ACORDAR: A Test Collection for Ad Hoc Content-Based (RDF) Dataset Retrieval

Tengteng Lin
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
tengtenglin@smail.nju.edu.cn

Qiaosheng Chen
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
qschen@smail.nju.edu.cn

Gong Cheng
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
gcheng@nju.edu.cn

Ahmet Soylu
OsloMet – Oslo Metropolitan University
Oslo, Norway
Norwegian University of Science and Technology
Gjøvik, Norway
ahmet.soylu@oslomet.no

Basil Ell
Bielefeld University
Bielefeld, Germany
University of Oslo
Oslo, Norway
basile@ifi.uio.no

Ruoqi Zhao
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
191098346@smail.nju.edu.cn

Qing Shi
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
qingshi@smail.nju.edu.cn

Xiaxia Wang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
xxwang@smail.nju.edu.cn

Yu Gu
The Ohio State University
Columbus, Ohio, USA
gu.826@osu.edu

Evgeny Kharlamov
Bosch Center for Artificial Intelligence
Renningen, Germany
University of Oslo
Oslo, Norway
evgeny.kharlamov@de.bosch.com

ABSTRACT

Ad hoc dataset retrieval is a trending topic in IR research. Methods and systems are evolving from metadata-based to content-based ones which exploit the data itself for improving retrieval accuracy but thus far lack a specialized test collection. In this paper, we build and release the first test collection for ad hoc content-based dataset retrieval, where content-oriented dataset queries and content-based relevance judgments are annotated by human experts who are assisted with a dashboard designed specifically for comprehensively and conveniently browsing both the metadata and data of a dataset. We conduct extensive experiments on the test collection to analyze its difficulty and provide insights into the underlying task.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval effectiveness**; *Presentation of retrieval results*; *Resource Description Framework (RDF).*

KEYWORDS

ad hoc dataset retrieval, dataset search, test collection, RDF, dataset browsing

ACM Reference Format:
Tengteng Lin, Qiaosheng Chen, Gong Cheng, Ahmet Soylu, Basil Ell, Ruoqi Zhao, Qing Shi, Xiaxia Wang, Yu Gu, and Evgeny Kharlamov. 2022. ACORDAR: A Test Collection for Ad Hoc Content-Based (RDF) Dataset Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531729

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531729

1 INTRODUCTION

With increasingly many datasets registered on open data portals and public on the Web (e.g., [37, 38, 43]), a lot of research attention [4] has been given to the task of *ad hoc dataset retrieval*: answering a keyword query with a ranked list of datasets. Pioneers along this direction include the Google Dataset Search system [1, 17] and the NTCIR-15 test collection for ad hoc dataset retrieval [21].

Motivation. As the de facto paradigm, the above-mentioned data portals, dataset search engines, and evaluation efforts focus on *metadata*. As illustrated in Figure 1, the metadata of a dataset includes its title, description, author, etc. Despite its usefulness and ease of use for ad hoc dataset retrieval, its limitations are explicit. Systems only using metadata cannot support queries pointed to the *content* of a dataset [26], i.e., the data itself, while such queries are

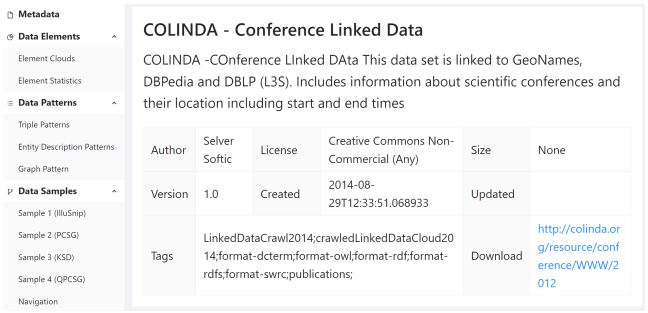


Figure 1: Right: an example of metadata. Left: tab groups provided by our dashboard for browsing a dataset.

frequent [5]. For example, although the COLINDA dataset[1] is a good answer to the query "conferences in France", its metadata in Figure 1 could not match "France" and hence it may not be retrieved by metadata-based techniques, while this keyword appears frequently in its data. Metadata is also weak in providing signals for dataset ranking [1] and in satisfying post-retrieval user needs such as data exploration and analysis [14]. To address these limitations, content-based dataset retrieval systems have recently emerged [2, 51, 52], exploiting both the metadata and data of a dataset for retrieval. However, their effectiveness has not been thoroughly evaluated due to the lack of content-based test collections.

Our Work. We fill the gap by building ACORDAR, a test collection for ad hoc content-based dataset retrieval. To this end, observe that creating content-oriented dataset queries and making content-based relevance judgments are both non-trivial because, compared with small-sized metadata, the data may be too large for a human annotator to browse without a tool support. Therefore, as a prerequisite for building the test collection, we develop a dashboard integrating a variety of summarization and visualization methods for browsing a dataset, its data in particular. With its help, we build and publish a test collection[2] containing 10,671 relevance judgments involving 493 queries over 31,589 datasets collected from 543 data portals. Our contributions in this paper include:

- to the best of our knowledge, the first public test collection for ad hoc content-based dataset retrieval,
- evaluation results of four standard retrieval models in three configurations on the test collection,
- empirical insights into the difficulty of the test collection and the usefulness of metadata and data for retrieval, and
- the design of a comprehensive dashboard for conveniently browsing datasets and a user study of its effectiveness.

The current implementation of the dashboard supports datasets in RDF format[3] and is extensible. We prioritize RDF because it is a standard model for data interchange on the Web and has been widely used for representing knowledge graphs [18] such as DBpedia and Wikidata, on which a wide range of IR tasks have been defined and extensively studied, e.g., keyword-based search [10, 39] and exploration [40, 41], question answering [47], and document enrichment [28]. Accordingly, the current version of the test collection is restricted to datasets having an RDF version.

[1]http://www.colinda.org/
[2]https://github.com/nju-websoft/ACORDAR
[3]https://www.w3.org/RDF/

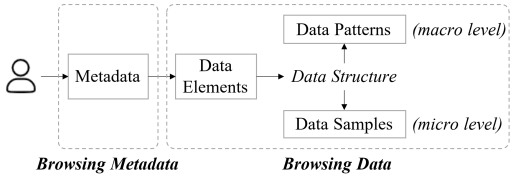


Figure 2: A typical process of browsing a dataset.

Outline. We introduce the design of the dashboard in Section 2, describe the construction of the test collection in Section 3, present experiments in Section 4, compare with related work in Section 5, and discuss limitations and future work in Section 6.

2 A DASHBOARD FOR BROWSING DATASETS

To build a test collection for ad hoc content-based dataset retrieval, we need a tool to help human annotators create content-oriented dataset queries and make content-based judgments about the relevance of a dataset to a query. These annotation activities rely on convenient access to both the metadata and data of a dataset, which is crucial to the efficiency of the annotation process and to the quality of the annotations because a dataset, unlike a document which can be easily read, is often both too complex (e.g., having a graph structure) and too large (e.g., containing millions of edges) for an annotator to browse plainly. To support these activities, we design and develop a comprehensive dashboard for browsing a dataset from multiple views to satisfy a variety of possible needs that an annotator may have in the annotation process.

Figure 2 presents a typical process of browsing a dataset from multiple views that are supported by our dashboard. Given a dataset, a user can start by browsing its metadata which is created by the dataset publisher and is supposed to offer a good starting point for beginners. Then the user moves to data browsing. Since data may have a complex structure, the user can browse data progressively [45]—scanning its elements before investigating its structure. To explore the data structure which may be huge, the user can browse its summarized macro-level patterns, micro-level samples, or both. All these browsing activities are supported by different *tabs* in the dashboard, thus allowing flexible mixtures of activities beyond the anticipated process presented in Figure 2.

Specifically, the dashboard provides four *tab groups* corresponding to the above-mentioned browsing targets: Metadata, Data Elements, Data Patterns, and Data Samples, as illustrated in Figure 1. Below we describe the design of each tab group.

2.1 Metadata

Datasets are published with metadata which is easy to be understood. Some common fields in the metadata of a dataset include its title, description, author, license, last updated date, etc.

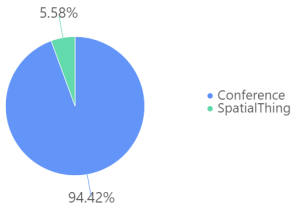
Metadata is the first tab group in the dashboard. It comprises only a single tab where a table trivially listing all the fields in the metadata is presented. Observe that different dataset publishers may

(a) Class cloud. (b) Property cloud. (c) Entity cloud. (d) Literal cloud.

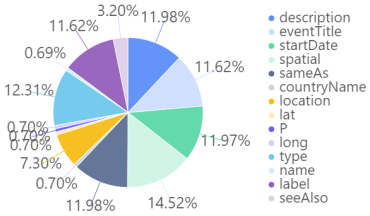
Figure 3: An example of element clouds.

Number of triples	100396
Number of classes	2
Number of properties	14
Number of entities	28368
Number of literals	40666

(a) Element counts.



(b) Distribution of classes in entities.



(c) Distribution of properties in RDF triples.

Figure 4: An example of element statistics.

adopt different metadata formats. To present them in a uniform way, we manually unify the fields in known popular metadata formats. Figure 1 illustrates this presentation for the COLINDA dataset.

2.2 Data Elements

Structured data may not be easy to be understood. To help annotators understand data better, we organize data presentations in a meaningful order that shows gradually more structures, and gradually more complex ones. As a starting point, only data elements are extracted and summarized but their structures are hidden to avoid overwhelming annotators at an early stage.

Data Elements is the second tab group in the dashboard. Since for a large dataset it would be difficult, if not impossible, to show all the elements of the data, in this tab group we provide two tabs offering complementary overviews of data elements: Element Clouds and Element Statistics. The former presents significant concrete elements of the data, and the latter gives their statistics.

Element Clouds. This tab extracts significant elements from the data and visualizes them as a tag cloud, which has been popularly used to visualize a distribution by significance. For an RDF dataset, the elements of the data are resources which, according to RDF Schema, fall into four categories: literals, classes, properties, and other resources which are commonly called entities. Significant elements in different categories are visualized in different tag clouds where tag size represents significance. Significance is measured by frequency, i.e., the number of RDF triples where a class, property, or entity is included, or the number of literals where a word appears. Figure 3 illustrates these clouds for the COLINDA dataset.

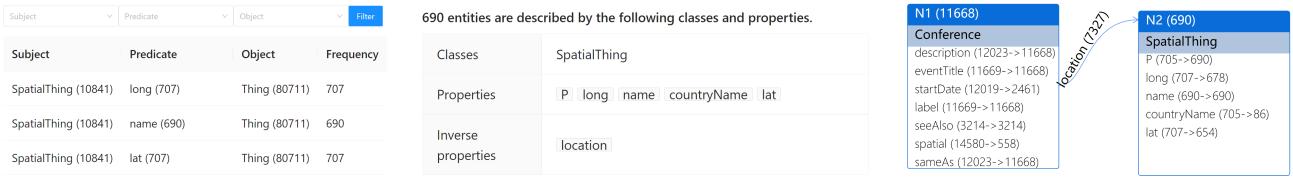
Element Statistics. This tab presents statistics about data elements. For an RDF dataset, a table listing the number of RDF triples in the data and the number of elements in each category is shown. The distribution of classes in entities and the distribution of properties in RDF triples are visualized as pie charts. Figure 4 illustrates these visualizations for the COLINDA dataset.

2.3 Data Patterns

Data structure is about how data elements are organized. For a large dataset it would be unrealistic to show all the structures, which is also unnecessary as structures often repeat themselves. Such patterns are mined from the data as an outline of the structures.

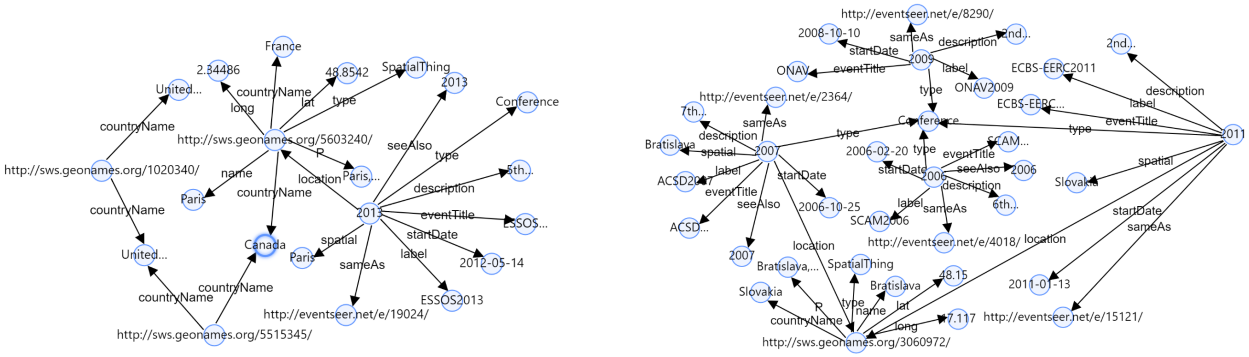
Data Patterns is the third tab group in the dashboard. Patterns can be defined at different granularities of data [3, 9, 11, 54]. To help annotators learn data better, for an RDF dataset we provide three tabs offering data patterns at increasing granularities: Triple Patterns, Entity Description Patterns, and Graph Pattern. Basically, a triple pattern represents an aggregate of similar RDF triples, an entity description pattern aggregates triple patterns about the same entity, and a graph pattern aggregates entity description patterns and their relations.

Triple Patterns. This tab implements ABSTAT [46] to mine triple-level patterns from the data. For an RDF triple $\langle s, p, o \rangle$, its pattern is a triple $\langle x, p, y \rangle$ where x and y are minimal classes of s and o, respectively; minimization is based on the class hierarchy in the ontology. We reproduce the user interface in [36] to show a table listing triple patterns in descending order of frequency, i.e., the number of RDF triples conforming to a pattern. In each triple pattern, after each class and property the number of its instances is shown in parentheses. To explore a large number of triple patterns, annotators can filter them by putting constraints on each position of a triple pattern using the corresponding text box which features autocomplete. Figure 5(a) illustrates this user interface for the COLINDA dataset.



(a) Triple patterns. (b) Entity description pattern. (c) Graph pattern.

Figure 5: An example of data patterns.



(a) Sampled by IlluSnip. (b) Sampled by PCSG.

Figure 6: An example of data samples.

Entity Description Patterns. This tab follows [49, 50] to mine entity-level patterns from the data. For an entity, its entity description pattern (EDP) consists of three sets: a set of all its classes, a set of all its properties, and a set of all its inverse properties which have it as a value. EDPs are listed in descending order of frequency, i.e., the number of entities conforming to an EDP. For each EDP its composition and frequency are presented. Figure 5(b) illustrates the presentation of an EDP in the COLINDA dataset.

Graph Pattern. This tab implements RDFQuotient [15] to mine graph-level patterns from the data. Based on the concept of quotient graph, entities are grouped according to their classes, properties, and relations (i.e., entity-value properties) to others. The result is a graph where nodes represent entity groups and edges represent relations between entity groups. We reproduce the visualization in [15] which draws this graph in the style of ER diagram. Within each node the number of entities in the group, their classes, and their properties are listed from top to bottom. After each property the number of its instances and the number of its distinct values appear in parentheses connected by an arrow. After each relation the number of its instances is shown in parentheses. Figure 5(c) illustrates this visualization for the COLINDA dataset.

2.4 Data Samples

In parallel to data patterns, representative samples of the data are extracted as a preview of the data structures in a possibly huge dataset [44]. Despite incompleteness, a data sample gives annotators the opportunity of seeing a snippet of the actual data.

Data Samples is the fourth tab group in the dashboard. There are methods for sampling different subsets of data having different characteristics [12, 16, 27, 30, 50]. To give annotators more flexibility, for an RDF dataset we provide four tabs offering data samples extracted by different state-of-the-art methods: Sample 1 (IlluSnip), Sample 2 (PCSG), Sample 3 (KSD), and Sample 4 (QPCSG). While the first and second samples are generated for general purposes, the third and fourth samples are dynamically generated to be biased towards a given keyword query. The latter is crucial to one kind of annotation activity the dashboard aims at supporting, i.e., annotators make content-based judgments about the relevance of a dataset to a query. Query-biased samples can help annotators easily locate some relevant parts of the data—if any. Besides these automatically computed samples, we also provide a Navigation tab where annotators can be guided to freely navigate to any portion of the data.

Sample 1 (IlluSnip). This tab implements IlluSnip [8] to extract, from the data, a representative subset of RDF triples about a set of interconnected entities which collectively have the most frequent classes, properties, and central positions (in terms of PageRank) in the RDF graph. For efficiency reasons we implement an approximate version of IlluSnip [29]. The extracted RDF triples are visualized as a node-link diagram where each entity node can be clicked to be expanded to explore its neighbors in the original RDF graph. Figure 6(a) illustrates this diagram for the COLINDA dataset.

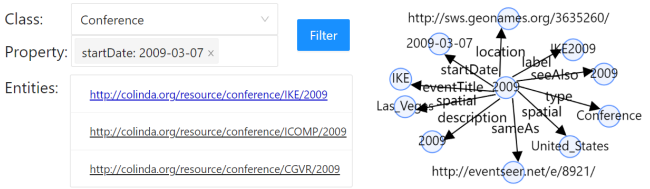


Figure 7: An example of data navigation.

Sample 2 (PCSG). This tab implements PCSG [49] to extract, from the data, a representative subset of RDF triples about sets of interconnected entities which collectively have the most frequent EDPs. The extracted RDF triples are visualized and interactable in the same way as in the first tab. Figure 6(b) illustrates this visualization for the COLINDA dataset.

Sample 3 (KSD). This tab implements KSD [48] which is an extension of IlluSnip where the extracted RDF triples also match the most keywords in a given query. We omit to illustrate this tab.

Sample 4 (QPCSG). This tab implements QPCSG [49] which is an extension of PCSG where the extracted RDF triples also match the most keywords in a given query. We omit to illustrate this tab.

Navigation. This tab implements a standard faceted browsing interface, allowing annotators to filter the entities in the data by putting constraints on their classes and properties using drop-down menus. Each entity can be clicked to be visualized in a node-link diagram and expanded to explore its neighbors in the original RDF graph. Figure 7 illustrates this interface for the COLINDA dataset.

3 TEST COLLECTION

To build a natural and representative test collection for ad hoc dataset retrieval, we collected real datasets from popular open data portals. Each dataset was processed to be accessible via the dashboard described in Section 2, and then human annotators were invited to create content-oriented dataset queries and make content-based judgments about the relevance of a dataset to a query with the help of the dashboard. The test collection is referred to as ACORDAR, short for **A**d hoc **CO**ntent-based **R**DF **DA**taset **R**etrieval. It is available under the Apache License. Relevance judgments are stored in TREC's qrels format[4] for ease of use.

3.1 Datasets

In CKAN,[5] DataPortals.org,[6] Open Data Portal Watch,[7] and Socrata,[8] we identified 1,131 unique data portals where 540 (48%) were accessible at the time of experimentation. These data portals collectively indexed 111,017 RDF datasets. For 31,589 RDF datasets (28%), their dump files were successfully downloaded and parsed by Apache Jena,[9] and our test collection is based on these datasets; failures were mainly due to broken links. Table 1 presents the distribution of data portals where these datasets were collected. Most of them are open government data portals.

[4]https://trec.nist.gov/data/qrels_eng/
[5]https://ckan.org/
[6]http://dataportals.org/
[7]https://data.wu.ac.at/portalwatch/
[8]https://dev.socrata.com/
[9]https://jena.apache.org/

Table 1: Source Distribution of Datasets

Data Portal	#Datasets	%
data.gov	8,700	27.54%
dati.gov.it	2,993	9.47%
data.cityofnewyork.us	1,172	3.71%
performance.smcgov.org	997	3.16%
data.wa.gov	942	2.98%
internal.open.piercecountywa.gov	821	2.60%
data.medicaid.gov	819	2.59%
opendata.utah.gov	774	2.45%
data.oregon.gov	608	1.92%
datahub.smcgov.org	576	1.82%
Others	13,187	41.75%
Total	31,589	100.00%

Table 2: Size Distribution of Datasets

	Min	Max	Mean	Median	Total
#Triples	3	62.8 M	9.9 K	2.0 K	312.2 M
#Classes	0	153	0.9	1	29.3 K
#Properties	1	668	18.5	11	584.4 K

A dataset might be associated with multiple dump files. For each dataset we merged all its dump files and removed redundant RDF triples. After deduplication the total number of RDF triples in the test collection was reduced from 317.8 M to 312.2 M. Table 2 presents the size distribution of the datasets in the test collection, which differ greatly in both data size (i.e., number of triples) and schema size (i.e., numbers of classes and properties). These datasets contain a median of 2.0 K RDF triples, being too large to be browsed directly, which justifies the necessity of providing tools like the dashboard described in Section 2. The total numbers of unique classes and properties that occur in these datasets are 29.3 K and 584.4 K, respectively, indicating a diverse and representative test collection.

3.2 Queries

Despite the availability of dataset queries published by previous studies [5, 19, 21], we found that those queries were mainly derived from posts published on online forums looking for datasets. While representing real information needs, they are considered "hard queries" since they were posted on online forums exactly because their posters could not find relevant datasets. This was confirmed by our preliminary experiment with those queries on the datasets in our test collection. To avoid building a trivial test collection containing few query-dataset pairs judged to be relevant, we need to solicit queries in other ways.

The queries in the test collection were solicited in two ways. First, we invited human annotators to browse a dataset using the dashboard described in Section 2 and then create a synthetic query to which the dataset could be judged relevant, thus ensuring at least a minimum number of relevant query-dataset pairs. Second, we

Table 3: Examples of Content Summaries and Queries

	Content Summary	Query
1	A dataset to record Maryland's land protection policies including their goals and definitions. For example, one policy's goal is nitrogen reduction to upgrade 18 wastewater treatment plants to ENR standards.	nitrogen reduction plan in Maryland
2	A dataset recording some programs to educate children of different ages. The target population include children, young adults, late adolescence and so on. Some programs are to create values, and some other programs are for their well-being, second interest, etc. The dataset records program descriptions and their ratings of how effective for the target people.	education programs and target population
3	A dataset recording the salary of senior civil servants, including their salary and profession.	salary of senior civil servants
4	This dataset describes chinook status in Washington state, as well as salmon populations.	chinook stock
5	This dataset mainly describes information about introduction narrative for career paths. It describes all kinds of career paths, about what kinds of tasks one should complete, how much one can earn by doing this job and so on.	career paths, introduction of each job, salary

observed that many queries in TREC's Ad hoc Test Collections[10] are about general topics and have great potential for finding relevant datasets. So those queries were reused for our test collection. Below we describe the solicitation process in detail.

Synthetic Queries. We recruited 9 college students having a background in RDF as annotators to create content-oriented dataset queries. Each annotator was assigned at most 30 random datasets. The assigned datasets were controlled to contain at least a minimum amount of data being readable by a large population, i.e., more than 2% of the data elements were checked[11] to have an English label. Given a dataset the annotator was asked to create a keyword query such that the dataset could be judged relevant to it. To encourage content-oriented (as opposed to metadata-oriented) dataset queries, the annotator was instructed to take two steps to create a query. In the first step, the annotator employed the dashboard to comprehensively browse the dataset and write a summary of the data describing its main content in 10-500 English words. The annotator was thus obliged to look into the data rather than superficially scanning the metadata for creating a query. In the second step, the annotator extracted a few keywords from the summary to form a query that was neither overly specific—with the intent of finding this particular dataset, nor overly general—potentially matching a broad range of datasets. We received 251 queries from all the annotators. We checked each query and asked its annotator to revise it if the query expressed a very vague intent. Table 3 exemplifies some data summaries and queries.

[10]https://trec.nist.gov/data/test_coll.html
[11]https://pypi.org/project/pyenchant/

Table 4: Length Distribution of Queries (#Keywords)

	Min	Max	Mean	Median
Synthetic Queries	1	14	4.1	4
TREC Queries	1	20	3.8	3
All Queries	1	20	3.9	3

TREC Queries. We imported all the 450 ad hoc topics used in the English Test Collections of TREC 1–8,[12] i.e., those numbered 1–450. For each topic its title field was extracted as a query.

Table 4 presents the length distribution of the queries in the test collection. Synthetic queries containing an average of 4.1 keywords are generally longer than TREC queries containing 3.8 keywords.

3.3 Pooling

Annotating complete relevance judgments for all the 31,589 datasets and 701 queries in the test collection would be infeasible. We followed common practice in IR evaluation to use the pooling method with a number of standard retrieval models to reduce the number of required relevance judgments.

For pooling we indexed both the metadata and data of each dataset in the test collection. Specifically, Apache Lucene[13] was used to construct an inverted index with eight fields representing a dataset. For the metadata, four fields that often contain human-readable information to match keyword queries were indexed: title, description, author, and tags. For the data, all the elements in four categories were indexed: literals, classes, properties, and entities. Based on the index we implemented four standard retrieval models measuring the similarity between a query and an indexed dataset:

- TF-IDF based cosine similarity,
- BM25F,
- Fielded Sequential Dependence Model (FSDM), and
- Language Model using Dirichlet priors for smoothing (LMD).

All these models rely on field weights representing possibly different degrees of importance in scoring. To tune the weight of each field in each model we constructed a pre-validation set consisting of 522 query-dataset pairs where the queries are disjointed with those in the test collection, and each pair was manually annotated using a graded relevance scale of 0–2 with 0 meaning irrelevant, 1 meaning partially relevant, and 2 meaning highly relevant. Based on the pre-validation set, grid search was performed to tune each weight from 0 to 1 in 0.1 increments and an optimal setting in terms of NDCG@10 was adopted for each model.

For each query in the test collection, we employed each retrieval model to fetch 10 top-ranked datasets and then took the union of the retrieved datasets over all the models. For 13 TREC queries no dataset was retrieved, since their keywords could not be matched by any dataset in the test collection. The remaining 688 queries collectively produced 15,038 query-dataset pairs.

3.4 Relevance Judgments

We, again, invited the 9 human annotators who participated in creating synthetic queries in Section 3.2 to make content-based

[12]https://trec.nist.gov/data/topics_eng/index.html
[13]https://lucene.apache.org/

Table 5: Relevance Distribution of Query-Dataset Pairs

	#Queries	#Q.-D. Pairs	Rel. Scale	#Q.-D. Pairs (%)
Synthetic Queries	241	5,303	0: none	3,140 (59.2%)
			1: partial	1,233 (23.3%)
			2: high	930 (17.5%)
TREC Queries	252	5,368	0: none	3,802 (70.8%)
			1: partial	1,129 (21.0%)
			2: high	437 (08.1%)
All Queries	493	10,671	0: none	6,942 (65.1%)
			1: partial	2,362 (22.1%)
			2: high	1,367 (12.8%)

judgments about the relevance of a dataset to a query. Given a query-dataset pair the annotator was asked to judge relevance after comprehensively browsing both the metadata and data of the dataset by exploiting the dashboard described in Section 2. Relevance was given using a graded scale of 0–2 with

- 0 representing irrelevant,
- 1 representing partially relevant, and
- 2 representing highly relevant.

Resembling the definition used by TREC, we suggested the following definition of relevance to be used by all the annotators.

> If you were writing a report or developing an application on the subject of the topic and would use the information contained in the dataset in the report or application, then the dataset is relevant.

To ensure annotations of high quality, each query-dataset pair was assigned to two independent annotators for relevance judgments. If their annotations were identical, this consensus annotation would be taken. Otherwise, such a query-dataset pair would be assigned to a third annotator for calculating a major vote. If the three annotations were different from each other, a partial relevance (i.e., 1) representing their average would be taken. Overall, the inter-annotator agreement measured by Krippendorff's α is 0.59, indicating a fairly acceptable level of reliability.

For 10 synthetic queries and 185 TREC queries, all their retrieved datasets were judged irrelevant and hence they were removed from the test collection. Table 5 presents the relevance distribution of the remaining 10,671 query-dataset pairs involving 493 queries. Synthetic queries are generally associated with more partial or highly relevant datasets (40.8%) than TREC queries (29.2%).

3.5 Training, Validation, and Test Sets

To allow future evaluation results on the test collection to be comparable with each other, we specified and released a split of the test collection into training, validation, and test sets. Specifically, the query-dataset pairs in the test collection were partitioned by query into 5 approximately equally sized subsets $P_0, \ldots, P_4$ to support 5-fold cross-validation. The partition was randomized but stratified with respect to the proportions of synthetic and TREC queries in each subset. For $0 \leq i \leq 4$, the i-th fold should use $P_i, P_{(i+1)\%5}, P_{(i+2)\%5}$ as the training set, $P_{(i+3)\%5}$ as the validation set, and $P_{(i+4)\%5}$ as the test set. Evaluation results should then be aggregated from the test sets in all the 5 folds.

4 EXPERIMENTS

We conducted two experiments. The first experiment provided insights into our test collection. The second experiment analyzed the usefulness of our dashboard in building the test collection.

4.1 Dataset Retrieval

We evaluated the effectiveness of a number of standard retrieval models on the test collection to analyze its difficulty and provide insights into the ad hoc content-based dataset retrieval task.

4.1.1 Experimental Setting. The four retrieval models used for pooling in Section 3.3 were evaluated on the test collection: TF-IDF, BM25F, FSDM, and LMD. To analyze the usefulness of metadata and data for ad hoc dataset retrieval, each model has three configurations using different sets of fields to represent a dataset:

- by default, using all the eight metadata and data fields,
- `[m]`: using only the four metadata fields, and
- `[d]`: using only the four data fields.

We followed Section 3.5 to perform 5-fold cross-validation. Since all the above retrieval models are unsupervised, in each fold the training set was ignored and the validation set was used to tune the field weights in each model. The tuning process resembled the one described in Section 3.3 using grid search and NDCG@10.

The evaluation metrics used on the test sets are NDCG@5, NDCG@10, MAP@5, and MAP@10. When calculating MAP, graded relevance in human annotations was converted to binary relevance by treating both highly and partially relevant as relevant.

4.1.2 Experimental Results. Table 6 presents the evaluation results of each retrieval model in each configuration averaged over synthetic queries, over TREC queries, and over all the queries.

Comparison between Retrieval Models. In the default configuration, FSDM consistently outperforms the other retrieval models on both synthetic and TREC queries in terms of all the four evaluation metrics. BM25F is better than LMD on synthetic queries while LMD slightly surpasses BM25F on TREC queries. They both noticeably exceed TF-IDF. In the `[m]` configuration using only metadata fields, BM25F outperforms the other retrieval models.

Difficulty of Queries. Table 7 aggregates the evaluation results of all the retrieval models in the default configuration. The mean values over all the queries are moderate, indicating that the test collection is neither trivial nor too difficult for current techniques, and there is much room for novel models for ad hoc content-based dataset retrieval. According to the values, TREC queries generally seem more difficult than synthetic queries. It is consistent with the relevance distribution in human annotations described in Section 3.4. This is no surprise as TREC queries were not originally designed for the dataset retrieval task, while synthetic queries were created for specific datasets. Figure 8 presents the NDCG@10 distribution of all the retrieval models in the default configuration on each query. The mean values are almost evenly distributed over the queries, indicating that the test collection contains queries at all levels of difficulty. Table 8 exemplifies the five easiest and five hardest synthetic queries. Figure 9 aggregates the results by query length. There is no explicit correlation between NDCG@10 and

Table 6: Mean Evaluation Results of Each Retrieval Model in Each Configuration

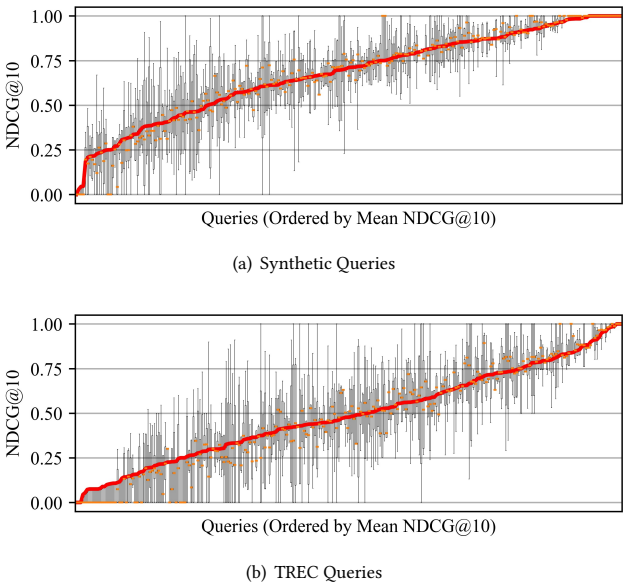
	NDCG@5	NDCG@10	MAP@5	MAP@10
Synthetic Queries				
TF-IDF	0.6158	0.6293	0.3409	0.4560
TF-IDF [m]	0.5603	0.5766	0.3081	0.4161
TF-IDF [d]	0.2367	0.2376	0.1241	0.1455
BM25F	0.6611	0.6868	0.3780	0.5103
BM25F [m]	0.6171	0.6150	0.3481	0.4494
BM25F [d]	0.2768	0.2720	0.1729	0.1889
FSDM	**0.7348**	**0.7193**	**0.4430**	**0.5434**
FSDM [m]	0.6117	0.6015	0.3530	0.4325
FSDM [d]	0.3104	0.3131	0.1801	0.2109
LMD	0.6437	0.6654	0.3764	0.4927
LMD [m]	0.5108	0.5207	0.2967	0.3775
LMD [d]	0.3004	0.3037	0.1775	0.2031
TREC Queries				
TF-IDF	0.4066	0.4649	0.2358	0.3417
TF-IDF [m]	0.3923	0.4306	0.2290	0.3230
TF-IDF [d]	0.1473	0.1568	0.0766	0.0955
BM25F	0.4513	0.4932	0.2642	0.3645
BM25F [m]	0.3969	0.4390	0.2264	0.3209
BM25F [d]	0.1584	0.1696	0.1058	0.1226
FSDM	**0.4579**	**0.5156**	**0.2791**	**0.3806**
FSDM [m]	0.3644	0.3947	0.2044	0.2742
FSDM [d]	0.1918	0.2105	0.1169	0.1422
LMD	0.4537	0.4992	0.2789	0.3748
LMD [m]	0.3651	0.3967	0.2138	0.2896
LMD [d]	0.1819	0.2030	0.1071	0.1328
All Queries				
TF-IDF	0.5088	0.5452	0.2871	0.3976
TF-IDF [m]	0.4743	0.5019	0.2676	0.3685
TF-IDF [d]	0.1910	0.1963	0.0998	0.1199
BM25F	0.5538	0.5877	0.3198	0.4358
BM25F [m]	0.5045	0.5250	0.2859	0.3838
BM25F [d]	0.2163	0.2196	0.1385	0.1550
FSDM	**0.5932**	**0.6151**	**0.3592**	**0.4602**
FSDM [m]	0.4853	0.4958	0.2770	0.3516
FSDM [d]	0.2497	0.2606	0.1478	0.1758
LMD	0.5465	0.5805	0.3266	0.4324
LMD [m]	0.4363	0.4573	0.2543	0.3325
LMD [d]	0.2398	0.2523	0.1415	0.1672

Table 7: Mean Evaluation Results of All the Retrieval Models in the Default Configuration

	NDCG@5	NDCG@10	MAP@5	MAP@10
Synthetic Queries	0.6638	0.6752	0.3846	0.5006
TREC Queries	0.4424	0.4932	0.2645	0.3654
All Queries	0.5506	0.5821	0.3232	0.4315

query length, indicating that query length might not be a decisive factor in the difficulty of this task.

Usefulness of Metadata and Data. Table 9 aggregates the evaluation results of all the retrieval models by configuration over all



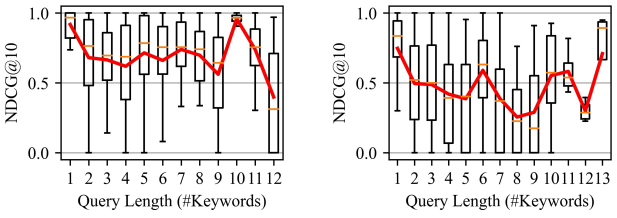
(a) Synthetic Queries

(b) TREC Queries

Figure 8: Distribution of (box-plot) and mean (curve) NDCG@10 of all the retrieval models in the default configuration on each query.

Table 8: Easiest and Hardest Synthetic Queries

	Easiest Queries	Hardest Queries
1	IEEE conferences	immunocompetent people
2	Finnish municipalities	Austin's park development comments
3	Civil Penalties of Columbia Riverkeeper	education expenditure
4	Southampton airport	senior manager of data and assessment
5	comments for clean fuels program	Public Act



(a) Synthetic Queries

(b) TREC Queries

Figure 9: Distribution of (box-plot) and mean (curve) NDCG@10 of all the retrieval models in the default configuration over all the queries of each length.

Table 9: Mean Evaluation Results of All the Retrieval Models in Each Configuration over All the Queries

	NDCG@5	NDCG@10	MAP@5	MAP@10
Default	0.5506	0.5821	0.3232	0.4315
[m]	0.4751	0.4950	0.2712	0.3591
[d]	0.2242	0.2322	0.1319	0.1545

the queries. The default configuration outperforms the other configurations, indicating that metadata and data are both useful for ad hoc content-based dataset retrieval. For example, NDCG@10 drops considerably by 0.0871 after excluding data fields, showing the usefulness of content in ad hoc dataset retrieval.

4.2 Dataset Browsing

We also evaluated the effectiveness of each tab in our dashboard in browsing datasets for building the test collection. As a preliminary user study, it would benefit future research in dataset browsing.

4.2.1 Experimental Setting. Each human annotator was invited to complete a post-experiment questionnaire rating and commenting on the usefulness of each tab in browsing datasets for creating queries and for making relevance judgments. Rating was given using a graded scale of 0–2 with

- 0 representing rarely useful for the annotations,
- 1 representing moderately useful for the annotations, and
- 2 representing frequently useful for the annotations.

4.2.2 Experimental Results. Figure 10 presents the rating distribution of each tab. For both query creation and relevance judgments, all the tabs were rated frequently or moderately useful by some annotators. Metadata received the highest ratings, followed by Element Clouds and Sample 1 (IlluSnip). Sample 2 (PCSG) and Navigation are also relatively useful (mean ≥ 1) for query creation.

We identified several representative comments on the less useful tabs. Element Statistics was not considered more useful than Element Clouds for the annotation tasks. For exploring the data structure, the tabs visualizing data patterns were not favored since they are less intuitive than data elements and samples.

4.3 Main Findings and Discussion

Below we summarize and discuss our empirical findings.

Our test collection contains queries at all levels of difficulty and in general the difficulty is moderate and suitable for evaluation. It remains to be seen what, except query length, determines the difficulty of an ad hoc dataset retrieval task.

All the tested models benefit from incorporating data fields into retrieval, supporting recent research on content-based dataset retrieval. The models display varying degrees of competence to exploit metadata and data. BM25F and FSDM achieve the best results on metadata and data fields, respectively, suggesting future research on specialized retrieval models for a hybrid of metadata and data.

Besides metadata, data-oriented tabs in our dashboard including element clouds, data samples, and navigation are helpful for users to comprehend retrieved datasets and judge relevance, while such capabilities are currently under-exploited in deployed systems.

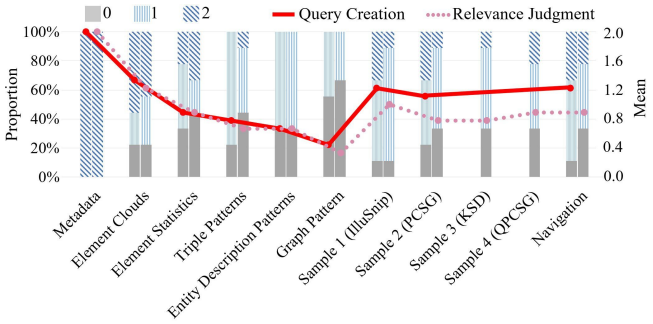


Figure 10: Distribution of (bar) and mean (curve) ratings of each tab over all the annotators.

5 RELATED WORK

5.1 Test Collections for Dataset Retrieval

There have been several test collections for ad hoc dataset retrieval released in the literature. Among others, bioCADDIE-2016 [13] is one for the biomedical domain, containing 15 queries created by instantiating templates resembling search questions collected from potential users of dataset retrieval. Another is BEF-China [31] built in a similar manner for the biodiversity domain. The NTCIR-15 (English) test collection [21] is built on open government datasets and contains 192 free-form queries translated from information needs mined from online forums. Table 10 compares these test collections with the one built in this paper. Our test collection is distinguished by its content-based nature: not only the queries are created in a content-oriented fashion, but also the pooling and relevance judgments of query-dataset pairs are based on both metadata and data, while these activities in building previous test collections mainly rely on metadata. Therefore, our test collection is more suitable for supporting the trending research on content-based dataset retrieval [2, 32, 51, 52].

There are some research directions that may not be referred to as dataset retrieval but are potentially related. Ad hoc table retrieval is one such task. There have been a number of test collections for this task [7, 42, 53] where retrieval often relies on the content of a table. However, it is arguable whether a table can be viewed as a dataset. Different from the datasets considered in this paper and in other test collections for ad hoc dataset retrieval, a table is not always self-contained but usually exists as part of and is contextualized by a webpage. Therefore, such Web tables are rarely registered as datasets on open data portals.

5.2 Dataset Retrieval Systems

Existing data portals, including Google Dataset Search [1], mainly support ad hoc dataset retrieval based on metadata. The research community have recognized the usefulness of data in dataset retrieval and leveraged data to infer missing domains [34] or generate schema labels [6] to enhance retrieval models. There are also efforts to mine inter-dataset relationships from their data for dataset recommendation [33] and combination [32]. As to RDF datasets, there have been several ad hoc content-based retrieval systems. For example, LODAtlas [35] indexes classes and properties in the

Table 10: Comparison between Test Collections for Ad Hoc Dataset Retrieval

	bioCADDIE-2016 [13]	BEF-China [31]	NTCIR-15 (English) [21]	ACORDAR (This Paper)
Domain	biomedical	biodiversity	government	open-domain
#Datasets	795.0 K	372	46.6 K	31.6 K
Data Format	not mentioned	not mentioned	Excel, CSV, PDF, XML, JSON, RDF, text	RDF
#Queries	15	14	192	493
Query Form	template-based question	template-based question	free-form	free-form
Query Creation	imitating user questions	imitating user questions	translated from user needs	extracted from content summaries
#Query-Dataset Pairs	20.2 K	5.2 K	10.5 K	10.7 K
Basis for Pooling	metadata	n/a (pooling all datasets)	metadata + table headers + entities	metadata + data
Basis for Relevance Judgments	metadata + linked articles	expert knowledge	webpage (metadata)	dashboard (metadata + data)

data to match keyword queries, and visualizes graph-level data patterns [15]. CKGSE [51, 52] extracts data snippets [48] to facilitate relevance judgments. Their counterpart for tabular datasets is Auctus [2] which indexes data summaries and presents data samples. However, the effectiveness of these systems has not been systematically evaluated due to the lack of test collections for ad hoc content-based dataset retrieval. Our work fills this gap.

6 LIMITATIONS AND FUTURE WORK

While ACORDAR represents the first evaluation effort to support the trending research on content-based, as opposed to metadata-based, ad hoc dataset retrieval, its current version is restricted to datasets containing RDF data. Beyond this standard and popular model for data exchange on the Web, there are numerous datasets published in other formats such as CSV. Therefore, a natural next step for extending the test collection would be to include other types of data, which in turn requires extending our dashboard for browsing datasets. This is not deemed a research challenge but more of a feasible engineering work thanks to the high availability of methods for summarizing and visualizing data of major types.

When building ACORDAR, the standard yet shallow retrieval models used for pooling may find insufficient datasets and miss those that are only implicitly relevant to a query. Dense models could partially address this issue, but questions would arise as to whether and how they can effectively scale to a huge amount of data in a dataset, establishing an attractive research direction.

We also anticipate reusing some tabs in our dashboard to enhance existing content-based dataset search engines.

Moreover, an important future work for us is to adjust ACORDAR to industrial and in particular to manufacturing settings with production RDF datasets and knowledge graphs since we whiteness an explosion thereof [23, 56], e.g., in Bosch [20, 55], Siemens [24, 25], Equinor [22], and other large production companies.

ACKNOWLEDGMENTS

This work was partially supported by the NSFC (62072224), by the H2020 projects Dome 4.0 (Grant Agreement No. 953163), OntoCommons (Grant Agreement No. 958371), and DataCloud (Grant Agreement No. 101016835), and by the SIRIUS centre: Norwegian Research Council project No 237898. The authors would like to thank all the annotators.

REFERENCES

[1] Dan Brickley, Matthew Burgess, and Natasha F. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *WWW 2019*. 1365–1375. https://doi.org/10.1145/3308558.3313685

[2] Sonia Castelo, Rémi Rampin, Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A Dataset Search Engine for Data Discovery and Augmentation. *VLDB J.* 14, 12 (2021), 2791–2794.

[3] Sejla Cebiric, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2019. Summarizing semantic graphs: a survey. *VLDB J.* 28, 3 (2019), 295–327. https://doi.org/10.1007/s00778-018-0528-3

[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *VLDB J.* 29, 1 (2020), 251–272. https://doi.org/10.1007/s00778-019-00564-x

[5] Jinchi Chen, Xiaxia Wang, Gong Cheng, Evgeny Kharlamov, and Yuzhong Qu. 2019. Towards More Usable Dataset Search: From Query Characterization to Snippet Generation. In *CIKM 2019*. 2445–2448. https://doi.org/10.1145/3357384.3358096

[6] Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D. Davison. 2020. Leveraging Schema Labels to Enhance Dataset Search. In *ECIR 2020, Part I*. 267–280. https://doi.org/10.1007/978-3-030-45439-5_18

[7] Zhiyu Chen, Shuo Zhang, and Brian D. Davison. 2021. WTR: A Test Collection for Web Table Retrieval. In *SIGIR 2021*. 2514–2520. https://doi.org/10.1145/3404835.3463260

[8] Gong Cheng, Cheng Jin, Wentao Ding, Danyun Xu, and Yuzhong Qu. 2017. Generating Illustrative Snippets for Open Data on the Web. In *WSDM 2017*. 151–159. https://doi.org/10.1145/3018661.3018670

[9] Gong Cheng, Cheng Jin, and Yuzhong Qu. 2016. HIEDS: A Generic and Efficient Approach to Hierarchical Dataset Summarization. In *IJCAI 2016*. 3705–3711.

[10] Gong Cheng, Shuxin Li, Ke Zhang, and Chengkai Li. 2020. Generating Compact and Relaxable Answers to Keyword Queries over Knowledge Graphs. In *ISWC 2020, Part I*. 110–127. https://doi.org/10.1007/978-3-030-62419-4_7

[11] Gong Cheng, Daxin Liu, and Yuzhong Qu. 2021. Fast Algorithms for Semantic Association Search and Pattern Mining. *IEEE Trans. Knowl. Data Eng.* 33, 4 (2021), 1490–1502. https://doi.org/10.1109/TKDE.2019.2942031

[12] Gong Cheng, Danyun Xu, and Yuzhong Qu. 2015. Summarizing Entity Descriptions for Effective and Efficient Human-centered Entity Linking. In *WWW 2015*. 184–194. https://doi.org/10.1145/2736277.2741094

[13] Trevor Cohen, Kirk Roberts, Anupama E. Gururaj, Xiaoling Chen, Saeid Pournejati, George Alter, William R. Hersh, Dina Demner-Fushman, Lucila Ohno-Machado, and Hua Xu. 2017. A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 bioCADDIE dataset retrieval challenge. *Database J. Biol. Databases Curation* 2017 (2017), bax061. https://doi.org/10.1093/database/bax061

[14] Auriol Degbelo. 2020. Open Data User Needs: A Preliminary Synthesis. In *Companion of WWW 2020*. 834–839. https://doi.org/10.1145/3366424.3386586

[15] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. 2020. RDF graph summarization for first-sight structure discovery. *VLDB J.* 29, 5 (2020), 1191–1218. https://doi.org/10.1007/s00778-020-00611-y

[16] Kalpa Gunaratna, Amir Hossein Yazdavar, Krishnaprasad Thirunarayan, Amit P. Sheth, and Gong Cheng. 2017. Relatedness-based Multi-Entity Summarization. In *IJCAI 2017*. 1060–1066. https://doi.org/10.24963/ijcai.2017/147

[17] Alon Y. Halevy, Flip Korn, Natalya Fridman Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing Google's Datasets. In *SIGMOD 2016*. 795–806. https://doi.org/10.1145/2882903.2903730

[18] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4 (2021), 71:1–71:37. https://doi.org/10.1145/3447772

[19] Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. 2018. Characterising Dataset Search Queries. In *WWW 2018*. 1485–1488. https://doi.org/10.1145/3184558.3191597

[20] Elem Güzel Kalayci, Irlán Grangel-González, Felix Lösch, Guohui Xiao, Anees ul Mehdi, Evgeny Kharlamov, and Diego Calvanese. 2020. Semantic Integration of Bosch Manufacturing Data Using Virtual Knowledge Graphs. In *ISWC 2020, Part II*. 464–481. https://doi.org/10.1007/978-3-030-62466-8_29

[21] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2021. A Test Collection for Ad-hoc Dataset Retrieval. In *SIGIR 2021*. 2450–2456. https://doi.org/10.1145/3404835.3463261

[22] Evgeny Kharlamov, Dag Hovland, Martin G. Skjæveland, Dimitris Bilidas, Ernesto Jiménez-Ruiz, Guohui Xiao, Ahmet Soylu, Davide Lanti, Martin Rezk, Dmitriy Zheleznyakov, Martin Giese, Hallstein Lie, Yannis E. Ioannidis, Yannis Kotidis, Manolis Koubarakis, and Arild Waaler. 2017. Ontology Based Data Access in Statoil. *J. Web Semant.* 44 (2017), 3–36. https://doi.org/10.1016/j.websem.2017.05.005

[23] Evgeny Kharlamov, Yannis Kotidis, Theofilos Mailis, Christian Neuenstadt, Charalampos Nikolaou, Özgür L. Özçep, Christoforos Svingos, Dmitriy Zheleznyakov, Yannis E. Ioannidis, Steffen Lamparter, Ralf Möller, and Arild Waaler. 2019. An ontology-mediated analytics-aware approach to support monitoring and diagnostics of static and streaming data. *J. Web Semant.* 56 (2019), 30–55. https://doi.org/10.1016/j.websem.2019.01.001

[24] Evgeny Kharlamov, Theofilos Mailis, Gulnar Mehdi, Christian Neuenstadt, Özgür L. Özçep, Mikhail Roshchin, Nina Solomakhina, Ahmet Soylu, Christoforos Svingos, Sebastian Brandt, Martin Giese, Yannis E. Ioannidis, Steffen Lamparter, Ralf Möller, Yannis Kotidis, and Arild Waaler. 2017. Semantic access to streaming and static data at Siemens. *J. Web Semant.* 44 (2017), 54–74. https://doi.org/10.1016/j.websem.2017.02.001

[25] Evgeny Kharlamov, Gulnar Mehdi, Ognjen Savkovic, Guohui Xiao, Elem Güzel Kalayci, and Mikhail Roshchin. 2019. Semantically-enhanced rule-based diagnostics for industrial Internet of Things: The SDRL language and case study for Siemens trains and turbines. *J. Web Semant.* 56 (2019), 11–29. https://doi.org/10.1016/j.websem.2018.10.004

[26] Laura M. Koesten, Emilia Kacprzak, Jenifer Fay Alys Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour. In *CHI 2017*. 1277–1289. https://doi.org/10.1145/3025453.3025838

[27] Junyou Li, Gong Cheng, Qingxia Liu, Wen Zhang, Evgeny Kharlamov, Kalpa Gunaratna, and Huajun Chen. 2020. Neural Entity Summarization with Joint Encoding and Weak Supervision. In *IJCAI 2020*. 1644–1650. https://doi.org/10.24963/ijcai.2020/228

[28] Shuxin Li, Zixian Huang, Gong Cheng, Evgeny Kharlamov, and Kalpa Gunaratna. 2020. Enriching Documents with Compact, Representative, Relevant Knowledge Graphs. In *IJCAI 2020*. 1748–1754. https://doi.org/10.24963/ijcai.2020/242

[29] Daxin Liu, Gong Cheng, Qingxia Liu, and Yuzhong Qu. 2019. Fast and Practical Snippet Generation for RDF Datasets. *ACM Trans. Web* 13, 4 (2019), 19:1–19:38. https://doi.org/10.1145/3365575

[30] Qingxia Liu, Gong Cheng, Kalpa Gunaratna, and Yuzhong Qu. 2021. Entity summarization: State of the art and future challenges. *J. Web Semant.* 69 (2021), 100647. https://doi.org/10.1016/j.websem.2021.100647

[31] Felicitas Löffler, Andreas Schuldt, Birgitta König-Ries, Helge Bruelheide, and Friederike Klan. 2021. A Test Collection for Dataset Retrieval in Biodiversity Research. *Res. Ideas Outcomes* 7 (2021), e67887. https://doi.org/10.3897/rio.7.e67887

[32] Michalis Mountantonakis and Yannis Tzitzikas. 2020. Content-based Union and Complement Metrics for Dataset Search over RDF Knowledge Graphs. *ACM J. Data Inf. Qual.* 12, 2 (2020), 10:1–10:31. https://doi.org/10.1145/3372750

[33] Angelo Batista Neves, Rodrigo G. G. de Oliveira, Luiz André P. Paes Leme, Giseli Rabello Lopes, Bernardo Pereira Nunes, and Marco A. Casanova. 2018. Empirical Analysis of Ranking Models for an Adaptable Dataset Search. In *ESWC 2018*. 50–64. https://doi.org/10.1007/978-3-319-93417-4_4

[34] Masayo Ota, Heiko Mueller, Juliana Freire, and Divesh Srivastava. 2020. Data-Driven Domain Discovery for Structured Datasets. *Proc. VLDB Endow.* 13, 7 (2020), 953–965. https://doi.org/10.14778/3384345.3384346

[35] Emmanuel Pietriga, Hande Gözükan, Caroline Appert, Marie Destandau, Sejla Cebiric, François Goasdoué, and Ioana Manolescu. 2018. Browsing Linked Data Catalogs with LODAtlas. In *ISWC 2018*. 137–153. https://doi.org/10.1007/978-3-030-00668-6_9

[36] Renzo Arturo Alva Principe, Blerina Spahiu, Matteo Palmonari, Anisa Rula, Flavio De Paoli, and Andrea Maurino. 2018. ABSTAT 1.0: Compute, Manage and Share Semantic Profiles of RDF Knowledge Graphs. In *ESWC 2018 Satellite Events*. 170–175. https://doi.org/10.1007/978-3-319-98192-5_32

[37] Dumitru Roman, Vladimir Alexiev, Javier Paniagua, Brian Elvesæter, Bjørn Marius von Zernichow, Ahmet Soylu, Boyan Simeonov, and Chris Taggart. 2022. The euBusinessGraph ontology: A lightweight ontology for harmonizing basic company information. *Semantic Web* 13, 1 (2022), 41–68. https://doi.org/10.3233/SW-210424

[38] Yulin Shen, Ziheng Chen, Gong Cheng, and Yuzhong Qu. 2021. CKGG: A Chinese Knowledge Graph for High-School Geography Education and Beyond. In *ISWC 2021*. 429–445. https://doi.org/10.1007/978-3-030-88361-4_25

[39] Yuxuan Shi, Gong Cheng, and Evgeny Kharlamov. 2020. Keyword Search over Knowledge Graphs via Static and Dynamic Hub Labelings. In *WWW 2020*. 235–245. https://doi.org/10.1145/3366423.3380110

[40] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Evgeny Kharlamov, and Yulin Shen. 2021. Efficient Computation of Semantically Cohesive Subgraphs for Keyword-Based Knowledge Graph Exploration. In *WWW 2021*. 1410–1421. https://doi.org/10.1145/3442381.3449900

[41] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Jie Tang, and Evgeny Kharlamov. 2021. Keyword-Based Knowledge Graph Exploration Based on Quadratic Group Steiner Trees. In *IJCAI 2021*. 1555–1562. https://doi.org/10.24963/ijcai.2021/215

[42] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Canim. 2020. Web Table Retrieval using Multimodal Deep Learning. In *SIGIR 2020*. 1399–1408. https://doi.org/10.1145/3397271.3401120

[43] Ahmet Soylu, Óscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Tom Blount, Francisco Yedro Martínez, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Dumitru Roman. 2022. TheyBuyForYou platform and knowledge graph: Expanding horizons in public procurement with open linked data. *Semantic Web* 13, 2 (2022), 265–291. https://doi.org/10.3233/SW-210442

[44] Ahmet Soylu, Martin Giese, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Dmitriy Zheleznyakov, and Ian Horrocks. 2017. Ontology-based end-user visual query formulation: Why, what, who, how, and which? *Univers. Access Inf. Soc.* 16, 2 (2017), 435–467. https://doi.org/10.1007/s10209-016-0465-0

[45] Ahmet Soylu, Evgeny Kharlamov, Dmitriy Zheleznyakov, Ernesto Jiménez-Ruiz, Martin Giese, Martin G. Skjæveland, Dag Hovland, Rudolf Schlatte, Sebastian Brandt, Hallstein Lie, and Ian Horrocks. 2018. OptiqueVQS: A visual query system over ontologies for industry. *Semantic Web* 9, 5 (2018), 627–660. https://doi.org/10.3233/SW-180293

[46] Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. 2016. ABSTAT: Ontology-Driven Linked Data Summaries with Pattern Minimalization. In *ESWC 2016 Satellite Events*. 381–395. https://doi.org/10.1007/978-3-319-47602-5_51

[47] Yawei Sun, Lingling Zhang, Gong Cheng, and Yuzhong Qu. 2020. SPARQA: Skeleton-Based Semantic Parsing for Complex Questions over Knowledge Bases. In *AAAI 2020*. 8952–8959.

[48] Xiaxia Wang, Gong Cheng, and Evgeny Kharlamov. 2019. Towards Multi-Facet Snippets for Dataset Search. In *PROFILES & SEMEX 2019*. 1–6.

[49] Xiaxia Wang, Gong Cheng, Tengteng Lin, Jing Xu, Jeff Z. Pan, Evgeny Kharlamov, and Yuzhong Qu. 2021. PCSG: Pattern-Coverage Snippet Generation for RDF Datasets. In *ISWC 2021*. 3–20. https://doi.org/10.1007/978-3-030-88361-4_1

[50] Xiaxia Wang, Gong Cheng, Jeff Z. Pan, Evgeny Kharlamov, and Yuzhong Qu. 2021. BANDAR: Benchmarking Snippet Generation Algorithms for (RDF) Dataset Search. *IEEE Trans. Knowl. Data Eng.* Early Access (2021), 1–14. https://doi.org/10.1109/TKDE.2021.3095309

[51] Xiaxia Wang, Tengteng Lin, Weiqing Luo, Gong Cheng, and Yuzhong Qu. 2021. Content-Based Open Knowledge Graph Search: A Preliminary Study with OpenKG.CN. In *CCKS 2021*. 104–115. https://doi.org/10.1007/978-981-16-6471-7_8

[52] Xiaxia Wang, Tengteng Lin, Weiqing Luo, Gong Cheng, and Yuzhong Qu. 2022. CKGSE: A Prototype Search Engine for Chinese Knowledge Graphs. *Data Intell.* 4, 1 (2022), 41–65. https://doi.org/10.1162/dint_a_00118

[53] Shuo Zhang and Krisztian Balog. 2018. Ad Hoc Table Retrieval using Semantic Similarity. In *WWW 2018*. 1553–1562. https://doi.org/10.1145/3178876.3186067

[54] Liang Zheng, Yuzhong Qu, Jidong Jiang, and Gong Cheng. 2015. Facilitating Entity Navigation Through Top-K Link Patterns. In *ISWC 2015, Part I*. 163–179. https://doi.org/10.1007/978-3-319-25007-6_10

[55] Baifan Zhou, Yulia Svetashova, Seongsu Byeon, Tim Pychynski, Ralf Mikut, and Evgeny Kharlamov. 2020. Predicting Quality of Automated Welding with Machine Learning and Semantics: A Bosch Case Study. In *CIKM 2020*. 2933–2940. https://doi.org/10.1145/3340531.3412737

[56] Baifan Zhou, Yulia Svetashova, Andre Gusmao, Ahmet Soylu, Gong Cheng, Ralf Mikut, Arild Waaler, and Evgeny Kharlamov. 2021. SemML: Facilitating development of ML models for condition monitoring with semantics. *J. Web Semant.* 71 (2021), 100664. https://doi.org/10.1016/j.websem.2021.100664