DATA:SEARCH'18 – Searching Data on the Web

Paul Groth
Elsevier Labs
Amsterdam, The Netherlands
p.groth@elsevier.com

Laura Koesten
The Open Data Institute; University
of Southampton
London, United Kingdom
laura.koesten@theodi.org

Philipp Mayr
GESIS Leibniz-Institute for the Social
Sciences
Cologne, Germany
philipp.mayr@gesis.org

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

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

ABSTRACT

This half day workshop explores challenges in data search, with a particular focus on data on the web. We want to stimulate an inter-disciplinary discussion around how to improve the description, discovery, ranking and presentation of structured and semi-structured data, across data formats and domain applications. We welcome contributions describing algorithms and systems, as well as frameworks and studies in human data interaction. The workshop aims to bring together communities interested in making the web of data more discoverable, easier to search and more user friendly.

KEYWORDS

Data search, dataset retrieval, human data interaction, web of data

1 ORGANISING COMMITTEE

- Paul Groth, Elsevier Labs, The Netherlands
- Laura Koesten, The Open Data Institute, United Kingdom
- Philipp Mayr, GESIS Leibniz-Institute for the Social Sciences, Germany
- Maarten de Rijke, University of Amsterdam, The Netherlands
- Elena Simperl, University of Southampton, United Kingdom

2 MOTIVATION AND RELEVANCE

2.1 Background

As more and more data becomes available on the web, searching for it becomes an increasingly important, timely topic [6]. 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 [3, 12]. This data fuels many novel applications, for example fact checkers and question answering systems, and enables advances in machine learning and AI.

Just like any other resources 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 [2] and from a user experience point of view s [10], as related areas such as document search.

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 [4]. Conducting research to explore dataset search outside these constraints is both important and timely for a venue such as SIGIR. The aim of the workshop is 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. We will seek contributions and encourage interactions to discuss how principles, techniques and experiences could be applied across research fields that so far have mostly pursued related data search questions in isolation.

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.

We see a large space for discussion and future research in the development of federated data discovery and search technologies, which leverages the most recent advances in information retrieval, Semantic Web and databases, and is mindful of human factors.

2.2 Recent developments

Dataset search and discovery has emerged in a range of complementary disciplines. Kunze and Auer [11] introduce dataset retrieval

as a specialization of information retrieval, however, they restrict their scope to the process of returning relevant RDF datasets.

Several architectures have been proposed to support discoverability of web data. The Linked Data community has put forward a set of principles and technologies to publish data in a form that makes it easy for applications to find and reuse it. Publishers are encouraged to define links between datasets, which can be used, alongside de-referencing URIs, to find additional data. The Linked Open Data Cloud, as well as large knowledge graphs such as Wikidata and DBpedia are prime examples of this approach.

A related development are data portals, for example in open government and some scientific domains, and data sharing networks, such as Kaggle and data.world. Data portals are centralized repositories where an institution, or multiple institutions, provide access to their datasets. They classify the datasets according to pre-defined categories, support basic keyword and faceted search features, and present dataset results via short descriptions, metadata and sometimes visualizations. In a search log analysis of open data portals, Kacprzak et al. [7] found that queries issued on data portals differ from those issued to web search engines in their length and structure. Data sharing networks use a social network paradigm to help people discover new datasets and engage with data publishers and users.

Dataset search might be construed as just another type of entity search, like expert finding [1] or product search [14]. However, Thomas et al. [13] show that dataset repositories have poor search 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 appropriate. They describe an alternative, based on inferring types of data and indexing columns as a unit, and demonstrate some improvements in early success 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 [4].

Data requires context to create meaning [5], to *make sense of it*. This is dependent on people's data literacy, technical skills and prior knowledge. Kelly [8] also shows that individuals vary significantly in terms of cognitive makeup, prior knowledge and behavioral dispositions. While this applies to search for all information sources generally, literature suggests unique characteristics when the information source is structured data. In user studies with social scientists, Kern and Mathiak [9] found that the quantity and quality of metadata are far more critical in dataset search than in literature search, where convenience is most important. For empirical social scientists, the choice of research data was found to be more relevant than the choice of literature; therefore they were willing to put more effort into the retrieval process. In a mixed methods study describing the information seeking process for structured data, Koesten et al. [10] combined in-depth interviews with data professionals and a search log analysis of a large open governmental data portal. They note that finding data is challenging, even for data professionals who are familiar with state of the art tools, and that data search is often exploratory and complex. Evaluation criteria for datasets in a search scenario showed unique characteristics – the importance of context alongside information about provenance and methods for collection and analysis emerged as key factors,

which help professionals determine whether a dataset is relevant and useful for their purposes [10].

3 THEME AND PURPOSE

The objective of this workshop is to bring together researchers and practitioners interested in advancing data search on the web. This 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 aim 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. We are interested in approaches to analyze, characterize and discover data sources. We want to facilitate a discussion around data search across formats and domain-specific applications.

We envision the workshop as a forum for researchers and practitioners from various disciplines to come together and discuss common challenges and identify synergies for joint initiatives.

3.1 Topics

DATA:SEARCH'18 will seek application-oriented papers, as well as more theoretical papers, position papers and empirical studies.

The workshop proposes a multidisciplinary discussion on the following themes, with a focus on search and discovery of RDF, CSV, JSON and other structured and semi-structured data sources:

- Analyzing behavioral traces during data search
- Approaches to personalization and contextualization in dataset search
- Data summarization
- Dataset representation for retrieval (standards, models, workarounds)
- Decentralized and distributed architectures and algorithms in data search
- Deep linking of datasets
- Entity recognition in datasets
- Evaluation of dataset search tools and algorithms
- Fusing, cleaning, ranking and refining dataset search results
- Information seeking behavior for data (interactive data retrieval)
- Data indexing and profiling approaches
- Learning to rank for data search
- Query routing taking into account relevance, quality and profiles of distributed datasets
- Retrieval models for data search
- Scalability and performance of distributed data queries
- Search results presentation for datasets
- Visual and speech interfaces to datasets
- Semantic dataset search
- Usability of data portals and data discovery tools
- User modeling for data search
- Systems and user studies in data search in vertical domains, including transport, geospatial data, science, weather etc.

We are interested in contributions using a variety of methods. This can include, for example, user studies, lab experiments, system-based evaluations, but also experiments using gamification and crowdsourcing.

4 WORKSHOP FORMAT

- Proposed duration: Half day workshop

Following an introduction by the organizers, the workshop will open with a keynote from a distinguished academic or practitioner who will give participants an overview of recent developments around data search. During the workshop, we will have brief presentations of the papers submitted to the workshop as well as a working session. Paper presentations will follow a lightning talk format where most of the time will be dedicated to questions and comments from the audience. For the working sessions the audience will be invited to join round tables on different topics, defined based on accepted submissions as well as on statements of interest by authors of accepted papers. Topics will be published on the website on the day of the workshop, and the audience will be invited to join one of the round tables. The workshop will wrap up with a summary of discussions from the round tables, collecting calls to action for future activities and a short reflection on the workshop.

We encourage full papers, short papers as well as position papers. A maximum number of 8 pages in ACM format is considered appropriate for the aims of the workshop. We are expecting to accept up to 12 papers. A current version of the workshop homepage can be seen here: https://profiles-datasearch.github.io/2018

5 TENTATIVE SCHEDULE

The tentative schedule looks as follows:

09:00–09:15 Introduction by organizers

09:15–09:45 Keynote

The workshop will be open with a keynote from a distinguished practitioner who will give participants an overview of recent developments around data search. We are considering speakers such as:
- Abraham Bernstein (University of Zürich)
- Andreas Rauber (Vienna University of Technology)
- Edgar Meij (Bloomberg)
- Jamie Callan (CMU)

09:45–10:30 Lightning presentations of contributed papers

10:30–11:00 Coffee break

11:00–12:30 Round tables

12:30–13:00 Summary of discussions from the round tables, call to action for future activities, wrap up of the workshop

6 SELECTION PROCESS

We will follow a single-blind process with at least two reviewers per paper. Papers will be evaluated according to their significance, originality, technical content, style, clarity, relevance to the workshop, and likelihood of generating discussion. Workshop participation will be open to all interested registered parties.

7 ORGANIZERS

Datasets as information objects are situated at the intersection of several disciplines – information retrieval, semantic web, user interaction, and library science. Discovery of, and access to datasets, is an emerging shared interest of academic and industrial researchers. The team behind this workshop proposal represents all of these angles and interests, both in the organizers and in the proposed program committee.

7.1 Co-chairs

Dr Paul Groth is Disruptive Technology Director at Elsevier Labs. He has done research at the University of Southern California and the Vrije Universiteit Amsterdam. His research focuses on intelligent systems for dealing with large amounts of diverse contextualized knowledge with a particular focus on web and science applications. This includes research in data provenance, data science, data integration and knowledge sharing. Paul is co-author of "Provenance: an Introduction to PROV" as well as over a 100 peer-reviewed publications. He has chaired multiple international events including Beyond the PDF 2, The International Semantic Web Conference, and the International Provenance and Annotation Workshop.
More info: http://pgroth.com. Elsevier Labs; email: p.groth@elsevier.com

Laura Koesten is a Maria Curie Skłodowska fellow, doing her PhD at the Open Data Institute and at the University of Southampton in the UK. She is part of WDAqua, a European Union's Horizon 2020 initiative to advance state of the art Question Answering. Her research interests are Human Computer Interaction, Interactive Information Retrieval with a focus on dataset retrieval, Open Data and Semantic Interfaces. In her PhD she is looking at ways to improve Human Data Interaction in IIR systems. She publishes at CHI and has a background in Human Factors, with an MSc degree from Loughborough University.
The Open Data Institute; University of Southampton, UK; Email: laura.koesten@theodi.org

Dr Philipp Mayr is a deputy department head and a team leader at the GESIS - Leibniz-Institute for the Social Sciences department Knowledge Technologies for the Social Sciences (WTS). Philipp Mayr received his PhD in applied informetrics and information retrieval from the Berlin School of Library and Information Science at Humboldt University Berlin in 2009. To date, he has been awarded substantial research funding (PI, Co-PI) from national and European funding agencies. Philipp Mayr has published in top conferences and prestigious journals in the areas informetrics, information retrieval and digital libraries. His research interests include: interactive information retrieval, scholarly recommendation systems, data retrieval, non-textual ranking, bibliometric and scientometric methods, applied informetrics, science models in digital libraries, knowledge representation, semantic technologies, user studies, information behavior.
More info: https://philippmayr.github.io/. GESIS - Leibniz-Institute for the Social Sciences, Email: philipp.mayr@gesis.org

Prof Maarten de Rijke is professor of Computer Science at the University of Amsterdam. He is a member of the Royal Dutch Academy of Arts and Sciences (KNAW), a former director of Amsterdam Data Science, a collaborative network involving 600+ data scientists in the Amsterdam area, and the founding director of ICAI, the Innovation Center for Artificial Intelligence. His research is situated at the interface of AI and IR, and focused on learning to rank, semantic search, and autonomous environments for information interaction. More info: https://staff.fnwi.uva.nl/m.derijke/. University of Amsterdam. Email: derijke@uva.nl

Prof Elena Simperl is professor of Computer Science at the University of Southampton. Her research interests include knowledge engineering, Social Web technologies, and crowdsourcing. She has contributed and led over 20 national and European research projects and authored more than 100 scientific publications, and chaired the European Semantic Web Conference (ESWC) in 2011 and 2012 and International Semantic Web Conference (ISWC) in 2016. She was vice-president of STI International until 2016 and the director of the ESWC summer school series. She has co-chaired more than 15 workshops, including the series on Theory and Practice of Social Machines (SOCM) at WWW, Crowdsourcing the Semantic Web at ISWC and Ontology Engineering in a Data Driven World at EKAW. More info: http://elenasimperl.eu/. University of Southampton, UK; Email: e.simperl@soton.ac.uk

7.2 Tentative Programme Committee

A preliminary list of PC members is as follows:

- Alexander Kotov (Wayne State University)
- Arjen de Vries (Radboud University Nijmegen)
- Arno Scharl (Modul University Vienna)
- Axel Polleres (Vienna University of Economics and Business)
- Chaomei Chen (Drexel University)
- ChengXiang Zhai (University of Illinois, Urbana-Champaign)
- Ed Hovy (CMU)
- Emanuel Sallinger (University of Oxford)
- Eva Méndez (Open research data)
- Jie Tang (Tsinghua University)
- Jürgen Umbrich (WU Vienna)
- Krisztian Balog (University of Stavanger)
- Kuansan Wang (Microsoft)
- Laura Dietz (University of New Hampshire)
- Liudmila Ostroumova Prokhorenkova (Yandex)
- Max Froumentin (GDS, UK Government Digital Service)
- Michael Gubanov (University of Texas, San Antonio)
- Natalya Noy (Google)
- Nicolas Torzec (Yahoo)
- Norman Paton (University of Manchester)
- Peter Haase (Metaphacts)
- Philipp Cimiano (Universität Bielefeld)
- Stefano Modafferi (University of Southampton)
- Steffen Lohmann (Fraunhofer IAIS)

8 RELATED WORKSHOPS

Some of the organizers will be involved in the delivery of a workshop on data profiling and search in April 2018 at the Web Conference.[1] That workshops is a follow-up of a workshop that has been run for many years at the European and International Semantic Web Conference, which focused on semantic techniques to enrich datasets to help with tasks such as discovery, description and sensemaking of entity-centric data on the web.[2]

While we see many synergies between DATA:SEARCH'18 and those events, the focus of DATA:SEARCH'18 is on search. We aim to gain a better understanding of the extent to which techniques, methods and lessons learned from document retrieval broadly construed could apply to data-centric contexts, and explore in more depth the differences between the two areas from a technical and interaction perspective.

REFERENCES

[1] 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.

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

[3] 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). https://idc-emea.app.box.com/s/k7xv0u3gl6xfvq1rl667xqmw69pzk790.

[4] 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.

[5] Brenda Dervin. 1997. Given a context by any other name: Methodological tools for taming the unruly beast. *Information seeking in context* 13 (1997), 38.

[6] 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).

[7] Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. *A Query Log Analysis of Dataset Search.* Springer International Publishing, Cham, 429–436. https://doi.org/10.1007/978-3-319-60131-1_29

[8] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

[9] Dagmar Kern and Brigitte Mathiak. 2015. Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?. In *Research and Advanced Technology for Digital Libraries*, Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla (Eds.). Springer International Publishing, Cham, 197–208.

[10] 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. https://doi.org/10.1145/3025453.3025838

[11] Sven R. Kunze and Soren Auer. 2013. Dataset Retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing.* https://doi.org/10.1109/ICSC.2013.12

[12] 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. https://doi.org/10.1145/2872518.2889386

[13] 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. https://doi.org/10.1145/2838931.2838941

[14] 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.

---

[1] https://profiles-datasearch.github.io/2018/
[2] http://data4urbanmobility.l3s.uni-hannover.de/index.php/en/2017/11/07/profiles-workshop-iswc-2017/