Report on the DATA:SEARCH'18 workshop – Searching Data on the Web

Laura Koesten
The Open Data Institute/
University of Southampton, UK
laura.koesten@gmail.com

Philipp Mayr
GESIS – Leibniz Institute
for the Social Sciences, DE
philipp.mayr@gesis.org

Paul Groth
Elsevier Labs, NL
p.groth@elsevier.com

Elena Simperl
University of Southampton, UK
e.simperl@soton.ac.uk

Maarten de Rijke
University of Amsterdam, NL
derijke@uva.nl

Abstract

The increasing availability of structured data on the web makes searching for data an important and timely topic. This report presents the motivation, output, and research challenges of the second DATA:SEARCH workshop which was held in conjunction with SIGIR 2018, in Ann Arbor, Michigan. This workshop explored challenges in data search, with a particular focus on data on the web. The aim was to share and establish links between different perspectives on search and discovery for different kinds of structured data, which can potentially inform the design of a wide range of information retrieval technologies. The DATA:SEARCH workshop tries to bring together communities interested in making the web of data more discoverable, easier to search and more user friendly.

1 Motivation

As an increasing amount of data becomes available on the web, searching for it becomes an increasingly important, timely topic (Gregory et al., 2018). The web hosts a whole range of new data species, published in structured and semi-structured formats - from web markup using schema.org and web tables to open government data portals, knowledge bases such as Wikidata and scientific data repositories (Cattaneo et al., 2015; Lehmberg et al., 2016). This data fuels many novel applications, for example fact checkers and question answering systems, and enables advances in machine learning and AI.

Data search and discovery has emerged in a range of complementary disciplines. Just like any other resource on the web, data benefits from network effects - it becomes more useful, and creates more value, when it is discoverable. And yet, despite advances in information retrieval, the Semantic Web and data management, data search is by far not as advanced, both technologically (Cafarella et al., 2011) and from a user experience point of view (Koesten et al., 2017), as related areas such as document search. Recently, Google has introduced dataset search in beta, an initiative to use the schema.org markup language to index datasets[1] and make them discoverable[2]. In Table 1, we present a subjective overview of typical information retrieval aspects by emphasizing the differences between "classical" document retrieval and dataset retrieval.

Table 1: Overview of the current situation in "classic" document retrieval and dataset retrieval.

Aspects	**Document retrieval**	**Dataset retrieval**
Availability of corpora	high	medium
Reproducibility	medium	low
Accessibility	medium	low
Available Retrieval Systems	high	medium
Ranking features/models	high	low
Research on interfaces (e.g. recommendation)	high	low
User studies	high	low

Most approaches to user-centric data search are domain-specific or have been created with certain task contexts, data schemas or data formats in mind (Dai et al., 2017). Conducting research to explore dataset search outside these constraints is an important and timely topic for a venue such as SIGIR. The aim of this workshop was to be a venue to present and exchange ideas and experiences for discovering and searching all types of structured or semi-structured datasets and to discuss how concepts and lessons learned from academic search, entity search, digital libraries, and web search could be transferred to data search scenarios.

The opportunities to share and establish links between different perspectives on search and discovery for different kinds of data are significant and can inform the design of a wide range of information retrieval technologies, including search engines, recommender systems and conversational agents.

Dataset search might be construed as just another type of entity search, like expert finding (Balog et al., 2012) or product search (Van Gysel et al., 2016). However, Thomas et al. (2015) show that dataset repositories present relative poor search results over and inside tables. It is difficult for a user to tell from a repository's portal whether a useful dataset is available, and this problem is only likely to get worse. Thomas et al. demonstrate that the naïve approach of full-text search is not necessarily appropriate. They describe an alternative, based on inferring types of data and indexing columns as a unit, and demonstrate some early improvements, especially when long captions are not available. New retrieval models are needed, models, moreover, that can be optimized with limited training and/or interaction data (Dai et al., 2017; Carevic et al., 2018). In this workshop we are interested in approaches to analyze, characterize and discover data sources and the aim was to facilitate a continuing discussion around data search across formats and domain-specific applications.

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

[2] https://toolbox.google.com/datasetsearch

2 Introduction

This workshop at SIGIR 2018 includes looking at the specifics of data-centric information seeking behavior, understanding interaction challenges in data search on the web, and analyzing the cognitive processes involved in the consumption of structured data by users. At the same time, we aimed to discuss architectures and technologies for data search - including semantics and information retrieval for structured and semi-structured data (e.g., ranking algorithms and indexing), in particular in the context of decentralized and distributed systems such as the web.

The workshop was kicked off by Laura Koesten with a general overview of the motivation, a short overview of the previous Profiles and DATA:SEARCH workshop at the Web Conference 2018 and open research challenges. The previous, and first DATA:SEARCH workshop was held at the Web Conference, 2018 together with the Profiles workshop, which focuses on dataset profiling and federated search for linked data. The first edition of the workshop featured a panel discussion on the topic ”Do we need a Google for data, and how would it look like?”. (Panelist were: Paul Groth, Aidan Hogan, Jeni Tennison, Stefan Dietze and Natasha Noy.) Some of the emerging themes in the panel discussion were the importance of quality metadata and the challenge of getting to quality metadata, everywhere but especially in a web search context. The panelist further discussed the fact that if we compare data search to traditional document search it can still be seen in it's infancy. Google search for documents has been trained for years and we do not yet have similar feedback loops for data search. The existing functionalities in search for data currently influences publishing practices and search strategies (e.g. Koesten et al. (2017)). This influences our ability to improve data search, as we can't rely on logs to make us understand how people would search for data if they were not restricted by current systems, functionalities and result presentation which are mostly tailored towards textual sources. In the introduction participants were urged to think about the potential units of interest within data search (data points, datasets, data packages), which is explained in more detail in the discussion section. They were further asked to think about whether and how traditional IR approaches can be applied to data search, or whether we need new and more suitable retrieval models specifically for structured data.

3 Workshop program

3.1 Keynote

Krisztian Balog gave a keynote titled **Table Retrieval and Generation** in which he described tables as complex information objects, which contain and summarize existing information in a structured form - which he argues can for some information needs be the desired unit of retrieval. He discussed three studies conducted within that space. One described work published at the Web Conference this year about the problem of *ad hoc table retrieval* in which a ranked list of tables is returned for a keyword query (Zhang and Balog, 2018a). Secondly he presented another variant of this task, referred to as *query-by-table*, the input is not a keyword query, but an incomplete table. Tables can be ranked much like documents, by considering the words contained in them. Their main research objective is to move beyond lexical matching and improve table retrieval performance by incorporating semantic matching. They achieved that by representing tables and queries in multiple semantic spaces (employing both discrete sparse and continuous dense vector representations). Thirdly, Balog introduced a method for answering keyword queries with tables that are generated "on the fly." In this case, results tables are not available as retrievable units, but are assembed dynamically by first identifying the entities and their attributes that should be included in the table, and then finding the values of those attributes. They use a table corpus and a knowledge base as data sources (Zhang and Balog, 2018b).

3.2 Presentation + Lightning talks

The paper **Recognizing Quantity Names for Tabular Data** (Yi et al., 2018) was presented by Yang Yi and describes how common units of measurements (quantity names) in numerical columns in CSV files can be abstracted to identify relevant units based on features extracted from the column. They identified five common categories (length, time, weight, percent and currencies), from which percent was the most common in their datasets extracted from data.gov. She described how they assigned each column to a class label corresponding to a quantity name and so treat the problem as a multi-class classification task. They describe features based on column name and content and show how these are used to predict quantity names for columns in tables.

Lightning talks

Discussing data search queries - Emilia Kacprzak gave an overview of results from her PhD work on dataset search. She presented results from initial studies they conducted with queries and data requests collected from governmental open data portals. She showed the difference in length between issued queries and queries generated by crowd workers based on data requests. She also highlighted the directions emerging from their studies, focusing on temporal and geospatial information. She concluded that dataset search needs indexing and ranking practices tailored to this source.

Searching beyond datasets in the Social Sciences - Philipp Mayr discussed the state of the art in data set retrieval at GESIS. The GESIS Search system[3] consists of curated social science datasets (mainly surveys and longitudinal data) and an linking infrastructure which connects datasets with publications and other materials. Data is retrievable via a Elasticsearch Index. He described a next release of the system that will connect more fine-grained parts of the datasets like survey questions and variables (e.g. to reuse/refind certain questions from a survey which has been used in an other study). The lightning talk ended with the statement that Google-like searching for dataset is just a starting point and much more advanced retrieval facilities and interfaces are needed.

Searching for datasets - Brian Davison discussed the need for new interfaces specifically for datasearch and pointed to variety of web search modalities, suggesting that similar modalities may be needed for dataset search as well. The possibility of being able to query by column name, adding context to the query and a richer description of the dataset returned by the engine before download was discussed.

[3] https://search.gesis.org/

Scientific table search using keyword queries - Jamie Callan presented work on table retrieval for scientific publications in which each table was presented as an XML document (paper title, paper abstract, table caption, referring sentences, footnotes, row header, column header, cell values) which can then be queried using standard IR techniques. The tables are described using context from the scientific publication and not just by it's content (Gao and Callan, 2017).

4 Discussion and Research Challenges

There was a lively discussion in the second half of the workshop which covered the following topics:

Units of interest:
Retrieving tables versus retrieving knowledge from a set of tables, or from within a table
One of the points of discussion was the potential units of interest within data search (outside of bespoke databases). This can range from (1) searching within tables, (2) searching for whole tables or spreadsheets (datasets) or (3) searching for whole dataset packages. Searching within tables (1) results from an information need for one or more particular data points - e.g. as an answer to a question. Searching for datasets (2) and dataset packages (3) presents challenges in terms of dataset summarization, quality metadata, recommender systems. Retrieval models for each of these could be different. For each of these 3 units of interest we can discuss retrieval on the web, for a specific domain (e.g. scientific table retrieval has received more attention) or withing closed systems (e.g. on a data portal). Currently it seems that people define this space on an ad hoc basis when discussing one of these units of interest which suggests a need for a better defined definition of the potential problem areas within this space.

Traditional Information Retrieval versus new approaches The discussion highlighted the fluid boundary between traditional document retrieval and data search, however there was a consensus that there is a large space for future research to understand how existing approaches can be tailored to data search, as well as in developing new models specifically for data search. Standard text retrieval models can be used for structured data. One of the lightning talks described how they modeled a table as a multi field document which essentially represents the table as text then treated it as standard structured document retrieval problem. However, also other IR methods which can take advantage of the structure in the data can be applicable.

Querying for data and interfaces We also spoke about keyword queries similar to web search and more complex querying methods. The role of faceted search versus keywords versus completely new approaches (querying with a table). The trade-off between complex search interfaces taking advantage of the structure of the data to simple keyword based search boxes that we are used to - which might be the goal for data search as well. Result presentation for data was briefly discussed, emphasizing the importance of presenting more of the content of the dataset in a search scenario and developing a better understanding of selection criteria in dataset search.

Tasks and users Another topic was the tasks that people do with data and that in reality there is not a lot research exploring people's information needs with data, especially when it comes to general web search. It is yet to be seen whether traditional information seeking models are directly applicable to searching for structured data on the web. Tasks and connected information needs have shown to be different in exploratory studies as mentioned in the motivation section. From a user perspective we might not care whether an information need is satisfied with structured data or with textual documents - so it might be worth thinking about approaches which do not force users to actively make this kind of differentiation.

Venue and direction of the workshop The question if whether SIGIR is the right venue for this emerging topic was discussed and the participants strongly agreed that, although some of the challenges in data search are focused on human interaction with data and with a system, that SIGIR is the appropriate venue for the DATA:SEARCH workshop.

Research challenges
A broad range of methods and insights are important to enable the discovery of, and access to, data published on the web, including:

- analyzing contextual information for datasets, including mentions of datasets
- browsing and query support for structured and semi-structured data
- inference and data enrichment systems
- learning to match for datasets
- learning to rank datasets
- mining direct links between documents, datasets or data records
- summaries and descriptions of datasets targeting users or search engines
- concepts and methods to present data and entity-centric results.

Workshop proceedings can be downloaded under: `http://ceur-ws.org/Vol-2127/`

5 Conclusion + future directions

There was a clear interest in continuing the data search workshop as an event with several pointers to recent activity in this space. PhD topics are being started on the topic, grant proposals submitted and there is a clear scope for a lot of research in the identified topics.

One of the identified gaps was the availability of easy to use datasets to use for experiments. This led to the plan of providing a dataset for a common challenge for the next DATA:SEARCH workshop. There is a mailing list for the DATA:SEARCH workshop which you can join by sending an email to: laura.koesten@gmail.com.

6 Acknowledgements

We thank all the PC members for their reviews. A list of their names can be found on the workshop's website (https://datasearch-ws.github.io/2018/). We thank all presenters and participants for their contribution.

References

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise retrieval. *Foundations and Trends in Information Retrieval* 6, 2–3 (August 2012), 127–256.

Michael J. Cafarella, Alon Halevy, and Jayant Madhavan. 2011. Structured Data on the Web. *Commun. ACM* 54, 2 (Feb. 2011), 72–79. DOI:http://dx.doi.org/10.1145/1897816.1897839

Zeljko Carevic, Sascha Schüller, Philipp Mayr, and Norbert Fuhr. 2018. Contextualised Browsing in a Digital Library's Living Lab. In *Proceedings of JCDL 2018*.

Gabriella Cattaneo, Mike Glennon, Rosanna Lifonti, Giorgio Micheletti, Alys Woodward, Marianne Kolding, Angela Vacca, Carla La Croce, and David Osimo. 2015. European Data Market SMART 2013/0063, D6 - First Interim Report. (October 2015).

Zhuyun Dai, Yubin Kim, and Jamie Callan. 2017. Learning To Rank Resources. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 837–840.

Kyle Yingkai Gao and Jamie Callan. 2017. Scientific Table Search Using Keyword Queries. *CoRR* abs/1707.03423 (2017).

Kathleen Gregory, Helena Cousijn, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2018. Understanding Data Retrieval Practices: A Social Informatics Perspective. *arXiv preprint arXiv:1801.04971* (2018).

Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. 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 (CHI '17)*. ACM, New York, NY, USA, 1277–1289.

Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*. 75–76.

Paul Thomas, Rollin M. Omari, and Tom Rowlands. 2015. Towards Searching Amongst Tables. In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*. 8:1–8:4. DOI:http://dx.doi.org/10.1145/2838931.2838941

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2016. Learning latent vector spaces for product search. In *CIKM 2016: 25th ACM Conference on Information and Knowledge Management*. ACM, 165–174.

Yang Yi, Zhiyu Chen, Jeff Heflin, and Brian D. Davison. 2018. Recognizing Quantity Names for Tabular Data. In *Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop* *on Data Search (DATA:SEARCH'18) Co-located with (ACM SIGIR 2018), Ann Arbor, Michigan, USA, July 12, 2018.* 68–73.

Shuo Zhang and Krisztian Balog. 2018a. 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.* 1553–1562.

Shuo Zhang and Krisztian Balog. 2018b. On-the-fly Table Generation. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 595–604. `DOI:http://dx.doi.org/10.1145/3209978.3209988`