Searching Data Portals – More Complex Than We Thought?

Laura M Koesten
The Open Data Institute; Univ. of Southampton
UK
laura.koesten@theodi.org

Jaspreet Singh
L3S Research Center, Hannover
Germany
singh@l3s.de

ABSTRACT

The amount of data published openly on the web is increasing rapidly. Most people either use web search or specialised data portals, which are repositories of datasets, to search for data. Most data portals today use similar faceted search interfaces. In this paper we focus on how a large governmental data portal in the UK supports users in conducting complex search tasks involving data. Based on a previous interview study with users of the portal we constructed a typical complex work task. In this work, we analyzed how the current system supports users during this task and subsequently identify problems with the interface. Based on this we discuss potential research directions to improve interfaces for complex data related search tasks.

1 INTRODUCTION

We live in the age of data-driven decision making where we take action based on insights gathered from a collection of relevant datasets. A dataset in our scenario refers to structured information collected by an individual or organisation and distributed in a standard format, for instance CSV files containing bus timetables collected by the local administration. Today, more than a million datasets have been made available by governments worldwide [7, 10]. The Web Data Commons project extracted no less than 233 million web tables containing structured data from HTML pages in 2015 [6], earlier studies estimated the amount of structured data on the web to be over one billion sources in February 2011 [1].

With this increase in availability, searching for data is becoming more important. One of the primary ways to search for data on the web is through data portals, which are repositories of datasets. The European Data Portal [1] indexes, to date, 629, 476 datasets published by regional and national authorities in EU countries; the official US government data portal[2] covers 193, 976 and the UK portal [3] covers 37, 079 published datasets to date.

Data search presents many challenges, as ideas and tools from web search cannot (yet) be directly applied [9]. Using conventional web search engines is not ideal, as these have been designed primarily for documents, not data [1]. This has led to the creation of document surrogates for datasets which are indexed by search engines. These usually consist of a textual description and related metadata presented for human consumption.

[1]https://www.europeandataportal.eu/data/en/dataset
[2]http://www.data.gov
[3]https://data.gov.uk/data/search

CHIIR 2017 Workshop on Supporting Complex Search Tasks, Oslo, Norway.

Searching for Data is Complex

When done in a work context, the search for data is often complex. A previous interview study with data professionals across a wide range of domains and skill sets [5] suggests that, in the majority of cases, searching for data shows characteristics of an exploratory or complex search task. That involves multiple queries, iterations and refinement of the original information need, as well as complex cognitive processing.

Data professionals, who are the primary users of such portals – often engage in tasks which involve more than one dataset and a sequence of queries to fulfill their information need. For example, tasks requiring datasets often involve trying to understand changes in data over time; or collecting different sources to make informed decisions based on relationships between them.

There are several aspects that add to the complexity of search tasks for data. In contrast to document search, users need skills to access and download data; interpret different or limited formats the data might be available in; and understand connected licences and metadata. Furthermore, data requires context to create meaning [2], to *make sense of data*. In contrast to searching for digital objects, such as e.g. physical artifacts in a digital library, datasets contain information within them which can be used to contextualize them and so support a search process. We currently rely on metadata, which varies in quality and availability. However, we argue that utilising the original data to enrich metadata can provide relevant indexable content which would make data search more effective.

Decisions about the amount of context provided with the data are made by data publishers or by those designing data portals; interface design plays a key role in representing the context [4]. For example the UI of the UK governmental data portal [4], as shown in Figure 1, shows the format and the publishing organization for each dataset in the result list.

Motivating example From discussion with experts and users of the portal we created an exemplary task:

You work for the local council in the city of York (UK) and you have been given the task to decide the top 3 areas in which to advertise NHS (National Health Services) health checks. These are checks recommended above a certain age by the NHS in the UK. An area that should be prioritized would be one were many people are eligible, but haven't participated before.

We know from previous studies that people experience difficulties in finding datasets [5]. In this paper, we focus on such complex

[4]https://data.gov.uk/data/search

search tasks and illustrate how search user interfaces on current data portals support such tasks.

We highlight drawbacks of the current search interface, such as snippets, dataset previews, and missing links between datasets. Following that we give possible directions for improvement and further research.

Search results are displayed similar to web search, with a title and short snippet. Furthermore, metadata including the data publisher (e.g. Public Health England), topical category (e.g. Society) and format (e.g. CSV) are displayed. Clicking on a result takes the user to a page that contains the textual description and metadata. Some pages also include a dataset preview by displaying some portion of the raw data.

2 SEARCH USER INTERFACE

Many data portals on the web offer a similar faceted-search user interfaces. The search results are displayed using the ten blue links paradigm found in web search. To highlight the drawbacks of such interfaces, we select the UK governmental data portals' search interface (Figure 1) as a typical example [5]. The interface consists of a standard query bar and a series of facets to further filter results. Clicking on a search result takes the user to a preview page (Figure 2) that contains the textual description and metadata. Some pages also include a dataset preview by displaying a portion of the raw data. In this section we use the example task described in the introduction as a means to substantiate our claims.

A. Search Result Display . Search results are displayed similar to web search, with a title and short snippet. Further metadata is presented, including the data publisher (e.g. Public Health England), topical category (e.g. Society) and format (e.g. CSV). An individual search result's display should provide the user with sufficient information to judge the relevance, quality and usability of the results [5]. For our task we issued an initial query "NHS health check" which returned 1,233 results. The format, data publisher and the frequency of updates are displayed along with the title and first three lines of the description for each result. The system also provides a set of facets which aid the user in browsing the results. Due to the lack of a geographical facet however, we refined our query to "NHS health check York" and got 19 results. Subsequently, to judge relevance we found that there was no indication as to which part of the textual description of a dataset matched the query – title, description or metadata, as can be seen in Figure 1 in the third search result. This would help assess the relevance of a search result as it gives an indication of the context in which a given query term is used in the dataset. Little information about the granularity of the data is available, which makes it hard to judge whether the level of aggregation of the dataset is suitable for the task. Additionally it was not apparent what type of data could be found in the dataset - geographical, time series, demographics, etc. This requires the user to download and open each dataset individually to get more details. Extending the existing facets to such, more content oriented, facets would support complex search tasks.

B. The dataset preview page, can be accessed by clicking on a dataset. This page shows specific metadata as seen in (Figure 2), however it did not give us an indication of the content of the dataset. On this page, next to a download link the user can click on the button "details". This leads to a preview of the dataset which shows the headers and a sample of rows (for CSV files). This further gives an indication of relevance and quality as it is the only item on the page which exposes the actual data. While the preview is useful, it does not offer a comprehensive overview of the content of the dataset, nor support for interpretation of the content. We argue that an overview has potential to be more meaningful by exposing more information about the dataset to the user. An overview would e.g. give the range of values per column or the distinct locations mentioned in the data and give information about the structural profile of the dataset.

The additional metadata displayed on the preview page is not helpful in selecting datsets for our task. E.g. further categories of information are shown such as "last updated" and "date updated" - the difference is not clear and this information is also only available for some dataset packages.

C. Limited support of discovering links between datasets In complex search tasks users commonly need to link multiple datasets together. In our task we link NHS data with demographic data; for this we need to download the datasets and discover methods of making connections manually. The system provides no recommendations or links to similar or complimentary datatsets. Links also offer the possibility to understand how datasets are related to each other. A visualisation of links between datasets would reduce cognitive load of the user and aid discovery. For our task, datasets that contain aggregated information about health checks in the area York could link to each other, or data referring to specific locations could be linked to geospatial boundary data.

3 FUTURE PERSPECTIVES

We envision an interface which is tailored to data search. Such an interface would better support users in (i) evaluating the relevance, quality and usability of a dataset result (ii) getting a suitable overview of the dataset and (iii) finding related datasets.

For **search result display** we believe that presenting a brief overview of the actual content of the dataset would support users in assessing whether a result is relevant for their task. We found that many datasets have descriptions that are incomplete or short – this suggests space for future research to automatically generate better descriptions of datasets. We believe that novel methods to generate query-driven snippets both from the content and human generated description of a dataset would be highly beneficial. For instance, by displaying *headers, entities and/or summarising statistics* of a relevant column or field alongside each search result. We propose to capture these as additional metadata that can also be utilised for faceted browsing and indexed to improve ranking efficiency. To judge data quality, we propose visual or textual indicators on the interface, backed up by automatically computed metrics, user-generated reviews and annotations or reuse statistics.

In the **dataset preview page** interactive visualisations could be used, which allow users to choose their area of interest within a larger dataset as well as providing a comprehensive overview of the content. Filtering, sorting and exploring different views of the data on demand are recommended for such tasks. **The discovery and exploration of links** should be supported by interfaces by

[5] https://data.gov.uk/data/search

visualising connections between different datasets or data points, and possibly represent data within a network - to make a user understand its meaning within the context of other data. This could also be used to create recommendation systems for datasets based on reuse or on datasets which were downloaded together, as well as on content or structure of the dataset. Furthermore, we intend to experiment with search interface paradigms that go beyond ten blue links. We are planning to draw on techniques used in semantic web technologies such as e.g. Cluster Maps [3] to provide graph based visualizations that display connections between search results and other related datasets. [8] recommends 3 high level capabilities for data exploration tools to support sensemaking: visual and interactive data exploration, data enrichment through recommendation systems and data cleaning functionalities. Our suggestions can be seen as a step in that direction.

Acknowledgements This work is supported by the European Union Horizon 2020 program under the Marie Sklodowska-Curie grant agreement No. 642795.

REFERENCES

[1] Michael J Cafarella, Alon Halevy, and Jayant Madhavan. 2011. Structured data on the web. *Commun. ACM* 54, 2 (2011), 72–79.
[2] Brenda Dervin. 1997. Given a context by any other name: Methodological tools for taming the unruly beast. *Information seeking in context* 13 (1997), 38.
[3] Christiaan Fluit, Marta Sabou, and Frank Van Harmelen. 2006. Ontology-based information visualization: toward semantic web applications. In *Visualizing the semantic web*. Springer, 45–58.
[4] Saul Greenberg. 2001. Context as a dynamic construct. *Human-Computer Interaction* 16, 2 (2001), 257–268.
[5] Laura M Koesten, Emilia Kacprzak, Tennison Jenifer, 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.
[6] 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. DOI: http://dx.doi.org/10.1145/2872518.2889386
[7] McKinsey Global Institute: James Manyika, Michael Chui, Peter Groves, Diana Farrell, Steve Van Kuiken, and Elizabeth Almasi Doshi. 2013. Open data: Unlocking innovation and performance with liquid information. (2013).
[8] Kristi Morton, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. 2014. Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment* 7, 6 (2014), 453–456.
[9] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.
[10] Barbara Ubaldi. 2013. Open Government Data. (2013).

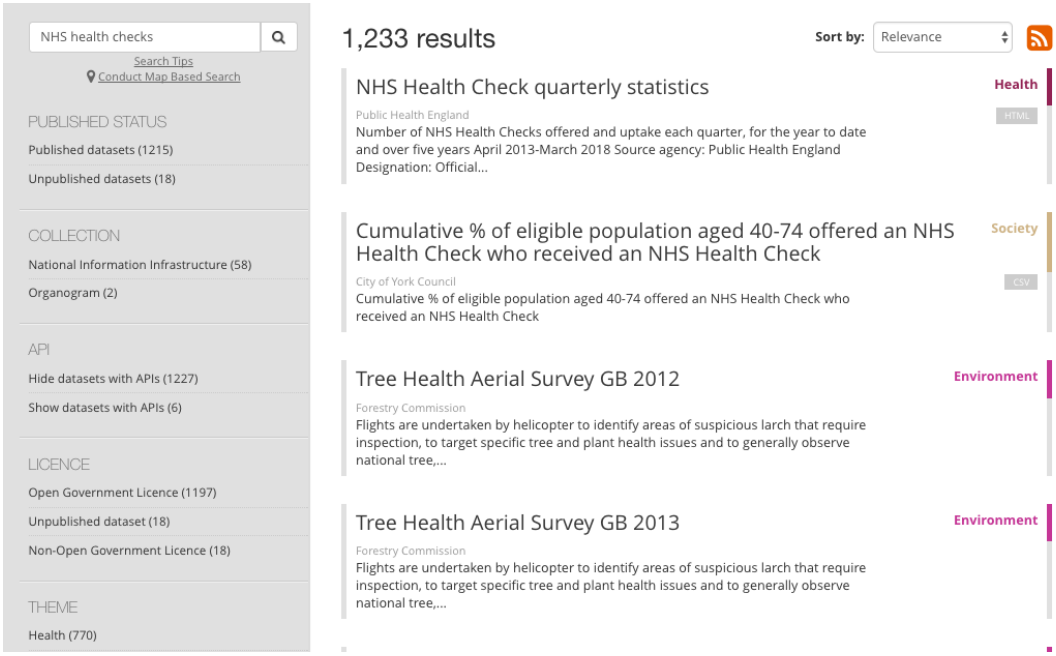

Figure 1: Search UI: The interface consists of a standard query bar and a series of facets to further filter results. Search results are displayed similar to web search, with a title and short snippet. Furthermore, metadata including the data publisher (e.g. Public Health England), topical category (e.g. Society) and format (e.g. CSV) are displayed. This interface is used at data.gov.uk/data/search, one of the largest European open data portals.

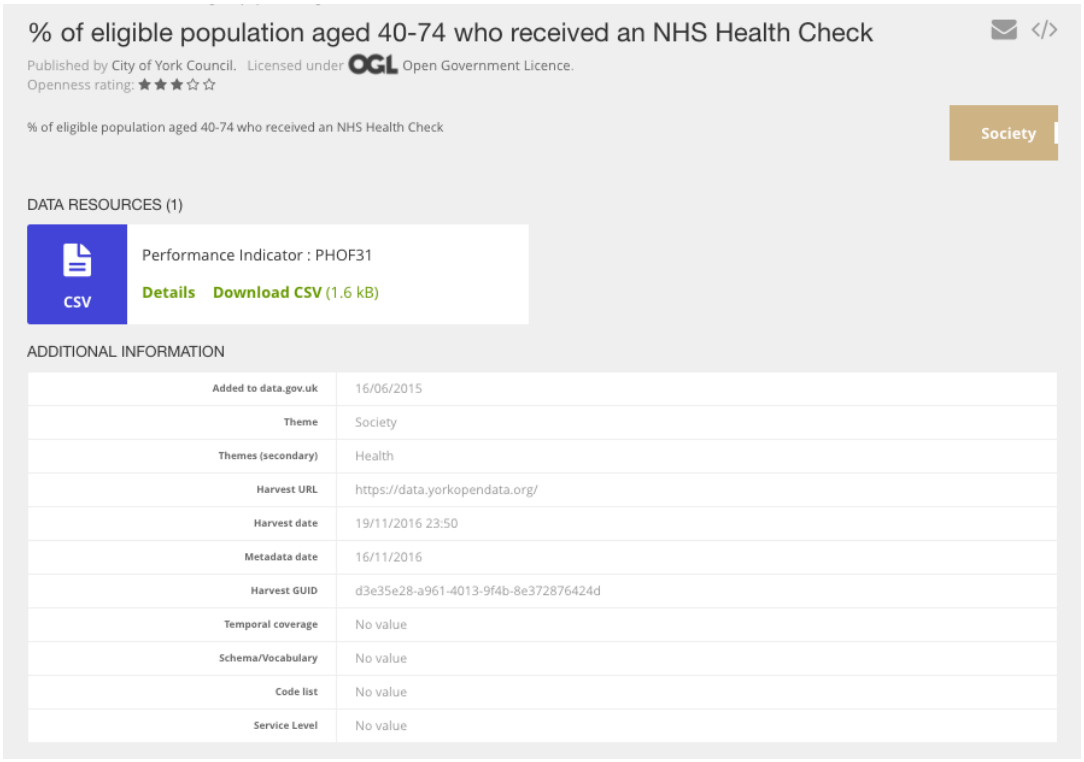

Figure 2: The dataset preview page can be accessed by clicking on a dataset. This page shows some metadata: format, publishing organisation and date, licence, and an openness rating, topic tags on the portal, the harvest URL and date.