content* Configuration on the *Restricted Collection*. Differences greater than 0.01 are reported in bold. † ($p \leq 0.05$) and ‡($p \leq 0.01$) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.1790	0.1795	0.1007	0.1160
	Reproduced	0.1689	0.1795	0.0934	0.1126
	Difference	**-0.0101**	0.0000	-0.0073	-0.0034
BM25F	Original	0.1983	0.1994	0.1319	0.1451
	Reproduced	0.1835†	0.1897	0.1166†	0.1322†
	Difference	**-0.0148**	-0.0097	**-0.0153**	**-0.0129**
FSDM	Original	0.2420	0.2428	0.1517	0.1721
	Reproduced	0.1718‡	0.1664‡	0.1121‡	0.1207‡
	Difference	**-0.0702**	**-0.0764**	**-0.0396**	**-0.0514**
LMD	Original	0.2279	0.2320	0.1412	0.1617
	Reproduced	0.2213	0.2314	0.1329	0.1563
	Difference	-0.0066	-0.0006	-0.0083	-0.0054

However, to substantiate these observations, we conducted a paired t-test comparing the Metadata Configuration and Full Configuration runs. The analysis revealed that, for several measures and retrieval models, including nDCG@5, MAP@5, and MAP@10 using TF-IDF, as well as MAP@5 using LMD, there was no statistically significant difference ($p \leq 0.01$) between the two configurations. Consequently, metadata plays a crucial role in influencing performance, as content-based approaches have only a marginal impact.

To further investigate the role of data, we carried out two experiments that leverage the content of the datasets. Specifically, we developed a re-ranking method focusing on features extracted from the structure of each dataset, and we explored an alternative indexing strategy considering only the top 20 nodes with the highest betweenness centrality as content information.

As a re-ranking method, we trained a Random Forest Regressor to predict the score of each document based on features such as the number of classes, the number of connections in the graph, and the number of connected vertices. The predicted score is then used to re-rank the datasets retrieved by the standard pipeline.

Table 7 reports the retrieval performances of the re-ranking method using BM25F, TF-IDF, and LMD as retrieval models and considering the Restricted Collection. Compared to the results presented in Section 4 for the Restricted Collection, we see a statistically significant gain ($p \leq 0.01$) for the Metadata and Full Configurations across all measures, except for nDCG@10 and MAP@10 for the LMD with the Full Configuration. Concerning the Content Configuration, performances are slightly worse, with no statistical difference between the re-ranked runs and the runs presented before in the Content Configuration except for the nDCG@5 using TF-IDF and the nDCG@5 and nDCG@10 using LMD.

Table 5. Original and reproduced results for the *Full* Configuration on the *Complete Collection*. Differences greater than 0.01 are reported in bold. † ($p \leq 0.05$) and ‡($p \leq 0.01$) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.5088	0.5452	0.2871	0.3976
	Reproduced	0.4801‡	0.5101‡	0.2711†	0.3711‡
	Difference	**-0.0287**	**-0.0351**	**-0.0160**	**-0.0265**
BM25F	Original	0.5538	0.5877	0.3198	0.4358
	Reproduced	0.5366†	0.5600‡	0.3077†	0.4119‡
	Difference	**-0.0172**	**-0.0277**	**-0.0121**	**-0.0239**
FSDM	Original	0.5932	0.6151	0.3592	0.4602
	Reproduced	0.4740‡	0.4376‡	0.2762‡	0.3122‡
	Difference	**-0.1192**	**-0.1775**	**-0.0830**	**-0.148**
LMD	Original	0.5465	0.5805	0.3266	0.4324
	Reproduced	0.4958‡	0.5343‡	0.2890‡	0.3916‡
	Difference	**-0.0507**	**-0.0462**	**-0.0376**	**-0.0408**

The second strategy involved using the 20 nodes with the highest betweenness centrality as an alternative to the data fields proposed in the reference paper. This experiment investigates the impact of reducing the content of each dataset only to the most influencing node, which often acts as a bridge from one part of a graph to another. Replacing the proposed content information with a distinct field containing only the top twenty nodes with the highest betweenness centrality yielded comparable performance results in both the Content and the Full Configurations. We compared these results with the runs without the inclusion of boosting weights to ensure the fairness of the evaluation since we were utilizing a novel field for which boost weights were not available.

In Table 8, we present the results for the Content Configuration, where it becomes evident that the distinction between indexing all graph nodes and solely the top twenty nodes is minimal and lacks statistical significance across all measures when using BM25F and MAP@5 with LMD ($p \leq 0.01$). However, it is worth noting that there is a decline in performance for TF-IDF and LMD. This experiment underscores that incorporating information into the index for a maximum of one hundred thousand nodes or just twenty central nodes has a comparable impact on retrieval. There is even a slight performance improvement in certain instances, although not statistically significant. This analysis highlights the potential for optimizing content indexing, as it appears to be unnecessary to index all nodes within RDF graphs. Simultaneously, it underscores the relatively limited role of content in the ACORDAR collection compared to metadata.

6 Conclusions

We analyzed ACORDAR, the first test collection for Ad-hoc Content-Based RDF Dataset Retrieval, focusing on the reusability of the collection and repro-

Table 6. Original and reproduced results for the *Full* Configuration on the *Restricted Collection*. Differences greater than 0.01 are reported in bold. † ($p \leq 0.05$) and ‡($p \leq 0.01$) indicate statistical difference with a paired t-test.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Original	0.4949	0.5096	0.3040	0.3878
	Reproduced	0.4764†	0.5039	0.2951	0.3879
	Difference	**-0.0185**	-0.0057	-0.0089	+0.0001
BM25F	Original	0.5232	0.5407	0.3275	0.4188
	Reproduced	0.5200	0.5523	0.3233	0.4272
	Difference	-0.0032	**+0.0116**	-0.0042	+0.0084
FSDM	Original	0.5567	0.5594	0.3610	0.4362
	Reproduced	0.4717‡	0.4445‡	0.2991‡	0.3368‡
	Difference	**-0.0850**	**-0.1149**	**-0.0619**	**-0.0994**
LMD	Original	0.5210	0.5390	0.3340	0.4184
	Reproduced	0.4896‡	0.5294	0.3083‡	0.4059†
	Difference	**-0.0314**	-0.0096	**-0.0257**	**-0.0125**

Table 7. Results of the re-ranking strategy on the *Restricted Collection*. † ($p \leq 0.05$) indicates statistical difference with a paired t-test compared with Tables 2, 4 and 6.

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	Metadata	0.5464†	0.5204†	0.3481†	0.4122†
	Content	0.1418	0.1325	0.0826	0.0898
	Full	0.5575†	0.5416†	0.3541†	0.4280†
BM25F	Metadata	0.5775†	0.5539†	0.3665†	0.4399†
	Content	0.1712	0.1643	0.1052	0.1145
	Full	0.6133†	0.6020†	0.3897†	0.4788†
LMD	Metadata	0.5051†	0.4851†	0.3172†	0.3768†
	Content	0.1660†	0.1598†	0.1019	0.1110
	Full	0.5495†	0.5445	0.3428†	0.4194

ducibility of the experimental results presented in the reference paper. Through a reproducibility-oriented methodology, we explored the critical role of Data Search in enabling efficient retrieval. Improving data access and discovery enhances data portals and digital libraries. It benefits scholarly communication, for instance, facilitating integration between datasets and publication in scholarly graphs as, for instance, OpenAIRE [14] and its curated releases [4].

Concerning the quality of the collection, some concerns arise from using non-permanent URLs to release the content of each dataset. This approach suffers from unstable URLs, resulting in 17.6% of the datasets not being available, primarily due to broken links, just one year after the collection's publication. Datasets without any indexable content have a sizable impact on retrieval performance. Thus, we also compared our reproduced systems with the original in

[14] https://www.openaire.eu/

Table 8. Results of using only 20 nodes as content information compared with the Content Configuration without boosting on the *Restricted Collection*. ‡ indicates statistical difference with a paired t-test ($p \leq 0.01$).

		NDCG@5	NDCG@10	MAP@5	MAP@10
TF-IDF	All nodes	0.1519	0.1652	0.0864	0.1041
	Top 20	0.0822‡	0.0941‡	0.0502‡	0.0603‡
	Difference	-0.0697	-0.0711	-0.0362	-0.0438
BM25F	All nodes	0.0926	0.1132	0.0540	0.0689
	Top 20	0.1134	0.1213	0.0716	0.0810
	Difference	0.0208	0.0081	0.0176	0.0121
LMD	All nodes	0.1448	0.1616	0.0819	0.1000
	Top 20	0.0960‡	0.1095‡	0.0623	0.0727‡
	Difference	-0.0488	-0.0521	-0.0196	-0.0273

a restricted environment, considering only the datasets for which we could parse the content. We conclude that to enhance data discovery, test collections must be released with accessible content, perhaps using permanent links, i.e., Digital Object Identifiers (DOIs).

Regarding reproducibility, thanks to the ACORDAR 2.0 repository, the experimental results were successfully reproduced for BM25F, TF-IDF, and LMD. Indeed, when we did not consider the empty datasets, our runs presented no statistical difference from the originals for most measures. On the other hand, we failed to reproduce the results for FSDM although we used the implementation by the ACORDAR authors. The reference paper and the published GitHub repository provide little details about their experimental setting, especially concerning the indexing phase and FSDM.

The focal point of the ACORDAR test collection is that authors showed that including the content of each dataset provides more effective retrieval systems. To further check on this statement, we investigated the impact of data on the experimental results. We showed that indexing the content only marginally improved performance compared with indexing only metadata. We saw that a re-ranking strategy improves retrieval for the Metadata and the Full Configurations but is ineffective for the Content Configuration. Using only the top 20 nodes with the highest betweenness centrality as content information, we achieved similar results in the Full Configuration, showing that metadata is still the primary means for retrieving datasets. Thus, novel dataset collections and techniques leveraging the datasets' graph structure are needed to understand the content's further impact on the dataset retrieval task.

Bibliography

[1] Benjelloun, O., Chen, S., Noy, N.F.: Google Dataset Search by the Numbers. In: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12507, pp. 667–682. Springer (2020), https://doi.org/10.1007/978-3-030-62466-8_41

[2] Brickley, D., Burgess, M., Noy, N.F.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: WWW 2019: Proceedings of The World Wide Web Conference, San Francisco, CA, USA, May 13-17, 2019. pp. 1365–1375. ACM (2019), https://doi.org/10.1145/3308558.3313685

[3] Castelo, S., Rampin, R., Santos, A.S.R., Bessa, A., Chirigati, F., Freire, J.: Auctus: A dataset search engine for data discovery and augmentation. Proc. VLDB Endow. **14**(12), 2791–2794 (2021), https://doi.org/10.14778/3476311.3476346

[4] Irrera, O., Mannocci, A., Manghi, P., Silvello, G.: A novel curated scholarly graph connecting textual and data publications. ACM J. Data Inf. Qual. **15**(3), 26:1–26:24 (2023). https://doi.org/10.1145/3597310, https://doi.org/10.1145/3597310

[5] Kacprzak, E., Koesten, L.M., Ibáñez, L., Simperl, E., Tennison, J.: A query log analysis of dataset search. In: Web Engineering - 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10360, pp. 429–436. Springer (2017), https://doi.org/10.1007/978-3-319-60131-1_29

[6] Kato, M.P., Ohshima, H., Liu, Y., Chen, H.O.: A test collection for ad-hoc dataset retrieval. In: SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2450–2456. ACM (2021), https://doi.org/10.1145/3404835.3463261

[7] 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: CHI 2017: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017. pp. 1277–1289. ACM (2017), https://doi.org/10.1145/3025453.3025838

[8] Kunze, S.R., Auer, S.: Dataset retrieval. In: 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013. pp. 1–8. IEEE Computer Society (2013), https://doi.org/10.1109/ICSC.2013.12

[9] 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: SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Infor-

mation Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 2981–2991. ACM (2022), https://doi.org/10.1145/3477495.3531729

[10] Megler, V.M., Maier, D.: Are data sets like documents?: Evaluating similarity-based ranked search over scientific data. IEEE Trans. Knowl. Data Eng. **27**(1), 32–45 (2015), https://doi.org/10.1109/TKDE.2014.2320737

[11] Wang, X., Lin, T., Luo, W., Cheng, G., Qu, Y.: CKGSE: A prototype search engine for chinese knowledge graphs. Data Intell. **4**(1), 41–65 (2022), https://doi.org/10.1162/dint_a_00118

[12] Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. SIGIR Forum **51**(2), 268–276 (2017), https://doi.org/10.1145/3130348.3130377

[13] Zhiltsov, N., Kotov, A., Nikolaev, F.: Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In: SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015. pp. 253–262. ACM (2015), https://doi.org/10.1145/2766462.2767756