

# Dataset search: a survey

Adriane Chapman · Elena Simperl · Laura Koesten · George Konstantinidis · Luis-Daniel Ibez-Gonzalez · Emilia Kacprzak · Paul Groth

Received: date / Accepted: date

**Abstract** Generating value from data requires the ability to find, access and make sense of datasets. There are many efforts underway to encourage data sharing and reuse, from scientific publishers asking authors to submit data alongside manuscripts to data marketplaces, open data portals and data communities. Google recently beta released a search service for datasets, which allows users to discover data stored in various online repositories via keyword queries. These developments foreshadow an emerging research field around dataset search or retrieval that broadly encompasses frameworks, methods and tools that help match a user data need against a collection of datasets. Here, we survey the state of the art of research and commercial systems in dataset retrieval. We identify what makes dataset

search a research field in its own right, with unique challenges and methods and highlight open problems. We look at approaches and implementations from related areas dataset search is drawing upon, including information retrieval, databases, entity-centric and tabular search in order to identify possible paths to resolve these open problems as well as immediate next steps that will take the field forward.

## 1 Introduction

Data is increasingly used in decision making: to design public policies, identify customer needs, or run scientific experiments [49,130]. For instance, the integration of data from deployed sensor systems such as mobile phone networks, camera networks in intelligent transportation systems (ITS) [80] and smart meters [8]) is powering a number of innovative solutions such as the city of London’s oversight dashboard [19]. **Datasets are increasingly being exposed for trade within data markets [15,56] or shared via open data portals [1,63,77,3,4,5] and scientific repositories [2,10].** Communities such as Wikidata or the Linked Open Data Cloud [3] come together to create and maintain vast, general-purpose data resources, which can be used by developers in applications as diverse as intelligent assistants, recommender systems and search engine optimization. The common intent is to broaden the use and impact of the millions of datasets that are being made available and shared across organizations [24,111,139]. This trend is reinforced by advances in machine learning and artificial intelligence, which rely on data to train, validate and enhance their algorithms [120]. In order to support these uses, we must be able to search for datasets. Searching for data in principled ways has been

A. Chapman  
University of Southampton  
E-mail: adriane.chapman@soton.ac.uk

E. Simperl  
University of Southampton  
E-mail: E.Simperl@soton.ac.uk

L. Koesten  
The Open Data Institute  
E-mail: laura.koesten@theodi.org

G. Konstantinidis  
University of Southampton  
E-mail: G.Konstantinidis@soton.ac.uk

L. Ibez-Gonzalez  
University of Southampton  
E-mail: L.D.Ibanez-Gonzalez@soton.ac.uk

Emilia Kacprzak  
The Open Data Institute  
E-mail: emilia.kacprzak@theodi.org

Paul Groth  
University of Amsterdam  
E-mail: p.groth@uva.nl

researched for decades [34]. However, many properties of datasets are unique, with interesting requirements and constraints. There are many open problems across dataset search, which the database community can assist with.

Currently, there is a disconnect between what datasets are available, what dataset a user needs, and what datasets a user can actually find, trust and is able to use [24, 120, 124]. Dataset search is largely keyword based over published metadata, whether it is performed over crawls across the web [52, 122] or within organizational holdings [63, 77, 128]. There are several problems with this approach. Available metadata may not encompass the actual information a user needs to assess whether the dataset is fit for a given task [82]. Search results are returned to the user based on filters that were appropriate for web-based information, but do not always transfer well to datasets [54]. These limitations impact the use of the retrieved data - machine learning can be unduly affected by the processing that was performed over a dataset prior to its release [125], while knowing the original purpose for collecting the data aids interpretation and analysis [140]. In other words, in a dataset search context, approaches need to consider additional aspects such as data provenance [27, 53, 64, 87, 101, 142], annotations [67, 93, 144], quality [116, 131, 148], granularity of content [81], and schema [9, 20] to effectively evaluate a dataset's fitness for a particular use. The user does not have the ability to introspect over large amounts of data, and their attention must be prioritized [13]. In other cases, a user's need may require integrating data from different sources to form a new dataset [48, 116]. Furthermore, using a dataset is constrained by licenses and terms and conditions, which may prohibit such integration, especially when personal data is involved [102].

In order to realize the full potential of the datasets we are generating, maintaining and releasing, there is more research that must be done. Dataset search has not emerged in isolation, but has built on foundational work from other related areas. In Section 2, we outline the basic dataset search problem, and provide a quick review of the sub-areas that have influenced dataset search. Current commercial dataset search offerings are outlined in Section 3 while Section 4 provides a survey of dataset search research. Finally, Section 5 provides a synopsis of open problems in dataset search as well as related research that could be applied. Section 6 highlights a possible route to take steps to advance the field.

## 2 Background





To understand the fundamental problem of dataset search, we define a dataset. The concept of dataset is abstract, admitting several definitions depending on the particular community [24, 111]. There is a large body of work discussing the nature of data and its relation to practice and reuse [24, 25]. From a statistical point of view, the statistical data and metadata exchange initiative (SDMX) [123] defines a dataset as *'a collection of related observations, organized according to a predefined structure'*. This definition is shared by the DataCube vocabulary, which adds the notion of a *'common dimensional structure'* [134]. Meanwhile, the Organization for Economic Co-operation and Development (OECD), citing the US bureau and census, uses *'any permanently stored collection of information usually containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for multiple survey instances'* [123]. The Data Catalog Vocabulary [95] includes a dataset class, defined as a *'collection of data, published or curated by a single agent, and available for access or download in one or more formats.'* Finally, for the MELODA (MEtric for reLeasing Open DAta) initiative, a dataset is a *'group of structured data retrievable in a link or single instruction as a whole to a single entity, with updating frequency larger than a once a minute'* [98]. For the purposes of this paper, we will use the following definition:

**Definition 1 (Dataset)** *A collection of related observations organized and formatted for a particular purpose.*

Thus, dataset search involves the discovery, exploration, and return of datasets to an end user. We note two very distinct types of dataset search in this work. In what we will call "basic" dataset search, the set of related observations were organized for a particular purpose, and then released for consumption and reuse. We see this pattern of interaction within individual data repositories, such as Figshare [128], Dataverse [10], Elsevier Data Search [2], Open Data Portals [1, 63, 77, 3, 4, 5] and global searches such as DataMed [122] or Google Dataset Search [51]. A basic search, using any of these services is discussed in Example 1. Alternatively, a dataset search may involve a set of related observations that are organized for a particular purpose by the searcher themselves. This pattern of behaviour is particularly marked in Data Lakes [47, 117], data markets [15, 56], and tabular search [88, 151]; Example 2 illustrates this kind of data search.

New York City			A		
ALLIANCE ENERGY	239 10TH AVE	New York	NY	1000	
EASTSIDE SERVICE STATION	253 E 2ND ST	New York	NY	1000	
BP	21 E HOUSTON ST	New York	NY	1001	
FREDERICK BP	2040 FREDERICK DOUGLASS BLVD	New York	NY	1002	
ORLANDO TEJEDA	3225 BROADWAY	New York	NY	1002	
RIVER DRIVE CAR WASH AND GAS	673 W 125TH ST	New York	NY	1002	
SHELL	1599 LEXINGTON AVE	New York	NY	1002	
GETTY	348 E 106TH ST	New York	NY	1002	
MOBIL ON THE RUN	2165 AMSTERDAM AVE				
BROADWAY MOBIL	3740 BROADWAY				
GETTY	89 SAINT NICHOLAS PL				
COCO 4633	3936 10TH AVE				
HESS 32517	401 W 207TH ST				
BP	2326 1ST AVE				
SHELL	2276 1ST AVE				
BP	255 E 125TH ST				
EASTSIDE GAS	1890 PARK AVE				
HESS 32215	502 W 45TH ST				
145TH STREET MOBIL	150 W 145TH ST				
SHELL	232 W 145TH ST				
NEW YORK GETTY	119 W 145TH ST				
HESS 32520	120 W 145TH ST				
SHELL	1855 1ST AVE				
ADAMS GAS STATION	248 BAY ST				
STATEN ISLAND GETTY	1201 VICTORY BLVD				
7-ELEVEN	1252 FOREST AVE				
LIBERTY GAS	745 PORT RICHMOND AVE				
FOREST AND RICHMOND CI	1810 FOREST AVE				
HESS 32581	2121 FOREST AVE				
FOREST GULF	2151 FOREST AVE				
BP	1098 RICHMOND RD				

Tweets			B		
	NYC GAS @NYC_GAS	30m			
	RT @mks1188: 30 minute gas line at Shell on Long beach road and Merrick road near South Nassau #ligas				
	Expand				
	NYC GAS @NYC_GAS	32m			
	#nycgas #brooklyngas RT @juanguzman5422: 7-Eleven 301 65th & 3rd Brooklyn, NY 11220 gas now				
	Expand				
	NYC GAS @NYC_GAS	42m			
	#siopen RT @william_Nitka: Mobil station on Richmond Ave & Arthur Kill in Staten Island has gas. Minimal line. Regular only. #sigas				
	Expand				
	Axis Of Oversteer @AxisOfOversteer	2h			
	Gas line not bad at all @45th street Hess on 10th av in NYC. Maybe 10 min.				
	Retweeted by NYC GAS				

**Fig. 1** Datasets about gasoline availability in New York City in the week after Hurricane Sandy in 2012. (a) The American Automobile Association (AAA) created a structured dataset twice post-Sandy by phoning every gas station in the NYC area. It is complete, easy to use (CSV), accurate, clean, and was out of date by the time it was released. (b) The second dataset is a collection of tweets to NYC\_GAS. It is incomplete, requires Natural Language Processing (NLP) techniques to use, is dirty with respect to place names and addresses, but is up to date and timely throughout post-hurricane clean-up efforts.

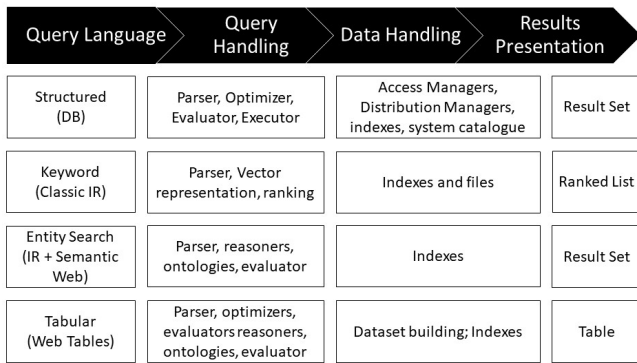
**Example 1 (Basic dataset search)** *Imagine you want to write an article on how Hurricane Sandy impacted the gasoline prices in New York City in the week after the incident. Consider the two datasets shown in Figure 1. Dataset A is from the American Automobile Association (AAA) and dataset B is from Twitter, documenting the gasoline available for purchase in New York City in the week after Hurricane Sandy. The choice of which dataset to use depends on the specifics of the information need, potentially the purpose and requirements of algorithms or processing methods, as well as the user's tool-set and data literacy. In order to find the right dataset, a user must issue a query that will return datasets, not tuples, documents or corpora. Differences inherent in the datasets should alter their ranking. For instance, a user who requires easy-to-use data, with fewer restrictions on timeliness may feel that the AAA dataset is a better fit than the other one. A user who wishes to establish an accurate timeline of gas in NYC would have a different assessment. These two users have different purposes, and therefore would assess the datasets differently. Moreover, both users use the content (gasoline) as the initial inclusion require-*

*ment, but use very different criteria and metrics to rank the datasets.*

**Example 2 (Constructive dataset search)** *In order to better understand the needs of the city, for instance to deal with flooding, the Centro De Operacoes Prefeitura Do Rio in Rio de Janeiro, Brazil mashes-up 'traffic and public transport, municipal and utility services, emergency services, weather feeds, and information sent in by employees and the public via phone, internet and radio' [80]. Consider a simple scenario in which datasets on weather highlighting rain amounts that could trigger a flash flood are integrated on the fly with datasets on traffic volume and augmented with identification of emergency response services in order to create a dataset that highlights the current populations at risk during an event. A recent extension to RapidMiner highlights the opportunities inherent in creating a dataset, with additional examples [48].*

## 2.1 Overview of generic dataset search

Figure 2 contains a high-level view of the search process, as well as a mapping to other communities who



**Fig. 2** An abstract view of the search process, comprising of querying, query processing, data handling and results presentation. Examples of how the database, IR, semantic web and tabular search communities implement these steps is shown.

are active in search. We will use the generalized steps indicated to outline the generic dataset search process below. A general approach to providing search over datasets is to model the user interface over existing keyword based information retrieval search systems where a user poses a query and a ranked list of existing datasets is returned. Indeed, a majority of data repositories provide this form of interface.

**Querying.** In the case of dataset search, a query is typically a keyword or Contextual Query Language (CQL) expression. Figure 3 shows the search interface for the UK government’s Open Data portal [5]. In addition to the keywords search box, the “Filter by” boxes allow the user to subset the data according to categories pre-identified by the repository.

**Query Handling.** The keywords, and any categories, submitted by the user are used to search over the metadata published about a dataset. Based on the metadata similarity to the search terms, a result set is produced.

**Data Handling.** In preparation for querying, the dataset owners must populate the metadata about their dataset. For instance, the dataset publisher supplies information such as title, description, language, temporal coverage, etc.; DCAT [95] is the W3C standard for interoperability of catalogues, and contains a representation and vocabulary for datasets. Additional metadata, such as summarizations [81, 106, 144] could also be contributed. Unfortunately, the creation and maintenance of this metadata is currently resource intensive.

**Results Presentation.** Search Engine’s Results Pages (SERPs) for dataset search currently follow a traditional *10 blue links* paradigm, as can be seen on many data portals [5, 10, 63, 77] as well as the Google Dataset search [52]. Basic filtering options, as can be seen in Figure 3, are sometimes available for faceted search within specific portals. Clicking on a search result takes the

## Search results

**Fig. 3** Dataset search engine result page for the UK government’s Open Data portal, data.gov.uk.

user to a preview page that contains metadata, such as information about the publisher, publishing data, licensing, etc. (see DCAT [95]). If available the preview page also contains a textual description. Many data portals will also include a preview by displaying a portion of the raw data or a visualization of particular patterns.

## 2.2 Common Search Architectures

As with searches in databases, IR and the semantic web, searches for datasets can be local, e.g. within a single repository [2, 10, 128, 117]. In a similar manner to a distributed database, given a query  $Q$  and a set of datasets (the *sources*), the query engine first selects the datasets relevant to the query [121, 133] and then chooses between different approaches: aggregating the datasets locally, using distributed processing as in Hadoop [143], or a federated approach [107].

The dataset search problem can be addressed at various levels. Services such as Google Dataset Search [52] and DataMed [122] crawl across the web and facilitate a global search across all distributed resources. These approaches use tags found in schema.org [57] or DCAT [95] to structure and identify the metadata considered important for datasets. However, the problem also exists at a local level, including open government portals such as data.gov.uk [5], organizational data lakes [117], scientific repositories such as Elsevier’s [2] and data markets [15, 56]. Across all these systems, users are attempting to discover and assess datasets for a particular purpose. Supporting them requires frameworks, methods and tools that specifically targets data as its input form and the specific information needs of data professionals.



## 2.3 Search sub-communities

Search has been addressed across many computer science sub-disciplines, such as databases, information retrieval, entity-centric search and tabular search. Figure 2 contains an overview of the high-level steps, and how each sub-discipline implements them. While dataset search is a subject in its own right, with distinct challenges and characteristics, it shares commonalities and draws upon insights from all these areas. In this section, we provide a very brief review of the focus and tools each community uses. We focus specifically on sub-areas in which the type of object returned is the same as the underlying data, e.g. a result set of data from a database of data, or a document from a corpus of documents. We neglect approaches such as question-answering [86] in which extra reasoning and manipulation of the returned result set is possible.

### 2.3.1 Entity-centric search

The task of entity search is to provide information about a specific named-entity (e.g. person, place, organization, ...) [17]. For a comprehensive view of entity-oriented search we refer the reader to [16]. Here, we introduce the work from the semantic web community in this space as it focuses on entities represented in data and not derived from text.

The semantic web community has worked towards creating machine-understandable graph-based representations of data [62]. It proposes languages, models and techniques to publish data online in the form of entities, properties, literals, and, most importantly, links to other resources. These links facilitate search and exploration of a global decentralized data space, similar to browsing and navigation on the web. The World Wide Web Consortium (W3C) settled on the Resource Description Format (RDF) as a standard model for representing and exchanging data about resources, which can refer to conventional web content as well as entities in the offline world such as people, places and organizations, identified by *International Resource Identifiers* (IRIs). Properties link entities or attach attributes to them. By reusing and linking IRIs, publishers signal that they hold data about the same entity, therefore enabling queries such as *Who holds data about England?* and *What do you know about England?* across multiple resources without any additional integration effort.

Realising the web of data requires several steps: lifting existing data into the semantic web, commonly referred to as linked (open) data [12], defining vocabularies and schemas to describe data in RDF and connecting to other datasets. For example, the Linked Open

Vocabulary portal<sup>1</sup> lists 635 such vocabularies, and provides search and exploration capabilities to find commonly used IRIs, to assist publishers in reusing them to facilitate data interpretation and interlinking.

Interlinking comprises two complementary problems. First is *entity resolution*: given two or more datasets, identify which entities and properties are the same. A general framework of entity resolution is described [33]. It covers the design of *similarity metrics* to compare entity descriptions, and the development of *blocking* techniques to group roughly similar entities together, in order to not apply more expensive similarity metrics to entities that have a low chance of ultimately resolving to the same entity. Finally, more recent efforts have tried *iterative* approaches, where discovered matches are used as input for computing similarities between further entities. The second part of interlinking is referred to as *link discovery*, where given two datasets, one has to find properties that hold between their entities. Properties can be equivalence or equality, as in entity resolution, or domain specific such as 'part-of' [62].

Information that is interlinked in this way allows for entity-centric searching, by identifying entities in the query and where they have similar matches in the data [150, 152].

### 2.3.2 Information Retrieval

IR systems can be broadly classified in web and document search engines, and engines for other types of entities (images, people, etc), called *vertical search engines*. The web and document engines use a number of statistical techniques to compute the relevance of a word (coming from an user query) to a document.

Vertical search engines are specifically tailored to the characteristics of the resources. For example, an email search has unique sets of resources for each users in addition to specific metadata such as sender and receiver addresses, topic or timestamp in order to judge the relevance [7]. Due to the specificity and limited scope of resources, vertical search engines often offer greater precision, utilize more complex schemas to match specific searching scenarios, and tend to support more complex user tasks [91, 147, 135].

### 2.3.3 Databases

The classic pipeline for search within a database begins with a structured query. Once a query is issued, the classic pipeline includes: parsing the query; creating an evaluation plan; optimizing the plan; executing

<sup>1</sup> <https://lov.linkeddata.es>

the plan utilizing appropriate indexes and catalogues. Other sources can provide greater details on each of these steps.

#### 2.3.4 Hidden/deep web

The *hidden*, or *deep*, web refers to datasets that lie “behind” web forms typically written in HTML [61,96], and ranging from medical research data to financial information and shopping catalogues. To access data behind a form a user needs to insert input text and submit the form, in order to be directed to a web page presenting the appropriate dataset [79]. It has been estimated that the data stored in hidden databases is an order of magnitude larger than the so-called *surface* web data, i.e., the data directly accessed by web crawlers [61,96].

There have been two main approaches to searching for data on the deep web. The first uses more traditional techniques to build vertical search engines, whereby semantic mappings are constructed between each website and a centralized mediator tailored to a particular domain. Structured queries are posed on the mediator and redirected as appropriate queries over the web forms using the mappings. Kosmix [115] (later transformed into WalmartLabs.com) was such a system presenting vertical engines for a large number of domains, ranging from health, and scientific data to car and flight sales. Other approaches to build such vertical search engines and automatically align different web forms, learn the forms’ possible inputs, and create centralized mediated forms [61]. A second group of approaches tries to generate the resulting web pages, usually in HTML, that come out of web form searches. Google has proposed a method for such *surfacing* of deep web content by automatically estimating input to several millions of HTML forms, written in many languages and spanning over hundreds of domains, and adding the resulting HTML pages into the Google search engine index [96]. The form inputs are stored as part of the indexed URL, and when users click on a search result they are directed to the result of the (freshly submitted) form.

#### 2.3.5 Tabular Search

In many cases, users are not interested in finding one specific dataset but instead are interested in extending or filling out an existing dataset, usually in the form of a table. Imagine a table with a series of columns and a series of rows, broadly the aim is to add additional relevant rows, columns or to fill in missing cell values. Thus, the input to the process is a table and the corresponding output is an enriched table. [145] identified three core tasks in the augmentation of tables.

1. Augmentation by attribute name - given a populated table and a new column name (i.e. attribute), populate the column with values. This is also referred to *table extension* elsewhere [28]. One can see this as finding tables which can be joined.
2. Attribute discovery - given a populated table, discover new potential column names.
3. Augmentation by example - given a populated table where some values are missing, fill in the missing values. This often referred to as *table completion* in the literature [150]. This task is like finding tables which can be unioned.

We refer to the combination of table extension, table completion, and attribute discovery as *tabular search*. This highlights that the query itself is a table and lends itself to the information retrieval perspective where the challenge is to answer the latent information need of the user. It is important to distinguish this task from *table search* which is the discovery of tables given a keyword search. Table search is a sub-task of dataset search.

**Table extension** [89] is divided into constrained and unconstrained table extension. Constrained table extension is essentially the augmentation by attribute name previously defined. Unconstrained table extension is also the addition of additional columns to a table but with no predefined label for the attribute. One can think of this as attribute discovery followed by constrained table extension.

A common technique to perform table extension is to discover existing tables through table similarity - in particular by measuring *schema similarity* [40]. Indeed, table extension was introduced by [28] where they defined a special operator EXTEND that would discover similar web tables to the given input table. Similarity here is computed with respect to the schema of the table. The values of the most similar table are then used to populate the input table’s additional column. The Infogather system [145] uses a similar approach but instead of just calculating the direct similarity between the input table and potential augmenting tables it also takes into the account the neighborhood around the potential augmenting tables. These indirect tables provide ancillary information that can be better suited for augmentation than the tables with the highest similarity to the input tables. Of interest, [40] have discovered that with respect to web tables there seems to be a latent link structure between tables. Recent work in table similarity has shown that semantic similarity using embedding approaches can improve performance over syntactic similarity measures [151].

**Table Completion** also relies heavily on table similarity as the mechanism for finding potential values that can be added to a table. [150] defines the notion of row

population, which adds additional rows to a table. For simplicity, we view this as a type of table completion in which the values to be complete form an additional row. Even more broadly, one could provide a set of columns as a query and have the system fill in the remaining rows [113].

The task of table completion can be seen as entity set completion where the goal is to complete a list given a set of seed entities [40, 150]. This task is relevant for a number of other tasks beyond table completion, including entity search [18] and knowledge base completion [39]. The completion of rows is similar to the broad problem of imputation and dealing with incomplete data [99]. Specific work in the context of the web has looked at performing imputation through the use of external data [6, 92, 126]. Much of that work has used web tables as the data source.

### 3 Current implementations

There are many functioning versions of dataset search in production today. In this section, we break down the set of dataset search services that exist according to their focus and how they deal with datasets.

#### 3.1 The Google Model

**Google’s systems.** In 2016, Google introduced *Goods*, an enterprise dataset search system, to manage datasets originating from different departments within the company with no unified structure or metadata [59]. In this catalog, related datasets are clustered based on the structure of the dataset or gathering frequency. Members of a group then become a single entry in the catalog. This helps to structure the catalog and also reduces the workload of metadata generation and schema computing. Within the Goods system each dataset entry has an overview of the dataset presented on a profile page. Using this profile, users can judge the dataset’s usefulness to their task. Keyword queries are then laid on top of this structure, producing a ranked result list of datasets as an output. Search functionality was built based on an inverted index of a subset of the datasets metadata. **In the absence of the information on the importance of each resource, [59] propose to rank the datasets based on heuristics over the type of a resource, precision of keyword match, if the dataset is used by other datasets and if the dataset contains an owner-sourced description.**

Following this work, in 2018 Google introduced a vertical web search engine tailored towards datasets on the web [51]. This system uses schema.org [57], which

is a schema for describing structured data on the web, and is applicable across a wide variety of data formats. It can be used as markup to describe structured content (e.g. tables within web pages) or as a metadata schema describing specific data with a defined list of metadata attributes. Google crawls the web for all the datasets described with use of schema.org *Dataset* class and collects all the metadata provided to describe a given resource. They further build the search capabilities on top of this metadata with additional information such as PageRank score of a page which contains metadata describing a given dataset [52].

**Open data portals.** Like Google, the open data portals [1, 3, 4, 5, 63, 77] provide search over the metadata of available datasets. The most popular platform in the governmental open data domain is CKAN [1]. CKAN is built using Apache Solr<sup>2</sup>, which uses Lucene to index the documents. **In this scenario, the documents are the datasets’ metadata provided by the publishers.** CKAN integrates the DCAT metadata schema which is an RDF vocabulary facilitating interoperability between data catalogs published on the web [95]. The main difference is that the open data portals do not need to crawl to collect this metadata. The open data portals catalogue their resources into pre-specified categories such as filetype, geographic region, etc. In addition metadata descriptions according to standards such as DCAT, which defines attributes such as title, description, language or licence [95] are also maintained. Despite the search functions provided by such catalogues, it is often not possible for an ordinary user to find relevant pieces of information quickly. This can be caused by: non-intuitive or limited data descriptions; misleading naming conventions; incorrect assignment of categories to datasets; the users lack of in-depth knowledge of the subject; or simply because the search is only conducted over the metadata records provided by the publishing bodies rather than the data itself [54]. The metadata describing datasets is often incomplete or outdated, as maintaining it is frequently manual and expensive. In many cases the metadata does not describe the full potential of the data, so some relevant datasets may not be presented as a result of a query simply because appropriate keywords were not used in the description.

In addition to DCAT [95] and Schema.org [57], other efforts were introduced to accommodate the most popular data format on the web. For example the ‘CSV on the Web’ working group has developed a standard for expressing useful metadata about tabular resources and CSV files specifically [127]. Their goal is to provide a uniform way of ensuring consistency of data types

<sup>2</sup> <http://lucene.apache.org/solr/>

and formats (e.g. uniqueness of values within a single column) for every file, which can provide basis for validation and prevent potential errors.

### 3.2 The Adding Value Model

In order to be more useful to a specific set of end users, many domains have also adopted strategies to effectively curate the contents of their search results for their specific end users. The searches for domain-specific datasets have corollaries to both the vertical web search engine provided by Google and the in-house searches of the Open Data portals. For instance, DataMed, a biomedical search engine uses a suite of tags, DATS, to allow a crawler to automatically index scientific datasets for search [122]. The Open Contracting Partnership released a Open Contracting Data Standard that identifies information needed about contracts to allow their crawler to access and catalogue contracting datasets [110]. On the other hand, data repositories like Elsevier [2], Figshare [128], Dataverse [10], and many Open Data portals [5,63,77], have no need of crawling, and primarily search over metadata contained within their purview.

The common theme of current dataset search strategies, both on the web and within the boundaries of a repository, is the reliance on dataset publishers tagging their data with appropriate information in the correct format. Because current dataset search only uses the metadata view of a dataset, it is imperative that these metadata descriptions are correct and maintained. Other, domain-specific solutions function in similar ways.

In aid of better searches, there are several attempts at monitoring and working over Open Data portals to provide a meta-analysis. For instance, the Open Data Portal Watch [104,105] currently watches 261 open data portals. Once a week, the metadata from all watched portals is fetched, the quality of the metadata computed, and the site updated to allow an cohesive search across the open data. Similarly, the Open Data Monitor reviews open data portals, and identifies where to search for information, in addition to assisting data owners successfully open their data [4].

### 3.3 The Constructive Dataset Model

Many private companies have understood that data is a commodity that can be effectively monetized. Some companies, such as Thomson Reuter have been collecting data to create datasets for sale for decades<sup>3</sup>. However, companies such as OpenCorporates uses public

data sources, with provenance, to gather information on legal entities. This dataset is then made publicly available<sup>4</sup>. Similarly, Researchably compiles information from scientific publications and makes interest-specific datasets for sale to biotech companies<sup>5</sup>. In all of these cases, the data exists in a scattered manner, and the company provides value by gathering, organizing and releasing it as a constructed dataset.

**Data Markets** exist as a way for organizations to realize value for their data [15,54,56]. While the user is able to download the entire dataset from a data market, it is also possible to access subsets of the data as needed to construct a dataset.

## 4 Survey of Dataset Search Research

This section surveys the current work related to dataset search. To organize it, we utilize the headings from Figure 2.

### 4.1 Querying

**Creating queries.** Users interact with datasets in a different manner than they interact with documents [78]. While this study is limited to social scientists, it indicates that users have a higher investment in the results, and are thus willing to spend more time searching. Moreover, the relationship of the dataset to the task at hand may play a larger role in dataset search; e.g. two datasets about cars could fit within a user's ability to understand and utilize, but may have very different results when Data-centric tasks can be categorized into two categories: (1) *Process-oriented tasks* used to produce an end analysis and (2) *Goal-oriented tasks* used in a machine learning process [82]. While the boundaries between the two categories are somewhat fluid and the same user might engage in both types of tasks, the primary difference between them lies in the 'user information needs', i.e. the details users need to know about the data in order to interact with it effectively. For process-oriented tasks, aspects such as timeliness, licenses, updates, quality, methods of data collection and provenance have a high priority. For goal-oriented tasks, intrinsic qualities of data such as coverage and granularity play a larger role. As yet, beyond the user filtering by certain characteristics, there is no way to state the task needs in the query. There has not yet been a movement away from keywords and CQL to query datasets.

<sup>3</sup> <https://www.thomsonreuters.com/>

<sup>4</sup> [opencorporates.com/](https://opencorporates.com/)

<sup>5</sup> <https://www.researchably.com/>



**Query Types.** As stated earlier, most queries for datasets use keywords or CQL over the metadata of the dataset. A formal query language that supports dataset retrieval does not yet exist. Instead, specific query interfaces are created for the underlying data type, e.g. [70] provides a SQL interface over text data and [104] for temporal and spatial data. Current implementations provide platform specific faceted search to allow basic filtering for categories such as publisher, format, license or topics (for instance [5]).

## 4.2 Query Handling

As stated in Section 2, most dataset searches operate over the dataset’s metadata. Unfortunately, low metadata quality (or missing metadata) affects both the discovery and the consumption of the datasets within Open Data Portals [131]. The success of the search functionality depends on the publishers knowledge of the dataset and the quality of the descriptions they provide.

Moving away from just searching over the metadata, [129] use the data type and column information for mapping columns in a query to the underlying table columns, while [113] allow keyword queries over columns. Similarly, [58] describe how to map structured sources into a semantic search capability. This is taken further in [151] by providing the ability to pose a keyword query over a *table*.

## 4.3 Data Handling

While the “handling” that typically needs to occur for dataset search at the moment is collection and indexing of metadata, there is research in additional data handling that can improve the effectiveness of search.

**Quality and Entity Resolution.** There are several efforts dealing with metadata quality [105,131]. One solution proposed to tackle the metadata quality problem include cross-validating metadata by merging feeds from identified entities [65]. Using self-categorized information [85] as facets is another. Attempts to better represent the underlying data [22] do have an affect on search. This includes better links with others data [43].

In the context of constructive dataset search, the Mannheim Search Join Engine [88,89] and WikiTables [21] use a table similarity approach for table extension but also look at the unconstrained task. In both cases, a similarity ranking between the input and augmentation tables is used to decide which columns should be added. Interestingly, the Mannheim system also consolidates

columns from different potential augmentation tables before performing the table extension.

**Summarization and Annotations.** To help both search and user understanding, summarizations and annotations are additional metadata that can be generated about the underlying dataset [81]. For instance, [102] deal with the problem that the underlying dataset cannot be exposed, but good summaries may help the user undertake the task of data access. Meanwhile, [93] use annotations to help support searching over data types and entities within a dataset, while [73] provide better labeling for numerical data in tables.

## 4.4 Results Presentation

**Ranking Datasets.** There are several works that look at ranking datasets. Of the most basic, after performing a keyword query over tables, a ranking on the returned tables is attempted [151]. In a more advanced method, [132] use an unsupervised learning approach to identify topics of database that can then be used in ranking. Finally, [90] rank datasets containing continuous information.

**Interactions.** Interactive query interfaces allow ad-hoc data analysis and exploration. Facilitating users exploration changes the fundamental requirements of the supporting infrastructure with respect to processing and workload [71]. Choosing a dataset greatly depends on the information provided alongside it. A number of studies indicate that standard metadata does not provide sufficient information for dataset reuse [81,106]. Recent studies have discussed textual ([81,129]) or visual [138] surrogates of datasets that aim to help people identify relevant documents and increase accuracy and/or satisfaction with their relevance judgments.

There has been additional research in how to help users interact with datasets for better understanding. For instance, there is the *many-answer problem*: users struggle to specify exact queries without knowing the data and their need to understand what is available in the whole result set to formulate and refine queries [94]. Currently dataset search is mainly performed over metadata, so the users understanding of what the dataset contains before download is limited by the quality, comprehensiveness and nature of metadata. A number of frameworks or SERP designs have been proposed as research prototypes for data search and exploration, such as TableLens ([114], DataLens [94], the relation browser [97] for sensemaking with statistical data, or summarization approaches of aggregate query answers in databases [136]. Navigational structures can support the cognitive representation of information [118] and we see a large space to explore interfaces that allow

more complex interaction with datasets such as sophisticated querying [69] (e.g. taking a dataset as input and searching for similar ones) or being able to follow links between entities in datasets.

Interaction characteristics for dataset search have been subject to several recent human data interaction studies. Moving beyond search as a technological problem, [54] show that there are also social considerations that impact a user when searching. In a comparison between document retrieval and dataset retrieval, [78] show that users are more reliant on metadata when performing dataset search. While looking at dataset users of varying abilities [26] show that the amount of tool support can impact a user's ability to effectively discover and use a dataset. Finally, in a framework for Human Interaction with Structured data [82] discuss three major aspects that matter to data practitioners when selecting a dataset to work with: *relevance*, *usability* and *quality*. Users judge the relevance of datasets for a specific task based on the dataset's scope (e.g. geographical and temporal scope) [104,75], basic statistics about the dataset such as counts and value ranges, and information about granularity of information in the data [81]. The documentation of variables and the context from which the dataset comes from also play a key role. Data quality is intertwined with a user's assessment of "fitness for use" and depends on various factors (dimensions or characteristics) such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability and verifiability [81]. Provenance is a prevalent attribute to judge a dataset's quality as it gives an indication of the authoritativeness, trustworthiness, context and original purpose of a dataset, e.g. [81,101]. In order to judge a dataset's usability for a given task, the following attributes have been identified as important: format, size, documentation, language (e.g. used in headers or for string values), comparability (e.g., identifiers, units of measurement), references to connected sources, and access (e.g. license, API) [81]. These are attributes independent of a dataset's content or topical relevance which can influence whether a user is actually able to engage with a dataset.

## 5 Open problems

In this survey, we have organized the literature into a framework that reflects the high-level steps necessary to implement a dataset search system. We have considered current research explicitly targeting dataset search challenges. In this section, we discuss several cross-cutting themes that need to be explored in greater detail to advance dataset search.

Issues of discoverability of open data were recognized by the European Commission which oversees the process of the data publishing within Europe. In 2011 they defined six barriers that challenge the reuse and true openness of data, which also apply to dataset search [44]:

- A lack of information that certain data actually exists and is available
- A lack of clarity of which public authority holds the data
- A lack of clarity about the terms of re-use
- Data which is made available only in formats that are difficult or expensive to use
- Complicated licensing procedures or prohibitive fees
- Exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

In addition to these challenges, we identify several additional problems that need attention.

### 5.1 Query languages: moving beyond keywords

Existing dataset search systems, whether it is Google's Dataset Search or vertical engines such as those used within data repositories, reuse query languages and concepts from information retrieval. Information needs are expressed via keyword queries, or, in the case of faceted search, via a series of filters modelled after metadata attributes such as domain, format or publisher. Studies in tabular search point to the need for alternative interfaces, which allow users to start their search journey with a table and then add to it as they explore the results. In addition to having different ways to capture information needs, it would also be beneficial to provide query languages that are able to combine information adaptively across multiple tables. This would be especially useful for tasks such as specifying data frames or generating comprehensive data-driven reports [55].

This connects dataset search to the area of text databases [70] and the deep web. However, much of that work has looked at verticals instead of search across datasets coming from multiple domains. The problem here is to be able to identify relevant tables for the input query, join them appropriately, and do subsequent query processing.

Existing research has primarily focused on structured queries (SQL, SPARQL) over the metadata of the datasets, without considering the actual content of the dataset. There is thus a need for richer query languages that are able to go beyond the metadata of datasets and are supported by indexing systems. Our understanding of the level of expressiveness of these languages is

still fairly limited. The W3C CSV on the Web working group [127] has made a proposal for specifying the semantics of columns and values in tables, but the approach requires mappings, which are typically specified manually.

#### 5.1.1 Entity-centric search building blocks

Entity-centric search naturally fits within the needs of dataset search. Datasets themselves are often built up of entities, and as such need the ability to specify as a query an entity, set of entities, or type of entity. Moreover, the notion of similarity [151] among entities should be expanded so that the entities themselves are not the focus of the match, but the number of similarities within the dataset.

#### 5.1.2 Database building blocks

Querying datasets will likely require new adaptations to query languages and methods. In addition to the exploration of a structured query language that can operate over datasets natively, other mechanisms to define queries should be explored. For instance, the overlap of programming languages and database query languages in which programming language concepts are used to define queries over databases with different levels of capabilities [35] or over MapReduce frameworks [45], could be one such rich area to explore.

#### 5.1.3 Tabular Search building blocks

Tabular search provides an interesting view on the potential query language requirements for dataset search, where instead of keywords, the input is a table itself. This also makes novel user interfaces possible, for example, to provide assistance during the creation of spreadsheets [149].

### 5.2 Query handling: Differentiated access

Most dataset search systems today either work within the confines of a single organization or on publicly available datasets that publish metadata according to a specified schema. However, there is demand to be able to pool information stemming from different organizations, for example, to be able to build cohorts for health studies from across clinical studies [36, 102]. Providing such *differentiated access* is critical for the emerging notion of *data trusts*,<sup>6</sup> which provide the legal, technical and

operational structures to share data between organizations.

We must facilitate an organizational as well as technical space to share data between both public and private entities. Thus, there are critical issues to be solved with respect querying over datasets with differing legal, privacy and even pricing properties. Without being able to search over these hidden datasets, access to a majority of data will be prevented. Here, aspects of using the provenance of data could be leveraged at query time [142]. We note that this is not just an issue for private data. Public data also has different properties (e.g. licenses) that users want to effectively integrate in their searches.

At an implementation level, further investigation into integrating security techniques in the query handling process is necessary. For example, searching over encrypted datasets [84, 14] or using digests to minimize disclosure while still enabling search [102]. All of this must be done while also considering that the demands of reuse may change the underlying requirements and bottlenecks of query processing [46].

#### 5.2.1 Information Retrieval building blocks

In the context of dataset retrieval the basic concepts supporting general web search are not sufficient, which indicates a need for more targeted approach for dataset retrieval, treating it as a unique vertical [28, 50].

#### 5.2.2 Database building blocks

The relational algebra that underpins our processing within a database [34], has no equivalent yet in dataset search. Recently, Apache released information about the query processing system used for many of the Apache products including Hive and Storm, and [20] investigated how the relational algebra can be applied to data contained within the various data processing frameworks in the Apache suite. Alternatively, other recent work in query processing attempts to handle non-relational operators via adaptive query processing [76].

Techniques such as those found in [112] suggest using a hybrid version of approximate query processing over samples and precomputation. Solutions such as ORCHESTRA [68] that were built to manage shared, structured data with changing schemas, cleaning, and queries that utilize provenance and annotation information (discussed in more detail below) need to be adapted to the dataset search problem. Other work from the probabilistic database area could also be of assistance. For instance [42] calculates the top-k results for queries

<sup>6</sup> <https://theodi.org/article/what-is-a-data-trust/>

over a probabilistic database by taking into account the lineage of each tuple. This usage of provenance to influence the overall ranking of the end result could inform dataset ranking.

Focusing on constructive dataset search, in which datasets are generated on-the-fly based on a user's needs and query, the work in data integration is particularly important. Querying sources in an integrated fashion [60, 83] becomes a foundational component of constructive dataset search.

### 5.3 Data handling: extra knowledge

In order to support the differentiated access and advanced exploratory interfaces articulated above, dataset search engines will need to become more advanced in their ingestion, indexing and cataloging procedures. This problem divides into two areas: incorporation of external knowledge in the data handling process and better management and usage of dataset-intrinsic information.

**Incorporating external knowledge**, whether through the use of domain ontologies, external quality indicators or even unstructured information (i.e. papers) that describe the datasets, is a critical problem. A concrete example of this problem: many datasets are described through code books that are written in natural language. These datasets are nearly useless without integration of external information about the codebooks themselves.

**Utilizing dataset-intrinsic information**, is necessary to more fully capture the richness of each dataset, and allow users to express a richer set of criteria during search. Within this space, there are open problems related to data *pre-processing*. How to do quality assessment on the fly? What kinds of indexes around quality need to be created? Moving beyond quality, in general, the automatic creation and maintenance of metadata that describes datasets is difficult. Users rely up on metadata to chose appropriate datasets. Open problems for metadata include:

1. identifying the metadata that is of highest value to users w.r.t. datasets;
2. tools to automatically create and maintain that metadata;
3. automatic annotation of dataset with metadata - linking them automatically to global ontologies.

In addition to pre-processing, current dataset search systems primarily rely on information retrieval architectures (e.g. indexing into Elasticsearch) to index and perform queries. Here, lessons learned from database architectures should be applied. This is particularly the case as we have seen the importance of lessons learned

from relational query engines being applied in the case of distributed data environments [11]. Thus, we think an important open problem is what the most effective *architectures* are for dataset search systems.

#### 5.3.1 Entity-centric search building blocks

One can apply the Linked Data paradigm to solve dataset search by converting datasets to RDF and following the full cycle, as described in [85]. However, for data publishers, it is often still very expensive to execute the full cycle. Furthermore, there is debate on whether certain datasets should have an RDF representation at all, as their original formats are perhaps more suited to the tools that are required for them (e.g. geospatial datasets). A middle-ground solution is to consider datasets as resources and encode only their description in RDF, for example, using the Data Catalog Vocabulary (a W3C recommendation) [95]. Then, the Linked Data cycle can then be applied to these descriptions, ultimately enabling the querying of datasets. The main challenge is the generation and maintenance of these descriptions, with some works tackling the problem of extracting specific properties from specific formats, like [104] for extracting spatio-temporal properties, and [74] for identifying the numerical properties in CSV tables.

#### 5.3.2 Database building blocks

As noted in [13], users do not have the 'attention' to introspect deeply into large and changing datasets. Instead, we can draw upon several areas of research from the database community, including data profiling and data quality.

Naumann's recent survey [103] provides a good overview of data profiling activities based on how data-users approach the task, and what resources are available for it. Of particular note for dataset search is the work on outlier detection [41, 94] as a way to provide indications to an end-user about the scope, spread and variety of a dataset during search. In particular, we note the techniques found in [153] are interesting for dataset search in that they split a large dataset into many smaller datasets and create an approximate representation of it for more accurate sampling of these sub-pieces. Finally, [47] establishes a tool that can comb through semi-structured log datasets to pull information into multi-layered structured datasets. All of these techniques may aid users in exploring and making sense of dataset. Given that a dataset is by definition a collection of pieces, imputation of missing pieces needs greater scrutiny. As discussed in Section 4, imputation efforts are underway [6, 22, 92, 126] but draw heav-



ily from web techniques. The imputation methods from the data management community should be considered.

The work on profiling contains expressions of data cleanliness and coverage, completeness and consistency. These properties are classic data quality metrics, and help the user form a picture of whether the data is fit for use. Automatic understanding of data quality in order to either populate metadata or answer metadata queries in a lazy manner will require techniques that can automatically determine complex datatypes such as [146]. Currently, though, the research in each of these areas has been focused on its relationship to describing or working within a specific artifact, not as a component for a search. To do this, the structures and content for each area need to be computable in a timely manner and presented in a way that can be taken advantage of by a search system. For instance, data quality is a traditionally resource expensive task that is often domain-specific. Generic, albeit possibly less accurate methods must be developed to compute data quality estimations that can be accessed and used during search [31, 100].

In order to facilitate understanding of the contents of a dataset, summarization can be used, as done in [108] over probabilistic databases. Provenance, another tool that could help users understand a dataset, has an unsolved problem of moving across granularity levels. A tuple within a dataset may have provenance associated with it, as may the table, and the entire dataset itself. The challenge is in understanding how the aggregation of tuple-provenance would affect the search results compared to dataset-provenance. Finally, using annotations to improve the data [67] will be needed. Interesting extensions could include using user feedback to facilitate ranking of datasets based on the searcher’s criteria, or utilizing the context under which the annotations were created to change how annotations impact ranking.

### 5.3.3 Hidden/Deep Web building blocks

An inherent challenge in dataset search over the web is to be able to identify particular resources as datasets of interest (and ignore, for example, natural language documents). This challenge will be also present in any forthcoming approach in searching for datasets on the deep web. Moreover, any such approach will build on some combination of the two main directions for surfacing deep web data. Building vertical engines for the hidden web has the difficulties of pre-defining all interesting domains, identifying relevant forms in front of datasets on the web and investigating automatic (or semi-automatic) approaches to create mappings; a task which seems extremely hard on a web scale. Hence, learning/computing web form inputs might be the op-

tion of choice. Nevertheless, in cases where there are complex domains that involve many attributes and involved inputs, e.g., airline reservations, when the datasets change frequently, e.g., financial data, or when forms use the http POST method [96] virtual integration remains an attractive direction.

### 5.3.4 Tabular Search building blocks

The majority of work in tabular search addresses web tables, not uploaded datasets. These tables have the benefit of generally being better described and often general-knowledge related, e.g., column names are human readable and not codes, or the tables are embedded in larger documents (e.g. HTML tables). In addition, a majority of work treats what are termed ‘entity-centric tables’, which are tables in which each row represents a single entity. Datasets can be much more general, for example, containing multiple tables in one file.

## 5.4 Result presentation: interactivity

As previously discussed, existing data search systems follow similar approaches to search showing a ranked list of search results with some additional faceted searching in place. At a tactical level, ranking approaches specifically tailored to dataset search should be developed. Importantly, this should take into account the kinds of rich indexes suggested in the prior section. Here, the challenges are that typical approaches to improving ranking from information retrieval such as learning to rank are difficult given that many data search engines do not have the kind of level of user traffic needed for learning to rank algorithms [132]. In addition, the integration of dataset search and entity search is an important open problem. For example, when searching for a chemical could you also display associated data and what that data should be.

Beyond standard search paradigms, supporting conversational search over data and embedding search into the actual data usage process deserves significant attention, particularly since dataset search is often needed in the context of a variety of tasks [124].

### 5.4.1 Information Retrieval building blocks

As pointed out by Cafarella et al. [28] structured data on the web is similar to the scenario of ranking of millions of individual databases. Tables available online contain a mixture of structural and related content elements which cannot easily be mapped to unstructured text scenarios applied in general web search. Tables lack the incoming hyperlink anchor text and are

two-dimensional - they cannot be efficiently queried using the standard inverted index. For those reasons PageRank-based algorithms known from general web search are not applicable to the same extent to the dataset/table search, particularly as tables of widely-varying quality can be found on a single web page.

Search for datasets is often complex and shows characteristics of exploratory search tasks, involving multiple queries, iterations and refinement of the original information need, as well as complex cognitive processing [82]. There are many possible reasons that users have diverse interaction styles, from context and domain specificity [54] to uncertainty in the search workflow itself [26]. It is important to note that users have different interaction styles with respect to 'getting the data'. These interactions range from question answering to "data return" to exploration [54,82]. From an interaction perspective, dataset search is not as advanced as web or document search. Contextual or personalized results, which are common on the web [137] are practically non-existent for dataset search. Additionally, dataset search relies on limited metadata instead of looking at the dataset itself. While many classifications for information seeking tasks exist [23], there is no widely used classification of dataset information seeking tasks yet.

#### 5.4.2 Database building blocks

Provenance [27,53,64,142] is likely to be a key element in assisting the user in choosing a dataset of interest. Until now, provenance has been used to facilitate trust in an artifact [37,38] or automatically estimate quality [66]. New methods must be developed to facilitate translation of this large graph into a format that a user who is evaluating whether or not to use a dataset can interpret and utilize [29]. The logic and possible new operators behind dataset search will open up new areas for determining why and why not to consider provenance of the dataset query results themselves [30,64,87].

The presentation of data models has been a topic in database literature [69] as well as exploration strategies of result spaces beyond the 10 blue links paradigm. For instance, the use of sideways and downwards exploration of web table queries by [32]. Challenges and directions for search results presentation and data exploration as part of the search process are discussed on a mostly speculative basis in literature, and include representing different types of results in a manner that express the structure of the underlying dataset (tables, networks, spatial presentations,etc) [69].

An overview of search results can enhance orientation and understanding of the information provided [118], which allows to get an awareness of the dataset result space as a whole. Making a large set of possible results more informative to the user has been explored for databases [136]. At the same time being able to investigate the dataset on a column, row and cell level to match both process and content oriented requirements on the search result can be necessary [113,127].

Within the scope of constructive dataset search, the work of [141] is essential to appropriately annotate and cite the results of queries.

In the next section, we discuss one foundation that is crucial for addressing these open problems, benchmarks.

## 6 The Road Forward: Benchmarks

One of the most widely recognized problems of dataset search is the lack of benchmarks. For instance, the Bio-CADDIE project, which attempts to index for discovery scientific datasets, has a pilot project to recommend appropriate datasets to users based on similar topic, size, usage, user background and context [72]. In order to do this, the pilot participants are creating a topic model across scientific articles, and using user query patterns to identify similar users. While this is an interesting start, and acknowledges that there are a myriad of overlapping concerns that impact dataset search, from content through user's ability, there is no way yet to measure whether the solution works. For this, a clear benchmark is needed. In this section we will outline the state of the art with respect to the evaluation of different parts of the dataset search pipeline, which were discussed earlier in this work.

Step one is identifying the set of metrics that are appropriate to dataset search. Do they mimic the online and offline metrics of information retrieval? At first blush, session abandonment rate, session success rate and zero result rate from information retrieval online metrics appear relevant, while click-through rate may need some adjustment for the context of datasets. Meanwhile, most of the offline metrics, from the set of precision-based metrics, to recall, fall-out, discounted cumulative gain, etc. are obviously still necessary.

However, there are dataset-specific metrics that may need to be considered. For instance, "completeness" could be an interesting new metric to consider. Many tasks involving datasets require the stitching of several datasets to create a whole that is fit for purpose. Is the right set, that creates a "complete" offering returned? How do we measure that the appropriate set of datasets

for a given purpose were returned. For instance, in the context of information retrieval on an Open Data Platform, [75] found that some user queries require multiple datasets which are equally relevant in opposition to a ranked result list of resources with single resource per rank. The question of how such result list should be returned to the user remains open, and creates an interesting case within benchmark creation.

The availability of benchmarks upon which solution across the query processing pipeline for dataset search can be tested is essential. Any benchmark created for dataset search needs to, explicitly or implicitly, highlight the relationships that exist between the user, the task at hand and the properties of the dataset or its metadata. Unlike classic web retrieval, there are added dimensions for dataset search. It is no longer enough for a user to find the information appropriate; for dataset search, the user and the specific task requirements must be satisfied. The result list presented to the user must be understandable and explorable, due to the added complexity of interpreting and using data.

Several benchmarks have already been created that cover tasks related to dataset search. These benchmarks include: managing RDF datasets [109]; information retrieval over Wikipedia tables [151]; assignment of semantic labels to web tables [119]. Further efforts in this area needed in order to truly understand and make progress on the underlying technology.

## 7 Conclusions

The topic of data-driven research will only grow; we are at the start of a journey in which datasets are used for analysis, decision making and resource optimization. Our current needs for Dataset Search require us to give due attention to this problem. The current state-of-the-art is focused on tuple, document or webpage. Datasets are an interesting entity to themselves with some properties shared with documents, tuples and webpages, and some unique to datasets.

In this work, we highlight that dataset search can be achieved through two different mechanisms: 1. issue query, return dataset; 2. issue query, build dataset. However, dataset search itself is in its infancy. Techniques from many other fields, including databases, information retrieval, and semantic web search can be applied towards the problem of dataset search. The creation of an initial service, Google Dataset Search, that allows for automatic indexing of datasets, and Google-style search over that indexed information marks this problem as important. Moreover, it highlights the research that still needs to be performed within the dataset retrieval domain, including: formal query language(s),

dealing with social and organizational restrictions when processing a query, providing additional information to support query processing, facilitating user exploration and interaction with a result set made up of datasets. This is an exciting time with respect to dataset search, in which there is a high need for datasets of all sorts, combined with burgeoning tools for dataset search, like Google Dataset Search, that provide the necessary infrastructure. However, further research is needed to fully understand and support dataset search.

## References

1. CKAN (2018). URL <https://ckan.org/>
2. Elsevier scientific repository (2018). URL <https://datasearch.elsevier.com/>
3. Linked open data cloud (2018). URL <https://www.lod-cloud.net/>
4. Open data monitor (2018). URL <https://www.opendatamonitor.eu>
5. Uk open data portal (2018). URL <https://data.gov.uk/>
6. Ahmadov, A., Thiele, M., Eberius, J., Lehner, W., Wrembel, R.: Towards a hybrid imputation approach using web tables. In: Big Data Computing (BDC), 2015 IEEE/ACM 2nd International Symposium on, pp. 21–30. IEEE (2015). DOI <https://doi.org/10.1109/BDC.2015.38>
7. Ai, Q., Dumais, S.T., Craswell, N., Liebling, D.: Characterizing email search using large-scale behavioral logs and surveys. In: Proceedings of the 26th International Conference on World Wide Web, WWW '17, pp. 1511–1520. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017). DOI [10.1145/3038912.3052615](https://doi.org/10.1145/3038912.3052615)
8. Alahakoon, D., Yu, X.: Smart electricity meter data intelligence for future energy systems: A survey. IEEE Transactions on Industrial Informatics **12**(1), 425–436 (2016). DOI [10.1109/TII.2015.2414355](https://doi.org/10.1109/TII.2015.2414355)
9. Alexe, B., ten Cate, B., Kolaitis, P.G., Tan, W.C.: Designing and refining schema mappings via data examples. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011), pp. 133–144. Athens, Greece (2011)
10. Altman, M., Castro, E., Crosas, M., Durbin, P., Gannett, A., Whitney, J.: Open journal systems and data-verse integration—helping journals to upgrade data publication for reusable research. The Code4Lib Journal **30** (2015)
11. Armbrust, M., Xin, R.S., Lian, C., Huai, Y., Liu, D., Bradley, J.K., Meng, X., Kaftan, T., Franklin, M.J., Ghodsi, A., et al.: Spark sql: Relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1383–1394. ACM (2015). DOI <https://doi.org/10.1145/2723372.2742797>
12. Auer, S., Böhmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., Van Nuffelen, B., Stadler, C., Tramp, S., Williams, H.: Managing the life-cycle of linked data with the lod2 stack. In: International semantic Web conference, pp. 1–16. Springer (2012). DOI [https://doi.org/10.1007/978-3-642-35173-0\\_1](https://doi.org/10.1007/978-3-642-35173-0_1)

13. Bailis, P., Gan, E., Rong, K., Suri, S.: Prioritizing attention in fast data: Principles and promise. In: Conference on Innovative Dataset Research (CIDR) (2017)
14. Bakshi, S., Chavan, S., Kumar, A., Harganekar, S.: Query processing on encoded data using bitmap. *Journal of Data Mining and Management* **3** (2018)
15. Balazinska, M., Howe, B., Koutris, P., Suciu, D., Upadhyaya, P.: A Discussion on Pricing Relational Data, pp. 167–173. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). DOI 10.1007/978-3-642-41660-6\_7
16. Balog, K.: Entity-oriented search. Springer (2018)
17. Balog, K., Meij, E., de Rijke, M.: Entity search: Building bridges between two worlds. In: Proceedings of the 3rd International Semantic Search Workshop, SEM-SEARCH '10, pp. 9:1–9:5. ACM, New York, NY, USA (2010). DOI 10.1145/1863879.1863888. URL <http://doi.acm.org/10.1145/1863879.1863888>
18. Balog, K., Serdyukov, P., de Vries, A.P.: Overview of the trec 2010 entity track. In: TREC (2010)
19. Batty, M.: Big data and the city. *Built Environment* **42**, 321–337(17) (2016). DOI 10.2148/benv.42.3.321
20. Begoli, E., Camacho-Rodríguez, J., Hyde, J., Mior, M.J., Lemire, D.: Apache calcite: A foundational framework for optimized query processing over heterogeneous data sources. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pp. 221–230. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3190662
21. Bhagavatula, C.S., Noraset, T., Downey, D.: Methods for exploring and mining tables on wikipedia. In: Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, pp. 18–26. ACM (2013). DOI <https://doi.org/10.1145/2501511.2501516>
22. Bischof, S., Harth, A., Kmpgen, B., Polleres, A., Schneider, P.: Enriching integrated statistical open city data by combining equational knowledge and missing value imputation. *Journal of Web Semantics* **48**, 22 – 47 (2018). DOI <https://doi.org/10.1016/j.websem.2017.09.003>
23. Blandford, A., Attfield, S.: Interacting with information. *Synthesis Lectures on Human-Centered Informatics* **3**(1), 1–99 (2010)
24. Borgman, C.L.: The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* **63**(6), 1059–1078 (2012). DOI 10.1002/asi.22634
25. Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press (2015)
26. Boukhefifa, N., Perrin, M.E., Huron, S., Eagan, J.: How data workers cope with uncertainty: A task characterisation study. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pp. 3645–3656. ACM, New York, NY, USA (2017). DOI 10.1145/3025453.3025738
27. Buneman, P., Chapman, A., Cheney, J.: Provenance management in curated databases. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD '06, pp. 539–550. ACM, New York, NY, USA (2006). DOI 10.1145/1142473.1142534
28. Cafarella, M.J., Halevy, A., Wang, D.Z., Wu, E., Zhang, Y.: Webtables: Exploring the power of tables on the web. *Proceedings of the VLDB Endowment* **1**(1), 538–549 (2008). DOI 10.14778/1453856.1453916
29. Chapman, A., Blaustein, B.T., Seligman, L., Allen, M.D.: Plus: A provenance manager for integrated information. In: 2011 IEEE International Conference on Information Reuse Integration, pp. 269–275 (2011). DOI 10.1109/IRI.2011.6009558
30. Chapman, A., Jagadish, H.V.: Why not? In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, SIGMOD '09, pp. 523–534. ACM, New York, NY, USA (2009). DOI 10.1145/1559845.1559901
31. Chapman, A.P., Rosenthal, A., Seligman, L.: The challenge of quick and dirty information quality. *J. Data and Information Quality* **7**(1-2), 1:1–1:4 (2016). DOI 10.1145/2834123
32. Chirigati, F., Liu, J., Korn, F., Wu, Y.W., Yu, C., Zhang, H.: Knowledge exploration using tables on the web. *Proceedings of the VLDB Endowment* **10**(3), 193–204 (2016). DOI 10.14778/3021924.3021935
33. Christophides, V., Efthymiou Vasilis, S.K.: Entity Resolution in the Web of Data. Morgan & Claypool (2015)
34. Codd, E.F.: Relational completeness of data base sublanguages. Citeseer (1972)
35. Costa Seco, J., Ferreira, P., Loureno, H.: Capability-based localization of distributed and heterogeneous queries. *Journal of Functional Programming* **27**, e26 (2017). DOI 10.1017/S095679681700017X
36. Cui, L., Zeng, N., Kim, M., Mueller, R., Hankosky, E.R., Redline, S., Zhang, G.Q.: X-search: an open access interface for cross-cohort exploration of the national sleep research resource. *BMC Medical Informatics and Decision Making* **18**(1) (2018). DOI 10.1186/s12911-018-0682-y
37. Curcin, V., Fairweather, E., Danger, R., Corrigan, D.: Templates as a method for implementing data provenance in decision support systems. *Journal of Biomedical Informatics* **65**, 1 – 21 (2017). DOI <https://doi.org/10.1016/j.jbi.2016.10.022>
38. Dai, C., Lin, D., Bertino, E., Kantarcioglu, M.: An approach to evaluate data trustworthiness based on data provenance. In: W. Jonker, M. Petković (eds.) *Secure Data Management*, pp. 82–98. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
39. Dalvi, B.B., Cohen, W.W., Callan, J.: Websets: Extracting sets of entities from the web using unsupervised information extraction. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pp. 243–252. ACM, New York, NY, USA (2012). DOI 10.1145/2124295.2124327
40. Das Sarma, A., Fang, L., Gupta, N., Halevy, A., Lee, H., Wu, F., Xin, R., Yu, C.: Finding related tables. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 817–828. ACM (2012). DOI <https://doi.org/10.1145/2213836.2213962>
41. Dong, B., Wang, H.W., Monreale, A., Pedreschi, D., Giannotti, F., Guo, W.: Authenticated outlier mining for outsourced databases. *IEEE Transactions on Dependable and Secure Computing* (2017). DOI <https://doi.org/10.1109/TDSC.2017.2754493>
42. Dylla, M., Miliaraki, I., Theobald, M.: Top-k query processing in probabilistic databases with non-materialized views. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 122–133 (2013). DOI 10.1109/ICDE.2013.6544819
43. Ellefi, M.B., Bellahsene, Z., Dietze, S., Todorov, K.: Dataset recommendation for data linking: an intensional approach. In: *International Semantic Web Conference*, pp. 36–51. Springer (2016)
44. European Commission, D.A.: Commission's open data strategy, questions and answers. Memo/11/891 (2011)
45. Fegaras, L.: An algebra for distributed big data analytics. *Journal of Functional Programming* **27**, e27 (2017). DOI 10.1017/S0956796817000193



46. Galakatos, A., Crotty, A., Zraggen, E., Binnig, C., Kraska, T.: Revisiting reuse for approximate query processing. *Proceedings of the VLDB Endowment* **10**(10), 1142–1153 (2017). DOI 10.14778/3115404.3115418
47. Gao, Y., Huang, S., Parameswaran, A.: Navigating the data lake with datamaran: Automatically extracting structure from log datasets. In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, pp. 943–958. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3183746
48. Gentile, A.L., Kirstein, S., Paulheim, H., Bizer, C.: Extending rapidminer with data search and integration capabilities. In: H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, C. Lange (eds.) *The Semantic Web*, pp. 167–171. Springer International Publishing, Cham (2016)
49. Gohar, M., Muzammal, M., Rahman, A.U.: Smart tss: Defining transportation system behavior using big data analytics in smart cities. *Sustainable Cities and Society* **41**, 114 – 119 (2018). DOI <https://doi.org/10.1016/j.scs.2018.05.008>
50. Gonzalez, H., Halevy, A.Y., Jensen, C.S., Langen, A., Madhavan, J., Shapley, R., Shen, W., Goldberg-Kidon, J.: Google fusion tables: Web-centered data management and collaboration. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10*, pp. 1061–1066. ACM, New York, NY, USA (2010). DOI 10.1145/1807167.1807286
51. Google: Blog: Google dataset search (2018). URL <https://ai.googleblog.com/2018/09>
52. Google: Google dataset search (2018). URL <https://developers.google.com/search/docs/data-types/dataset>
53. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: *Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '07*, pp. 31–40. ACM, New York, NY, USA (2007). DOI 10.1145/1265530.1265535
54. Gregory, K., Groth, P.T., Cousijn, H., Scharnhorst, A., Wyatt, S.: Searching data: A review of observational data retrieval practices. *CoRR abs/1707.06937* (2017)
55. Groth, P.T., Scerri, A., Jr., R.D., Allen, B.P.: End-to-end learning for answering structured queries directly over text. *CoRR abs/1811.06303* (2018)
56. Grubenmann, T., Bernstein, A., Moor, D., Seuken, S.: Financing the web of data with delayed-answer auctions. In: *Proceedings of the 2018 World Wide Web Conference, WWW '18*, pp. 1033–1042. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2018). DOI 10.1145/3178876.3186002. URL <https://doi.org/10.1145/3178876.3186002>
57. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. *Communications of the ACM* **59**(2), 44–51 (2016). DOI 10.1145/2844544
58. Gupta, S., Szekely, P., Knoblock, C.A., Goel, A., Taheriyani, M., Muslea, M.: Karma: A system for mapping structured sources into the semantic web. In: E. Simperl, B. Norton, D. Mladenic, E. Della Valle, I. Fundulaki, A. Passant, R. Troncy (eds.) *The Semantic Web: ESWC 2012 Satellite Events*, pp. 430–434. Springer Berlin Heidelberg, Berlin, Heidelberg (2015)
59. Halevy, A., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang, S.E.: Goods: Organizing google's datasets. In: *Proceedings of the 2016 International Conference on Management of Data*, pp. 795–806. ACM (2016)
60. Halevy, A.Y.: Answering queries using views: A survey. *The VLDB Journal* **10**(4), 270–294 (2001)
61. He, B., Patel, M., Zhang, Z., Chang, K.C.C.: Accessing the deep web. *Communications of the ACM* **50**(5), 94–101 (2007)
62. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool (2011)
63. Hendler, J., Holm, J., Musialek, C., Thomas, G.: Us government linked open data: Semantic.data.gov. *IEEE Intelligent Systems* **27**(3), 25–31 (2012). DOI 10.1109/MIS.2012.27
64. Herschel, M., Diestelkämper, R., Ben Lahmar, H.: A survey on provenance: What for? what form? what from? *The VLDB Journal* **26**(6), 881–906 (2017). DOI 10.1007/s00778-017-0486-1
65. Heyvaert, P., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Merging and enriching dcat feeds to improve discoverability of datasets. In: *International Semantic Web Conference*, pp. 67–71. Springer (2015)
66. Huynh, T., Ebden, M., Fischer, J., Roberts, S., Moreau, L.: Provenance network analytics: An approach to data analytics using data provenance. *DATA MINING AND KNOWLEDGE DISCOVERY* (2018). DOI 10.1007/s10618-017-0549-3
67. Ibrahim, K., Du, X., Eltabakh, M.: Proactive annotation management in relational databases. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15*, pp. 2017–2030. ACM, New York, NY, USA (2015). DOI 10.1145/2723372.2749435
68. Ives, Z.G., Green, T.J., Karvounarakis, G., Taylor, N.E., Tannen, V., Talukdar, P.P., Jacob, M., Pereira, F.: The orchestra collaborative data sharing system. *SIGMOD Rec.* **37**(3), 26–32 (2008). DOI 10.1145/1462571.1462577
69. Jagadish, H.V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., Yu, C.: Making database systems usable. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007*, pp. 13–24 (2007). DOI 10.1145/1247480.1247483
70. Jain, A., Doan, A., Gravano, L.: Sql queries over unstructured text databases. In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 1255–1257. IEEE (2007)
71. Jiang, L., Rahman, P., Nandi, A.: Evaluating interactive data systems: Workloads, metrics, and guidelines. In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, pp. 1637–1644. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3197386
72. Jiang, X., Qin, Z., Vaidya, J., Menon, A., Yu, H.: Pilot project 2.1 - data recommendation using machine learning and crowdsourcing (2018)
73. Kacprzak, E., Giménez-García, J.M., Piscopo, A., Koesten, L., Ibáñez, L.D., Tennison, J., Simperl, E.: Making sense of numerical data-semantic labelling of web tables. In: *European Knowledge Acquisition Workshop*, pp. 163–178. Springer (2018)
74. Kacprzak, E., Gimnez-Garca, J.M., Piscopo, A., Koesten, L., Ibez, L.D., Tennison, J., Simperl, E.: Making Sense of Numerical Data - Semantic Labelling of Web Tables. In: C. Faron Zucker, C. Ghidini, A. Napoli, Y. Toussaint (eds.) *Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science*, pp. 163–178. Springer International Publishing (2018)

75. Kacprzak, E., Koesten, L., Ibez, L.D., Blount, T., Tennison, J., Simperl, E.: Characterising dataset search - an analysis of search logs and data requests. *Journal of Web Semantics* (2018). DOI <https://doi.org/10.1016/j.websem.2018.11.003>
76. Kaftan, T., Balazinska, M., Cheung, A., Gehrke, J.: Cuttlefish: A lightweight primitive for adaptive query processing. *CoRR* **abs/1802.09180** (2018)
77. Kassen, M.: A promising phenomenon of open data: A case study of the chicago open data project. *Government Information Quarterly* **30**(4), 508 – 513 (2013). DOI <https://doi.org/10.1016/j.giq.2013.05.012>
78. Kern, D., Mathiak, B.: Are there any differences in data set retrieval compared to well-known literature retrieval? In: S. Kapidakis, C. Mazurek, M. Werla (eds.) *Research and Advanced Technology for Digital Libraries*, pp. 197–208. Springer International Publishing, Cham (2015)
79. Khare, R., An, Y., Song, I.Y.: Understanding deep web search interfaces: A survey. *ACM SIGMOD Record* **39**(1), 33–40 (2010)
80. Kitchin, R.: The real-time city? big data and smart urbanism. *GeoJournal* **79**(1), 1–14 (2014). DOI [10.1007/s10708-013-9516-8](https://doi.org/10.1007/s10708-013-9516-8)
81. Koesten, L., Simperl, E., Kacprzak, E., Blount, T., Tennison, J.: Everything you always wanted to know about a dataset: studies in data summarisation. *CoRR* **abs/1810.12423** (2018)
82. Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E.: The trials and tribulations of working with structured data: -a study on information seeking behaviour. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, CO, USA, May 06–11, 2017., pp. 1277–1289 (2017). DOI [10.1145/3025453.3025838](https://doi.org/10.1145/3025453.3025838)
83. Konstantinidis, G., Ambite, J.L.: Scalable query rewriting: a graph-based approach. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 97–108. Athens, Greece (2011)
84. Kumar, A., Hussain, M.: Secure query processing over encrypted database through cryptdb. In: P.K. Sa, S. Bakshi, I.K. Hatzilygeroudis, M.N. Sahoo (eds.) *Recent Findings in Intelligent Computing Techniques*, pp. 307–319. Springer Singapore, Singapore (2018)
85. Kunze, S.R., Auer, S.: Dataset Retrieval. In: *2013 IEEE Seventh International Conference on Semantic Computing*, pp. 1–8 (2013)
86. Kwok, C.C.T., Etzioni, O., Weld, D.S.: Scaling question answering to the web. *ACM Transactions on Information Systems* **19**(3), 242–262 (2001). DOI [10.1145/502115.502117](https://doi.org/10.1145/502115.502117)
87. Lee, S., Khler, S., Ludscher, B., Glavic, B.: A sql-middleware unifying why and why-not provenance for first-order queries. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 485–496 (2017). DOI [10.1109/ICDE.2017.105](https://doi.org/10.1109/ICDE.2017.105)
88. Lehmberg, O., Bizer, C.: Stitching web tables for improving matching quality. *Proceedings of the VLDB Endowment* **10**(11), 1502–1513 (2017). DOI [10.14778/3137628.3137657](https://doi.org/10.14778/3137628.3137657)
89. Lehmberg, O., Ritze, D., Ristoski, P., Meusel, R., Paulheim, H., Bizer, C.: The mannheim search join engine. *Journal of Web Semantics* **35**, 159 – 166 (2015). DOI <https://doi.org/10.1016/j.websem.2015.05.001>. URL <http://www.sciencedirect.com/science/article/pii/S157082681500030X>. Semantic Web Challenge 2014
90. Li, J., Deshpande, A.: Ranking continuous probabilistic datasets. *Proc. VLDB Endow.* **3**(1-2), 638–649 (2010). DOI [10.14778/1920841.1920923](https://doi.org/10.14778/1920841.1920923). URL <http://dx.doi.org/10.14778/1920841.1920923>
91. Li, X., Liu, B., Yu, P.: Time sensitive ranking with application to publication search. In: *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pp. 893–898. IEEE (2008)
92. Li, Z., Sharaf, M.A., Sitbon, L., Sadiq, S., Indulska, M., Zhou, X.: A web-based approach to data imputation. *World Wide Web* **17**(5), 873–897 (2014)
93. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment* **3**(1), 1338–1347 (2010)
94. Liu, B., Jagadish, H.V.: Datalens: making a good first impression. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, pp. 1115–1118 (2009). DOI [10.1145/1559845.1559997](https://doi.org/10.1145/1559845.1559997)
95. Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (dcat). *W3C Recommendation* **16** (2014). URL <https://www.w3.org/TR/vocab-dcat/#class-dataset>
96. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep web crawl. *Proceedings of the VLDB Endowment* **1**(2), 1241–1252 (2008)
97. Marchionini, G., Haas, S.W., Zhang, J., Elsas, J.: Accessing government statistical information. *Computer* **38**(12), 52–61 (2005). DOI [10.1109/MC.2005.393](https://doi.org/10.1109/MC.2005.393)
98. MELODA: Meloda dataset definition (2018). URL <http://www.meloda.org/dataset-definition/>
99. Miao, X., Gao, Y., Guo, S., Liu, W.: Incomplete data management: A survey. *Frontiers of Computer Science* pp. 1–22 (2018)
100. Missier, P., M. Embury, S., Mark Greenwood, R., D. Preece, A., Jin, B.: Quality views: Capturing and exploiting the user perspective on data quality. In: *Proceedings of the 32Nd International Conference on Very Large Data Bases*, pp. 977–988. VLDB Endowment (2006)
101. Moreau, L., Groth, P.T.: Provenance: An Introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool Publishers (2013). DOI [10.2200/S00528ED1V01Y201308WBE007](https://doi.org/10.2200/S00528ED1V01Y201308WBE007)
102. Mork, P., Smith, K., Blaustein, B., Wolf, C., Samuel, K., Sarver, K., Vayndiner, I.: Facilitating discovery on the private web using dataset digests. *Int. J. Metadata Semant. Ontologies* **5**(3), 170–183 (2010). DOI [10.1504/IJMSO.2010.034042](https://doi.org/10.1504/IJMSO.2010.034042)
103. Naumann, F.: Data profiling revisited. *SIGMOD Rec.* **42**(4), 40–49 (2014). DOI [10.1145/2590989.2590995](https://doi.org/10.1145/2590989.2590995)
104. Neumaier, S., Polleres, A.: Enabling spatio-temporal search in open data. Tech. rep., Department für Informationsverarbeitung und Prozessmanagement, WU Vienna University of Economics and Business (2018)
105. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. *J. Data and Information Quality* **8**(1), 2:1–2:29 (2016). DOI [10.1145/2964909](https://doi.org/10.1145/2964909). URL <http://doi.acm.org/10.1145/2964909>
106. Nguyen, T.T., Nguyen, Q.V.H., Weidlich, M., Aberer, K.: Result selection and summarization for web table search. In: *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pp. 231–242. IEEE (2015)

107. Oguz, D., Ergenc, B., Yin, S., Dikenelli, O., Hameurlain, A.: Federated query processing on linked data: a qualitative survey and open challenges. *The Knowledge Engineering Review* **30**(5), 545–563 (2015)
108. Orr, L., Balazinska, M., Suciu, D.: Probabilistic database summarization for interactive data exploration. *Proceedings of the VLDB Endowment* **10**(10), 1154–1165 (2017). DOI 10.14778/3115404.3115419
109. Pan, Z., Zhu, T., Liu, H., Ning, H.: A survey of rdf management technologies and benchmark datasets. *Journal of Ambient Intelligence and Humanized Computing* **9**(5), 1693–1704 (2018). DOI 10.1007/s12652-018-0876-2. URL <https://doi.org/10.1007/s12652-018-0876-2>
110. Partnership, O.C.: Open contracting data standard (2015). URL <http://standard.open-contracting.org/latest/en/>
111. Pasquetto, I.V., Randles, B.M., Borgman, C.L.: On the reuse of scientific data. *Data Science Journal* **16** (2017). DOI 10.5334/dsj-2017-008
112. Peng, J., Zhang, D., Wang, J., Pei, J.: Aqp++: Connecting approximate query processing with aggregate precomputation for interactive analytics. In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, pp. 1477–1492. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3183747
113. Pimplikar, R., Sarawagi, S.: Answering table queries on the web using column keywords. *Proceedings of the VLDB Endowment* **5**(10), 908–919 (2012). DOI 10.14778/2336664.2336665
114. Pirolli, P., Rao, R.: Table lens as a tool for making sense of data. In: *Proceedings of the workshop on Advanced visual interfaces 1996*, Gubbio, Italy, May 27–29, 1996, pp. 67–80 (1996). DOI 10.1145/948449.948460
115. Rajaraman, A.: Kosmix: High-performance topic exploration using the deep web. *Proceedings of the VLDB Endowment* **2**(2), 1524–1529 (2009). DOI 10.14778/1687553.1687581
116. Rekatsinas, T., Dong, X.L., Srivastava, D.: Characterizing and selecting fresh data sources. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14*, pp. 919–930. ACM, New York, USA (2014). DOI 10.1145/2588555.2610504
117. Reynolds, P.: DHS Data Framework DHS/ALL/PIA-046(a). Tech. rep., US Department of Homeland Security (2014)
118. Rieh, S.Y., Collins-Thompson, K., Hansen, P., Lee, H.: Towards searching as a learning process: A review of current perspectives and future directions. *J. Information Science* **42**(1), 19–34 (2016). DOI 10.1177/0165551515615841
119. Ritze, D., Lehmberg, O., Bizer, C.: Matching HTML tables to dbpedia. In: *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, WIMS 2015*, Larnaca, Cyprus, July 13–15, 2015, pp. 10:1–10:6 (2015). DOI 10.1145/2797115.2797118
120. Roh, Y., Heo, G., Whang, S.E.: A survey on data collection for machine learning: a big data - AI integration perspective. *CoRR abs/1811.03402* (2018)
121. Saleem, M., Ngomo, A.N.: Hibiscus: Hypergraph-based source selection for SPARQL endpoint federation. In: *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Crete, Greece, May 25–29, 2014. Proceedings*, pp. 176–191 (2014). DOI 10.1007/978-3-319-07443-6\_13
122. Sansone, S.A., Gonzalez-Beltrn, A., Rocca-Serra, P., Alter, G., Grethe, J., Xu, H., Fore, I., Lyle, J., E. Gururaj, A., Chen, X., Kim, H., Zong, N., Li, Y., Liu, R., Burak Ozyurt, I., Ohno-Machado, L.: Dats, the data tag suite to enable discoverability of datasets. *Scientific Data* **4** (2017). DOI 10.1038/sdata.2017.59
123. SDMX: Sdmx glossary. Tech. rep., SDMX Statistical Working Group (2018)
124. Stonebraker, M., Ilyas, I.F.: Data integration: The current status and the way forward. *IEEE Data Eng. Bull.* **41**(2), 3–9 (2018)
125. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. *CoRR abs/1312.6199* (2013)
126. Tang, Y., Wang, H., Zhang, S., Zhang, H., Shi, R.: Efficient web-based data imputation with graph model. In: *International Conference on Database Systems for Advanced Applications*, pp. 213–226. Springer (2017)
127. Tennison, J.: CSV on the web: A primer. W3C note, W3C (2016). [Http://www.w3.org/TR/2016/NOTE-tabular-data-primer-20160225/](http://www.w3.org/TR/2016/NOTE-tabular-data-primer-20160225/)
128. Thelwall, M., Kousha, K.: Figshare: a universal repository for academic resource sharing? *Online Information Review* **40**(3), 333–346 (2016). DOI 10.1108/OIR-06-2015-0190
129. Thomas, P., Omari, R.M., Rowlands, T.: Towards searching amongst tables. In: *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015*, Parramatta, NSW, Australia, December 8–9, 2015, pp. 8:1–8:4 (2015). DOI 10.1145/2838931.2838941
130. Townsend, A.: Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia. W.W. Norton & Company, Inc. (2013)
131. Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment and evolution of open data portals. In: *2015 3rd International Conference on Future Internet of Things and Cloud*, pp. 404–411 (2015). DOI 10.1109/FiCloud.2015.82
132. Van Gysel, C., de Rijke, M., Kanoulas, E.: Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems* **36**(4), Article 38 (2018)
133. Vidal, M.E., Castillo, S., Acosta, M., Montoya, G., Palma, G.: On the Selection of SPARQL Endpoints to Efficiently Execute Federated SPARQL Queries. In: A. Hameurlain, J. Kng, R. Wagner (eds.) *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXV, Lecture Notes in Computer Science*, pp. 109–149. Springer Berlin Heidelberg, Berlin, Heidelberg (2016)
134. W3C: The rdf data cube vocabulary (2014). URL <https://www.w3.org/TR/vocab-data-cube/t>
135. Weerkamp, W., Berendsen, R., Kovachev, B., Meij, E., Balog, K., de Rijke, M.: People searching for people: analysis of a people search engine log. In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, Beijing, China, July 25–29, 2011, pp. 45–54 (2011). DOI 10.1145/2009916.2009927
136. Wen, Y., Zhu, X., Roy, S., Yang, J.: Interactive summarization and exploration of top aggregate query answers. *Proceedings of the VLDB Endowment* **11**(13), 2196–2208 (2018). DOI 10.14778/3275366.3275369
137. White, R.W., Bailey, P., Chen, L.: Predicting user interests from contextual information. In: *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, Boston, MA, USA, July 19–23, 2009, pp. 363–370 (2009). DOI 10.1145/1571941.1572005

138. Wiggins, A., Young, A., Kenney, M.A.: Exploring visual representations to support datafire-use for interdisciplinary science. Association for Information Science & Technology (2018)
139. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). DOI 10.1038/sdata.2016.18
140. Woodall, P., Wainman, A.: Data quality in analytics: Key problems arising from the repurposing of manufacturing data. In: Proceedings of the International Conference on Information Quality (2015)
141. Wu, Y., Alawini, A., Davidson, S.B., Silvello, G.: Data citation: Giving credit where credit is due. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pp. 99–114. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3196910
142. Wylot, M., Cudré-Mauroux, P., Hauswirth, M., Groth, P.T.: Storing, tracking, and querying provenance in linked data. *IEEE Trans. Knowl. Data Eng.* **29**(8), 1751–1764 (2017). DOI 10.1109/TKDE.2017.2690299
143. Wylot, M., Hauswirth, M., Cudr-Mauroux, P., Sakr, S.: RDF Data Storage and Query Processing Schemes: A Survey. *ACM Comput. Surv.* **51**(4), 84:1–84:36 (2018)
144. Xiao, D., Bashllari, A., Menard, T., Eltabakh, M.: Even metadata is getting big: Annotation summarization using insightnotes. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15, pp. 1409–1414. ACM, New York, NY, USA (2015). DOI 10.1145/2723372.2735355
145. Yakout, M., Ganjam, K., Chakrabarti, K., Chaudhuri, S.: Infogather: Entity augmentation and attribute discovery by holistic matching with web tables. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12, pp. 97–108. ACM, New York, NY, USA (2012). DOI 10.1145/2213836.2213848
146. Yan, C., He, Y.: Synthesizing type-detection logic for rich semantic data types using open-source code. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pp. 35–50. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3196888
147. Yu, P.S., Li, X., Liu, B.: Adding the temporal dimension to search - A case study in publication search. In: 2005 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2005), 19–22 September 2005, Compiegne, France, pp. 543–549 (2005). DOI 10.1109/WI.2005.21. URL <https://doi.org/10.1109/WI.2005.21>
148. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A Survey. *Semantic Web* **7**(1), 63–93 (2016)
149. Zhang, S.: Smarttable: Equipping spreadsheets with intelligent assistancefunctionalities. In: The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval, SIGIR '18, pp. 1447–1447. ACM, New York, NY, USA (2018). DOI 10.1145/3209978.3210219
150. Zhang, S., Balog, K.: Entitables: Smart assistance for entity-focused tables. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017, pp. 255–264 (2017). DOI 10.1145/3077136.3080796
151. Zhang, S., Balog, K.: Ad hoc table retrieval using semantic similarity. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23–27, 2018, pp. 1553–1562 (2018). DOI 10.1145/3178876.3186067
152. Zhang, S., Balog, K.: On-the-fly table generation. In: The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval, SIGIR '18, pp. 595–604. ACM, New York, NY, USA (2018). DOI 10.1145/3209978.3209988
153. Zhang, X., Wang, J., Yin, J.: Sapprox: Enabling efficient and accurate approximations on sub-datasets with distribution-aware online sampling. *Proceedings of the VLDB Endowment* **10**(3), 109–120 (2016). DOI 10.14778/3021924.3021928