UNIVERSITY OF SOUTHAMPTON

A User Centred Perspective on Structured Data Discovery

by

Laura Koesten

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Engineering, Science and Mathematics

Electronics and Computer Science

November 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Laura Koesten

Structured data is becoming critical in every domain and its availability on the web is increasing rapidly. Despite its abundance and variety of applications, we know very little about how people find data, understand it, and put it to use.

This work aims to inform the design of data discovery tools and technologies from a user centred perspective, aiming to better understand how we can support people in finding and selecting data that is useful for their tasks. We approached this by advancing our understanding of user behaviour in structured-data discovery through a mixed-methods study looking at the work flow of data practitioners when searching for data.

From that we present a framework for structured data interaction describing data-centric tasks, search strategies, as well as an in-depth characterisation of selection criteria in data search. We identified textual summaries as a main element that supports the decision making process in information seeking activities for data.

Based on these results we conducted a mixed-methods study to identify attributes that people consider important when describing a dataset. This enabled us to better define criteria for textual summaries of datasets for human consumption. We designed a set of template questions to help guide the summary writing process and conducted an online study to validate the applicability of dataset summaries in a dataset selection scenario.

The findings of this work revealed unique interaction characteristics in information seeking for structured data. Our contributions can inform the design of data discovery tools, support the assessment of datasets and help make the exploration of structured data easier for a wide range of users.

Contents

List of Figures

List of Tables

Research Thesis: Declaration of Authorship

Print name: Laura Koesten

Title of thesis: A User Centred Perspective on Structured Data Discovery

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

- The Trials and Tribulations of Working with Structured Data – a Study on Information Seeking Behaviour. Laura Koesten, Emilia Kacprzak, Jeni Tennison, Elena Simperl. Proceedings of ACM CHI Conference on Human Factors in Computing Systems, CHI 2017.
- Everything You Always Wanted to Know about a Dataset: Studies in Data Summarisation. Laura Koesten, Elena Simperl, Emilia Kacprzak, Tom Blount, Jeni Tennison. International Journal of Human-Computer Studies. 2019
- A User Centred Perspective on Structured Data Discovery. Laura Koesten. The Web Conference 2018, PhD Symposium.
- Searching Data Portals – More Complex Than We Thought? Laura M Koesten, Jaspreet Singh. Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2017. Workshop: Supporting Complex Search Tasks.
- Characterising dataset search – An analysis of search logs and data requests. Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, Elena Simperl; Journal of Web Semantics
- A Query Log Analysis of Dataset Search. Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibánñez, Elena Simperl, Jeni Tennison. Proceedings of the International Conference on Web Engineering, ICWE 2017.
- Learning when searching for web data. Laura Koesten, Emilia Kacprzak, Jeni Tennison. Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval, SIGIR 2016. Workshop: Search as Learning (SAL).

Signature:

Date:

Acknowledgements

I want to thank all the amazing people that I was lucky to meet during this PhD. They were kind, they believed in me and they gave me chances I am very grateful for.

To name a few:

Tom Heath, who convinced me to do a PhD and did not fire me when I repeatedly asked him to.

Jeni Tennison, for her wisdom and patience and for being there when I needed her. She welcomed me to the ODI with open arms, showed me that you can be yourself and successful at the same time, and that all lists need to correspond to Fibonacci numbers.

Of course, Elena Simperl, without whom none of this would have been possible. She became my supervisor and role model, and over time also a colleague, friend and boss. But more importantly, who became somebody who I truly admire.

My family for their support and acceptance of not just changing career, but also leaving the country and doing things that might not have always made sense from a distance.

My dear friends (at home, as well as those I met during this journey), without whom I could not have finished this. They were patient, looked out for me and reminded me of what's really important when I needed it.

Also a massive thanks to the participants of the experiments and the reviews we received over time.

Chapter 1

Introduction

With the rise of data science, millions of datasets have been published on the web, in institutional repositories, online marketplaces, on social networks or by individual publishers; in sectors from science and finance to marketing and government, amongst others (Noy et al., 2019; Verhulst and Young, 2016).

For instance by 2013 one million data sets were already made available on governmental data portals worldwide, many with an open license (Cattaneo et al., 2015). In Europe there are more than 940,000 datasets published by regional and national authorities in over 35 EU countries, indexed in September 2019 by the European Data Portal[1]. Datasets are increasingly available on scientific repositories (Kindling et al., 2017) or on specialised repositories for particular types of data. For instance the site Data Planet[2] lists to date no less than 6.2 billion statistical datasets, many of which are public. Geoportals are increasingly used to support national spatial data infrastructure (Maguire and Longley, 2005). These numbers are indicative of a general trend that can be observed in many domains. At the same time, the demand for and spending on financial, economic, and marketing data provided by vendors such as Bloomberg or Thomson Reuters continues to increase (Burton-Taylor, 2015) and last but not least there are large numbers of datasets shared within organisations (e.g. (Halevy et al., 2016)).

Despite its abundance and applications, there is a gap between available datasets, the datasets users need, and the datasets a user can find, understand and is able to use (Gregory et al., 2019a,b; Koesten et al., 2017; Swamiraj and Freund, 2015). When narrowing down the requirements for using web data from a user perspective, the first barrier is identifying whether relevant data exists, and if so, where to find it. Therefore this work is concerned with how people are discovering and understanding data, be that entire datasets or individual data points within a dataset. While the focus is on the discovery process, we aim to consider the context in which data search takes place,

[1] https://www.europeandataportal.eu/data/en/dataset
[2] https://www.data-planet.com/

which can determine the success of a search. For instance, once an adequate dataset is found, it needs to be accessed in order to be used. Access itself can be a question of the provision of certain formats, of an understanding of the connected licences and of the skill set of the person trying to access it. Only once a dataset is found and accessible, people can start to think about how to find meaning in the data itself. The usefulness of the dataset depends on the persons' specific information need, alongside other factors that we discuss in this work.

Given the variety of applications, we still know very little about how people find data, understand it, and put it to use. Research on dataset search is a relatively new area and efforts to this day have focused primarily on technical processes with the development of standards (e.g. CSV on the web[3]), community guidelines (e.g. Share-PSI[4] and domain specific data search engines (e.g. in scientific data repositories such as Elsevier's[5]). Google has recently released dataset search in Beta[6], using its schema.org markup language to index datasets alongside text documents, images and products[7]. This is a step to overcome the siloed nature of domain-specific data portals and to make datasets on the web easier to discover, which underlines the timeliness of this work.

1.1 Motivation

As the use of data-driven technology is growing there has been a huge surge in available data that can potentially be reused by others (Cattaneo et al., 2015; Manyika et al., 2013). Structured and semi-structured data in particular, which refers to data that is organised explicitly, for example as spreadsheets, web tables, databases and maps, has become critical in most domains (Manyika et al., 2013). We use such data to improve services, design public policies, generate business value, advance science, and make more informed decisions (Lavalle et al., 2011; Verhulst and Young, 2016).

However, no matter how impressive the volume might be, they tell only half of the story. The other half is that having data available does not always mean that data is easy to discover or that it can be used purposefully (Erete et al., 2016).

To illustrate this we present an example of the complexity of information seeking tasks for data, informed by tasks as described by participants of the study presented in Chapter 3: *Imagine a data journalist writing an article about the runway expansion at London's main airports in the UK. As part of her research, the journalist will look for factual evidence to substantiate her story, in the form of reports, news on similar topics, as well as data about the economic, social, and environmental ramifications of the project,*

[3] https://www.w3.org/TR/tabular-data-primer/
[4] https://www.w3.org/2013/share-psi
[5] https://datasearch.elsevier.com/
[6] https://toolbox.google.com/datasetsearch
[7] http://schema.org/Dataset

arguing for or against expansion plans. A large share of the relevant data is already available online, published by governmental agencies, researchers, and other journalists. However, finding it is not always straightforward. The journalist could use regular search engines, fact checking services[8]*, and other channels in the same way she does when looking for less structured kinds of information. She might also know of the existence of a particular data catalogue, which offers access to collections of data resources released by one or several organisations or she might try Google's dataset search. She might even query the API (Application Programming Interface) of a trusted data provider, or crawl the web looking for bespoke data snippets. Once she has identified several matches, her next step would be to explore the most promising among them and rank them according to how relevant they are to the narrative of her article. Depending on the tools used to discover the data, this step might involve downloading data files; working with different formats (e.g., CSV, XML, HTML, RDF, relational tables); choosing between several versions of a dataset; alongside sensemaking tasks such as establishing what exactly a dataset covers (its 'attributes' or 'schema'), and how accurate, complete, or up-to-date the data is.* This example illustrates the unique characteristics of searching for structured data opposed to searching for textual documents from a user perspective.

We aim to better understand scenarios like this one in order to inform the design of data tools and technologies that offer the same level of support and quality of user experience that we are used to from the web. Other examples of stories including data-centric tasks can be found, for instance, on the Data Blog of the Guardian newspaper[9].

While structured data is becoming critical across domains and many professional roles (Mayer-Schönberger and Cukier, 2013) the number of people engaging with data professionally is growing as well. In 2014, more than six million data professionals were spread across all industries in the EU and this number is increasing (Cattaneo et al., 2015). These are people who collect, store, manage and analyse data as their primary activity in their job. They come from various backgrounds and use a mix of skills for data-based decision making and to build services on top of data. We use the term data professional in a slightly broader fashion – to refer to any person who uses data in their job, even if it is not their primary activity. With the increase of data being available, searching for data becomes more common, including among audiences that are less data literate.

As data is used by people with different skill sets, across domains and different roles this results in diverse needs from tools that support them in working with data (Choi and Tausczik, 2017; Erete et al., 2016; Muller et al., 2019). We use structured data in various ways – from consulting official statistics to running scientific experiments, finding travel routes, creating maps, predicting elections, or designing better products. There is a large spectrum of data-centric tasks and the more data becomes available the more likely it is that tasks and users will become more varied.

[8] http://factcheckeu.org or https://fullfact.org
[9] https://www.theguardian.com/data

Understanding user needs for data discovery and exploration tools is the basis for designing them. We know little about which information to display for which task and how to best support the discovery and selection processes. Making sense of data is dependent on the ability to put pieces of information into context and to understand connections between them (Albers, 2015).

This work aims to understand how the discovery of new data sources can be improved by supporting users in assessing their fitness for use when selecting data sources from a pool of search results. This can be on the web, or equally in an enterprise setting. The findings can be used to inform the design of data discovery tools to make searching for structured data easier for a wider range of users. We focus on structured data, which is typically grouped into datasets, many of which are published on the web (for instance in data catalogues available on, or referenced from, data portals).

To search for datasets can be difficult as the structure of data catalogues varies. Moreover, search is dependent on inconsistent metadata and to search within the data itself is rarely possible. The concept of a dataset has slightly different definitions depending on domains and communities (Pasquetto et al., 2017). In this work, a 'dataset' refers to structured or semi-structured information collected by an individual or organisation, which is distributed in a standard format, for instance as CSV files. In the context of search, it refers to the information object (a dataset) returned by a search algorithm in response to a user's search query.

Selecting data from a pool of results after a search is dependent on the information exposed to the user and influenced by the design of the interface. Some information about the data is explicitly available as metadata, some is implicitly available through the context the data is situated in, for instance the publishing institution or what the data has been used for in the past. These are examples of information that already exists and could be used to aid result selection. We want to understand what information we should present together with search results and how that will support discovery and selection, as well as understanding the content of datasets. We approach this by learning from related areas of research such as general web search, Interactive Information retrieval as well as Information Seeking theories and Sensemaking. We also take a look at existing approaches for data search and exploration interfaces, although the majority of these focus on supporting exploration and collaboration with data rather than focusing on a retrieval context.

The term Human Data Interaction (HDI) has been recently introduced by Crabtree and Mortier (2015) to refer to how people engage specifically with personal data and tackle privacy issues. Elmqvist (2011) proposed a broader definition that includes the manipulation, analysis, and sensemaking of data. In our work, we consider the whole interaction process and its context. This interaction can cover information needs for data in which people are looking for an answer to a question (which can be a single data

point) and those in which they are interested in an entire dataset or multiple datasets. In both cases we assume the user intends to find a dataset. For example, someone could be trying to find out the number of schools in a given post-code and would need to extract that piece of information from a larger dataset that contains all entries for all schools in a country in, say, 2017. Alternatively, one could be studying how the number of schools across different areas has changed over time, which would involve processing and aggregating several versions of the same dataset, published year after year. Both types of tasks have an element of data search, evaluation, and exploration of data sources, taking into account different data properties; our work considers them equally. The findings of this work can advance the field of Human Data Interaction by identifying areas for research and further improvement, in particular around a more rigorous evaluation of the user experience in data search.

1.2 Approach

This work aims to increase our understanding of the following high-level question:

How can we help people select data that is useful for their task?

We define *usefulness* in Chapter 2 in Section 2.3.5. Literature describes user centric models for information seeking in web search, which are described in more detail in Chapter 2. However, less research can be found which focuses on data as the information source and the differences in the information seeking process for data as opposed to web or document search. Pfister and Blitzstein (2015) present the data science process, which is centered around activities to collect relevant data, explore it to make sense of it, and finally build an analysis model to draw conclusions from it. Based on these models we discuss the process of working with data from a user perspective with five pillars: *tasks, search, evaluate, explore and use.*

In pillar one (*tasks*) a user defines the task according to a goal. In pillar two (*search*) the search process is started, in pillar 3 (*evaluate*) the sources of data returned as results are evaluated and selected and in pillar 4 (*explore*), data sources are downloaded and opened and a users tries to understand the data and its context, do some exploratory data analysis, to then be able to perform the task intended in the beginning (pillar 5). This model was used as the basis for the initial study reported in Chapter 3. The assumption is that this process is not a linear one and that it involves multiple iterations and backwards movements between pillars for many data-centric tasks.

FIGURE 1.1: Simplified linear representation of the five pillars of the workflow with data.

1.2.1 Study 1: Working with structured data

We defined a number of sub-questions in our initial study, which is described in detail in Chapter 3:

- How do people currently search for data?
- What are the characteristics of information seeking tasks for data?
- How do people evaluate data that they find?
- What types of work tasks do people do with data?
- How do people explore data that they have found?

1.2.1.1 Methods

After an extensive literature review we conducted a mixed-methods studies, informed by Bryman (2006) for the purpose of this work. This initial study aimed to shed light on how data practitioners look for data online, with a focus on a qualitative component using in-depth interviews with twenty data professionals from various backgrounds. To supplement the in-depth interviews, we analysed a unique dataset of search logs of a large open government data portal. This gave us a less obtrusive way to learn about the behaviour of data search users (Jansen, 2006), of which our interviewees were a subset (17 out of 20 participants mentioned they used this portal to search for datasets).

1.2.1.2 Outcomes

Key findings from the initial study showed that search is a major issue for data professionals in their work with data and that searching for data is more often than not exploratory and complex. We found that finding and selecting a dataset can be difficult for people as they generally do not think they have enough information about the content of a dataset to make an informed decision. We further found that selection criteria in dataset search has unique characteristics. For instance the underlying methodology of data collection is of high importance when selecting a dataset. We also found that concepts such as data quality or usability are important and but that their definition is complex and their role in selecting a dataset is inherently task dependent. We discuss

the complexity of data-centric search tasks, how they differ from search tasks for documents and how commonly used interfaces on governmental data portals support users in the selection process of data in Chapter 2. Our findings furthermore showed that the majority of textual summaries of data are perceived to be of low quality and limited usefulness.

1.2.2 Study 2 + 3: Dataset summaries

Based on the findings from Study 1 we conducted two studies which examine textual summaries of datasets, which are a commonly used element in data search scenarios and help people to understand the content of a dataset. This mixed-methods approach is described in detail in Chapter 4.

The main research question addressed in this study was:

- What data attributes do people choose to mention when summarising a dataset to others?

1.2.2.1 Methods

For Study 3 we conducted a task-based lab experiment in which 30 participants described and summarised datasets in a writing task. Subsequently, we conducted a crowdsourcing study for Study 4 in which we replicated the lab experiment with a larger variety of datasets; and asked crowdworkers to rate the dataset summaries according to perceived quality. This allowed us to get a better understanding of the influence of the underlying dataset and of differences in participants and settings on the resulting summary. We collected 150 long and 150 short summaries from the lab experiment and 250 crowdsourced summaries and analysed these qualitatively and quantitatively.

1.2.2.2 Outcomes

The findings of this study have shown that textual summaries were laid out according to common structures; they contain four main information types and cover a set of dataset features. By identifying attributes that people consider important when they are describing a dataset we get insights into what a meaningful summary should contain. The most prevalent attributes across all summaries were: a high-level, one-phrase summary, describing the topic of the dataset (subtitle), explicit references to dataset headers, and the geographical scope of the data at different levels of detail.

Resulting from this study and from an additional analysis of dataset search diaries that gave insights into selection criteria in dataset search, we created a template for the

creation of dataset summaries that can be used as guidance in the summary writing process.

1.2.3 Study 4

We then tested the effect of dataset summaries in a search scenario to better understand how to support people in a dataset selection process (described in Chapter 5).

The research questions addresses in this study:

- Does the dataset presentation mode (text summary, table summary, preview, title) affect search time, performance, user behaviour and experience?
- Does the presentation mode of a dataset summary (as text or as a table) affect search time, performance, user behaviour and experience?

1.2.3.1 Methods

To validate the summaries in an actual data search scenario we conducted a follow-up online study with three mock-up tasks in which participants were asked to search for data. In a between-subjects design we compared four versions of search engine results pages (SERPs) of which two contained a summary based on the template from the prior study for a set of performance, interaction, and experience based metrics.

1.2.3.2 Outcomes

We found that merely displaying the title together with the publisher and format of the data is not perceived as enough information in order to make an informed selection of a dataset out of a pool of search results. Our results might further suggest that dataset summaries presented as a structured table could be perceived to be most useful for the dataset search tasks we tested. However these results were not statistically significant and would need to be validated in further research. We discuss that summary tables (potentially together with a preview of the data) could be an interesting direction to explore further.

1.3 Contributions

The main contributions of this work lie in a better understanding of user behaviour in data discovery. This includes *categorisation of data-centric work tasks, a better understanding of search strategies in data search, in-depth knowledge on selection criteria in*

data search, insights into exploration activities for datasets, an analysis of the composition of dataset summaries created by different types of participants as well as insights into interactive information retrieval experimentation for data search (discussed in detail in Chapter 6).

This allows us to address the main high-level question of this work on how to help people select data that is relevant and useful for their task.

Framework for Human Structured Data Interaction.

We synthesise the findings on information seeking for datasets in a framework for Human Structured Data Interaction, initially based on the findings from Chapter 3 and refined throughout this work. This presents a novel perspective on the conceptualisation of data-centric tasks, search and exploration strategies in Chapter 3, as well as an in-depth analysis of selection criteria in data search in both Chapters 3 and 4. This framework aims to help system designers and publishers of data understand what activities people do when searching for and engaging with datasets.

Template for dataset summaries.

We consolidate the findings from Study 2 and 3 in a template to guide the dataset summary writing process. We propose guidelines for the creation of textual summaries of datasets for human consumption. These can, if validated in further research, be a potential solution to support users in the selection process of a dataset out of a pool of search results. In Chapter 4 we present an initial prototype solution of how to semi-automatically guide users through the summary writing process based on the findings of this work.

We translate these contributions into actionable insights into how people could be supported in selecting datasets from a pool of search results and how to further advance Human Data Interaction research for dataset search:

- A framework for Human Structured Data Interaction that can be used as guidance for data publishers, data portals and tool designers of data discovery tools
- A proposed template to create user centred dataset summaries
- Initial suggestions how to present these summaries in a data search scenario
- Specifications for a data summarisation tool as well as an initial prototype
- A detailed discussion of potential directions for future research in data search

1.3.1 Publications

This work has led to a number of peer-reviewed publications.

- **The Trials and Tribulations of Working with Structured Data – a Study on Information Seeking Behaviour.** Laura Koesten, Emilia Kacprzak, Jeni Tennison, Elena Simperl. Proceedings of ACM CHI Conference on Human Factors in Computing Systems, CHI 2017.

  Personal contribution: lead author of this publication. Included study design, set-up and execution of the qualitative study described in this publication and in Chapter 3. Further includes qualitative analysis and write-up of the mixed-methods results. The other authors contributed in terms of discussions and planning of the study design, to the analysis of the search logs and feedback on the paper content.

- **Everything You Always Wanted to Know about a Dataset: Studies in Data Summarisation.** Laura Koesten, Elena Simperl, Emilia Kacprzak, Tom Blount, Jeni Tennison. International Journal of Human-Computer Studies. 2019

  Personal contribution: lead author of this publication. Included study design, set-up and execution of the experiments described in this publication and in Chapter 4. Further includes analysis and write-up. The other authors helped in terms of discussions and planning of the study design, as well as editing the paper content.

- **A User Centred Perspective on Structured Data Discovery.** Laura Koesten. The Web Conference 2018, PhD Symposium.

- **Searching Data Portals – More Complex Than We Thought?** Laura M Koesten, Jaspreet Singh. Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2017. Workshop: Supporting Complex Search Tasks.

  Personal contribution: lead author of this publication.

- **Characterising dataset search – An analysis of search logs and data requests.** Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, Elena Simperl; Journal of Web Semantics

  Personal contribution: analysis of data requests together with the lead author, collaborative write-up of this analysis, feedback and editing of the paper content.

- **Characterising Dataset Search Queries.** Emilia Kacprzak, Laura Koesten, Jeni Tennison, Elena Simperl; Companion of the The Web Conference 2018, Workshop: International Workshop on Profiling and Searching Data on the Web

  Personal contribution: planning of study design, set-up of the experiment and write-up of the paper together with the lead author.

- **A Query Log Analysis of Dataset Search.** Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Elena Simperl, Jeni Tennison. Proceedings of the International Conference on Web Engineering, ICWE 2017.

  Personal contribution: contributed to data analysis and write-up.

- **Learning when searching for web data.** Laura Koesten, Emilia Kacprzak, Jeni Tennison. Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval, SIGIR 2016. Workshop: Search as Learning (SAL).

  Personal contribution: lead author of this publication.

1.4 Thesis structure

This report is structured as follows: related work is summarised in Chapter 2. Chapter 3 presents our initial study on information seeking for structured data. We present two more in-depth studies on textual summaries of datasets, which are based on findings of the previous studies, in Chapter 4. Chapter 5 describes an online experiment to test the applicability of text based dataset summaries in a dataset search scenario. In Chapter 6 we discuss the findings and implications this work and lay out several potential future directions for research in HDI and user centred dataset search. Chapter 7 summarises the contributions of this work.

1.5 Terminology

Structured data
Data that is explicitly organised, for example in spreadsheets, web tables or relational databases.

Dataset
A dataset refers to structured information collected by an individual or organisation, distributed in a standard format (for instance CSV files). This is similar to the definition by Noy et al. (2019); however, in our scenario, we focus on alphanumeric data and exclude images due to their specific interaction affordances (Datta et al., 2008).

Data portal
Data portals are repositories of datasets which provide a central point of discovery of these datasets. Data portals often have a search functionality and represent one of the ways to search for data on the web.

Data publisher
Lawrence et al. (2011) define data publishing as the act of 'making data as permanently available as possible' on the web. In this context a data publisher is the person or institution who publishes data, usually online, for a group or the public to access.

Metadata
Metadata is data about a dataset. Metadata describes properties of the dataset, such as its title and description, provider, etc. Examples for metadata vocabularies are the Data

Catalogue Vocabulary (DCAT[10]) or Schema.org [11].

Data user

The person using the data. This work focuses on users with data-centric work tasks which come with an information need for structured data that requires locating and selecting datasets in the context of a task in which the data would be used.

Data search and data discovery

This refers to searching for data on the web, or on dedicated data portals. From the perspective of a data user, data search can refer to searching for datasets, as well as for a data point within a dataset. For the purpose of this work we consider search contexts where the results are datasets rather than individual data records published in a dataset. We do not include search within database structures in this definition (e.g. using SQL or SPAQRL).
While 'dataset search' and 'data discovery' are often used interchangeably, in the context of this work 'data discovery' includes all data search strategies (including human recommendation), as well as serendipitous discovery of datasets.

Snippet

The short summarising text that is returned by search engines. This helps users to make a decision about the relevance of a search result (Hearst, 2009).

Interactive Information Retrieval

Interactive Information Retrieval (IIR) is the field dedicated to studying and evaluating users' interacting with IR systems and information (Kelly, 2009). Also known as Human-Computer Information Retrieval (Marchionini, 2006b).

'User centred'

User centred design is a multidisciplinary approach to interactive system development that focuses specifically on making systems usable (ISO 9241-210:2010)[12]. This work presents a user centered perspective, which means the focus of interest is on the interaction process and the user experience in data search.

[10] https://www.w3.org/TR/vocab-dcat/
[11] http://schema.org/Dataset
[12] https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en

Chapter 2

Background

This chapter presents background literature in relation to this work. We describe how we search for data currently and where we see the differences between web search and data search. We discuss key concepts used in this work from the Information Seeking and Interactive Information Retrieval literature; and where we see our work situated in these areas of research.

We then describe factors influencing user experience in datasearch. We present these from the perspective of the user, such as different search types, the task context, sense-making and selection criteria for datasets. Furthermore we discuss factors from the perspective of the data publisher, such as features of the datasearch system, including the interface, SERPs, and the way overviews of datasets are presented.

2.1 Data search currently

As an increasing amount of data becomes available on the web, searching for it becomes an increasingly important topic. The web hosts a whole range of 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 (e.g. Wikidata[1]) and scientific data repositories (Cattaneo et al., 2015; Lehmberg et al., 2016; Kindling et al., 2017). Datasets are published in data markets (Balazinska et al., 2013; Grubenmann et al., 2018), open data portals (Hendler et al., 2012; Kassen, 2013; DataGovUk, 2018), scientific repositories (Elsevier, 2018; Altman et al., 2015; GESIS, 2018), or as part of general-purpose resources, such as Wikidata or the Linked Open Data Cloud[2] (Assaf et al., 2015).

[1] https://www.wikidata.org/
[2] https://www.lod-cloud.net/

Data search and discovery has emerged as a topic of research in a range of complementary disciplines. 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 (Gregory et al., 2017; Koesten et al., 2017), as related areas, such as document search.

Searching for structured data online is commonly done via general purpose web search engines or on domain specific data portals. A specific type of data portal that we discuss in more detail in this work are open data portals. They represent a point of free access to governmental and institutional data by cataloguing and archiving datasets allowing users to browse and search through repositories of these institutions.

Recently, Google has introduced dataset search[3] in beta, an initiative to use the `schema.org`[4] markup language to index datasets and make them discoverable through a general purpose search engine (Noy et al., 2019). There are a number of other ways to access data online, such as dedicated web crawlers and APIs which require specific skill sets. The focus of this work is on information seeking scenarios in which datatsets are discovered through a traditional type of search functionality, such as on Google, or on many data portals. We aim to discuss both professional data users as well as less data literate users who might search for and use data, but would not consider themselves as data professionals.

2.1.1 Example of a data search process

We consider scenarios in which users have an information need that requires data as the information source. Imagine you want to analyse trends in street crime rates in London over the past year. You are trying to find data that is relevant for this information need/task. You might enter a search query such as *'hate crime statistics 2018'* in the search box of a UK data portal (see Figure 2.1).

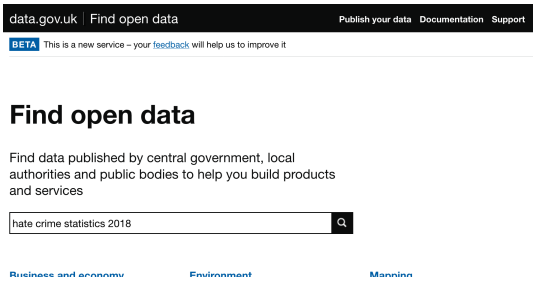


FIGURE 2.1: Data search interface on a data portal. Users can type keywords in the search field or browse the data collection by domain and other attributes.

[3] https://toolbox.google.com/datasetsearch
[4] https://schema.org/Dataset

The search results may look like in Figure 2.2. On the left hand side, you can find a classification of the results based on metadata attributes such as publisher, topic, format and license. You can use these facets to explore the collection of datasets or filter the results. On the right hand side, you can choose from a ranked list of datasets. Each dataset is presented via its metadata with title, publisher, last update and available file formats.

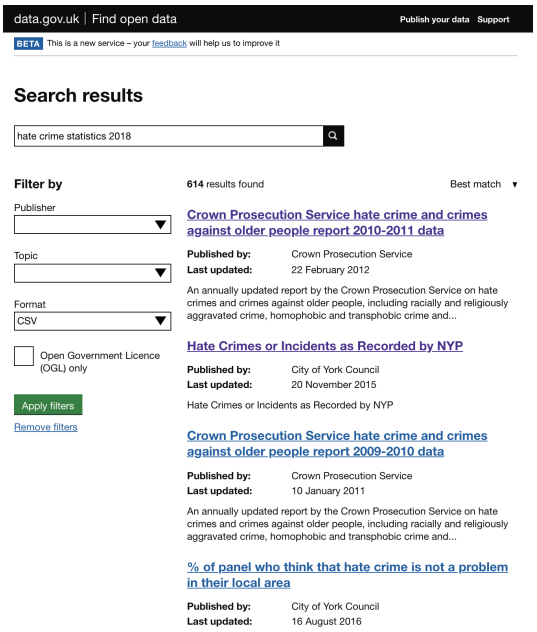


FIGURE 2.2: Example of a data search result list on the UK governmental open data portal. Below each search result is some metadata and a snippet describing the underlying dataset.

In most cases, the dataset is accompanied by a short text summary, discussed in detail in Chapter 4. You select one of the results to explore further, based on what else is on the list and on the information displayed in the summary. This commonly takes you to a new page, where you can also download the dataset to examine it in detail. Google dataset search[5] provides a similar experience to data portals. All search results are pointers to datasets, and the dataset preview page in their case is displayed together with each search result.

For other types of users, data search might start on a general purpose web search engine. Data search results on general-purpose web search engine have a similar look and feel (see Figure 2.3), although the hits are a mix of datasets and other types of sources. In this case, you might be able to tell from the result snippets which links refer to datasets, click on the results, and look for a download link, a table, or an API.

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

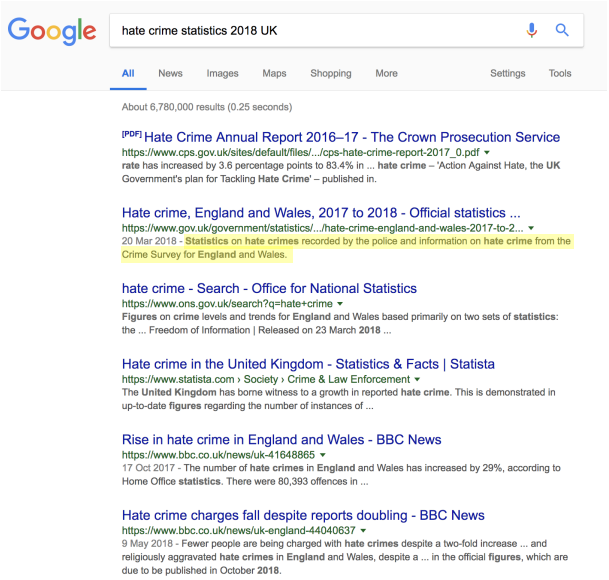


FIGURE 2.3: Example of a data search result list on Google. Underneath each search result is a a snippet describing the underlying website which may or may not contain data.

2.1.2 Data search in contrast to web search

Web search has evolved over time; before web search engines were developed, documents were recommended to people by other people. Manually created catalogues were used, which collected together content on the web that people recommended. First attempts at search found matching terms within these indexes, then techniques were developed that could rank results according to their relevance and from that more complex search algorithms evolved. Currently web search can already provide sophisticated contextual and personalised results by combining explicit queries with other, more implicit information (White et al., 2009; Sieg et al., 2007; Marchionini, 2008). In contrast to that techniques to support searching for data are far less advanced and do not currently provide the same user experience as we are used to from web search.

Data sources can be categorised according to their level of structure (unstructured, semi structured and structured data). While the boundaries between these categories are not always strictly defined, we typically broadly distinguish between unstructured, such as regular text; and structured data, which is explicitly organised, such as in relational databases, web tables, lists, and spreadsheets. This work discusses structured data in general, independent of format or data type. However, for our experimental studies we focused on tabular data (CSV files and spreadsheets) as they come with a low technical barrier and are so accessible to a wider range of users than more complex data types. Structured data is available on dedicated websites, including data catalogues hosting multiple datasets (e.g., scientific repositories, open government data platforms); back-end deep web databases (e.g., books data records on amazon.com) (Madhavan et al.,

2006); bespoke marketplaces (such as Microsoft Azure's[6]); and specialised data vendors (such as Bloomberg[7]) for financial market data. Data search has evolved heavily based on what we know about web search, however we believe that ideas and tools from web search cannot be directly applied. Searching for sources of structured data presents many challenges and datasets have unique characteristics which come with certain requirements or constraints. While this work focuses merely on the user experience in data search we give a short overview of some of the underlying technical differences to demonstrate the constraints within which data search can currently be approached.

Web search engines are based on algorithms which are designed to rank web pages (Page et al., 1998) and do not equally support the indexing of structured content. Some aspects of data discovery, such as data mining techniques, are very advanced but others, like techniques to integrate heterogeneous data published on the web, are not yet fully usable. Putting aside the question of links between resources, which is at the core of algorithms such as Page Rank, established Information Retrieval technologies are primarily designed to work on (unstructured) documents and less to fulfil the specific needs that people have when looking for data (Wilson et al., 2010; Cafarella et al., 2008). We currently know little about how we search for data, therefore we know little about how we ideally like datasets to be ranked for which tasks, and which signals are of particular importance in dataset search (Noy et al., 2019).

One characteristic of data search on the web or on data portals, is that we tend to use keyword search on metadata that is published together with the dataset (described in more detail in Section 2.4.2). Currently approaches hardly ever search through the datasets themselves to assess the relevance of those datasets to a query. Instead, we rely on metadata and descriptions of data in order to find datasets, which is inconsistent in quality and availability (Atz, 2014; Noy et al., 2019); that is one reason why search functions within data portals lack accuracy. We know that keyword-based matching works less well on structured and semi-structured data (Lopez-Veyna et al., 2012). For instance Thomas et al. (2015) show that dataset repositories have poor search over and inside tables and demonstrate that the naïve approach of full-text search is not appropriate. The majority of datasets do not have links between themselves or to web pages, which is another reason keyword search on such data has limitations (Lopez-Veyna et al., 2012).

While there is a significant volume of studies from the semantic web community (e.g. entity centric search (Balog et al., 2010), as well as the database and management community, there is to this day little research targeting general purpose dataset search on the web or on data portals. Cafarella et al. (2011) have discussed the complexity of surfacing structured data on the web. The majority of deep-web data is not crawlable, or requires special techniques (Cafarella et al., 2008). Initiatives such as Open Data

[6] https://azure.microsoft.com
[7] https://www.bloomberg.com/

Portal Watch[8] or the European Data Portal[9], which build meta-portals that crawl the web to offer integrated access to multiple data repositories, require manually curated mappings between the attributes and schemas used by different sites to describe their datasets and make them discoverable.

While we focus on selection criteria another interesting aspect of data search is specialised querying interfaces trying to take advantage of the structure inherent to the data to allow users to access it via different routes. A number of studies have discussed specific aspects of data search, such as querying interfaces for SPARQL queries for Linked Data (a semantic query language), SQL (Structured Query Language) or CQL (Contextual Query Language) interfaces; or querying support for temporal and geospatial queries (Neumaier and Polleres, 2018). However, the specifics of such interfaces are very dependent on the data format and are outside the scope of this work. We discuss search result presentation in more detail in Section 2.4.1.

In addition to these technological limitations, there is little research examining data search from a user perspective. We know that different information sources result in people searching and choosing results differently, as relevance depends on context. This has been shown in research on search verticals (which focus on a specific type of content), for instance for scientific publications (Li et al., 2008; Yu et al., 2005), people (Weerkamp et al., 2011) or products (Gysel et al., 2016; Rowley, 2000).

Wynholds et al. (2011) show that, in the context of digital libraries, seeking and using documents and data for research purposes has specific information needs, processes and required level of support. Kern and Mathiak (2015) conducted a user study with empirical social scientists specifically contrasting data search and literature search. They found that the quantity and quality of metadata is more critical in dataset search and that participants were willing to put more effort into the retrieval of research data than in literature retrieval, where convenience prevails.

Again, most of the things we know from the literature have been designed with other requirements in mind and the extent to which they apply to our scenario is yet to be fully investigated. In web search, a common denominator is the differentiation between simple (known-item) search and exploratory search (addressing more complex information needs) (Diriye et al., 2010; Albers, 2015; Marchionini, 2006a; Broder, 2002; Rose and Levinson, 2004), which we discuss in more detail in Section 2.3.1. Other models distinguish between question answering (Voorhees, 1999) and document retrieval (Gupta et al., 2016) to emphasise the specific challenges associated with identifying not just the document in which a specific piece of information could be found, but the actual answer to a users' query. We can find some parallels to this in relation to structured data. More specifically, people looking to find answers with data will sometimes require not just to

[8] https://data.wu.ac.at/portalwatch/about
[9] https://www.europeandataportal.eu/

be pointed to the right database, spreadsheet, or list that might contain the information they need, but to the record that answers their question. This has implications for how data search is implemented, as algorithms that focus on metadata, which are the de facto standard in existing portals, do not have this capability.

We believe searching for data should be an area of attention in its own right, rather than extrapolating results from traditional document search. We focus this work on the interaction of people with data, so this refers to the user's experience and ability to make sense of data that they find.

2.1.3 Dataset search versus database search

A somewhat related area of research is search and usability in databases. We describe where we see differences from a user perspective; we do not aim to cover the technicalities of different types of databases and connected data formats as this is outside the scope of this work.

One of the key differences in how users interact with databases is that it is usually possible to search on the data or row (tuple) level; whereas in dataset search, as we explore it in this work, this is usually not possible and we rely heavily on information about the data (metadata and textual descriptions of datasets). The result set returned is dependent on the query and the capabilities and specifications of the database; whereas we look at search for entire datasets which are not likely to conform to the same schema. While we also consider information needs for individual data points we focus on those scenarios that require finding a dataset that contains the desired data point - as this represents data search on the web currently. There is no predefined schema when we look at a dataset published on the web, that search functionalities could take advantage of, and, as mentioned in the previous section, no way to access the dataset's content for the retrieval system.

Different interfaces both for database querying or results presentation have been explored by a number of authors (e.g. (Catarci, 2000)). For instance, Wen et al. (2018) looked at a summarisation approach to aggregate query answers to show larger numbers of potentially relevant results that users can explore interactively. Liu and Jagadish (2009b) discuss a spreadsheet-like interface potentially suitable for non-technical users of databases. 'DataLens' by Liu and Jagadish (2009a) uses a zooming-in analogy by presenting clusters of result units in a database, always showing one representative result for each cluster. The idea is to learn what is in the data without actually seeing all the results. The examples mentioned are meant to be of exemplary nature and do not aim to present a comprehensive overview of the vast amount of database related research.

Jagadish et al. (2007) claim that user's expectations for interacting with databases are fundamentally different from expectations for the web: They expect to query the

database in a more sophisticated way, such as with semantically complex queries, taking advantage of the database schema, which is different from what keyword based interfaces afford. They further assume that users have significantly higher expectations on precision and recall, amongst other aspects.

Visual and form based querying interfaces have been proposed to assist users in building queries to avoid putting the burden of complex querying syntaxes (e.g. SQL/XQuery) on users (Jagadish et al., 2007). The benefits of keyword based querying interfaces have also been explored in database systems (e.g. (Agrawal et al., 2002)). However, they can go deeper into the structure of the data than is possible in data search on the web.

One interesting approach in this context is Saint-Paul et al. (2005)'s work, where the authors try to summarise content in an entire database by producing rows (tuples) with higher level content that represent clusters of content. This would allow users to access the content at different levels of granularity. An overview of work on database summarisation was done for instance by Roddick et al. (1999). We discuss summarisation approaches specifically for datasets in Section 4.2 in Chapter 4.

We believe to better understand result presentation for dataset search, there is value in looking at the database literature in more detail. However, this is not the main focus of this work. In addition, many approaches that might be interesting from an interface point of view are rather far from what is technically possible for a general purpose data search engine that returns datasets published on the web or searches through data repositories.

2.2 Information seeking

In this section we specifically discuss information seeking models and interactive information retrieval, as this work is situated in these areas of research. While the high-level steps in information seeking likely apply to all information sources, we focus on those aspects where we see potential unique characteristics of structured data that influence this process.

2.2.1 Information seeking models and interactive information retrieval

Information Seeking, as a discipline rooted in Library Sciences, looks primarily at how people seek information (Blandford and Attfield, 2010), placing the people and the finding activity at the focus. Information Retrieval (IR), more rooted in Computer Science, is concerned with the technologies that support finding and presenting of information (ibid.) and was traditionally focused on, e.g., developing algorithms that improve precision and recall in search processes. A rich body of information retrieval literature

explores how people select documents and determine their relevance to a given task or information need (Barry, 1994; Park, 1993; Schamber et al., 1990). Interactive Information Retrieval (IIR) studies users interacting with systems and information (Kelly, 2009). The focus here is whether people can use a system to retrieve information, whereas classic Information Retrieval is focused on whether a system can retrieve relevant documents (Kelly, 2009).

This work is based in IIR and information seeking, aiming to understand the specific characteristics of how people retrieve structured data as the source of information as opposed to other sources, such as textual documents or web pages.

The information seeking process is described by Marchionini and White (2007) as a set of actions done by users in a progressive and iterative manner. It starts with recognising an information need, involving cognitive activities such as formulating the problem sufficiently to take action, and expressing this information need in a search system. This expression is constrained by the search system. When the system delivers a response the user engages in the examination of results, which is the most time intensive activity (ibid.). The user either decides to reformulate the query, or to stop the search activity and use the found information. This process mostly mirrors the workflow with data described in Chapter 1 and the higher level steps described in many information seeking models. Within this information seeking process for data we primarily focus on the examination of datasets as search results by the user. Information seeking that is specifically centred around (structured) data is not often discussed in literature (Kern and Mathiak, 2015). There are a variety of information seeking models, standard ones (Sutcliffe and Ennis, 1998) consider the information need to be static (Hearst, 2009). Other, more recent models consider the dynamic nature of information seeking and take into account that the information need gets refined or changed based on the retrieved information, as suggested for example by Bates (1989). Fidel (2012) proposes an ecological approach to information seeking in which the emphasis is on the environment and context in which the search takes place. Similarly, we believe that the design of effective information systems needs to be aware of the complexities of the information space the search takes place in. This includes, for instance, the types and formats of data sources, how reliable they are, how often they change, and whether more datasets need to be brought together to complete a task.

A well-known model based in Information Seeking is the Information Search Process (ISP) introduced by Kuhlthau (2004), which describes different stages in the search process. Kuhlthaus' ISP understands information seeking from a user's perspective as a means to accomplish a goal. The process is described in six stages: task initiation; topic selection; pre-focus exploration; focus formulation; information collection; and search closure. When applying these stages to the process of searching for data we have *intiation* when the user realises the need for information. *Selection* as the point where the user decides to use data as the source of information. The stages of *exploration* and *focus*

formulation can be interpreted as the learning process which occurs while searching. The search gets refined, selection criteria become more obvious and potentially search strategies get changed based on what was learned in the search process. In *information collection and search closure* a user has not only found data, but has also performed an initial exploration of the datasets, to determine whether they are adequate for the task. As mentioned by Kuhlthau (2004) the user has a sense of availability of information at that stage and confidence increases with the feeling of performing a comprehensive search.

We believe that the collection phase is more complex when searching for data as opposed to documents. Factors such as format, licence or access to the data as well as the availability of meaningful documentation can complicate this process. In contrast to Kuhlthau (2004), we are using the term 'exploration' as a means to understand whether a search result, in our case a dataset, is relevant for the searcher's purposes. This interpretation is much more aligned with data science frameworks such as the one introduced by Pfister and Blitzstein (2015), which is centered around activities to collect relevant data, explore it to make sense of it, and finally build an analysis model to draw conclusions from it.

Resulting from a number of studies on information seeking (e.g. Adams and Blandford (2005)), Blandford and Attfield (2010) propose the *information journey framework*. It similarily describes phases of information gathering; from recognising an information need to acquiring, interpreting and validating that information to subsequently using this interpretation. Similar to the workflow with data that we proposed in Chapter 1, these phases are not necessarily seen as sequential (Blandford and Attfield, 2010). Marchionini and White (2007) describe a very similar set of activities, focusing a bit more explicitly on the examination of results. They mention that reviewing primary content is the activity in the information seeking process that takes the most time.

Based on these models and informed by the data science process described by Pfister and Blitzstein (2015), we discuss the information seeking process for data in the high-level categories *'task, search, evaluate, explore and use / refine'*. The main focus of this work is on search and evaluation activities, however to get a thorough understanding of user experience in data search we consider the entire information seeking activity. We assume that when applying information seeking models to an information need for data, the high level phases of information seeking are similar, as displayed in Figure 2.4. Nonetheless, the processes a user needs to work through, as well as the potential barriers, the required skills and the connected tasks differ, and therefore user needs for tools and interfaces are not the same as when searching for textual information. As for all linear abstractions of the information seeking process, this is usually an iterative process and can involve returning to previous activities at any point. The user learns during this process and refines her activities in the context of the task (Rieh et al., 2016).

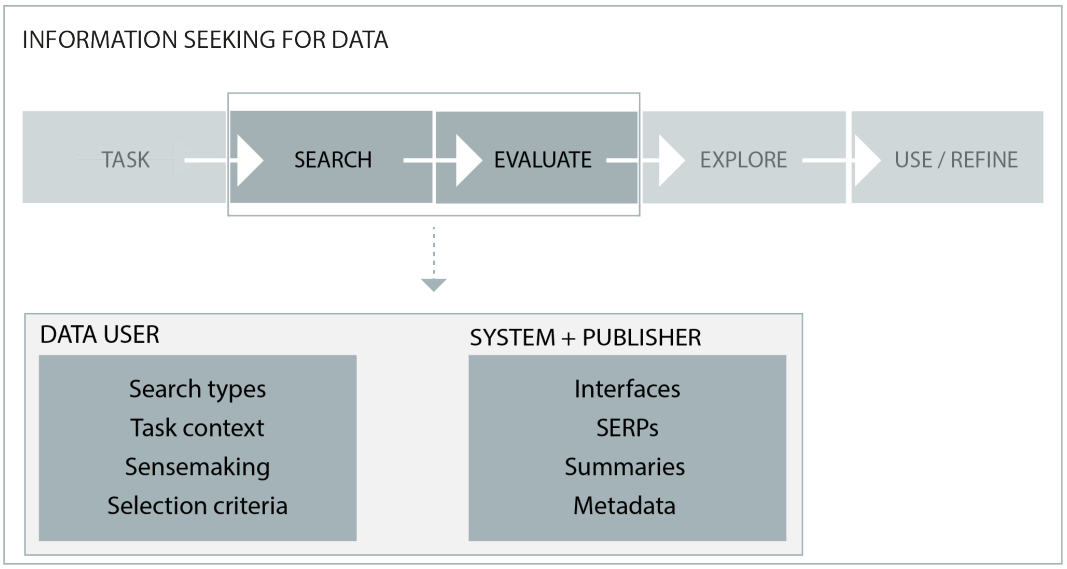


FIGURE 2.4: Simplified information seeking process for data and influencing factors.

Figure 2.4 lays out the structure of the remainder of this Chapter. We discuss the factors influencing user experience in data search from the perspective of the user and their context, in Section 2.3; as well as from the perspective of the data publisher and features of the system that is used for the data search activity in Section 2.4.

2.3 The user in data search

2.3.1 Search types: simple versus exploratory search tasks

Effective information retrieval systems must be based on an understanding of users and their tasks (Ingwersen, 1992). Search tasks range from simple (known-item) to more complex tasks and the interaction with a search interface depends on the type of task and the information need connected to the task (Baeza-Yates and Ribeiro-Neto, 2011). Marchionini (2006a) differentiates between lookup tasks and exploratory search. Lookup tasks involve fact retrieval, question answering and are commonly successful by displaying short, discrete information pieces (Baeza-Yates and Ribeiro-Neto, 2011). In our scenario this could be a value within a dataset that answers a question.

We can find some parallels to this in relation to structured data: people searching for data will sometimes need not just to be pointed to the right database, spreadsheet, or list, but to the record that answers their question. This has implications for how data search is implemented, as algorithms that focus on metadata (which are the de facto standard in current data search on the web) cannot support this type of query.

Searching data on the web currently is rarely a straight-forward lookup task, as described in Section 2.1. Finding data often involves open-ended and multifaceted search scenarios

involving comparing, synthesising and evaluating data, which is defined as exploratory search by Marchionini (2006a).

People engaging in exploratory search are often unfamiliar with the domain, unsure how to achieve their goal or also unsure of their goal (White and Roth, 2009). For example, a person might not know exactly what to search for because they are not familiar enough with the topic and the available resources. Sensemaking and learning are often inherent to this type of task (Diriye et al., 2010; Marchionini, 2006a) which are discussed in more detail in Section 2.3.3. Exploratory search often requires the consultation of additional information from external sources and results in *unsystematic steps through the information space* (White and Roth, 2009; Simon, 1973). Given the breadth of possible information needs and the potential vagueness of the user's goal, typical measures of success such as time, performance, or result accuracy cannot be directly applied in exploratory search (Wilson et al., 2009).

White and Roth (2009) draw parallels to exploratory data analysis (EDA) (Tukey, 1970); describing similar systematic learning mechanisms such as hypothesis formulation and testing. In EDA the data gets explored in different ways, usually without having a prior understanding of what the data contains, until a 'story' emerges (e.g. through patterns, outliers, or comparisons to other data) (Marchionini, 2006a; Baker et al., 2009).

Considering the types of context needed to use data meaningfully, as described in Section 2.3.4, we hypothesise that search for data displays characteristics of exploratory search tasks. The users goal and information need might differ, as the type of information available differs when being able to search for data directly (Wilson et al., 2010). We argue that this applies also when searching for datasets. The processes of retrieving data as the information source is likely to have unique characteristics. This affects every part of the retrieval process, both from the systems and the user side.

Keyword search has become the standard information retrieval and web search strategy (Hearst, 2009). However, the literature suggests that offering only keyword search options might not cater to all users' information needs (Wilson et al., 2009). For instance, exploratory search might require systems to support more diverse search strategies than simple keyword search (White and Roth, 2009), such as comparing, synthesising and evaluating (Marchionini, 2006a). Especially when users have imprecise goals, complex questions, not much pre-search knowledge or use systems with poor indexing (Wilson et al., 2010). Exploratory search often includes both focused searching as well as exploratory browsing activities (Marchionini, 2006a). Future research can show if these characteristics might have to be supported in specialised dataset search systems and interfaces or if different interaction patterns might be more useful.

2.3.2 Data-centric information seeking tasks

Data-centric search activities are likely to be dependent on the task people intend to use the data for (Ingwersen, 1992). We focus on work tasks, which are particular types of tasks with an assigned activity or unit of work, that needs to be completed to meet a pre-defined goal, as defined by Freund (2013). The work task provides the context and potentially the environment in which the search is conducted (Byström and Hansen, 2005) and co-defines so the evaluation criteria of when a task is considered successful, including relevance and usefulness assessments (Borlund, 2003). In that sense, work tasks motivate and frame information seeking tasks (Freund, 2013), which can be seen as subtasks, as they are connected to the primary purpose or goals of a work task (Byström and Hansen, 2005).

While there are a variety of information seeking tasks classifications in the literature, to this date there is no comprehensive taxonomy for data-centric work tasks and connected information seeking activities. This might reflect the rapidly changing work practices with data in the last decade. As mentioned in Chapter 1, Pfister and Blitzstein (2015) describe tasks connected to data science activities in their model of the data science process. Convertino and Echenique (2017) report subtasks in data analysis conducted by business users. Similarly, Boukhelifa et al. (2017) describe the subtasks 'dataworkers' engage in when encountering uncertainy in data analysis. None of the mentioned authors presented a deeper analysis of data-centric work tasks. Looking at information seeking tasks more generally, Li and Belkin (2008) propose a faceted classification of such tasks, in which one of the facets is the 'product of the task', which can be 'factual information' such as data, facts or similar. Gregory et al. (2017) describe tasks as 'data uses' drawing from a qualitative study with researchers from five different disciplinary areas. The types of task and the connected goals are known to influence information behaviour (Freund, 2013). This can have implications on the design of the querying interface, on search result presentation, as well as for ranking or recommendation algorithms. We believe that a classification for just data-centric tasks could be useful as a preliminary step to improve the user experience in dataset search, for instance by allowing to model the interaction with the IR system and tailor functionalities to specific task types. As described by Freund (2013) a user's perception of usefulness is dependent on the task, the information type and on the relationship between these two factors. The study described in Chapter 3 is a step towards describing data-centric work tasks of data professionals. However, the focus of this work is mostly on selection criteria for data search, looking at tasks only as the context for such a selection.

2.3.3 Sensemaking with structured data

Information seeking can be seen as part of a larger process referred to as sensemaking (Baeza-Yates and Ribeiro-Neto, 2011). Sensemaking is defined as the process of constructing meaning from information (Blandford and Attfield, 2010; Naumer et al., 2008). It has been studied by different disciplines, such as psychology, information seeking, HCI, and naturalistic decision making. Imagine the journalist mentioned in the introduction. She has identified a dataset she wants to use, but which she has never seen before. As a next step she will try to understand what that data means on various levels, e.g., as a whole dataset, the columns, headers, or other structural features, and its relation to the world and the context of her story. In this first moment, when looking at numbers she has never seen before, she will see 'data' in its original sense - numbers or words, which only turn into information when context is added (Albers, 2015). In Albers (2015)'s hierarchy of levels of information this 'information' is called knowledge only when it produces further understanding of a situation. We also refer to this process as data exploration.

In other words, sensemaking is an iterative process that involves linking of different pieces of information into one conceptual representation that constitutes an interpretation (Hearst, 2009; Russell, 2003). This process is heavily dependent on the individual person involved in it. The individuality of users is well recognised in literature in terms of their cognitive makeup, their prior knowledge about the topic and about searching, their motivation, as well as other factors (Kelly, 2009; Marchionini, 1995). Ingwersen and Järvelin (2005) describe the user as a constructive actor, with cognitive conceptions rather than as a passive receiver of information (White et al., 2009). Each user has had different physical, cognitive and affective experiences and behavioural dispositions (Kelly, 2009) as well as different levels of data literacy. This could mean tools that support search and understanding of data need to have a certain amount of flexibility. While this applies to all information seeking activities we believe that there are unique characteristics when the source of information is structured data.

In Information Seeking, sensemaking often refers to understanding documents or web pages, which can be more or less structured. Making sense of *structured* data has mostly been studied in connection to information visualisation, which can help to see patterns in data (e.g. (Furnas and Russell, 2005; ah Kang and Stasko, 2012; Marchionini et al., 2005)). For instance, when discussing how a certain type of analyst grasps large bodies of information, Stasko et al. (2008) presented a visual analytics system tailored for the sensemaking tasks of particular groups of data analysts. As part of their work on accessing government statistics, Marchionini et al. (2005) allowed users to explore relevant information from different perspectives and understand relationships within the data via agile display mechanisms. While there are other examples like these in the literature, they are not very well integrated into more mainstream productive environments and

data work practice. Furthermore, visualisations are often not available, especially when searching data on the web we mostly rely on metadata.

For example in their work on accessing statistical information, Marchionini et al. (2005) emphasise that people need context as well as means to reveal the story behind numbers, which should vary with the level of expertise of the user. Roth et al. (1992) describe scenarios in which some specific data can suddenly be critical in a particular situation when referring to accident scenarios in which a usually unimportant piece of data becomes the catalyst of events. But also on a more basic level – it can be unpredictable to gauge the importance of particular data as new events can always change the context (Woods et al., 2002). Some domains have developed specific data exploration systems which support their needs; we discuss these in more detail in Section 2.4.1.

In order to make sense of data we find we need to apply cognitive work to make it meaningful (Rasmussen, 1985). We do this based on the data we find and the information provided with it – the metadata. There is a large body of research concerned with the cognitive processes that take place during search, specifically on cognitive IR models (Belkin, 1984; Daniels, 1986), that can help to better understand sensemaking in information seeking. As mentioned before, data search differs from traditional document search as finding, accessing, understanding and using data requires specific, potentially additional, skills. The user's prior knowledge and experience with the domain or topic determine their ability to understand the data. Skills such as accessing, interpreting and critically assessing data are part of a user's data literacy (Calzada Prado and Marzal, 2013). Data usually requires additional context to be interpreted, as described in more detail in Section 2.3.4. Hence potentially more complex search interfaces are required, that offer different viewpoints to facilitate learning during the search process. Further research is needed to understand how people make use of data resources and progress from finding to understanding (Marchionini, 2006a).

The better the sensemaking process from data through information to knowledge is understood, the better systems we can create to facilitate learning from and through data search. As discussed earlier, our research, as much of the related work in data search and sensemaking with data, is based on the assumption that, in order to offer the best user experience, we cannot simply reuse or re-purpose principles, models, and tools that have been proposed for less structured sources of information – the results of our initial study described in Chapter 3, confirm this.

2.3.4 Context

Web pages often offer textual information and so provide curated and processed data, or information, that comes with context. Furthermore, web search engines are very advanced in providing additional context - they can provide contextual and personalised

results by combining explicit queries with implicit feedback, such as integrating the user's browsing behaviour into a ranking system (Sieg et al., 2007; White et al., 2009). Reusing data beyond its original purpose is known to be challenging (Wilkinson et al., 2016). Context is a necessary source of meaning (Dervin, 1997), and there is added complexity of context within data search due to the additional information required to create meaning from data as opposed to from text documents: data can have caveats attached to it that influence what it can be used for. The process of data collection, processing and cleaning that takes place before a dataset gets published can be very complex and is often not reflected in the dataset itself, nor in the documentation attached to it (Davies and Frank, 2013). The negotiation of potential biases, or the representativeness of a dataset, or the meaning of missing values often requires the user to access additional information and remains challenging (Baeza-Yates, 2018).

The ability to transform data to information and so make sense of it, is dependent on the context provided by the system as well as on the data literacy skills of the individual. This additional information can partly be provided through information about the data – metadata. Providing reference points with the data, or in the presentation of data, is necessary to enable the user to build relationships between different pieces of information, which is needed to understand complex information (Albers, 2015; Marchionini, 2006a). Similar to Hearst (2009) and Russell (2003), Rieh et al. (2016) describe the sensemaking process as creating knowledge structures between the data or information that has been acquired through the information seeking task. This is arguably facilitated by the context provided during the data search as well as its presentation. For instance, understanding geographical data can be easier when displayed on a map, and meaning can be attached to numbers if a range or a graph is presented that supports relating those numbers to reference points. The presentation of data influences sensemaking (Wilson et al., 2010), which has also been shown in work on uncertainty visualisations (e.g. Greis et al. (2018))

Decisions about the amount of context provided with the data are made by data publishers or by those designing the system with limited guidance on user needs in that space. At the same time interface design plays a key role in representing the context (Greenberg, 2001) and making it accessible to a wide range of users. Interfaces should enable discovery of connections between different data points, that represent data in a network to enable a user to understand its meaning within the context of other data. Publishing structured data as Linked Data can be seen as a partial realisation of this idea, as it provides a basis for interlinking data and so adding context (Bizer et al., 2009), however the majority of data on the web is not published as Linked Data.

Web data is heterogeneous and comes from different sources, it has different meanings attached to it. This presents interaction challenges that require thinking about the user, as well as about the underlying system and the design of the interface. An overview of search results can enhance orientation and understanding of the information provided

(Rieh et al., 2016). For data search, learning can be supported by allowing to zoom in and out of levels of data, allowing filtering and cross filtering (ibid.), rather than displaying one piece of content at a time, such as is done with a list of documents. Navigational structures can support the cognitive representation of information (ibid.) and there is a large space to explore interfaces that allow more sophisticated interaction with a datasets' context. We discuss current data interfaces in Section 2.4.1 to give an overview, but the aim of this work is to contribute to a better understanding of what type of context is useful in a dataset search scenario.

2.3.5 Selection criteria

As the focus of this work is to understand the selection process of data in a search scenario we discuss related literature on evaluation of search results, as well as selection criteria from a user perspective. We discuss data quality in a bit more detail as this is one aspect of selection criteria that has been studied directly in relation to structured data. Most other research in this area has focused on the retrieval of textual documents. To better understand how we can support users in selecting datasets that fit their information needs we need to understand the different dimensions which influence their decision.

While the concept of relevance was explored in a number of fields, including Communication, Philosophy and Psychology (Saracevic, 1996), we are focusing on the concept of relevance as applied in Information Retrieval and Information Seeking research. Within these areas relevance generally refers to the matching of a query to an information source (traditionally textual documents) (Belkin and Croft, 1992).

The applicability of using relevance as a criterion to evaluate single search results, such as in the Cranfield and TREC paradigm (Robertson, 2008), has been questioned for Interactive Information Retrieval contexts (Saracevic, 1996). It does not consider context related to the user and the use of the resource, as it is traditionally considered to be the property of the system. A number of authors have therefore defined relevance in a more situational context in which the social context and time dependence, amongst other factors, influence the nature of relevance. This definition acknowledges the importance of the user and so the subjective and fluid aspect of the concept of relevance (Schamber et al., 1990; Mizzaro, 1997). Both these relevance definitions are criticised, either for not taking the situation or the user into account, or not taking the system into account (Saracevic, 1996).

The concept of relevance changes across studies, authors and time and there a large amount of work has been dedicated to understanding how people assess relevance of documents (Barry, 1994; Park, 1993; Bales and Wang, 2005). Saracevic (1997) describe the complexities of using relevance as a criterion to assess the effectiveness of an information retrieval activity and discuss the notion of a 'system of relevances'. Bales and

Wang (2005) compiled 133 criteria in a review of 16 empirical studies on user-based relevance criteria of relevance studies. We see relevance as an equally broad concept, which we will discuss in more detail in Chapter 3, but believe that the specific criteria people apply to assess relevance (and its sub-concepts) is likely to have unique characteristics when searching for structured data.

Freund (2013) emphasises the context-dependency of selection criteria – relevance is dependent on the type of task that an information source is selected for. Different types of work tasks prompt different information behaviours and in most cases simply finding data does not resolve the information need of the user (ibid.). This might be different for fact-finding tasks, but only when access to the data is provided directly, such as in Question Answering scenarios. In most other cases the minimum activity that is required for a successful task is downloading and interpreting the data.

In the context of user evaluation in Chapter 5 we use usefulness as a concept, rather than relevance because it has been shown that relevance judgements do not necessarily correlate with perceived usefulness and user satisfaction (Mao et al., 2016). This is partially due to the fact that external assessors of relevance might not fully understand the information need as it is not their own 'real' task, and partially due to the simplified environment in which the assessments of relevance usually takes place (Mao et al., 2016). We use 'usefulness' as a metric, based on Cole et al. (2009)'s and Belkin et al. (2009)'s proposal of the concept for evaluation in Interactive Information Retrieval. Usefulness is a more general concept than relevance, which is the basis for measuring information retrieval systems (Cole et al., 2009; Belkin et al., 2009). Literature has shown that people are able to give usefulness judgements as intuitive assessments without having to understand the term technically (Cole et al., 2009). Within this work we see usefulness as the high-level assessment of whether a dataset contributes to accomplishing the overall information task. Taking into account the iterative and incremental nature of data-centric work tasks, as described in Chapter 3, we believe that it is necessary to extend evaluation criteria for this type of information retrieval to accommodate the specific characteristics of working with data.

A concept related to relevance and usefulness judgements is 'data quality'. A dataset is not considered useful if it does not fulfil the quality requirements attached to the task. Data on the web contains inconsistencies, misrepresented as well as incomplete information. Several authors have discussed quality dimensions for data, pointing out that quality for data could be more accurately defined than that for documents. Data quality is commonly described as 'fitness for use' for a certain application or use case (Knight and Burn, 2005; Wang and Strong, 1996). Data quality may depend on various factors (dimensions or characteristics) such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability and verifiability (Wang and Strong, 1996). Zaveri et al. (2016) analysed quality

dimensions with a focus on Linked Data and defined 4 core dimensions: accuracy, completeness, consistency and timeliness. A number of quality metrics for structured data have been proposed in the literature, such as for instance metrics for correctness of facts, adequacy of semantic representation and the degree of coverage (Zaveri et al., 2016; Batini et al., 2009). Umbrich et al. (2015) discuss quality assessments of open data portals, focusing on available metadata. They point out that low metadata quality (or missing metadata) affects both the discovery and the consumption of the datasets. In that sense, the quality of metadata can be seen as one aspect of data quality, but includes in itself a number of different concepts (such as completeness, accuracy or openness; amongst others suggested by Umbrich et al. (2015)).

Data on the web will always contain inconsistencies and incomplete information. In the context of this work we are interested in how to represent information about the dataset to users so they can make an informed decision when selecting a dataset from search results. Therefore the specific importance of different dimensions of data quality, how these influence perceived usefulness in different tasks are of interest in this context.

2.4 The system and the publisher in data search

2.4.1 User interfaces for data search and exploration

The conceptual models that are provided through the user interface determine the cognitive representation of a systems features for a user (White and Roth, 2009). For our scenario that means that the interfaces of data search or exploration tools impact what data users can discover and understand (Churchill, 2012). We discuss how search results for data are presented currently on data portals, and present work on specialised data interfaces for exploration, including a number of examples from literature.

2.4.1.1 Specialised data interfaces

One of the primary ways to search for data on the web is through data portals, which are repositories of datasets. In Section 2.1.1 we discuss how search results for data are presented currently, by looking at the interface of one of the largest open data portals[10] on an exemplary basis.

Data search often starts on a data portal with an interface as depicted in Figure 2.5, in step 1. Upon entering their query, users are presented with a compact representation of the results (step 2):

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

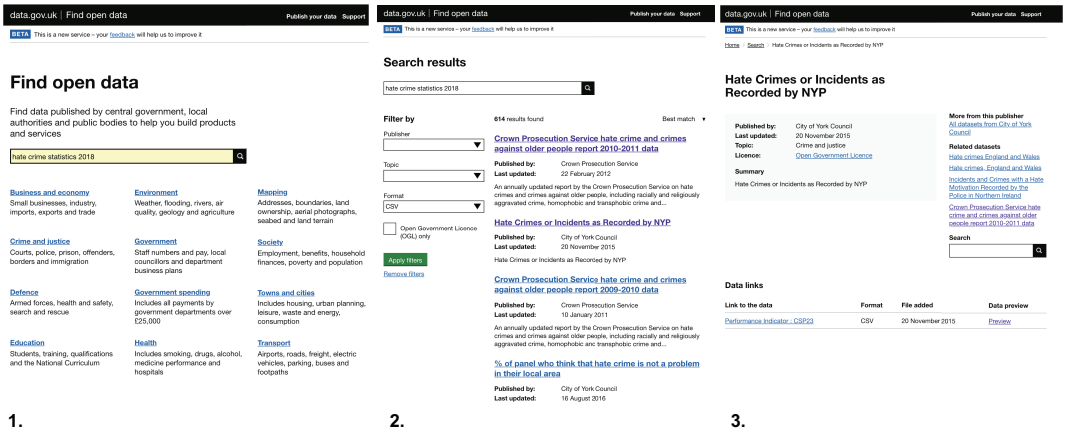


FIGURE 2.5: Search process on the UK governmental data portal

The results are displayed similar to web search, with a title and short snippet of text, following a traditional ten blue links paradigm. (This paradigm represents 10 search results in a list together with textual snippets of keywords from the document, and the corresponding URL of the associated search result (Hearst, 2009).) The interface consists of a standard query bar and a series of facets to further filter results. Additionally this includes for each dataset its metadata (title, publisher, publication date, format etc.). After selection of one the results to explore further, the user is commonly taken to a new page (see step 3), a dataset preview page with additional information, where the dataset can also be downloaded. 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. This can, in some cases, also include a data preview or a visualisation. Figure 2.5 shows this process for the UK governmental open data portal. Many data portals on the web offer a similar user interfaces. Data search results on general-purpose web search engines have a similar look and feel, although the hits are a mix of datasets and other types of sources. Interface design plays a key role in representing context (Greenberg, 2001), which we discuss in more detail in the following section.

In visualisations, data characteristics are assigned to visual cues such as position, size, shape or colour which provide users with a means to make sense of data (Card et al., 1999). To place data in context it needs to be situated in a conceptual space that illustrates relationships, events and contrasts that are meaningful in a field of practice (Woods et al., 2002). For instance Shneiderman (1996) proposes seven task types and seven data types as a type by task taxonomy of information visualisations. Tasks are: overview; zoom; filter; details-on-demand; relate; history; and extract. Data is categorised in: 1-; 2-; 3-dimensional data; temporal and multi-dimensional data; tree and network data. The data visualisation literature offers a variety of propositions and guidelines to facilitate sensemaking in visual data exploration tasks (e.g. Baker et al. (2009); Bertin (2010); Russell (2003)).

There can easily be too much information displayed on the screen, which can make the 'correct' results inaccessible (Woods et al., 2002). Yet reducing the amount of data displayed also potentially leads to problems – it is much harder to define what is useful to a user in data search than for traditional documents. There are simply more options and what might be relevant for a user depends on the task as well as the context (Woods, 1991). As one example of the discussion in literature, Woods et al. (2002) propose the creation of a coherent virtual perceptual field that aims at resembling objects and their relationships similar to a natural scene.

The problem of too many possible answers to a query has been discussed in the context of database search. Without knowing the data and having an understanding of the whole possible result set users struggle to formulate targeted queries (Liu and Jagadish, 2009a). Exploratory search interfaces have tried to tackle this issue by allowing users to interact with the result space as a whole, for instance Klouche et al. (2017) using a visual re-ranking approach on a relevance map. However, these approaches usually rely on document centred ranking strategies, not directly applicable to structured data.

Visual analytic tools have been explored by various authors – they support flexible visualisations to enable exploration by allowing the user to filter, sort and view different aspects of data, amongst other interactive options (Heer and Shneiderman, 2012; Keim et al., 2008; Zhang et al., 2012). Collaborative visualization tools provide for instance a shared network diagram (Balakrishnan et al., 2008) or a timeline (Ganoe et al., 2003) and facilitate additional communication through chats or comment threads (Heer et al., 2007), attentionaly ambient visualisations (Hajizadeh et al., 2013), and annotations (Ellis and Groth, 2004) that support users in building a shared mental model of the resource (Goyal and Fussell, 2016). In their paper describing Google fusion tables, a cloud-based data management and integration service, the authors report that map visualisations were found to be especially popular to display geospatial datasets (Gonzalez et al., 2010b).

Extensive work has been done about the design of search user interfaces for web pages or documents as the source of information (e.g. (Wilson, 2011)). Furthermore tools such as Tableau[11], Fusion tables (Gonzalez et al., 2010a) and Many Eyes (Viégas et al., 2007) are popular examples of services which support data exploration and analysis through visualisation (Morton et al., 2014). A large number of other, less popular, but more specialised interfaces to support the visual exploration of data have been developed.

Within this work we focus on understanding which content could facilitate an informed selection of a dataset. The visualisation of data can be helpful (amongst other types of presentation), however we still know little about how to present an overview of a dataset to users in a search scenario. Exploring the optimum representation of such content for

[11] https://www.tableau.com/

different data-centric tasks is outside the scope of this work, but would be an interesting direction of future research.

Examples of specialised interfaces and exploration systems for data.
In this section we discuss a number of approaches aiming to support sensemaking with data in an exploration context. We discuss those as they support users in understanding the content of a dataset or corpus of datasets which is relevant in a search scenario for data. One such approach was proposed by Marchionini et al. (2005) in their work on relation browsers for statistical information. These are interfaces using facets, containing sets of categories which are displayed graphically, so the user can view the information from different perspectives. Users can choose how the data is represented (e.g., a visualisation, spreadsheet, samples), which means that a wider set of people can potentially use the data. Relation browsers have been shown to work on some forms of structured data (e.g. for geospatial data), but have a limited number of possible facets that can be displayed (ibid.). Further research is needed to determine the applicability of the concept for different data-centric tasks.

One example of interactive visualisation research which aims to support sensemaking and information discovery is *Jigsaw*. This is a visual analytics system developed by Stasko et al. (2008), using entity recognition within texts and presents connections between entities across documents to the user. Although specifically developed for the exploration of documents it would be interesting to explore these principles for the exploration of datasets.

Another example is a prototype interface developed by Faisal et al. (2007) for the exploration of academic literature – Literature Knowledge Domain Visualizations (LKDViz). This was designed with the goal of providing multiple interactive features to explore and manipulate the visualisation according to personal preferences.

Spatial hypertext systems serve as another example of systems designed to support sensemaking. They include functions for finding and organising documents. The Garnet interface, developed for digital libraries, is an example which provides options to cluster documents in a spatial layout according to individual preferences and can so switch between supporting information seeking and structuring (Blandford and Attfield, 2010; Buchanan et al., 2004).

The above mentioned interfaces or systems are mostly designed for textual information, such as web pages or are designed to facilitate exploration of data in a specific environment. One technique specifically designed for visualising and making sense of data was presented by Pirolli and Rao, called *Tablelens*, which supports the interactive exploration of large tables (Pirolli and Rao, 1996). They take advantage of the inter-relatedness of table cells to present more content from a table in a single display supporting the human eye to easily detect patterns and correlations (Rao and Card, 1995). There are domain

specific systems and tools which support the exploration of data. For example in the financial domain the Bloomberg terminal[12] allows monitoring and analysing real time market data, among other functions, within a single system. However, this example of a relatively frictionless interaction is not yet realised for most other domains.

Morton et al. (2014) promote the development of data exploration tools that integrate three high level capabilities – visual and interactive data exploration, data enrichment through recommendation systems and data cleaning functionalities; to support the analysis of data in a so called visual analysis sensemaking cycle.

The interfaces and approaches mentioned in this section are not an exhaustive list, they are rather an exemplary overview of existing approaches and research in this area to illustrate the focus on textual documents and our limited knowledge of how to best support users through the design of the user interface in a search scenario.

2.4.2 Overviews of datasets

2.4.2.1 Search Results display

When searching on the web, we are used to being presented with a snippet, which is the short summarising text component that is returned by a search engine for each hit. This helps us make a decision about the relevance of the documents returned (Bando et al., 2010). Snippets adjust their content based on the user query to make selection more effective (Bando et al., 2010; Tombros et al., 1998). There are initial efforts that aim to do the same for dataset search (Au et al., 2016), but we are still very far from being able to provide the same user experience as in web search.

In information retrieval systems the user is usually presented with some kind of representation of the content of the search results. The page on which the results of a search activity are displayed is called SERP (Search Engines Results Page). In traditional IR the document snippet refers to the information that summarises the document and is a key part of the search interface (Baeza-Yates and Ribeiro-Neto, 2011). The information is usually a summary or surrogate of the information source and its quality effects the perceived relevance of the SERP and of the individual search results. We are used to a display of a short piece of text in document search, called the snippet, that is extracted from the document. These snippets can be influenced by the query as well as the personal search history of the user and commonly highlight the users query terms in the display. In specific types of search, such as over information collections (as in product search, search over scientific publications or music search) metadata are often displayed together with the search results, such as the publishing date or the author of the specific information piece (Baeza-Yates and Ribeiro-Neto, 2011).

[12] https://www.bloomberg.com/

Marchionini and White (2007) distinguish overviews, or 'surrogates', from metadata in that overviews are designed to support people to make sense of the information object before fully engaging with the object itself. Metadata can potentially serve a similar purpose, but is more often designed to support retrieval by the system and less targeted at human consumption, which is described in more detail in Section 2.4.3. As data search relies heavily on metadata (Noy et al., 2019) the displayed information either prompts the user to download the data or to search further – and is crucial in the interaction between people and data, potentially determining further engagement.

2.4.2.2 Textual dataset summaries

From a user's perspective, having a textual summary of the data can be very useful to get an overview of a dataset. Text is sometimes easier to digest than raw data or graphs (Law et al., 2005; van der Meulen et al., 2010), and richer in context than metadata. It can help people assess the relevance, usability and quality of a dataset for their own needs (Au et al., 2016; Bargmeyer and Gillman, 2000; Lehmberg et al., 2016; Tam et al., 2015). It can also improve data discovery, as search algorithms can match the text against keyword queries (Thomas et al., 2015).

In a review of related literature (Balatsoukas et al., 2009) concluded that textual metadata surrogates, if designed in a user-centred way, can help people identify relevant documents and increase accuracy and/or satisfaction with their relevance judgements. Several authors have shown that textual summaries perform better in decision making than graphs. For instance, Gatt et al. (2009) found in a evaluation of a system that summarises patient data that all users (doctors and nurses) perform better in decision making tasks after viewing a text summary with manually generated text versus a graph. These findings are confirmed by (Law et al., 2005; van der Meulen et al., 2010) in studies comparing textual and graphical descriptions of physiological data displayed to medical staff. Sultanum et al. (2018) emphasise the need to integrate textual summaries to get an overview of clinical documentation instead of relying on graphical representations. While we do not claim that text-based surface representations are superior to graphs, we believe they are, at a minimum, complementary to visualisations and accessible to a broad range of audiences, including less experienced users (Gatt et al., 2009; Sultanum et al., 2018).

We describe automatic and human summary creation in detail in Chapter 4, Section 4.2.

2.4.3 Metadata

When searching for data on the web we often rely on metadata. Most dataset search is built over metadata (as mentioned in Section 2.1.2), but, more relevant from a user

point of view, metadata is often the only representation of a dataset a user gets to see in order to select a dataset.

Metadata often uses domain-specific vocabularies and technical formats. Its creation and understanding is challenging (Mayernik, 2011; Edwards et al., 2011). High-level efforts for standardising data sharing practices, such as the FAIR data principles (Wilkinson et al., 2016) or the five stars of linked open data[13] promote efforts to increase findability and interoperability, amongst other aspects, to encourage the reuse of data.

Search engines can take advantage of standardised metadata about resources on the web; ideally metadata content would also refelect what users consider selection criteria for datasets. Several efforts have been undertaken to standardise metadata about datasets on the web. One of them is DCAT[14], an RDF vocabulary used to describe datasets on the web, to increase interoperability. It is for instance used by the CKAN platform, an open source data management system used by many governmental open data portals, and also indexed by Google dataset search (Noy et al., 2019). Primarily indexed by Google dataset search is currently Schema.org, an open vocabulary for structured data markup on the web (Guha et al., 2016). This means a set of attributes are suggested to describe resources on the web, amongst other things also for datasets[15]. However, to date we still know little on what metadata to capture to support a user's understanding of a dataset.

A related issue is the lack of concrete guidance for data publishers on how to produce meaningful descriptions, or summaries of datasets. Currently dataset summaries are created by people, often the data publishers, who might take metadata standards and community guidelines as a point of reference. Existing community guidelines for data sharing (which we discuss in more detail in Section 2.4.3), such as the W3C's Data on the Web Best Practices[16] or SharePSI focus on the machine readability of data. Textual descriptions are part of the standards, but guidelines for what should they contain are sparse.

This can be seen, for instance, in the W3C's Data on the Web Best Practices which is based on DCAT or, in a slightly different context, in the documentation of schema.org: a set of schemas for structured data markup on web pages. Instructions are formulated as follows:

> *(DCAT) description = free-text account of the dataset (rdfs:Literal)*
> *(schema.org) description = a description of the item (text).*

Existing metadata schemas might not be extensive or descriptive enough to support interaction patterns in dataset search. By investigating selection criteria in dataset search

[13] https://www.w3.org/DesignIssues/LinkedData.html
[14] https://www.w3.org/TR/vocab-dcat/
[15] http://schema.org/Dataset
[16] https://www.w3.org/TR/dwbp/

this work aims to increase our understanding of the importance of specific metadata attributes or to discover missing ones.

Dataset profiles.

Another related area of research aiming to provide overviews of datasets is data profiling. This refers to a wide range of methods to describe datasets, with a focus on their numerical or structural properties. Profiles can be merely descriptive or include analysis elements of a dataset (Naumann, 2014). Some approaches connect the dataset to other resources to add more context or to generate richer profiles, for example spatial or topical profiles (Fetahu et al., 2014; Shekhar et al., 2010). Most studies in this space work on datasets from a specific domain or on particular types of data such as graphs or databases. The result is not necessarily a human-readable text summary, but a reduced, higher-level version of the original dataset (Saint-Paul et al., 2005). Rich dataset profiles can potentially support the dataset retrieval process.

Chapter 3

Working with structured data

This chapter presents our scoping study, investigating information seeking behaviour for structured data. The results of this study can be used by data portal providers and designers of data discovery and exploration tools to help them understand what system capabilities their users need and appreciate, and how to design novel methods to describe, present and assess structured data.

3.1 Motivation

The objective of this study was to better understand peoples' interaction with structured data as part of their work. We examine the information seeking behaviour of people looking for new sources of structured data online, including the task context in which the data is going to be used, dataset search, and the identification of useful datasets from a range of possible candidates.

To study the requirements of our scenario outlined in Chapter 1, we followed a mixed-methods approach, combining semi-structured in-depth interviews that we analysed using thematic analysis, supported by a search log analysis from a large open governmental data portal.

As discussed in Chapter 1, this work is primarily motivated by the ubiquity of structured data on the web and its various applications, in particular in the context of initiatives such as Open Government Data (Ubaldi, 2013), which has already led to the publication of millions of structured datasets available on the web for everyone to access and reuse (Manyika et al., 2013).

The interviews focused on the information seeking activities of people who work with data as part of their daily jobs, such as scientists, data analysts, financial traders, IT developers, managers, and digital artists. We talked to 20 data practitioners in such

roles across a range of different domains. Based on self-reports, our participants were defined as people who worked with data in their jobs. However, that was not necessarily their only, or main work activity. The responses showed that people across different skill sets and professional backgrounds follow common workflows when engaging with structured data.

We envision this study to be useful in several ways: to the best of our knowledge, it is the first study of its kind that tries to shed light on how data practitioners look for data online, including a qualitative component with in-depth interviews and a quantitative analysis of a unique dataset that has not been explored so far. The resulting framework and guidelines will be used in our own work, but can also be used by open data portal providers, helping them understand what system capabilities their users need and appreciate, and how to design novel methods to automatically describe, present, assess, and rank structured data. At the same time, we believe that the findings of this study advance the field of Human Data Interaction by identifying areas for research and further improvement, in particular around a more rigorous evaluation of the user experience in data portals and in data search. In order to achieve that we can learn from related areas such as web search engines and user experience testing; improve the ranking and presentation of datasets matching a user's information need; as well as the automatic generation of rich (metadata) summaries for large datasets to help users inspect and understand them.

3.2 Methodology

This section outlines the methodology behind our study, addressing the following research questions:

- How do people currently search for data?
- What are the characteristics of information seeking tasks for data?
- How do people evaluate data that they find?
- What types of work tasks do people do with data?
- How do people explore data that they have found?

We used a mixed-methods approach, with a strong focus on the qualitative aspect to improve our understanding of how people work with data. To *expand* (Bryman, 2006) our findings around the specific question of dataset search we also analysed the search logs of a large open data portal quantitatively. We believe that using the logs alongside the qualitative data was meaningful for several reasons:

(1) the search logs were relevant to the study, as 17 out of 20 participants cited this specific portal as a tool they have used to search for datasets.
(2) the analysis of the search logs gave us a less obtrusive way to learn about the behaviour of data search users (Jansen, 2006), of which our interviewees are a subset.
(3) we used the quantitative insights only as a way to add more breadth to our enquiry, without making any claims about direct links between the two samples.

3.2.1 In-depth interviews

User-system interactions are influenced by factors that are not easily observable or measurable (Kelly, 2009). For this reason, and given the exploratory nature of this research, we focused the study on its qualitative element. We believe the rich data about interaction processes and workflows we were looking for was best provided by in-depth interviews.

3.2.1.1 Recruitment

Our sampling strategy was purposive to include a spread of sectors and a wide range of skill sets and roles. Participants were recruited via targeted emails and social media and asked to fill out an online scoping survey. Emails were sent to preselected relevant participants, namely members of the UKs governmental Open Data User Group (around 15 people). For social media recruitment we used the ODIHQ[1] account (at the time of the study over 31k followers, 5.3k impressions, 90 interactions, 20 retweets).

The survey helped us select the participants in a more targeted way. The criteria for inclusion were: people who are using data in their day-to-day work; and the diversity of domains, skills, and professional backgrounds. The scoping survey covered questions about the tools participants used, the type of tasks they performed with data, how often they interacted with new sources of data, and whether they searched for data. The survey can be seen in Appendix A.3. We tried to cover various domains and professional backgrounds to gain a broad overview and avoid unintended biases. The respondents identified as relevant at this stage were contacted via email to arrange a Skype or face-to-face interview of circa 45 minutes. We recruited 20 participants, who classified themselves as working with data. They are described below and listed in Table 3.1.

3.2.1.2 Description of participants

The sample consisted of $n = 20$ data professionals (17 male and 3 female). The majority were based in the UK ($n = 16$), with others working in Germany($n = 1$), USA($n = 1$),

[1] https://theodi.org/

France($n = 1$) or globally ($n = 1$). Their roles, as reported by the participants, are shown in Table 3.1.

P	*G*	*Role*	*Sector*
1	F	Crime and disorder data analyst	Public administration
2	M	Trainer for data journalists	Media & Entertainment
3	M	Data editor & journalist	Media & Entertainment
4	M	PhD researcher, social media analyst	Education
5	M	Senior research scientist	Technology & Telecoms
6	M	Data scientist	Technology & Telecoms
7	M	Lead technologist in data	Technology & Telecoms
8	M	Data consultant and publisher	Technology & Telecoms
9	M	Senior GIS analyst and course director	Geospatial/Mapping
10	M	Research and innovation manager	Public administration
11	M	Researcher	Transport & Logistics
12	M	Semantic Web PhD researcher	Science & Research
13	F	Project manager	Environment & Weather
14	M	Quantitative trader	Finance & Insurance
15	M	Data manager	Public administration
16	M	Head of data partnerships	Business Services
17	M	Lecturer in quant. human geography & Computation geographer	Science & Research
18	F	Data artist	Arts,Culture & Heritage
19	M	Associate professor	Health care
20	M	Business intelligence manager	Public administration

TABLE 3.1: Description of participants (P) with gender (G), their profession (Role) and sector they are working in (Sector)

They used both public (open) and proprietary data in different areas: environmental research, criminal intelligence, retail, social media, transport, education, geospatial information, smart cities, biological as well as financial data and sensor data. They reported to work with a large variety of common data formats.

Most interviewees stated that their tasks with data vary greatly and that the number of datasets they interact with – reportedly between two and 50 each week – fluctuates with the nature of their projects (mean of 13, median of 7). They reported acquiring the skills to work with data on the job ($n = 9$), from peers or self taught ($n = 15$), by *doing it or experimenting with data (P5)*, or through formal education ($n = 12$) (in particular on core pre-requisites such as the fundamentals of statistics) and professional training. All participants had searched for data for their work before, some on a daily basis ($n = 3$), others weekly ($n = 8$) and others less frequently. Half of the participants also reported using data that is new to them on a weekly basis ($n = 10$).

3.2.1.3 Data collection and analysis

We conducted semi-structured, in-depth interviews of circa 45 minutes, which were audio-recorded and subsequently transcribed. They were carried out via Skype or face-to-face for a period of six weeks in summer 2016.

The interviews were organised around the participants' data-centric tasks; the search for new data; and the evaluation and exploration of potentially relevant data sources, corresponding to the research questions. The interview schedule can be seen in Appendix A.4. They were analysed using thematic analysis (Robson and McCartan, 2016) using Nvivo, a qualitative data analysis package for coding. We applied two layers of coding to be able to look into the data at different levels of generality and from different viewpoints. As a primary layer, we used deductive categories mapping to stages of the data interaction process:

- *Tasks:* what people use data for
- *Search:* how and where to search
- *Evaluate:* what information they need about the data to decide what to select and whether the search results were useful
- *Explore:* how they explore and understand datasets

For each of these themes we applied a second layer of coding, in which we used an inductive approach (Campbell et al., 2013) to draw out further details. The resulting themes were then used to further categorise tasks and user needs, as can be seen in the findings. The coding was done by the author of this work, but to enhance reliability two senior researchers checked the analysis for a sample of the data.

3.2.2 Search logs

Search engines are routinely evaluated from a user's perspective, for example by studying click behaviour and eye tracking, as well as from the system perspective by using metrics such as precision and recall (Cleverdon et al., 1966). As pointed out earlier, most literature is centered around documents (news articles, web pages, text, etc.). State-of-the-art search evaluation metrics incorporate behavioural characteristics such as time spent on a page, clickthrough streams, and user activities such as tags, printing, and purchasing (Fox et al., 2005).

However, these metrics are not straightforward to apply when the results returned by the query are datasets. Some work has been done in understanding structured queries against online databases such as public endpoints of the Linked Open Data Cloud (Beek et al., 2014; Morsey et al., 2011). However, their aim was to understand which parts of

a dataset are more popular with users or how a relatively complex query language such as SPARQL is used, and not the interaction between users and structured data.

In this study, we take first steps in this direction. To supplement the in-depth interviews, we analyse the logs of a large open government data portal. This offers us a non-intrusive way to learn about the behaviour of data search users (Jansen, 2006).

We had access to the search logs generated via Google Analytics (GA) between 1st May 2015 and 30th April 2016 with a total of 100,970 queries, of which 52,824 were unique queries. These were identified by the fingerprint method provided by Open Refine (Pressac, 2016), as GA clusters only identical queries (The fingerprint method is more advanced). It first formats all query strings – changing all characters to lowercase, removing spaces and similar, and ordering query terms in alphabetical order – and then uses this structure to decide which queries are considered identical (Pressac, 2016)).

Over 80% of all portal users were identified by GA as being from the UK, with just over 26% of them users being located in London (GA collects this type of information). These users were responsible for 577,310 dataset downloads in the given time frame. 9.26% of users searched on the site from a mobile device (including tablets) and the rest from a desktop environment (90.74%). In similar web search statistics, more than half of all searches are made from a mobile environment (Sterling, 2015).

Google Chrome was the dominant desktop browser used to access the portal (50%) of users, 27% used Internet Explorer, 11% Firefox and Safari with 10%. Again, these results differ from web search, where Chrome is used in more than 70% (W3Schools.com, 2016). Governmental portals like `data.gov.uk` may be more popular in certain professions, for example with civil servants, who often have restrictions on the usage of new technology; this could be the reason for the high percentage of Internet Explorer users. However, the other differences we observed hint at areas that data tools and technologies providers might want to explore further.

In our analysis we considered the following data, which was readily available in the search log sample:

- (1) *Landing pages* - pages through which visitors entered the site
- (2) *Sessions* - the period of time a user is actively engaged with the page
- (3) *Exits* - how often users leave a page after viewing it
- (4) *Search refinements* - the total number of times a query is refined within a session
- (5) *Time after search* - the amount of time visitors spent on a site after getting the results of their query
- (6) *Search depth* - the number of pages visitors viewed

3.2.3 Ethics

The interview study was approved by the University of Southampton Ethical Advisory Committee (ERGO number 21909). Informed written consent was given by the participants. Access to Google Analytics to obtain data used for the search log analysis was kindly given to us by the UK governmental open data portal for research purposes. No personal information was used for the search logs analysis.

3.3 Findings

In this section we present the findings of our study, which are structured around the themes used in the interviews: tasks, search, evaluation and exploration.

3.3.1 Data-centric tasks

3.3.1.1 Taxonomy of data-centric tasks

When asked about their activities with data, participants reported a wide range of tasks[2], spanning from statistical analysis to using data as a material to create something – be that a service, a tool, or an artwork.

We categorised these activities into two broad categories:

- (1) *Process-oriented tasks* – people think of these tasks as doing something transformative with data
- (2) *Goal-oriented tasks* – people think of data as a means to an end.

Process-oriented tasks include: building tools with data; integrating data in a database; defining predictive statistical models; using data in machine learning processes, producing data; publishing data; visualising it.
Goal-oriented ones include: seeking the answer to a question; comparing datasets or data points; finding patterns; allocating or managing resources in a data-informed way. Table 5.2 presents examples of both task types. Most participants reported to have engaged in both process- and goal-oriented tasks at some point.

While the boundaries between the two categories are somewhat fluid, the primary difference between them lies in the 'user information needs' – that is, the details people need to know about the data in order to interact with it effectively. For process-oriented

[2]For the purpose of anonymity we do not include direct quotes describing participants' tasks; instead we present descriptions of the tasks in this section.

tasks, aspects such as timeliness, licences, updates, quality, methods of data collection and provenance were reported to have a high priority. For goal-oriented tasks, intrinsic qualities of data such as coverage and granularity were mentioned to play a bigger role.

Goal oriented tasks	*Process oriented tasks*
looking for an answer	publishing
identifying trends	linking
detecting events	visualising
comparing	mapping
establishing relations	input in statistical models
credibility analysis	building a tool
decision making	playing
evaluating interventions	producing
understanding change	integrating in e.g. a database
finding a story	processing

TABLE 3.2: Data-centric tasks as reported by the participants in this study

Another way to categorise tasks is based on the specific activities that involve data. Based on feedback from the participants, we identified five activities: (1) *linking*; (2) *analysing*; (3) *summarising*; (4) *presenting*; and (5) *exporting*.

These categories are not exclusive or exhaustive and the participants reported undertaking one or several of them, both in a process- and goal-oriented task context. However, as we will see later in the discussion, we believe they lead to different interaction requirements on a system such as a data catalogue or search engine.

Linking ($n = 14$) is about finding commonalities and differences between two or more datasets. From an interaction point of view, this requires capabilities to view datasets next to each other and be able to spot meaningful relationships. The other classes of activities usually concern only one dataset or datasets that have already been linked. In *analysis* tasks our participants mostly reported time series analysis ($n = 10$). In these tasks data is ordered by time; the aim is to identify trends, or detect and predict events. For instance, Chatfield (2016) looks at ways to carry out such tasks effectively. *Summarising* ($n = 11$) involves creating a more compact, meaningful representation of the data. This could be used to inspect a dataset or as a means to tell a story with data. As elaborated in the next section, data summaries raise questions related to the types of information that is useful in this context, the best approaches to obtain or generate them and the related user experience. *Presenting* ($n = 9$) includes activities that transform data into human-friendly formats, such as visualising them or producing textual descriptions of the data. Finally, *exporting* ($n = 11$) refers to all aspects around producing and publishing a dataset in a given format, including metadata.

3.3.1.2 Complex tasks

A common scenario reported by all interviewees was trying to answer high-level questions with data to explain change, establish causalities or understand behaviour. This is characteristic for the class of goal-oriented tasks introduced earlier. Gaining a better understanding of the problem area and adding different sources together gets people closer to answering their question. This often means breaking it down into more manageable sub-questions, which only become clear after using a particular dataset.

The range of activities can be very broad, including all five types discussed in our taxonomy. Examples are understanding why crime in a specific area has risen over the last 10 years or comparing user experience on the internet for people of differing socioeconomic status by considering multiple factors. Participants talked about their journey when searching for data in such cases, which often leads them from one dataset to the next. The following example illustrates this process: one participant discussed a project that aimed to create a map showing the locations of defibrillators in a geographic area, to identify where new defibrillators might be needed. He considered demographic data, as certain age groups are more prone to heart attacks. Furthermore, he tried to consider not only where that population lived, but also where they spend their time during the day. A meaningful interaction with a data catalogue such as data.gov.uk would allow data publishers and users to create links between such related datasets to facilitate browsing and exploration. This could include visualisation of links between datasets to aid discovery, or recommendations of datasets based on content or structure of the dataset.

3.3.1.3 The impact of data quality on task outcomes

Another aspect that emerged from the interviews was the use of data to increase accountability in decision-making. However, concerns were raised on the role of data quality in the process. It was clear from the interviews that people are often aware that they are not working with the ideal data for a particular task.

> *(P15) So although the data is a good quality, it's not really designed for my purpose so actually, for my purposes, there's quite a lot of uncertainty and quite a lot of risk in that, it's still the best data we have but it's that knowledge of how it was, why it was created, against how I want to use it.*

While this is a well-known challenge, the majority of the participants agreed that, as long as people are aware of the limitations of their data, and these limitations are clearly communicated (e.g., through documentation), it can be factored into decisions being made based on that data. The following comment is indicative of the views of the participants on the matter:

(P15) In the relationship between the data that's available and the decision that's made, we're saying this is the data we've used to make this decision, we're aware that it's not the best data but it's still the best we have so there is uncertainty there but are you happy with that level of uncertainty?

For data providers and the publishing platforms they use, this emphasises the importance of data governance and documentation, but also the need to consider different approaches to data quality and its context specific nature. This can extend to assessing a given quality dimension automatically, as well as the presentation of reviews and other annotations together with a dataset. We describe the different aspects considered relevant to assess the quality of data in Section 3.3.3 (Table 3.3.)

Reusing data in new contexts.

One participant discussed the idea of data as a material that can simply be shown in different forms and so reveal its content, or one that could also be used to create new things. Data in that context was used to *'expose invisible rhythms and invisible systems that are beyond the human sensory capacity (P18)'.* For some people, data is interesting in and of itself and not necessarily used as a means to an end. This could be seen as a process-oriented task which can focus on presenting as well as summarising.

At the same time, participants reported using data in a variety of different contexts, which has implications on the most effective ways to publish it to support wide reuse:

(P18) If data is a material - what can you do with it? Implicit in that is the idea that data can be used for anything, cause its really flexible [..] The things that can come out of the same set of data can be really different.

Participants reported to refine their data tasks during the information seeking activity, based on the specifics of the datasets they find, a characteristic of exploratory search (White and Roth, 2009).

3.3.2 Search for structured data

In this section we discuss how people search for datasets. Many participants (n = 17) had to regularly search for data in their work. All of these reported experiencing difficulties finding data in the past. Others ($n = 3$) were provided with either external or internally produced datasets or were mostly involved in collecting or publishing data rather than using someone else's. However, all participants had previously tried to search for data online when they did not know whether it existed or not. They reported using: web search engines (e.g. Google); bespoke websites (e.g,. portals, catalogs); recommendations from other people; and Freedom Of Information requests (FOI). More

than half ($n = 12$) of the interviewed practitioners said that they regularly struggle when trying to find the data they are looking for.

3.3.2.1 Searching on the web

A majority of participants ($n = 18$) reported often using Google to find data online, especially when they do not know which institution holds the specific dataset they are interested in. Some reported always using Google as their first search strategy ($n = 7$). When searching for datasets, people most commonly employ a keyword search that is slightly adapted towards data search. Most participants ($n = 17$) mentioned terms such as 'data', 'statistics', 'dataset', alongside keywords describing other aspects of the data, in particular its domain. Some of the respondents ($n = 7$) reported preferring using fewer keywords to make sure that the search engine returns a broad range of (perhaps less accurate) results, which they then filter manually.

3.3.2.2 Searching on portals

This section includes insights gained from interviews and analysis of the search logs. While 17 of the twenty people interviewed use that particular portal, the two samples do not directly compare. However, there is a clear overlap between our qualitative findings and the user behaviour represented in the log data. The numbers presented here are based on the search log analysis, as explained in Section 3.2.

Queries.

Only $2,462$ queries out of a total of $100,970$ contained the terms such as 'data' or 'dataset', which were mentioned by interviewees. This could point to the importance of specialised data search tools, which allow practitioners to use the same search strategies they are used to from the web. Furthermore, the search box on the portal was labelled *'Search for data'*. However, the format of the queries had other notable features, which hint at the qualities and characteristics of data that count for open data portal users.

555 (0.55%) queries mentioned a data format - we looked for all structured data formats supported by data.gov.uk: HTML, CSV, WMS, WCS, XLS, JSON, WFS, Esri REST.

5384 queries used a category offered on the portal (5.33%), which were environment, town, cities, city, mapping, government, society, health, governmental spendings, education, business, economy, transport, while a slightly larger share of $8,389$ (8.31%) queries contained a location. The latter was determined through keyword matching with lists of towns, counties, regions and countries, which we extracted manually from a subset of those queries issued at least 15 times. Such strategies were mentioned by some of the participants ($n = 8$) in our interview sample:

(P7) It was somewhat hierarchical [..], like 'transport' would be the sector and then I'd put in a specific thing that I'm looking for which could be in the summary or metadata, so 'driving licences' and then in this case I included a particular feature that I wanted in the dataset as well, so it was 'gender'.

The average query length in the logs was 2.44 words. This number refers to 99.9% of all queries – we excluded some outliers because of their length (more than 15 words) and popularity (query frequency less than 10). This length of query matches more or less the situation on the web in 1998 - as search technologies improved, people started writing more detailed queries (Taghavi et al., 2012).

The participants in the interviews noted using a similar approach for queries via Google. Furthermore, we found 22.09% of queries issued on the portal were subsequently refined, while 80% of sessions with search had only a single query. The majority of users do not refine their queries - either they found what they were looking for, or they are not confident the system can give them the desired result.

Sessions.

More than half of the search sessions (52.08%) were done by people who landed on the site through Google. Just under a quarter (24.26%) visited the site directly. This backs up some of the comments we received about people preferring to start their search on Google. We believe this may be due to people not knowing that the portal exists before they search; people having the link to a particular dataset preview page already; or a perceived poor search capability of the portal.

As noted earlier, the majority of interviewees ($n = 17$) mentioned having experienced difficulties when searching for data on the web either because the data was indeed unavailable or not easy to find with the existing tool support:

(P17) I often find on most of the websites that provide data [..] often the search function is pretty rubbish, so that's why I often find myself falling back to Google [..] The difference [from searching for web pages] is that it just takes a hell of a lot longer in my experience

In addition, participants ($n = 16$) described finding data as a complex, iterative, process:

(P17) I would get some things that looked really promising but weren't and then finally, through some kind of mysterious combination of search terms, I suddenly came across the dataset I'd been looking for the entire time

More than half of the search sessions were done by new users. The data contained 58.88% (122,822) new users and 41.12% (85,774) returning ones. Returning users showed a longer session duration (around 10 minutes versus six for newcomers) – we assume those who came back were more familiar with the site content and engaged more with the portal. The time spent on site after search was on average 3.03 minutes and the number of pages viewed after getting results for their query were around 2.39. Just under 23% of users issued a query and left the site immediately after retrieving their search results. From the interviews, we have learned of the complexities of the subsequent steps in evaluating, exploring and eventually deciding to use the data. We assume that the users exiting the site just after landing on the search results page were likely to do so because they could not find what they were looking for:

> *(P5) I think I would fall back to like a Google search which is more broad and has a lot better coverage, that someone else would have, I would find a page where someone would have discussed it, mentioned the corresponding dataset.*
>
> *(P10) I have to do an awful lot of either filtering or sorting through pages to find what I'm looking for.*

3.3.2.3 Human recommendation and FOI requests

The other two approaches reported by participants included asking colleagues or directly asking for data from institutions which are likely to hold it, for instance via Freedom of Information (FOI) requests ($n = 3$): requests to a public sector organisation to disclose a particular type of data, by virtue of the The Freedom of Information Act (GOV.UK, 2014).

Obtaining recommendations from colleagues or people working in the respective field who are likely to know about a dataset was very common ($n = 14$). The majority ($n = 15$) reported that they ask other people for data, either in their immediate environment or in public institutions:

> *(P13) I'm phoning people and asking them if they could give us their data.*
>
> *(P15) So generally I tend to go through contacts rather than searches.*

Another factor that sometimes affects the user experience ($n = 3$), especially in professions which work with official statistics, is the inability to know whether the data they are searching for actually exists or is simply difficult to find. FOI requests were deemed useful to handle such uncertainties.

> *(P3) [..] a lot of our articles are FOI based, that's where we get the data from [..] If we actively put out an FOI knowing what we want, we know the data's going to be good because we know what to ask for, we know how to get the information we want.*

3.3.3 Evaluation and exploration

Once some data of interest is found, participants ($n = 9$) reported spending a considerable amount of effort deciding whether to use it. They reported on a process consisting of two broad stages, each using different methods and data features: (1) a *first look* at the data to get an initial impression ($n = 19$ of participants); and (2) a subsequent more thorough *exploration*. In each of the two stages people reported to use slightly different types of information about the data in question.

Data properties.

Participants ($n = 16$) reported using different aspects of a dataset in order to decide whether it is useful for them. We grouped them into three categories, as listed in Table 3.3: (1) *relevance* to the task; (2) *usability*; and (3) *quality*.

Assess	*Information needed about*
Relevance	context, coverage, original purpose, granularity, summary, time frame
Usability	labelling, documentation of variables, license, access, machine readability, language used, format, schema, ability to share
Quality	collection methods, provenance, consistency of formatting / labelling, completeness, what has been excluded

TABLE 3.3: Information needs when selecting datasets

A number of participants ($n = 9$) raised issues around finding a relevant and usable dataset, e.g., the data might be relevant, but aggregated to an extent that it hides levels of specificity that they were hoping to obtain from it. For example, averaging deprivation of an area might not show pockets of deprivation within that area.

Another common theme ($n = 15$) was the struggle with inconsistencies of labelling and a lack of documentation, often even within the same institutions. An example was using multiple datasets because the organisational structure of the local authorities in a country, mean they end up publishing the same data for their respective regions differently. An interviewee noted:

> *(P6) Documentation is most frustrating, there's often data without documentation and fishing for this information is the hardest bit.*

Information about collection methods that would lead to a better understanding of the data was repeatedly mentioned ($n = 11$). Knowing more about the choices that were

made in the collection and initial processing was said to enable data practitioners to make a judgment of the impact these core data science activities had on their own analysis of the data and thus help mitigate the risk and uncertainty associated with reusing someone else's data:

> *(P10) I spend an awful lot of my time trying to match data with other people's data and in doing that, you may be spend more of the time researching the data than actually using the data*

One participant talked about difficulties in finding fully representative data. He explained that depending on what the data is about, the accuracy of the data can vary. One example was given about data relating to the epidemiology of snake bites – snake bites are much more likely to be recorded in urban areas then in rural areas. Therefore there is a paucity of data in rural areas and while there is data for urban areas, it is not necessarily representative of reality:

(P19) In Brasil, where there is data there are no snakes and where there are snakes there is no data.

The first look.

We compiled a list of actions participants reported in Table 3.4. People make judgments about whether the data is what they expected. They want to evaluate quality, establish trustworthiness and 'understand' the data. Many described *trying to get to know the data* ($n = 16$), by doing a basic visual screening and a resulting sense of relationship with the data they are using:

> *(P1) [..] look at the column headers and maybe literally read the first two rows of data just to get an idea of what's actually in it.*
>
> *(P10) [..] looking at the nuts and bolts of the very basic concepts of what that data is.*

First steps of data exploration:
scrolling through the data / basic visual scan
looking at headers
looking for obvious errors
looking at summarising statistics
filtering relevant data (pivot tables)
visualising to see trends / peaks
looking at the schema / format
understanding the semantics of the data
looking at documentation / metadata
checking the publication date and provenance
trying to understand the purpose of the original dataset

TABLE 3.4: First activities when engaging with a new dataset, as reported by participants

At this initial stage, people also reported looking for obvious errors such as missing data, inconsistency, out-of-range values or duplicated unique identifiers:

> *(P8) I'm looking at [..] the coverage of the data, so does it cover the geographic area I'm interested in? Or the time period that I'm interested in? And does it do that to the level of detail I need?*
>
> *(P16) Once I've had a quick scroll across, I would then probably look at a couple of records right the way across to see if I can make sense of it*

The majority of participants ($n = 18$) had experience of working with datasets they have little information about and were aware of its challenges:

> *(P17) There are a lot of people who radically underestimate the amount of work that needs to go into understanding the data in the first place, before you can even start doing research*

Exploration.

It emerged that people build a notion of quality and trust of the data during exploration, while being aware of the limitations of the methods they use to assess both:

> *(P14) Scrolling through the bottom right corner of Excel and going 'yeah, that's a lot of data', you haven't done anything. It doesn't actually tell you anything but I would*

definitely do that.

(P6) It's very difficult first when you download new data, to have a quick idea of what the data represents, a quick summary of the data.

Quality was reported to be associated with *clean* data ($n = 6$). However, several interviewees also mentioned that cleaning data can lead to deleting information that would be relevant for a different task. This is illustrated by one participant discussing a dataset showing the occupancy of local car parks. Due to malfunctioning sensors the car park can appear to be over-full. This can be easily checked (and corrected) by making sure the capacity of the car park is never more than 100%. However, if somebody wanted to use the data to understand the reliability of the sensors, these corrections would remove all relevant data. As discussed earlier, a preferred solution for such challenges is to have access to information about how data was collected and its initial purpose.

(P17) Helping people to understand what's in the data is incredibly useful and also what has been excluded from the data because obviously, a lot of the time, data gets cleaned before it gets put online for other people to work with so what were the cleaning choices? What constitutes a valid record? How did you filter out bad records?

Another critical dimension discussed by the participants was exploration tool support. After finding what looks like a promising dataset, people ($n = 12$) said they often have to download it to establish whether it is actually the data that they were looking for, due to poor description of the dataset:

(P17) Even when I've found what looks like the right dataset, I still have to download it and look at it because [..] they often give you a preview of the first few rows and that's like a nice starting point, although it doesn't deal very well with if you've got 24 tabs.

When discussing barriers when searching for data one of the main issues mentioned by most participants was the inability to judge whether a dataset is useful before downloading and exploring it. Participants discussed the concept of dataset descriptions (or summaries) that would help them assess its potential quickly.

(P7) I would usually expect some sort of summary description or meta data that goes with it so hopefully some of that and I would read the content in that human legible format.

(P5) Some high level description of the data [..] like a Read Me file that I would expect to accompany the data and then I try to understand the format because sometimes,

it seems to be simple but it's more complicated than what it looks like, there may be missing columns or there may be some issues and I try to understand what's the meaning of the fields, its structure..

In the financial sector, for instance, data is stored in binary form, but can be easily queried and displayed in human-readable formats. Similar tools that allow users to more easily filter relevant parts of a dataset would make working with data more efficient:

(P14) Honestly, if I need data, if I need a data dump in my current job, Bloomberg has a plug-in for Excel and I can get almost everything I want from there.

(P5) I think it's important to have tools that you let you explore the data in a very automatic way because that's a repetitive task, it will come up again and again.

3.4 Discussion of findings

In this section we elaborate on the implications of this study. We propose a framework that describes the interaction with structured data alongside design recommendations for data publishers and designers of data platforms such as catalogues and marketplaces.

3.4.1 Framework for Human Structured-Data Interaction

Based on the information seeking models and the data science process (e.g. (Kuhlthau, 2004; Pfister and Blitzstein, 2015) described in the background and on insights gained during this study, we understand the process of working with structured data as a five-pillars model: *(1) tasks*; *(2) search*; *(3) evaluation*; *(4) exploration*; and *(5) use*. The process is iterative by design and can involve returning to previous activities at any point - for example, the results of the exploratory data analysis (pillar 4) may lead the 'data workers' (Cattaneo et al., 2015) to consider evaluating other sources of data (pillar 3), or start a new search attempt (pillar 2). We characterised each pillar based on the descriptions of the participants by defining common categories for data-centric tasks (pillar 1) and identifying which data properties (both intrinsic, such as data attributes, granularity and errors; and extrinsic such as metadata, e.g., release date and licenses) people consider relevant for the subsequent three pillars (search, evaluation and exploration). A detailed description of pillar 3 *'evaluation'* can be seen in Table 3.3 in Section 3.3.3.

Similar to Yi et al. (2007), who introduced a taxonomy of tasks in information visualisation, we believe it is critical for system designers in data search to identify user task types in detail, and tailor functionalities accordingly. For this reason we proposed earlier

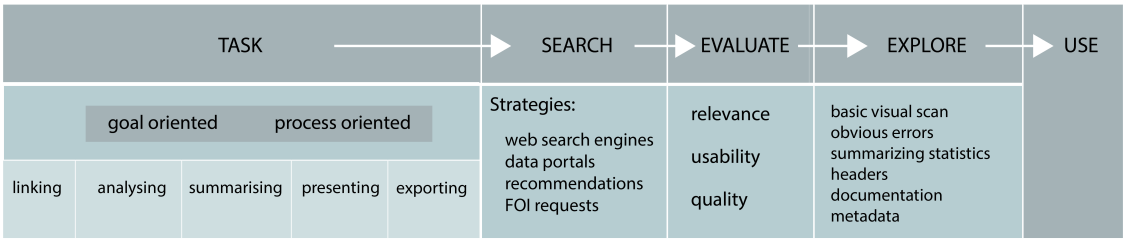


FIGURE 3.1: Framework for interacting with structured data

a taxonomy of data-centric tasks with two dimensions - one categorises tasks as either *process* or *goal-oriented*, the other differentiates between five core types of activities: *linking*; *analysing*; *summarising*; *presenting*; and *exporting*. This taxonomy, and the broader framework it is part of, are aimed to help system designers and publishers of data understand what people do when searching for and engaging with datasets and inform the decisions they make.

For an overview of the workflow through a search for data the framework in Figure 6.1 can be read from left to right, taking into account that this is usually an iterative process. This can be of interest for researchers e.g. in the area of information seeking, interested in the specificities of data search. It can be used as guidance to structure training for people learning how to find and use data to define how they should be lead through the process. Further, individual pillars can be prioritised, and the framework can be used to identify the area of interest within the workflow. For data portals the focus is likely to be Pillars 2 and 3 as they might concentrate on the data discovery aspect. Ranking will take into account those metadata attributes that people perceive as relevant for their decisions to use or ignore a dataset. For data publishers Pillar 3 can be of particular importance - for example to refine their metadata vocabulary to cover those bits of information that people need in their assessment. If someone is designing data exploration tools Pillars 1 and 4 can be of influence for their conceptualisation of user needs. A data catalogue designer will prioritise features such as data summaries and interlinking of datasets higher, as these activities are both common and important for data practitioners (Pillar 4), and will consider implementing social sharing and recommendation features (Pillar 2).

From interviews and the search logs analysis we learned that people experience difficulties finding datasets and that the information they need to evaluate their fitness of use is not always available or easy to interpret out of context. We found that participants expressed a need for more information about datasets before downloading them, such as for high level dataset summaries. Looking for data on the web emerged to be more often than not an exploratory search task, involving iterations and complex cognitive processing. In her work on the information seeking process, Kulthau concludes that uncertainty in the exploration stage indicates a space for system designers and intermediaries to intervene (Kuhlthau, 2004).

In the following section, we propose design recommendations for data discovery and exploration tools, as well as for data publishers and providers. We do not claim these to be exhaustive, or equally applicable to all types of data across our taxonomy of tasks and appreciate that in some instances they confirm insights from existing literature. However, this initial study strongly suggests that this space is by no means standardised and the user experience when interacting with the web of data leaves room for major improvements.

3.4.2 Design recommendations

Data portals and data publishers.

We established that in search users need to be supported in evaluating datasets according to their relevance, usability and quality. This could be achieved by providing *visual or textual indicators* of these aspects on the interface, backed up by automatically computed metrics or user-generated reviews and annotations. An interesting direction of future work would be to understand how all this additional information should be spread out across SERP, the dataset preview page and the dataset itself, to avoid overload.

As observed in our findings, data catalogues and similar platforms should allow *additional filtering* for the following types of information: location, provenance, format, licence, time frame and date, publishing date, location of publication and data schema. These filters would allow the user to direct the search process towards more desirable results. In addition, the search capability should be equipped to recognise *specific types of keywords*, such as dates to optimise the accuracy of their results. Providing this information together with search results would also be helpful for time series analysis tasks.

To support relevance assessments, we recommend also displaying information about the *granularity of the data*. One approach would be to display *headers, summarising statistics or textual summaries as well as previews of the data*, all of which could be provided alongside the search results. Furthermore, having more details on *how the original data was collected* emerged to be critical. While such reproducibility concerns are more common in some data-intensive fields (e.g., science, official statistics), it seems they always aid users develop a notion of trust in the data. While we appreciate that some dimensions of quality cannot be easily calculated automatically, hence creating an overhead for the data publisher in regards to documentation, there are is still room for some easy fixes such as *detecting empty fields and headers* in CSV files, as supported by tools such as Good Tables (International, 2016). This would improve the search experience for users across all types of tasks, in particular for data interlinking and exports.

We noted in our findings that many tasks people perform with data are complex and often span over multiple datasets. Being able to *identify and explore the links* between these datasets would help the user get a better understanding of availability and context. Approaches such as *Linked Data*, alongside tools that would discover relationships between datasets (perhaps available in different formats) would be a very useful addition to the services already offered by data platforms.

When designing tools for data exploration we recommend taking the different types of data-centric tasks into account. Our findings suggest that these result in specific requirements on functionality which could be explored further. For instance, *linking* requires tools which allow the comparison of two or more datasets. These should be able to highlight common attributes or visualise datasets side by side to facilitate understanding. Tools which support displaying data in different forms, such as creating interactive visualisations, textual descriptions of the data, or highlighting patterns would be beneficial for *presenting* and *summarising* tasks. For the latter, summarising statistics and representative samples as a preview of the data would be beneficial in allowing a *bird-eye view* ($P5$).

One approach for data sensemaking was proposed by Marchionini et al. (2005) in their work on relation browsers for statistical information. These are interfaces using facets, containing sets of categories which are displayed graphically, so the user can view the information from different perspectives. Users can choose how the data is represented (e.g., a visualisation, spreadsheet, samples), which means that a wider set of people can potentially use the data. Relation browsers have been shown to work on some forms of structured data, but have a limited number of possible facets. Further research is needed to determine the applicability of the concept for different data-centric tasks.

We learned that people are very interested in a history of the data. That means information about provenance, the processes of data collection, as well as subsequent choices made regarding, for instance, normalisation and cleaning. This is information that could be provided as documentation, capturing the legacy of the data, but also its reuse footprint, which could be displayed in a versioning system. We believe this could enable users who might struggle understanding an uncleaned version of the dataset to use a version which somebody else has already worked on. This would contribute towards the inclusiveness of data interaction tools.

Organisations publishing data should aim to support users better in evaluating the data according to its relevance, usability and quality in the context of a particular task. As mentioned earlier in this section, we recommend specific types of information to be made available and stored with the data. Collecting and managing this data should become a core part of the data governance process and the related overhead could be reduced by allowing for *manual annotations and additions* from users, *auto-generating metadata, dataset summaries or specific indicators* where possible and *cross-validation.*

For example, if a dataset is updated regularly, these updates could be matched with previous versions to provide consistent and machine-readable metadata.

3.5 Limitations

Every study, no matter how well it might have been conducted, has limitations. For the interviews, the majority of participants were male ($n = 17$) and working in the UK ($n = 16$). We interviewed a particular type of professional, though working in wide range of sectors and roles. As ours was meant as an initial study into engaging with structured data online, having a large number of participants per sector was less of a priority and some sectors and roles were not covered (Cattaneo et al., 2015). However, we were able to find common themes easily in the responses, which supports us in our belief that our sample size reached a saturation which allowed us to get a rich understanding of peoples' interactions with data (Green and Thorogood, 2013).

Regarding our findings concerning search, we expanded on the breadth of our study by including a second source of data based on a much larger sample. While we did not make any assumption about direct links between the interview and search log samples, the quantitative analysis in itself is novel and representative for a large category of tools people use to find datasets. However, search log analysis also comes with natural limitations: a biased sample of users from a single platform, albeit an important one, with a focus on public sector data, using keyword search (and not, for instance, browsing) to find the data that is right for them.

Finally, we recognise the problem with only using one author to code and interpret the data; the reliability and diversity of themes might have been different with more researchers. While we could not run any inter-rater reliability tests, we had two senior researchers overseeing the creation of themes in a sample of the data.

3.6 Summary

In this chapter we described a study that investigated how people engage with data in their daily work. By conducting in-depth interviews with data practitioners, we were able to obtain a better understanding of data-centric tasks, as well as about their search, evaluation and exploration strategies and the data qualities that influence these activities. Finding relevant data has emerged as particularly challenging, mostly due to the poor tool support and uncertainties around the availability of a given dataset for public use. To gain a better understanding of the requirements for better data search, we looked at search logs from within one of the largest open government data portal. Our findings and the design framework and recommendations that followed them can

inform the development of methods, interfaces and interaction models for core activities such as data search, evaluation and exploration. They can encourage the usage of data by people with different skill sets, for the variety of data-centric activities. Thus this study can be seen as a step towards understanding what usability of data interaction tools means.

We see a large space for future studies extending the current understanding of how people interact with data, which we discuss in detail in Chapter 6.

Chapter 4

Dataset summaries

In this chapter we report two complementary studies: an analysis of data-search diaries from 69 students, which offers insight into the information needs of people searching for data; and a larger summarisation study, with a lab and a crowdsourcing component with overall 80 participants, who produced summaries for 25 datasets. In each study we carried out a qualitative analysis to identify key themes and commonly mentioned dataset attributes, which people consider when searching and making sense of data.

4.1 Motivation

Our previous study, described in Chapter 3 has shown that, despite increased availability, data cannot be easily reused, as people still experience many difficulties in finding, accessing and assessing it. We discussed three major aspects that matter to data practitioners when selecting a dataset to work with: *relevance*, *usability* and *quality* (Koesten et al., 2017). For each of these aspects, people have to make sense of the content and context of a dataset to make an informed decision about whether to use it for their task.

Metdata is often limited and might not provide enough content to decide whether a dataset is useful for a task (Noy et al., 2019). One way to provide people with information about the dataset is in the form of a written dataset summary. Current dataset summaries vary greatly in terms of content, language and level of detail, which are often not fit for purpose (Neumaier et al., 2016; Koesten et al., 2017). From a user's perspective, having a textual summary of the data is therefore paramount: text is usually richer in context than metadata, and can be easier to digest than raw data or graphs (depending on the context and the quality of the representation) (Law et al., 2005; van der Meulen et al., 2010).

A review of related literature (Balatsoukas et al., 2009) concluded that textual metadata surrogates, if designed in a user-centred way, can help people identify relevant

documents and increase accuracy and/or satisfaction with their relevance judgements. Several authors have shown that textual summaries perform better in certain decision making tasks than graphs. For instance, Gatt et al. (2009) found in a evaluation of a system that summarises patient data that all users (doctors and nurses) perform better in decision making tasks after viewing a text summary with manually generated text versus a graph. These findings are confirmed by Law et al. (2005); van der Meulen et al. (2010) in studies comparing textual and graphical descriptions of physiological data displayed to medical staff. Sultanum et al. (2018) emphasise the need to integrate textual summaries to get an overview of clinical documentation instead of relying on graphical representations.

A textual summary can help people assess aspects of relevance, usability and quality of a dataset for their own needs (Au et al., 2016; Bargmeyer and Gillman, 2000; Lehmberg et al., 2016; Nguyen et al., 2015). It can potentially also improve data discovery, as search algorithms can match the text against keyword queries (Thomas et al., 2015).

In Chapter 2, we elaborated on existing practices and techniques to create text about structured data. In general, a good summary must be able to represent the core idea, and effectively convey the meaning of the source (Zhuge, 2016). In this study, we aim to understand what this means in a data context: what a dataset summary must capture in order to help data practitioners select the data to work with with more confidence. This is currently a gap in the human data interaction literature. The data engineering community, on the other hand, has created some standards and best practices for publishing and sharing data, including DCAT,[1] schema.org,[2] and SharePSI[3].

However, none of these initiatives offer any guidance on what to include in a dataset summary. Sometimes text summaries are generated automatically using so-called natural language generation (NLG) methods. These methods are commonly bootstrapped via parallel corpora of data and text snippets, but an extensive exploration of the qualities of these training corpora is missing (Wiseman et al., 2017). Overall, this leads to summaries that vary greatly in terms of content, language and level of detail, which are often not fit for purpose (Neumaier et al., 2016).

As described in Chapter 2, data search often starts on a data portal. Upon entering their query, users are presented with a compact representation of the results, which includes for each dataset its metadata (title, publisher, publication date, format etc.), a short snippet of text, and, in some cases, a data preview or a visualisation. Figure 4.1 shows an example. From a user's perspective, having a textual summary of the data is paramount: text is easier to digest than raw data or graphs (Law et al., 2005; van der Meulen et al., 2010), and is richer in context than metadata. It helps people assess the relevance, usability and quality of a dataset for their own needs (Au et al., 2016;

[1] https://www.w3.org/TR/vocab-dcat/
[2] http://schema.org/Dataset
[3] https://www.w3.org/2013/share-psi/bp/

Bargmeyer and Gillman, 2000; Lehmberg et al., 2016; Nguyen et al., 2015). It could also improve data discovery, as search algorithms can match the text against keyword queries (Thomas et al., 2015). No matter where the search journey starts, a textual description is often key to determine whether a dataset is fit-for-purpose, or if you need to continue the search. Our two studies aim to understand the characteristics of this crucial element in the interaction between people and data.

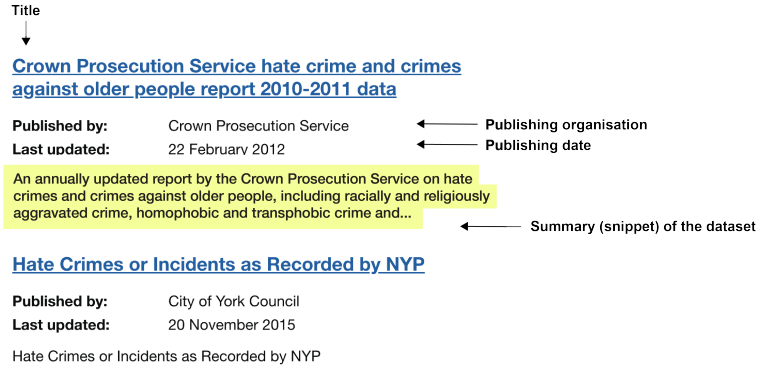


FIGURE 4.1: Example of a data search result list on the UK governmental open data portal. Next to title, publisher, domain and format, we see a textual description of the dataset.

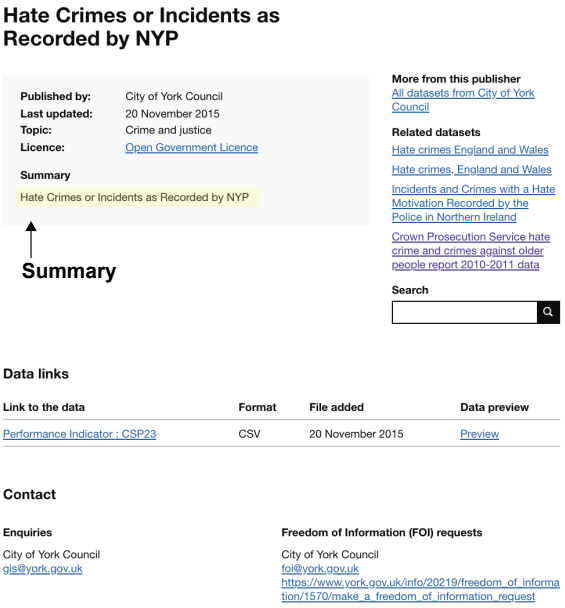


FIGURE 4.2: A dataset preview page on the UK governmental open data portal.

4.2 Related Work

4.2.1 Human-generated summaries of datasets

Summarising text is a complex and well-studied area of research in domains such as education, linguistics and psychology, amongst others (Yu, 2009). The cognitive processes triggered by this task, as studied in psychology, are described as involving three distinct activities: (i) *selection* (selecting which aspects of the source should be included in the summary); (ii) *condensation* (substitution of source material through higher-level ideas, or more specific lower-level concepts); and (iii) *transformation* (integrating and combining ideas from the source) (Bando et al., 2010).

Johnson defines a summary as a brief statement that represents the condensation of information accessible to a subject and reflects the central ideas or essence of the discourse (Hidi and Anderson, 1986). Describing or summarising something is a language activity and based in culture: the concepts, definitions and understandings developed in a community. Differences in cultural contexts can lead to misinterpretation of dataset content or to difficulties in developing a common understanding of a dataset summary. Constructing meaning from information – in our case the dataset and the accompanying summary – is always constructed by the reader, and is influenced by a variety of confounding factors.

Literature on text summarisation differentiates between writer-based summaries, which are summaries written for the writer herself, and reader-based summaries, which are written for an audience and usually require some planning (Hidi and Anderson, 1986). In this chapter we consider the latter.

Summarising structured or semi-structured data is inherently different to summarising free text. The complexity of constructing meaning from structured data (in contrast to text) has been discussed in the literature (Marchionini et al., 2005; Pirolli and Rao, 1996). Understanding data requires different cognitive work in order to contextualise it in relation to other information which makes it different to summarising natural language text (Gkatzia, 2016).

In their review of summary generation from text, Gambhir and Gupta (2017) point out the subjectivity of the task and the lack of objective criteria for what is important in a summary. Summary quality is suggested to depend on its purpose, focus and particular requirements of the task (Owczarzak and Dang, 2009). We focus on general purpose summaries due the variety of different data-centric tasks we found in the study described in Chapter 3 and to maximise the potential of a summary to be relevant for tasks of different nature. However, it would be interesting to explore task- or domain specific dataset summaries in future work.

4.2.2 Automatic summary generation

Automatically summarising text to accurately and concisely capture key content is an established area of research, with a wide range of techniques, most recently neural networks, employing language models of varying degrees of sophistication (Boydell and Smyth, 2007; Gambhir and Gupta, 2017; Reiter and Dale, 1997; Wiseman et al., 2017). Two broad approaches to summarisation are reported in the literature: (i) *extraction* (or intrinsic summarisation); and (ii) *abstraction* (or extrinsic summarisation).

Extractive approaches aim to create a summary by selecting content from the source they are summarising. Abstractive approaches aim to paraphrase the original source to provide a higher-level content representation (Boydell and Smyth, 2007). Research has focused more on extractive methods as abstractive methods are rather complex (Gambhir and Gupta, 2017). Based on existing studies in human data interaction, we believe that meaningful dataset summaries likely require abstractive elements including quality statements, descriptive statistics or topical coverage of a dataset (Gregory et al., 2017; Kern and Mathiak, 2015; Boukhelifa et al., 2017).

Automatic generation of summaries for data is a comparatively newer field, although there have been significant advances in this area (Wiseman et al., 2017). Current approaches tend to mostly work in closed domains and the complexity of performing these tasks is acknowledged in literature (Mei et al., 2015). Data-to-text generation has been explored in several areas, such as health informatics (Gatt et al., 2009; Scott et al., 2013), weather forecasts (Gkatzia et al., 2016; Sripada et al., 2004), finance (Kukich, 1983), sports reporting (Wiseman et al., 2017); as well as for different data formats, such as in graphs, databases and trend series (Bechchi et al., 2007; Cormode, 2015; Liu et al., 2014; Roddick et al., 1999; Sripada et al., 2003; Yu et al., 2007). Recognised subtasks in this space include: content selection (selecting what data gets used in the summary) and surface realisation (how to generate natural language text about the selected content) (Gkatzia, 2016).

Summaries produced with data-to-text generation methods are at the moment usually extractive rather than abstractive and tend to be merely textual representations of the dataset content, almost like a textual 'visualisation' (e.g.Wiseman et al. (2017)):

> *Extract taken from an automatically generated summary from Wiseman et al., 2017:*
>
> *The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday. Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here. Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers. Atlanta also dominated in the paint, winning the rebounding battle, 47 - 34, and outscoring them in the paint 58 - 26. The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets.*

Much of the work in automatic summary generation requires gold standards for evaluation (Bando et al., 2010). These corpora are typically created manually, but their

quality is uncertain and guidelines and best practices are largely missing (Gambhir and Gupta, 2017). Summary evaluation covers metrics computed automatically (e.g. BLEU, Rouge, etc.), human judgement or a combination of the two (Gambhir and Gupta, 2017; Owczarzak and Dang, 2009). A deep understanding of the best ways to run human evaluations, which criteria to use, the biases they create and so on on is not available - most studies use criteria such as accuracy, readability, coverage etc. but they are small-scale and not analysed in great detail. We believe this is partially due to a limited appreciation of what a meaningful summary should contain. Evidence for best-practice dataset summaries could lead to more meaningful evaluation methodologies in this space, by informing the design of evaluation benchmarks.

4.3 Methodology

We have undertaken two complementary studies in dataset selection and summarisation.

Based on the general lack of guidance, we focus this chapter on understanding the composition of meaningful summaries. There are very few studies that empirically evaluate any of the existing metadata standards in user studies - most efforts so far have concentrated on providing guidance for those who add information to a dataset, in many cases the data publishers. For the purpose of consistency, we refer to textual descriptions of datasets as *summaries*.

We explored the following research questions:

RQ1 What data attributes do people consider when determining the relevance, usability and quality of a dataset?

RQ2 What data attributes do people choose to mention when summarising a dataset to others?

As a starting point for the studies presented in this chapter, we make two assumptions: (i) textual summaries for datasets can be written by people without having an in-depth knowledge of data analysis and visualisation techniques; and (ii) summaries help data practitioners decide whether to use a dataset or not with more confidence. While we do not claim that text-based surface representations are superior to graphs, we believe they are, at a minimum, complementary to visualisations and accessible to a broad range of audiences, including less experienced users (Gatt et al., 2009; Koesten et al., 2017; Sultanum et al., 2018). Our findings support our assumptions. The crowdsourcing experiments showed that summaries can be created by people with basic data literacy skills who are not familiar with the dataset.

Building on insights from Chapter 3, the first study takes next steps to understand selection criteria in dataset search: we analysed 69 data-search diaries by students who were asked to document in detail how they go about finding and selecting datasets. The students wrote 269 diaries, which we analysed thematically starting from the framework from Chapter 3. This resulted in a *list of dataset selection attributes*.

In the second and larger study, we carried out a mixed-methods approach (Bryman, 2006) in which participants summarised datasets in a writing task. We conducted a task-based lab experiment and crowdsourcing experiments with overall 80 data-literate participants, who created a total of 360 summaries for 25 datasets. We analysed the summaries thematically to derive common structures in their composition, which led to a *list of dataset summary attributes*. We grouped these attributes into four main types of information: (i) *basic metadata* such as format and descriptive statistics; (ii) *dataset content*, including major topic categories, as well as geospatial and temporal aspects; (iii) *quality statements*, including uncertainty; and (iv) *analyses and usage ideas*, such as trends observed in the data.

We found a core set of attributes that were consistently prevalent in the two studies, across different datasets and participants. We used them to define a *template* to design more meaningful textual representations of data, which resonate with what people consider relevant when describing a dataset to others, and when trying to make sense of a dataset they have not used before.

Across the studies, we were able to identify core attributes that were prevalent for different datasets and participants. We compared them to existing metadata standards for data publication and sharing to understand existing gaps and design a summary template. Figure 4.3 gives an overview:

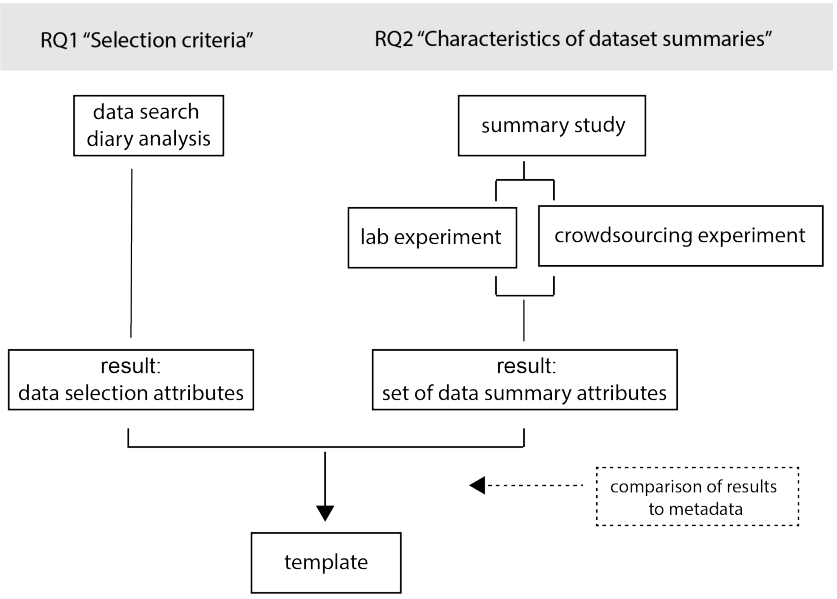


FIGURE 4.3: Overview of research methods and outcomes

4.3.1 Data-search diaries

In the first study we analysed data-search diaries (a user created record of their data search process) to get an in-depth understanding of the criteria that influence people's decisions to choose a dataset to work with. This also gave us insight into the kinds of information that need to be captured in dataset summaries to make them more useful for dataset sensemaking and selection.

Process.

We conducted a thematic analysis of 269 data-search diaries that were completed by 69 students[4] for a data science project within a university course.

Their task was to produce an online, magazine-style article (*a data story*), using at least two datasets to produce a minimum of three data visualisations that followed a narrative structure (such as for example in the Economists' Graphic Detail[5]). The participants were actively searching for datasets to work with and were instructed to write a diary entry for each data search task for two weeks. They were asked to find two to five datasets for their coursework. They were free to choose the topic of their project – there was hence no domain restriction to the datasets they could use or to the way they searched for the data.

The students were encouraged to document their data seeking behaviour directly after each search session and to reflect upon their data selection choices. The overall aim was to make them aware of the range of factors that come into play when looking for data, and of the importance of data sourcing for data science work.

We provided an online form with open-ended diary questions. The students self-selected when and what to report. For the purpose of our study, we focus on a subset of diary questions that concerned selection criteria for datasets:

- What do you need to know about a dataset before you select it for your task?
- What is most important for you when selecting a dataset for this task?
- What tells you that the data is useful and relevant for your task?
- What tells you that the data is good quality for your task?

Example data search tasks as described by the students include:

> *Example 1: I was looking for some datasets about tourism trends in Italy. I would like to find a few datasets which show how the tourism has changed in the last years.*
> *Example 2: Details on the molecular composition of Titan's atmosphere*

[4] MSc Data Science (n=49), MSc Computer Science (n=10), MSc Operational Research and Finance (n=6), MEng Computer Science (n=3), MSc Operational Research & Statistics (n=1)

[5] https://www.economist.com/graphic-detail/

Example 3: Finance data in UK across decades and years, categorised by gender, industry or region.

Analysis.

The free-text answers to these questions were analysed using thematic analysis (Robson and McCartan, 2016). Two researchers deductively coded the answers based on the framework for human structured-data interaction from Koesten et al. (2017), which defines *relevance*, *usability* and *quality* as general themes in dataset selection. As a second layer of coding we open-coded attributes emerging in each of these areas. This was done to obtain insight into how these high-level categories are operationalised by data searchers in practice. In this step, the coding was done by one researcher (the author of this work), but to enhance reliability two senior researchers checked the analysis for a sample of the data. The analysis resulted in a *list of data selection attributes*. As noted earlier, they helped us understand what kinds of information good summaries need to contain to aid data practitioners choose datasets with more confidence.

Ethics.

Responses were part of a university coursework at the University of Southampton. Participants consented to the data being used for research when joining the course. No personal data was analysed or reported.

4.3.2 Dataset summaries

4.3.2.1 Datasets: the $Set-5$ and $Set-20$ corpora

We used openly published datasets available as CSV files from three different news sources: *FiveThirtyEight*[6], *The Guardian*[7], and *Buzzfeed*[8]. We selected 'mainstream' datasets, understandable in terms of topic and content, excluding datasets with very domain-specific language or abbreviations. The datasets had to contain at least 10 columns and English strings as headers. The datasets varied across several dimensions: value types (strings, integers); topics; geospatial and temporal coverage; formatting of dates; ambiguity of headers, for example abbreviations; blank fields; formatting errors; size; and mentions of personal data.

The sample contained 25 datasets. We divided them into five groups, each containing: two datasets from *FiveThirtyEight*; two from *The Guardian* and one from *Buzzfeed*. One of these groups was our first corpus, $Set-5$. $Set-5$ was made of datasets $D1$ to $D5$,

[6] http://fivethirtyeight.com/
[7] https://www.theguardian.com/
[8] https://www.buzzfeed.com/news

which are described in more detail in Table 4.1. We used $Set-5$ in both experiments (see below). The remaining four groups of datasets (5 datasets per group, 20 in total) formed our second corpus, $Set-20$. $Set-20$ consisted of datasets $E1$ to $E20$ and was used only in the crowdsourcing experiment.

Working with $Set-5$ in both experiments allowed us to compare summaries generated by two different participant groups. $Set-20$ enabled us to apply our findings across a greater range of datasets (Characteristics of the $Set-20$ datasets can be seen in Appendix B.6). All datasets are available on GitHub[9].

Dataset	Topic	Example characteristics
$D1$	Earthquakes	>10k rows, 10 columns, dates inconsistently formatted, ambiguous headers, granular geospatial information
$D2$	Marvel comic characters	>16.000 rows, 13 columns, no geospatial information, many string values, limited value ranges, missing values, yearly and monthly values
$D3$	Police killings	>450 rows, 32 columns, contains numbers and text, geospatial information (long/lat as well as country, city and exact addresses), personal data, dates as year/month/day in separate columns, headers not all self-explanatory, some domain-specific language
$D4$	Refugees	192 rows, 17 columns, mostly text values, formatting inconsistencies, ambiguous headers, identifiers, geospatial information (continent/region/country), no temporal information
$D5$	Swine flu	218 rows, 12 columns, formatting inconsistencies, geospatial information (countries as well as long/lat), links to external sources, identifiers (ISO codes), some headers not straightforward to understandable

TABLE 4.1: Datasets in $Set-5$

4.3.2.2 Dataset summaries: Lab-based experiment

The objective of this experiment was to generate summaries of datasets written by data-literate people, who were unfamiliar with the datasets they were describing. Our assumption was that by asking people to summarise datasets unknown to them, they would create summaries that are relatable to a broad range of data users, and would be less biased in their descriptions than people who had been working with that data in the past, or had created it themselves. While this needs to be validated in future work, we further aimed to bring the initial exploration and scanning activity of datasets in this summary creation task through our task design. This was reported to be a crucial step in the assessment of a datasets' appropriateness for a given task by the participants in

[9] https://github.com/describemydataset/DatasetSummaryData2018

Chapter 3. Each participant was asked to summarise the datasets from $Set-5$. Having multiple summaries for the same datasets allowed for more robust conclusions.

Pilot.

We first conducted a pilot study with one dataset, six participants and different task designs. The aim was to get an understanding of the core task parameters, such as the time allocated to complete the task, and basic instructions about the length and format of the summaries. These parameters were important to constrain, as we wanted people to report only the most important features of datasets, rather than try to document everything they could see. The pilot dataset was on the topic of police searches in the UK; it contained 15 columns and 225 rows, contained missing values, geospatial information, temporal information (dates and age ranges), some inconsistencies in formatting, and domain specific language. We experimented with several task durations and varying restrictions on the number of words of the summaries. Before starting the study, we conducted an additional two pilots using the same 5 datasets as used in the study, to better understand feasibility and timings. We did not impose any restrictions on the participants' writing style (e.g., full text, short notes, bulleted lists). Based on the pilot, we decided to ask participants to write summaries of up to 100 words, with no time limit.

Recruitment.

We recruited participants who would be the primary target audience for textual data summaries, called *'data practitioners'* for the purpose of this work. This was defined in the call to participation as 'people who have been engaged in projects involving data'. While some of the subjects were very experienced in data handling, we chose not to restrict participation to formally trained data scientists, as the majority of people working with data are domain experts (Boukhelifa et al., 2017; Kern and Mathiak, 2015).

Our participants declared to have either received some training in using data or work with data as part of their daily jobs. Previous research has shown that this group are depending on summaries to select datasets with confidence (Gregory et al., 2018; Koesten et al., 2017). By contrast, most data scientists and engineers can more easily resort to a range of specialist techniques such as exploratory data analysis to make sense of new datasets.

We recruited participants through a call on social media via the author's institution. The call was published on the institution's website and a link to the call was posted on Twitter. The Twitter account had at the time of the study over $38.9k$ followers, with $23.025k$ impressions, 435 interactions and 91 retweets[10]. Our sample consisted of n=30 participants (19 male and 11 female), all based in the UK at the time of the study. Two

[10] https://support.twitter.com/articles/20171990

thirds of them were UK nationals (n=20) and had a Bachelor or Masters level education (n=26). All sessions were carried out between July and August 2016.

Process.

Respondents to the study call were contacted via email to receive an information sheet (as can be seen in Appendix B.3). We arranged a time for the experiment at the author's institution with those who volunteered to take part in the study. The task was formulated as follows: *We ask you to describe the datasets in a way that other people, who cannot see the data, can understand what it is about.*

Participants could open the CSV files with a software of their choice; we suggested MS Excel or Google Sheets. We asked them to describe all five datasets in up to 100 words each, one dataset at a time, in a text document. The order of the datasets was rotated to prevent potential order effects, due to fatigue or learning, that could influence the dataset summaries.

Analysis.

We collected 150 summaries, 30 per dataset. In our analysis we focused on the following aspects: (i) *form* (e.g. full sentences, bullet points, etc.) and *length* of the summaries; (ii) *high-level information types* people consider relevant for data sensemaking; and (iii) specific *summary attributes* (as shown in Figure 4.4).

To get a sense of the surface form of the summaries, we counted how many of them used full sentences, bullet points, tables or a mixture of the three. For their length we counted the number of words using the word-count feature in a text editor. (These descriptive findings are reported in Section 4.5.1.) To derive information types and their attributes, two of the authors independently analysed the data inductively to allow themes to emerge (Thomas, 2006). We ascribed open codes in an initial data analysis and explored the relationships between the codes in a further iteration (axial coding). We identified higher-level categories (*information types*) by examining properties that were shared across the codes. We adopted this approach because of the open nature of the research questions. As shown in Figure 4.4, the specific *summary attributes* that we present are codes that were drawn together within these general information themes. We aimed to identify the composition of summaries produced by our participants and understand the relative importance of particular attributes. Therefore we present our findings from two viewpoints: the higher level information types and the more granular summary attributes in Section 4.3.2. We used NVivo, a qualitative data analysis package for coding. In each of the two iterations, we cross-checked the resulting codes, refined them through discussions with two senior researchers, and captured the results in a codebook. We documented each code with a description and two example quotes. Two

senior researchers reviewed the conflict-prone codes based on a sample of the data. The unit of analysis was a summary (n=150) for the same dataset.

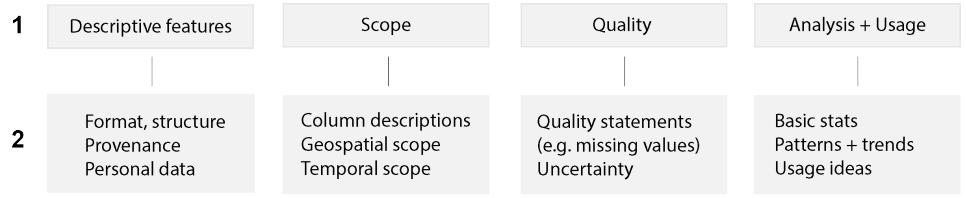


FIGURE 4.4: **Information types (1)** and emerging **summary attributes (2)** from the thematic analysis of the lab summaries, reflecting our coding process

Ethics.

The lab experiment was approved by the University of Southampton's Ethical Advisory Committee under ERGO Number 28636. Informed written consent was given by the participants prior to the experiment.

4.3.2.3 Dataset summaries: Crowdsourcing experiment

Following the lab experiment, we undertook a data summaries experiment on the crowdsourcing platform CrowdFlower (now Figure Eight)[11]. We used both dataset corpora, $Set-5$ and $Set-20$ and asked crowd workers to produce summaries of 50 to 100 words.

Through the crowdsourcing experiment we were able to reach out to a much larger number of participants to create summaries for more datasets. Using the five datasets from $Set-5$ in both experiments allowed us to compare the characteristics of summaries produced by data practitioners and the crowd.

Existing research suggests crowdsourcing platforms are a feasible alternative to the lab for our purposes. Previous studies have considered related tasks such as text writing (Bernstein et al., 2015); text summarisation (Borromeo et al., 2017; Marcu, 2000); and data analysis (Lin et al., 2013; Willett et al., 2013).

Recruitment.

Participants were crowd workers registered on CrowdFlower. We limited the experiment to *Level*2 crowd workers from native-English speaking countries[12].

Process.

[11] https://www.figure-eight.com/

[12] *Level*2 workers are workers who have reached a verified level of performance in their previous work.

Crowd workers had to describe five datasets (either the datasets $Set-5$ or one of the four groups from $Set-20$) in 50 to 100 words. The length of the summaries was informed by the lab experiment. The order of the datasets was rotated to prevent potential order effects due to fatigue or learning. We also included 12 short qualification question prior to the task, assessing basic reading, reasoning and data literacy skills to make sure workers have the capabilities to complete the task.

We used the same basic task description as in the lab: *We ask you to describe the datasets in a way that other people, who cannot see the data, can understand what it is about*, but included some additional information. Paid microtask crowdsourcing works well when the crowd is provided with a detailed description of the context of the task they are asked to complete. For this reason, we also showed participants step-by-step instructions, a picture of a dataset, and examples of corresponding summaries which were based on the outputs from the lab experiment.

Just like in the lab experiment, participants were free to structure their summaries as they saw fit. However, they were shown three examples, presented as text; a list; and a combination of list and text (the three summary representations seen in the results of the lab experiment). The minimum time allowed to summarise five datasets was 15 minutes; the maximum time was 60 minutes. Both settings were informed by the lab experiment.

The outputs were, for each worker, five textual summaries for five datasets. To minimise spam, we prevented copy-pasting of content and validated a random selection of ten words from each answer against an English language dictionary, requiring a 60% matching threshold to be accepted.

We recruited 30 crowd workers for $Set-5$ and 20 crowd workers for $Set-20$ (five workers per each group of five datasets from $Set-20$). Workers were allowed to do only one task i.e. summarise five datasets. They were paid $3.00 per task. From the lab we learned that the task duration is likely to be around 25 to 35 minutes, which was confirmed in an early pilot on CrowdFlower.

A screenshot of the CrowdFlower task is included in appendix B.5.

Analysis.

We collected a total of 250 crowdsourced summaries and manually excluded those which were obvious spam or off-topic. This resulted in 120 summaries for the five datasets in $Set-5$ (on average 24 summaries per dataset) and 90 summaries for the 20 dataset in $Set-20$ (between four and five summaries per dataset). We analysed: (i) the form and length of the summaries; and (ii) the summary attributes, grouped according to the information types identified in the lab experiment. On both accounts we used the same methods as in the lab experiments (see Section 4.3.2.2). We also looked at differences

between the two participant groups for the summaries of $Set-5$ and across datasets for all 25 datasets from the two corpora.

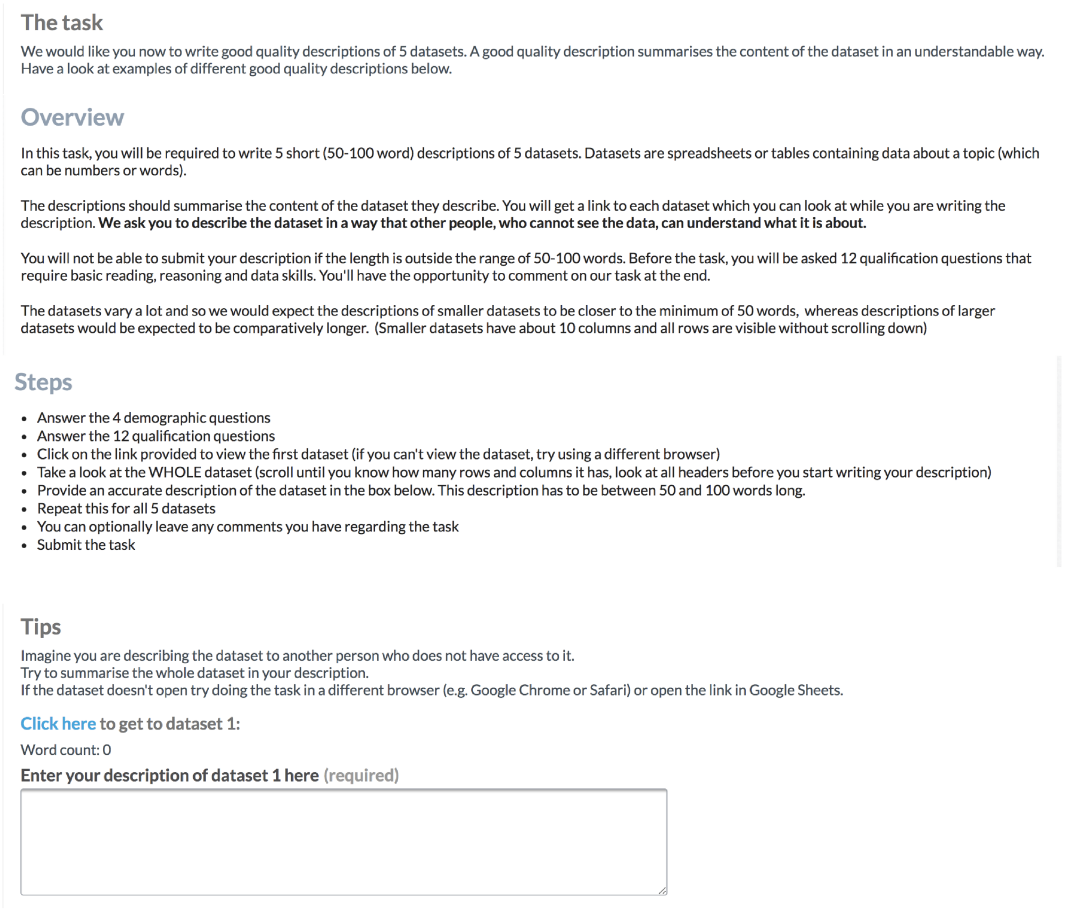


FIGURE 4.5: CrowdFlower task instructions in the crowdsourcing experiment

Ethics.

This experiment was approved by the University of Southampton's Ethical Advisory Committee under ERGO Number 29966. Consent was given by crowd workers previous to carrying out the task.

4.4 Findings: Data-search diaries

The analysis of the data-search diaries was performed to *complement* the results of the summary analysis (Bryman, 2006). In their diaries, the students explicitly answered questions about their thought processes and their rationales when selecting data to work with.

The data attributes emerging from this analysis are listed below, and analysed in more detail in the discussion when we compare the results of our studies with existing metadata standards. We grouped them according to the three high-level themes identified in

Chapter 3: *relevance*, *usability* and *quality* and describe the topics that emerged within these. Relevance refers to whether a dataset content is considered applicable to a particular task; e.g. is it on the correct topic. Usability refers to how suitable the dataset is considered, meaning practical implications of, e.g., format or license. Quality refers to anything that participants use to judge a datasets condition or standard for a task, such as e.g. completeness. Some of the attributes were mentioned by participants in the context of several themes, which emphasises their importance.

We present these as consolidated lists in Table 4.2. As mentioned earlier there is limited research investigating dataset specific selection criteria. We report on the prevalence of the individual attributes below, however we believe that these can only be seen as indicative, due to the limited number of participants and the specifics of the task. While we do not attempt to rank the attributes based on prevalence, they confirm and extend the findings from Chapter 3, which strengthens their validity.

Theme	Attribute	%
Relevance	Scope (topical, geographical, temporal)	36
	Granularity (e.g., number of traffic incidents per hour, day, week)	19
	Comparability	14
	Context (e.g. original purpose of the data)	11
	Documentation (e.g. understandability of variables, samples)	6
Usability	Format (e.g., data type, structure, encodings, etc.)	44
	Documentation (e.g. understandability of variables, samples)	11
	Comparability (e.g., identifiers, units of measurement)	11
	References to connected sources	6
	Access (e.g. license, API)	6
	Size	4
	Language (e.g. used in headers or for string values)	3
Quality	Provenance (e.g. authoritativeness, context and original purpose)	28
	Accuracy (i.e. correctness of data)	13
	Completeness (e.g. missing values)	13
	Cleanliness (e.g. well-formatted, no spelling mistakes, error-free)	9
	Methodology (e.g. how was the data collected, sample)	9
	Timeliness (e.g. how often is it updated)	6

Table 4.2: Findings on selection criteria for datasets, based on thematic analysis of the data-search diaries. Prevalence can be seen as indicative, but needs further validation.

4.4.1 Relevance

Two prevalent attributes were the *scope of the data* (in terms of what it contains) and its *granularity*. They were mentioned in 36% and 19% of responses, respectively. We hypothesise that students, by default, considered the content of the dataset to be an

important factor (due to the nature of their task), and therefore only a relatively low percentage of them mentioned this explicitly. The scope sometimes referred to the geographical area covered by the dataset, while the granularity described the level of detail of the information (e.g. street level, city level, etc.) Some participants mentioned *basic statistics* such as counts, averages and value ranges as a useful instrument to assess scope.

Interestingly, 14% of the diaries noted the relative nature of relevance (echoing discussions in the literature (Mizzaro, 1997)) and the need to consider multiple datasets at the same time to determine it. To a certain extent, this could be due to the nature of the task – students were free to choose the topic of the datasets and hence might have had a broader notion of relevance, which allowed them to achieve their goals by interchanging one dataset for another or through a combination of datasets. However, the relation to other sources was mentioned in other categories as well, which reinforces the need for tools that make it easy for data users to explore more than one dataset in the same time and to make comparative judgements. This is also in line with experience reports about data science projects in organisations – making complex decisions often involves working with several datasets (Koesten et al., 2017; Erete et al., 2016). Further attributes from the diaries suggest that a thorough assessment of relevance needs to include easily understandable variables, data samples for fast exploration, as well as insight into the context and purpose of the data.

4.4.2 Usability

To determine how usable a dataset is for their task participants mentioned a range of practical issues which, if all available in the desired way, would make working with a dataset frictionless: *format*, *size*, the *language* used in the headers or for text values, *units of measurement* and so on.

Format was the most prevalent attribute (44%), though *documentation* and the ability to understand the variables were perceived to impact usability as well (both at 11%).

The *size* of the dataset was mentioned primarily in the context of usability rather than as a basic descriptive statistic. This is probably due to the fact that students were mindful of the additional effort required to process large datasets.

The participants understood the importance of being able to integrate with other sources, for example through identifiers – 11% of the diaries mentioned this aspect explicitly. In their coursework, the students were asked to use at least two datasets and hence valued data integration highly. At the same time, using multiple datasets is not uncommon in most professional roles (Convertino and Echenique, 2017; Koesten et al., 2017). *Access* to the data was also mentioned in reference to APIs or licences, though only around 6% of the time. This low value is a function of our study - students were not looking to

source data to solve a fixed problem. Their search for data, documented in the diaries, happened while they were deciding on the topic of their project. If they could not find data for one purpose, they could adjust the project scope rather than having to tackle licensing or access fees.

4.4.3 Quality

Participants mentioned unique attributes such as *provenance* – in a broad sense of the term – that would allow judgements around the authoritativeness of the publisher and context of the data. This included information about the original purpose of the data, as well as questions of sponsorship of the research or data collection, and about other potential sources of bias. At 28% this attribute was ranked much higher than other quality dimensions such as *accuracy*, *completeness*, *timeliness* and *cleanliness*, which are in the focus of many quality repair approaches (Wand and Wang, 1996). The importance of provenance resonates with previous work in data quality (Ceolin et al., 2016; Malaverri et al., 2013); there is also a large body of literature proposing frameworks and tools to capture and use provenance, though their use in practice is not widespread (for example Stamatogiannakis et al. (2014); Simmhan et al. (2008)).

Some participants reported to be interested in details of the *methodology* to create and clean the data, including aspects such as the control group, whether a study had been done using randomised trials, confidence intervals, sample size and composition, etc. This is in line with Chapter 3, where we pointed out that awareness of methodological choices plays an important role in judging the quality of a dataset with confidence.

In the discussion in Section 4.6 we relate these different data selection attributes to the attributes extracted from the summaries, and compare them to existing guidelines for data publishing and sharing. We identify overlaps between the information needs of people searching for data, who are potential consumers of data summaries, and the information people choose when summarising an unfamiliar dataset.

4.5 Findings: Dataset summaries

We report on the main findings from the lab and crowdsourcing experiments, covering the three areas mentioned in the methodology in Section 4.3.2: (i) summary *form* and *length*; (ii) *information types*; and (iii) detailed *summary attributes*.

4.5.1 Form and length

In the lab experiment participants were not given concrete suggestions or examples for surface representation, yet most resulting summaries were presented as text using full

English sentences (64%). However, some (17.3%) were structured as a list or presented as a combination of text and lists (18.6%). In the crowdsourcing experiment participants were provided with examples of summaries using these three representations. Their summaries were structured as follows: (79%) used text, a few (7%) were structured as a list, and some (14%) presented a combination of the two.

The average lab summary was 98 words long (median of 103). By comparison, the crowd needed on average 63 words for the same datasets in $Set-5$. The average crowdsourced summary from Set-20 was 64 words long (median of 58). It would be interesting to explore how the length of the summaries impacts their perceived usefulness by readers or their potential information gain (Maxwell et al., 2017), particularly in the context of the summary template we propose in Section 4.6.

4.5.2 Information types

We identified four high-level types of information in the lab summaries, which we subsequently used to analyse summary attributes for all 360 summaries created in the two experiments. Our findings suggest these general categories to be dataset independent.

1. **Descriptive attributes** e.g. format, counts, sorting, structure, file-related information and personal data:

   (P1) The dataset has 468 rows (each representing one person who has been killed).
   (P7) No free text entries and character entries have a structured format. It contains no personal data.
   (P8) The header pageID appears to be a unique identifier.
   (P21) CSV in UTF encoding. Header and 110172 data rows. 10 columns.

2. **Scope of the data** which refers to the actual content of the dataset, through column descriptions such as headers or groupings of headers, or references to the geographic and temporal scope:

   (P1) For example, this includes details on the share of ethnicities, in each city, the poverty rate, the average county income, etc.
   (P14)Figures include number of confirmed deaths and proportion of cases per million people.
   (P2) Some columns have no particular meaning to a non-expert, e.g. columns named 'pop', 'pov', 'country-bucket', 'nat-bucket'.
   (P12) Each instance has specific details on the time, geographic location, earthquake's magnitude.

3. **Quality** which included dimensions such as errors, completeness, missing values and assumptions about accuracy, but also expressions of uncertainty and critique:

   (P1) The precision of the description varies wildly.
   (P14) A link (in some cases two) to the source of the data is provided for each country.
   (P7) It has column headers all in caps (apart from 'pageID'), which are mostly self-explanatory.
   (P8) Combination of personal data about person killed and demographic data, unclear if this is for area of killing.
   (P17) The data seems to be consistent and there aren't any empty cells.

4. **Analysis or ideas for analysis and usage** such as simple data analysis, basic statistics, highlights of particular values or trends within the data:

   (P5) The data does provide the method of how each individual has been killed which can provide an argument for police not using firearms in the line of duty.
   (P7) There is a significant amount of missing data in the 'state' column, but this information should be possible to infer from the 'longitude' and 'latitude' columns.
   (P18) The dataset shows that the greatest number of refugees originate from the Syrian Arab Republic.
   (P30) Killings took place all around America. The people who were killed mostly carried firearms.

Table 4.3 shows the percentage of summaries that contained each information type, split by dataset. The four types are not meant to define an exhaustive list - we consider them merely a reflection of the 150 lab summaries analysed and in Section 4.8 we discuss this limitation of the study. The types are also not exclusive - more than half of the summaries included all four types of information. *Analysis and usage* was the least frequent information type overall, though some attributes in this category were more popular than others. For example, as we will note later in this section, *basic statistics* were mentioned more frequently in the crowd-generated summaries than in the lab, while trends and ideas for further use were rather low overall, with the exception of some $Set-20$ datasets described by the crowd. We believe the main reason for this is the design of the task. In the lab experiment, the task description might have implied a focus on the raw data and on surface characteristics that could be observed through a quick exploration of the data rather than an extensive analysis. Crowdsourcing requires a higher level of detail in instructions, which included examples of summaries with, among other things, basic statistics.

We present all individual attributes associated with each of the four information types in the remainder of this section.

	Total	$D1$	$D2$	$D3$	$D4$	$D5$
Descriptive attributes	81	67	87	80	83	90
Scope of the data	99	97	100	100	100	100
Data quality	79	80	67	90	77	80
Analysis and usage	64	57	63	73	67	60

TABLE 4.3: Percentages of information types per dataset in $Set-5$, based on 150 lab summaries

4.5.3 Summary attributes

The summary attributes presented in this section represent a more granular analysis of the four high level information types.

Across all summaries the most prevalent attributes were: a subtitle, the datasets headers and information about the geographical scope of the dataset. In the following sections we present the identified attributes in detail without ordering them, as we believe that their actual importance based on prevalence would need to be validated in future work with a larger number of summaries and datasets.

Across the 360 summaries created in the two experiments we have identified the following attributes:

Summary attributes
Subtitle: A high-level one-phrase summary describing the topic of the dataset
Format: File format, data type, information about the structure of the dataset
Provenance: Where the dataset comes from, such as publisher, publishing institution, publishing date, last update
Headers: Explicit references to dataset headers
Groupings: Selection, groupings or abstraction of the headers into meaningful categories, key columns
Geographical: Geospatial scope of the data at different levels of granularity
Temporal: Temporal scope of the data at different levels of granularity
Quality: Data quality dimensions such as inconsistencies in formatting, completeness, etc.
Uncertainty: For example ambiguous or unintelligible headers or values, or unclear provenance
Basic statistics: For example, counts of headers and rows, size of the dataset, possible value ranges or data types in a column
Patterns/Trends: Simple analyses to identify highlights, trends, patterns, etc.
Usage: Suggestions or ideas of what the dataset could be used for

TABLE 4.4: Most frequent summary attributes, based on 360 summaries of datasets from $Set-5$ and $Set-20$.

Across the two experiments, a summary was commonly structured as follows: (i) a high-level *subtitle* describing the topic of the dataset; (ii) references to dataset *headers* (either the names of the headers or an abstraction of the headers such as a meaningful *grouping*); (iii) a *count* or other descriptive attribute such as possible values in a column; and (iv) *geographic* and *temporal scope*. Amongst other popular attributes were: *quality statements*; *provenance*; and, less frequently, ways to *analyse* or *use* the data.

Here is a summary that exemplifies this:

> *(P6) A list of people killed by US police forces in 2015. Data included is location of incident, police department, state, cause of death and whether the victim was wielding a weapon. Detailed and specific data with 34 columns. Useful for drawing parallels between criminal profiling and locations.*

Attributes that were mentioned in less than 10% of the lab summaries are not represented in this table. These include: mentions of personal data, license, methodology, funding organisation, and others.

Some summaries described the data by talking about the header row as an example:

> *(P16) Each row describes one of those 'earthquakes': lat, long, magnitude and location name.*

Percentages of these attributes over all summaries can be seen in Table 4.5, split by experiment and dataset corpus ($Set-5$ and $Set-20$). *Here we describe the most prevalent attributes across all summaries:*

Attributes such as *subtitle*, *geographical* and *temporal* scope and *headers* were present in a majority of summaries. *Format* was mentioned in more than half of the the $Set-5$ summaries and in 27% of the $Set-20$ summaries. *Basic statistics* were mentioned fairly often as well, in more than half of the $Set-5$ summaries and in 48% of the $Set-20$ summaries.

	L *Set* − 5 (N=150)	C *Set* − 5 (N=120)	C *Set* − 20 (N=95)
Subtitle	89	86	95
Format (*IT-1)*	61	52	27
Provenance (*IT-1)*	45	24	23
Headers (*IT-2)*	70	82	80
Groupings (*IT-2)*	49	70	69
Geographical (*IT-2)*	73	71	60
Temporal (*IT-2)*	58	55	56
Quality (*IT-3)*	55	23	21
Uncertainty (*IT-3)*	69	16	8
Basic statistics (*IT-4)*	56	74	48
Patterns/Trends (*IT-4)*	27	25	52
Usage (*IT-4)*	15	5	1

TABLE 4.5: Percentages of summaries created in the lab (L) and via crowdsourcing (C) that mention summary attributes. Darker fields have higher percentages. Numbers in brackets (N=) refer to the number of summaries analysed in each category. (IT= higher level Information Types, as presented in Section 4.5.2).

Table 4.6 elaborates on the distribution of summary attributes in the lab summaries (150 summaries in total, 30 per dataset). Across the five datasets analysed, *subtitle*, *format* and *headers* were mentioned consistently in more than 55% of the cases. *Basic statistics* and *quality* achieve slightly lower scores (47% and higher). We discuss differences in scores between datasets as well as the attributes that showed greater variation later in this section.

	$D1$	$D2$	$D3$	$D4$	$D5$
Format	60	63	57	60	67
Provenance	23	47	10	53	90
Subtitle	83	87	87	83	90
Headers	60	87	80	63	60
Groupings	13	40	70	53	30
Geographical	90	0	90	90	93
Temporal	87	37	87	33	47
Quality	57	50	47	60	60
Uncertainty	73	60	80	67	67
Basic Stats	53	73	63	47	43
Patterns/Trends	23	20	33	27	30
Usage	17	20	17	7	17

TABLE 4.6: Percentage of **lab summaries** containing respective attributes, per dataset from $Set-5$ (N=150). Darker fields have higher percentages.

	$D1$	$D2$	$D3$	$D4$	$D5$
Format	55	50	57	48	52
Provenance	20	23	0	17	62
Subtitle	75	95	91	70	100
Headers	75	86	83	78	86
Groupings	65	73	78	70	62
Geographical	95	0	91	78	90
Temporal	90	68	61	0	62
Quality	45	23	22	17	10
Uncertainty	35	9	17	4	14
Basic Stats	75	77	83	78	57
Patterns/Trends	30	32	17	26	14
Usage	0	9	4	4	5

TABLE 4.7: Percentage of **crowdsourced summaries** from $Set-5$ (N=120) containing respective attributes, per dataset. Darker fields have higher percentages.

Table 4.7 illustrates the distribution of attributes in the 120 summaries by the crowd for the datasets in $Set-5$ (24 summaries per dataset on average). The most prevalent attributes are slightly different than the ones observed in the lab: *subtitle*, *format* and *headers* remain important, but *basic statistics* are more consistently mentioned than in the other experiment. In the same time, the crowd focused on *groupings* of headers

as well, much more so the data practitioners who participated in the lab experiment – overall 70% of the crowd-generated summaries of $Set-5$ mentioned this attributes, compared to only 49% of the lab summaries (see Table 4.5); the scores for the individual datasets were between 62% and 78% on CrowdFlower and 13% and 70% in the lab.

The $Set-20$ summaries created by the crowd reinforce some of these trends (see Table 4.5). When looking at the 20 datasets from $Set-20$, *subtitle*, *geographical* and *temporal* scope and *headers* are mentioned in the majority of summaries, just like in the summaries of $Set-5$. *Groupings* seem to be popular among crowd workers across all datasets – 69% of the 95 $Set-20$ summaries mention them, compared to 70% of the 120 $Set-5$ summaries, but only 49% of the lab summaries. By contrast, $Set-20$ summaries showed greater variation in: *format* (27% vs 52% of the crowd-produced $Set-5$ summaries vs 61% for the lab-produced $Set-5$ summaries) and basic statistics (48% vs 74% of the crowd-produced $Set-5$ summaries vs 56% for the lab-produced $Set-5$ summaries). Aside the popularity of *groupings*, a second surprising result was the popularity of *patterns/trends* aspects – this attribute was mentioned in only 27% of the lab summaries and 25% of the crowd summaries for the same $Set-5$ datasets (see Table 4.5, but in 52% of the $Set-20$ summaries. This goes against our basic assumption that the task instructions suggested a focus on raw data and surface characteristics. Later in this section, we will examine the summaries that referred to *patterns/trends* to understand how this difference came about.

Differences between lab and crowd summaries.

The five datasets from $Set-5$ were used in both experiments. As noted earlier, for the lab experiment, our sample consisted of N = 150 summaries from 30 participants, while for the crowdsourcing experiment we used a sample of N = 120 summaries from 30 participants (spam answers were manually removed).

We compare the distributions of attributes in the two experiments shown in Table 4.5. For the five datasets in $Set-5$, *provenance* appears more frequently in the summaries created in the lab (45% vs 24%). We believe this to be due to the fact that participants were more data savvy and so placed a greater importance on where a dataset originates from. A similar trend was observed in the data-search diary, where participants were MSc students reading data science, computer science or statistics. We assume the same applies for *quality* statements (55% vs 23%) and ideas for *usage* (15% vs 5%), whose appreciation may equally require a certain level of experience with data work which was not given in the crowdsourcing setting. In the same time, the crowd appreciated attributes such as *groupings* of headers (21 point difference) and *basic statistics* (18 points) more. This demonstrates that the crowd had a fair level of data literacy and does not focus only on features that can be easily observed such as *subtitle*, *format* and

headers. As noted earlier, when looking at summaries for 20 other datasets, *groupings* remained popular, but *basic statistics* dropped to a lower level than in the lab (48%). We believe this calls for additional research to understand the relationship between the capabilities of summary authors and the aspects they consider important in describing datasets to others.

Looking at the distribution of summary attributes over $Set-5$ (Table 4.6), geospatial attributes, as well as provenance appear to have the highest dependency on the dataset. $D5$ differed from the other four datasets in the corpus by including an entire column titled *'Sources'*, displaying links to the source from which the values were taken from - this is likely the reason why 90% of the 30 data practitioners and 17 crowd workers mentioned it in their summaries. $D2$ similarly included a header called *'Page id'* pointing to the source of the data - this was less easy to spot by the crowd workers, who talked about *provenance* only 17% of the time.

We believe that geospatial attributes might in reality be more consistent for most datasets - four out of five datasets achieved consistently high scores in this category. $D2$ however, was set in a fictional universe and may therefore have prompted participants to discard any geospatial considerations.

Differences between $Set-5$ and $Set-20$ summaries.

The crowdsourcing experiment used two corpora: $Set-5$ with the same five datasets used in the lab (D) and $Set-20$ with 20 datasets (E). The reason to include a second corpus, albeit with fewer summaries per dataset (95 summaries in total, four to five summaries per dataset) was to explore how the main themes that emerged from the 270 summaries of $Set-5$ in total, generalise across datasets. To recapitulate, the total number of summaries in the crowdsourcing experiment from $Set-20$ is N=95, and from $Set-5$ is N=120, as described in Section 4.2.1.

Compared to the $Set-5$ crowd-generated summaries, $Set-20$ shows a higher prevalence of *subtitles* (95% vs 86%) and *patterns/trends* (52% vs 25%) and lower scores for *format*, *geographical* scope and *basic statistics* (see Table 4.5).

We looked at each of the 20 datasets from Set-20 to understand where these differences might come from. $Set-20$ contained a higher number of datasets with clearly identifiable *subtitles*, which explains the higher score. The datasets overall had fewer attributes representing *format* and *basic statistics*. Many $Set-20$ datasets either did not contain any *geographical* information or were clearly associated with a country or region that is not mentioned explicitly – for instance, $E10$ is about the UK's House of Commons, but there are no geospatial values in the dataset. The popularity of *patterns/trends* in $Set-20$ points to another dependency of summary content on the dataset – both in

the lab and on CrowdFlower, the summaries of the $Set-5$ datasets were consistent along this dimension. For instance, $E11$ explicitly mentions statistical content such as 'the median' as a header, other summaries with a high percentage of patterns/trends attributes tend to display clear trends or rankings and therefore afford quick judgements, for instance 'the country with the highest human development index'. The same counts for datapoints that stand out and get highlighted in a summary. For instance in the example of a dataset ($E10$) that contains salaries and expense claims from members of the British Parliament House of Commons which shows claims for a lawn mower, amongst other claims.

Just like the other summaries produced by crowd workers, *usage*, *provenance* and *quality* were not mentioned very often, which we believe is due to the level of data literacy in the experiment. In addition, we noted that $Set-20$ provenance was often not recorded when the context or origin of the dataset was very opaque – for example, $E4$ had mainly numerical values describing the elderly population worldwide – or in connection to uncertainty about the provenance – e.g. $E12$ was about US weather data, but did not make any reference to the source of the data.

4.5.3.1 Summary attributes in detail

In the previous section we presented a series of high-level findings across the two experiments and differences across datasets and participant groups. In this section, we discuss summary attributes individually and give additional details and example summaries. Summary quotes used throughout this section refer to $Set-5$. The total number of summaries from $Set-5$ in the lab study is N=150, and from $Set-5$ in the crowdsourcing experiment is N=120, which is what the respective percentages reported in this section refer to.

Format and file related information.

Format. The file format and references to the structure of the dataset were explicitly mentioned in more than 60% of all lab summaries and in about half of all $Set-5$ crowdsourced summaries. The mentions of file format or data type drop for $Set-20$ to 27%.

File related information. The summaries contained other attributes that described the file beyond its actual content, which refers to descriptive attributes as mentioned in the overview section on information types represented in the summaries. That included attributes such as: the type of values in a column; statements about the size of the file; mentions of licence (3% of the lab summaries and none of the crowdsourcing summaries); sorting of values; redundancies in the data; formatting; and unique identifiers. The

valuetype of a column was mentioned in 18% of all lab summaries, and in 23% for $Set-5$, 15% for $Set-20$.

There were also mentions of personal data in this category, as they describe a characteristic of the data rather than the data itself. Personal data was mentioned by 20% of the participants and mostly mentioned in connection to D3 which contained names of people in a police crime. We assume this is due to the fact that in the context of our task and the type of data we used (aside from $D3$), personal data was not a category that our participants were prompted to think of.

High-level subtitle.

Close to 90% of all summaries started with a high-level *subtitle* which gave the reader a quick first impression of what the dataset was about. In some cases *subtitle* referred to a key column $6-7$% or, more often, to the geospatial scope (48% of the lab summaries and 35% of the crowdsourced summaries), or to the temporal scope of the dataset (33% of the lab summaries, 17% of $Set-5$ crowdsourced summaries and 19% of $Set-20$).

(P1) This dataset, in csv format, describes police killings in what appears to be the USA in 2015.

(P11) Dataset of characteristics of Marvel comic book characters from the earliest published comics to around 2013.

(P13) This dataset describes the time, geographical location and magnitude of earthquakes in the United States.

Column descriptions.

A majority of summaries explicitly mentioned the headers of the dataset (70%). This was found to be consistent through all summaries done by the same participant – which points to the fact that this feature is not dependent on the underlying dataset. Out of those who mentioned headers explicitly ($N = 23$), the majority were consistent for all of their summaries (90%). About half of all summaries show some type of grouping or abstraction of the headers. Participants typically mention a selection of headers and group them according to meaningful categories, as can be seen below:

(P14) 34 variables, which comprehend personal information about the victim, place (inc. police department) of the incident, details about the incident, socio-demographic of the place.

(P15) Fields: Demographic data (name, age, gender, race), date (month, year, day), incident details (cause of death, individual armed status - categorical), county details (population, ethnicity), law enforcement agency, general reference data.

Similarly, a common strategy is the identification of a key column, which is the focus of the dataset:

(P11) For the victims, the metadata records their age, gender, ethnicity, address. The place and time of their death, as well as the cause of death and police force responsible are also recorded.

(P23) We are given useful information about each earthquake, specifically: latitude, longitude of the event, magnitude of the earthquake, a unique identifier for each earthquake called 'id', when the data was last updated, the general area the earthquake took place, the type of event it was, the geometrical data and if it took place in the US we are given the state it occurred in.

Some participants use the actual header name, others use a more descriptive version of the header. Many list the headers, together with qualifying information about them and/or possible values and ranges in a column.

(P30) It lists more than 15000 characters with their fictitious name and the real name in the comic. The data set records whether they are alive or dead characters, their gender, their characteristics (like: hair and eye colour). The data set records if the character has a secret identity [..] (and) whether the particular character has a negative or positive role.

Geographical information.

Geospatial aspects were very common in summaries across datasets and participant groups. In $Set-5$, the exception was $D2$, which described characters in a fictional world. They referred to different types of locations, including provenance (where the data comes from), coverage of the data itself (e.g. data from a particular region), and format, at varying levels of granularity. Summary authors often used higher-level descriptions of the relevant values, for example *'for most countries in the world'* or *'across the world'* to describe key columns with a wide range of country names.

(P17) The data goes down to country and includes country codes, the area and region.
(P1) Location, provided by latitude and longitude measurements.
(P2) Location, in latitude and longitude, but also in descriptive text about location

relative to a city.
(P7) Each observation refers to a unique country, using country codes.

Temporal information.

Temporal aspects were mentioned in connection to: time mentioned in the data, the publishing date, the last update and the time the data was collected, all at different levels of granularity. The numbers reported as here include only temporal attributes that refer to the temporal scope of the data itself and not to publishing date or last updates which were included in *provenance*.

Often summaries refer to both *'date'* and *'time'*, meaning the time of the day and the day that a particular event in the data occurred.

We found differences depending on the datasets: in the $Set-5$ lab summaries, for example, time was most often mentioned in relation to $D1$ and $D3$ (87%) and less often in connection to $D2$ and $D4$ ($<40\%$). $D3$ had three date columns separating day, month and year from each other which might prompt including this information in the summaries. $D1$ had high inconsistency in formatting dates and included two types of temporal information: when the earthquake took place and when the specific row was updated. $D1$ displayed a relatively high overlap between time and uncertainty (30% of all mentions of time were connected to uncertainty).This points to inconsistencies in formatting of dates in D1 and to potentially confusing headers called 'time' and 'updated', which show a mixture of dates and times. We assume this contributes to the varying prevalence of time in the summaries, which can be seen in Table 4.6. $D4$ on the other hand did not contain temporal information explicitly which explains the significantly lower percentage. This was reflected in the crowdsourced summaries for $D4$. $D2$ did contain temporal information (year and month), however it describes fictional comic characters which may lead to placing less importance on the temporal information represented in the data.

Temporal provenance. We further saw mentions of updates of the data, which we define as temporal provenance. This was present in 20% of all lab summaries and in 6% of the $Set-5$ and 12% for $Set-20$ crowdsourced summaries. It describes mentions of time that can be used to determine the relevance or quality of the data, such as:

(P30) The data set for confirmed cases of flu was last updated on 20/01/2010.
(P1) It is unclear whether this data is up to date, as there are no details on when this is from.

Quality statements and uncertainty.

Statements about uncertainty and quality were common in 70% of the lab summaries. Among the most popular words in this category were 'unclear' and 'missing'. The emerging themes connected to quality were features such as inconsistencies in formatting (e.g. dates), completeness, as well as statements about missing understandability (such as ambiguous or unintelligible headers or cells), as well as unclear provenance and authoritativeness of the source.

We further grouped uncertainty statements into six categories related to: completeness, precision, definitions, relations between columns, temporal and geospatial attributes, and methodology.

Completeness included statements about the representativeness, comprehensiveness and scope of the data, in addition to general statements about missing values:

> *(P4) Unclear how representative this list is of total population/whether this list is total population.*

> *(P13) The dataset appears to be missing data from some of the countries.*

Accuracy referred to inconsistencies in the data, for instance in units of measurements, or variations in the granularity of cell values.

> *(P13) The precision of the description varies wildly (eg. 23 km NE of Trona versus Costa Rica).*

Definitions were a common theme within uncertainty, such as unclear meaning of headers or identifiers, acronyms or abbreviations or other naming conventions. This seemed especially important for numerical values as there is often no further context given to a cell value or no information provided on what missing values mean:

> *(P24) Uncertainty what missing values mean was noted: This dataset is clear and is very dense although it is possible that the zero values in the set denote that the data could not be obtained.*

> *(P27) It's not clear how the 'magnitude' is measured, presumably it's the Richter scale but that isn't specified.*

Relations between columns, or dependencies between columns were mentioned within uncertainty.

(P1) It is unclear whether these are civilians who have been killed by police, or policemen who have been killed by, though I assume it is the former.

Temporal and geospatial attributes within uncertainty referred to unclear levels of aggregation or granularity of these attributes and potential ranges of values within a column. Furthermore, it seemed to be often unclear whether the data was up-to-date, and whether events in the data represent the time these were recorded or the time these happened. 19% of all mentions of uncertainty are connected to time and 28% to location:

(P14) All the data is related to 2015, although I do not know whether all the data about this year is contained in this dataset.

(P1) It is unclear to me whether these details are from the city, county, or state level.

Methodology: Uncertainty statements also presented questions related to methodology of data collection and creation. These covered aspects such as: *how were these numbers calculated, are they rounded, how was the data collected, what was the purpose of the data?* Some of these aspects refer to the provenance of the data and the importance of awareness of methodological choices during data creation was also found to be an explicit selection criteria in the results of the diary study.

Basic statistics.

Basic statistics about the dataset were one of the most prevalent features in the *analysis and usage* category (mentioned by 77% of all participants, with no significant differences in the occurrence per dataset. This included the number of rows, columns, or instances (such as the number of countries in the data). For instance: *'Size:* 468 *rows by* 32 *columns (incl. headers)'* or *'information on* 101,171 *earthquakes'*. Additionally, some summaries include the number of possible values which can be expected in a specific column, such as in this example for the header 'hair': *'HAIR - TEXT -* 23 *hair colours plus bald and no hair'*.

Possible values in a column were mentioned explicitly by 56.6% of all participants, most often in connection to D2. We assume that is because this dataset has a number of columns in which the range of values is limited. For instance headers referring to eye or hair colour or gender which have a limited number of possible entries:

(P20) The dataset also characterises whether the characters are good, bad, or neutral.

When there is a greater number of possible values these were presented through ranges or examples or by defining data types or other constraints for a column.

(P21) ID: Identity is secret/public/etc. ALIGN: Good/bad/neutral/-etc. EYE: Character's eye colour HAIR: Character's hair colour.

It is likely that the number of explicit mentions of possible values is under representing the importance of this category: As the participants were describing the dataset for someone else and in natural language we would assume that if the summary specifies e.g. 'age', there is no need to further explain this column presents the value type numbers as this would automatically be inferred, such as in a conversation between people. E.g. if there is a header called 'age', we expect the value type to be numerical.

4.6 Discussion of findings

We discuss the identified summary attributes, the results of the diary study and how these insights can inform the design of automatic summary creation. We compare our findings to existing metadata guidelines and detail the implications our results have on defining user centred dataset summaries. We conclude by discussing where we see the role of textual summaries, together with metadata, in the data discovery process.

4.6.1 Summaries attributes

We identified features that people consider important when trying to select a dataset *(RQ1)*, and when trying to convey a dataset to others *(RQ2)*, as can be seen in Table 4.8. Our findings address a gap in literature, relevant in the context of data publishing, search and sharing. We were able to see common structures and isolate different attributes that the summaries were made of *(RQ2)*, as can be seen in Figure 4.6. Summaries for the same dataset, created by different participants shared common attributes. We found a number of attributes tend to be less dependent on the underlying datasets, such as subtitle, format, headers and quality; whereas others tend to vary more depending on the data. Our findings allowed us to determine the composition and feasibility of general purpose dataset summaries, written solely based on the content of the dataset, without any further context.

Our findings suggests a range of datasets characteristics which people consider important when engaging with unfamiliar datasets. This analysis allows us to devise a template for the creation of text representations of datasets which is detailed in Section 4.7. Some of the attributes could be generated automatically, while others would still require manual input, for example from the dataset creator or from other users. We saw that all dataset summaries, as expected, explicitly describe the scope of the content in the dataset. Extracting content features directly from the dataset, and representing them as text is still subject of research, in particular in the context of extractive dataset summarisation

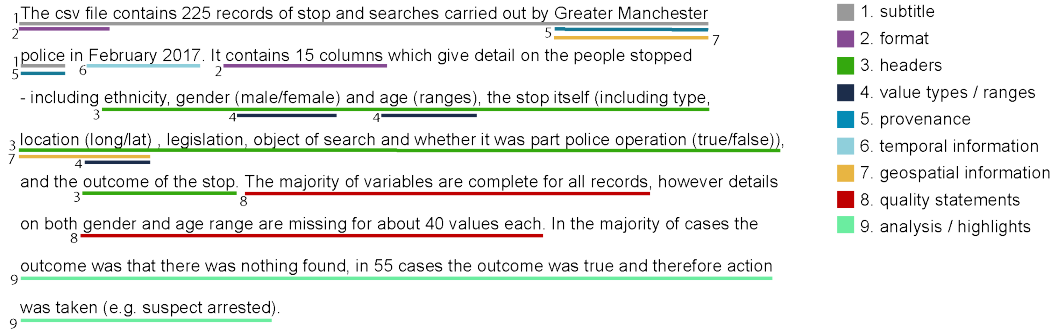


FIGURE 4.6: Example of an annotated dataset summary

(Ferreira et al., 2013) or semantic labeling of numerical data (Pham et al., 2016). Our findings can inform the design of these methods by suggesting parts of a dataset that matter in human data engagement.

At the same time, our analysis shows that most summaries also cover information that goes beyond content-related aspects, including groupings of headers into meaningful categories, the identification of key columns, and in some cases also the relationship between these and other columns in the dataset. These areas should be taken into account by data publishers when organising and documenting their data, and by designers of data exploration tools. For example, tools could highlight key columns and their relationships, or display structure overlaps that group headings in a relevant way. Furthermore, our summaries contained quality statements, some of which are complex as they refer to the potential context or use cases of the dataset; or an expression of uncertainty. We therefore conclude that purely extractive approaches will unlikely be able to produce useful text summaries of datasets that meet people's information needs.

While abstractive approaches to automatically generate summaries exist, we believe that the levels of abstraction and grouping needed for the creation of meaningful textual representations of data are not yet being realised. To be truly useful, a summary needs to be a combination of extractable features, combined with contextual information, human judgement, and creativity. This applies to selecting the right content to consider, as well as to representing this content in a meaningful way.

Comparing summaries created in a lab setting to those created in a crowdsourcing experiment gave us an understanding of the level of expertise or the *closeness* to the data that is needed to write a meaningful summary. It further gives insights into the feasibility of crowdsourcing as a potential method for dataset summary generation. We found dataset summaries can be produced using crowdsourcing, however, to fully reproduce summaries as they were created in the lab experiment crowdworkers could benefit from additional guidance, such as a template to support the summary writing process. We believe such a template would equally facilitate data publishers to write a comprehensive

and meaningful summary and is necessary for the development of automated dataset summarisation approaches.

Without this research, researchers and developers creating summaries would focus on obvious items such as column headers. This work demonstrates the importance of other aspects such as the grouping of headers, value types and ranges, information about data quality or usage suggestions – all attributes not commonly included in metadata. This highlights the difficult areas in fully automated approaches to summary creation. Understanding which attributes are considered important when selecting and describing datasets can focus future research efforts to deliver value to users. It can also be used to inform benchmark design for automated summary creation research.

4.6.2 Comparison to metadata standards and data search diaries

Table 4.8 shows a comparison between the results of the summary creation study, the outcomes of the analysis of data search diaries and current metadata standards. We can see that the attributes *basic stats, quality statements, patterns/trends and usage* are currently not represented in either of the two metadata schemas we discuss. Further differences include the grouping of column headers in meaningful semantic categories, the identification of a key column, and the importance of value types for the main columns.

We saw that many summaries, as well as the diary data suggest the usefulness of basic statistics about the dataset, such as the number of rows and columns, but also information on the possible values or ranges of important columns. These are potentially easy to extract from a dataset (e.g. Gatt et al. (2009)) but are not usually captured in standard metadata. In terms of geospatial and temporal attributes the main difference concerns the granularity of the information. Quality statements, initial analysis of the dataset content (patterns and trends) and ideas for usage are those attributes which are potentially complex to create but can be of great value in the selection process of datasets (Koesten et al., 2017). We believe that both provenance and methodology are under represented in the summaries due to the nature of the task and experiment design. Our work focuses on attributes people find important when selecting and describing datasets. However, whether the attributes should be represented in textual summaries or as structured metadata would be an interesting direction of future research.

Category	Summaries	Diary	Schema[13] and DCAT[14]
Format and file related info	file format, size of the file, personal data, last updated, license, unique identifiers	file format, API, access, unique identifiers, language, size	S: file format, license, identifier, url, D: size (bytes), format, identifier, language
Provenance	provenance: publisher, publishing organisation, temporal provenance: publishing date, last update, time of data collection geospatial provenance	publishing org (authoritative, reliable source), funding organisation (bias, independent source), original purpose (context)	S: author, contributor, producer, publisher, creator, editor, provider, source organisation D: contact point; publisher, landing page, sponsor, funder
Subtitle	high-level one phrase summary	title	S: main entity, about, headline D: theme, concept, keyword, title
Headers and **Groupings**	headers, selection and grouping of headers (+ explanation), key columns	headers, attributes/values and their meaning, value types (documentation)	S: variables measured
Geographical	geospatial scope (+ level of granularity)	location of publishing organisation, geospatial coverage (level of granularity)	S: location created, spatial coverage, content location D: spatial coverage, (spatial resolution - *pending evidence*)
Temporal	temporal coverage (+ level of granularity)	temporal scope, level of granularity, time of data collection (including time of the year), temporal provenance (time of publishing, up-to-date, maintained)	S: temporal coverage, content reference time, date created, date modified, date published D: temporal, temporal coverage, release date, update date, frequency of publishing, (temporal resolution - *pending evidence)*
Quality	*quality dimensions:* completeness consistency in formatting understandability (headers, acronyms, abbreviations) representativeness, coverage	*quality dimensions:* completeness, accuracy consistency in formatting, cleanliness understandability, clear provenance and authoritativeness of source	-
Basic statistics	ranges per column (possible values per column), counts of rows and columns, size possible value ranges and data types	units of measurement, upper/lower bounds to estimates,unique values for a column, comprehensiveness, range and variation, number of rows and columns	-
Patterns and Trends	analysis of the dataset content (patterns, trends, highlights)	-	-
Usage	ideas for usage	reasons not to use the dataset	-
Methodology	-	methods, control group, randomised trial, number of contributors, confidence intervals, sample and consideration of influencing factors, bias, sample time	S: measurement technique, variables measured

TABLE 4.8: Comparison of summary attributes to data-search diary and metadata standards (as per 5/2019). Summary = results from this study; Diary = Analysis of selection criteria in a data-search diary; (S) = Schema.org, (D) = DCAT (Data Catalog Vocabulary) – Attributes 'description' excluded

4.6.3 Making better summaries

In Chapter 3 we identified dataset relevance, usability and quality as critical to dataset search (Koesten et al., 2017). Relevance can be determined by having insights into what the dataset contains, and by analysing the data. Usability can be judged from the descriptive information in the summaries (such as format, basic stats, license, etc.). The quality and uncertainty statements expressed in the summaries deliver an assessment of dataset quality.

Individual attributes of the summaries could be generated using existing approaches, for instance from database summarisation methods some of which generalise column content into higher level categories, ideally describing the content in the column (Saint-Paul et al., 2005). Other approaches have tried to automatically identify the key column of a dataset (Ermilov and Ngomo, 2016; Venetis et al., 2011).

Granular temporal and location descriptions.

Among the results that confirmed existing best practices and standards were the prevalence of time and location in characterising datasets. These are commonly covered by existing metadata formats[15]. Our study has revealed a multitude of granularities in connection to these features, which are less well supported. The level of granularity of temporal or geospatial features of a dataset is crucial to understand its usefulness of a dataset for a particular task. This is reflected in the number of indications of these attributes in the summaries. Based on the results of this study we believe summaries should support users to determine whether a dataset has appropriate levels of aggregation for a given task.

Standard representations of quality and uncertainty.

Quality statements in the summaries included judgements on completeness, as well as assumed comprehensiveness of the data, errors and precision. Uncertainty statements referred to the meaning of concepts or values in the dataset (commonly including abbreviations and specialised terms) – which confirms findings in Koesten et al. (2017) – as well as unclear temporal or geographical scope of the data. Such statements illustrate the potential impact that good textual summaries and documentation can have for data users. W3C guidelines include completeness and availability as quality-related measures[16]. Our study shows that, especially in the more in-depth lab summaries, statements expressing uncertainty or sanctioning the quality of a dataset are very common. There is a body of research discussing how to best communicate uncertainty in visual

[15] http://schema.org/Dataset
[16] https://www.w3.org/TR/dwbp/

representations of data (for instance Boukhelifa et al. (2017); Kay et al. (2016); Simianu et al. (2016)). Understanding how to communicate uncertainty in textual representations of data, and furthermore, how this type of information impacts on the decisions of subsequent data users and on the ways they process the data, is comparatively less explored. Furthermore, previous research with data professionals has suggested that assessing data quality plays a role in selecting a dataset out of a pool of search results; studies such as Gregory et al. (2017); Wang and Strong (1996) have discussed the task-dependent and complex nature of quality. We assume that creating a more standardised way of representing uncertainty around datasets would be beneficial from a user perspective; related literature indicates that communicating uncertainty improves decision making and increases trust in everyday contexts (Joslyn and LeClerc, 2013; Kay et al., 2013).

Summary length.

One open question in the context of summary creation is the optimal length of a general purpose dataset summary. Regarding the effect of summary length – our study showed that the longer summaries produced in the lab experiment contained more qualitative statements which not only describe the data but judge the dataset for further reuse. This is not to say that, in all cases, the longer a summary, the better its quality. It is important to consider the likelihood that there is an optimal summary length, and surpassing this causes quality to decrease as the key elements of the summary become less accessible – which is an interesting area for future work. Determining snippet length in web search has been subject of numerous studies (for instance Cutrell and Guan (2007); He et al. (2012); Maxwell et al. (2017)), which generally suggest summary length influences relevance judgements by users. However, in this work we focus on summary content and not on summary length.

4.7 Dataset summary template

We present a template for user centred dataset summaries which can be incorporated into data portals, used by data publishers, and inform the development of automatic summarisation approaches.

Studies on text summarisation have found that people create better summaries when they are given an outline or a narrative structure that serves as a template, as opposed to having to create text from scratch (Borromeo et al., 2017; Kim and Monroy-Hernandez, 2016). Based on our findings, we propose such a template for text-centric data summaries. If used, it could improve current practices for manually written summaries, and potentially inform automatic data-to-text approaches as well.

Below we present the 9 questions that serve as the dataset summary template:

	TEMPLATE QUESTION	EXPLANATION
REQUIRED	**1. How would you describe the dataset in one sentence?**	What is the dataset about?
	2. What does the dataset look like?	File format, data type, information about the structure of the dataset
	3. What are the headers?	Can you group them in a sensible way? Is there a key column?
	4. What are the value types and value ranges for the most important headers?	Words/numbers/dates and their possible ranges
OPTIONAL	**5. Where is the data from?**	When was the data collected/published/updated? Where was the data published and by whom? (*required if not mentioned in metadata*)
	6. In what way does the dataset mention time?	What timeframes are covered by the data, what do they refer to and what is the level of detail they are reported in? (E.g. years/day/time/hours, etc.)
	7. In what way does the dataset mention location?	What geographical areas does the data refer to? To what level of detail is the area or location reported? (E.g. latitude/longitude, streetname, city, county, country, etc.)
	8. Is there anything unclear about the data, or do you have reason to doubt the quality?	How complete is the data (are there missing values)? Are all column names self explanatory? What do missing values mean?
	9. Is there anything that you would like to point out or analyse in more detail?	Particular trends or patterns in the data?

TABLE 4.9: Dataset summary template

These template items can be used as a checklist in the summary writing process. Our findings showed a dependency of attributes on the dataset content, mostly for temporal information, meaningful groupings of headers, provenance, basic stats and geospatial information (which may be an exception, as explained in the findings). Hence we suggest template questions number $1-4$ to be required, as they are generic attributes describing datasets. Number 5, a dataset's provenance, is usually provided in standard metatdata. Template questions number $6-9$ are considered to be optional in the summary, as they not necessarily applicable for all datasets. However, when applicable for a specific dataset questions number $5-9$ should be included in the dataset's summary.

The template focuses on attributes that can be inferred from the dataset itself, or on information that is commonly available in metadata, such as provenance. We do not include uncertainty about the dataset as a template question as the summaries have shown that uncertainty statements can refer to any of the categories of the template and is inherently dependent on the user.

We believe this template reflects the needs and expectations of data consumers, and can be adapted into current manual summarisation practices as a set of 'best-practice' guidelines, or by incorporating it directly into metadata standards. Initially each question could be translated into a semi-automatic questionnaire that extracts summary attributes, such as headers or basic statistics and guides the data publisher interactively through the summary writing process.

Use of this template could improve current practices for manually written summaries: the direct advantage is decreasing the burden for the publisher by reducing cognitive effort and contributing to standardising textual dataset summaries for datasets for the purpose of human consumption.

This template also has the potential to inform the development of automatic data-to-text approaches. The amount of support available to users could be increased through the use of machine learning techniques in data-to-text generation that are increasingly able to produce higher quality summarisation sentences, which could then be edited by the publisher.

4.7.1 From summaries to metadata

While our focus was on text summaries, the themes we have identified can inform the design of more structured representations of datasets, in particular metadata schemas as a primary form for automatically discovering, harvesting, and integrating datasets. Like any other descriptor, metadata is goal-driven, it is shaped by the type of data represented, but also by its intended use (Greenberg, 2010). Text summaries of data can be seen as metadata for consumption by people. They are meant to help people judge the usefulness of a dataset in a given context. Structured metadata, commonly in form of attribute value pairs, is potentially useful in this process as well; in fact, in the absence of textual summaries, people use whatever metadata they can find to decide whether to consider a dataset further. However, metadata records are primarily for machine consumption; they define a set of allowed attributes, use controlled vocabularies to express values of attributes, and are constrained in their expression by the need to be processable by different types of algorithms. This contrast is what makes text summaries of datasets so relevant for HCI – these are often the first 'point of interaction' between a user and a dataset (Koesten et al., 2017). Beyond that, we nevertheless believe that some of their most common content and structural patterns can inform the design of automatic metadata extraction methods, which in turn could improve dataset search, ranking, and exploration. For instance, knowing that the number of rows and headers in a datasets help users to determine a dataset's relevance, means these comparably easily extractable attributes could be included in automatic metadata extraction methods. Our results point to a number of attributes that could easily be extracted, but for which there is no standard form of reporting in general-purpose metadata schema. These include

descriptive attributes such as the mentioned numbers of rows and headers, possible value types and ranges, as well as different levels of granularity of temporal or geospatial information. A one-sentence summary, which has also been found to be useful by Yu et al. (2007) in a study on expert summarisation of time series data, or meaningful semantic groups of headers are more complex to create. Further complex features include the variety of elements which describe quality judgements and uncertainty connected to the data; and the identification of a key column.

4.7.2 Summary tool

The dataset summary template is an initial solution of how to guide a data publisher through the summary writing process. Based on our findings we propose an initial prototype of a dataset summarisation tool. The tool takes a CSV file as an input, performs simple pre-processing tasks and subsequently guides the data publisher with a questionnaire through the summary writing process. It then uses a rule-based template to produce a summary, both in a textual as well as a table representation (as used in the experiment in Chapter 5). The resulting summary can be edited by the writer and subsequently downloaded as either a text file, a table, a CSV file, or a JSON file. (Being able to edit and control the summary output is recommended in literature on text generation solutions (for instance Sripada et al. (2004).) The design, requirements and the interaction flow of the tool can be seen as part of the contributions of this thesis; the technical development however was done primarily by a developer. The requirements include the template question with corresponding sub-questions aimed to guide and support the summary creation process. It specifies the expected input mode for the answer (e.g: text input) and whether or not the answer is required.

Figures 4.7 to 4.9 provide an overview of the interactive questionnaire, a detailed description of the requirement as well as of each step of the summary writing process can be seen in Appendix B.7 as well as on a GitHub repository[17]. Questions marked with a red asterisk are required, others are optional.

[17] https://data-stories.github.io/data-summary/

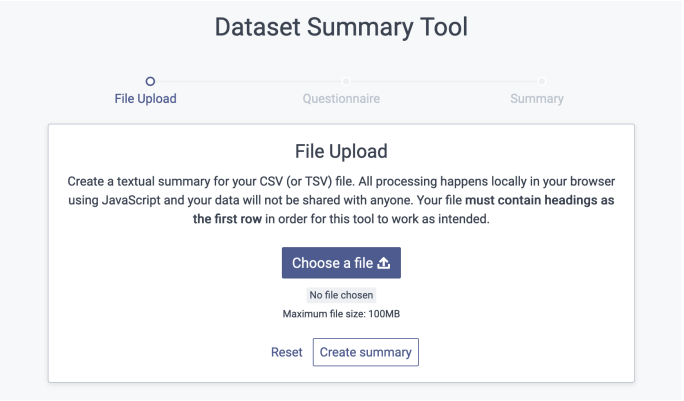


FIGURE 4.7: Dataset summary tool: Landing page

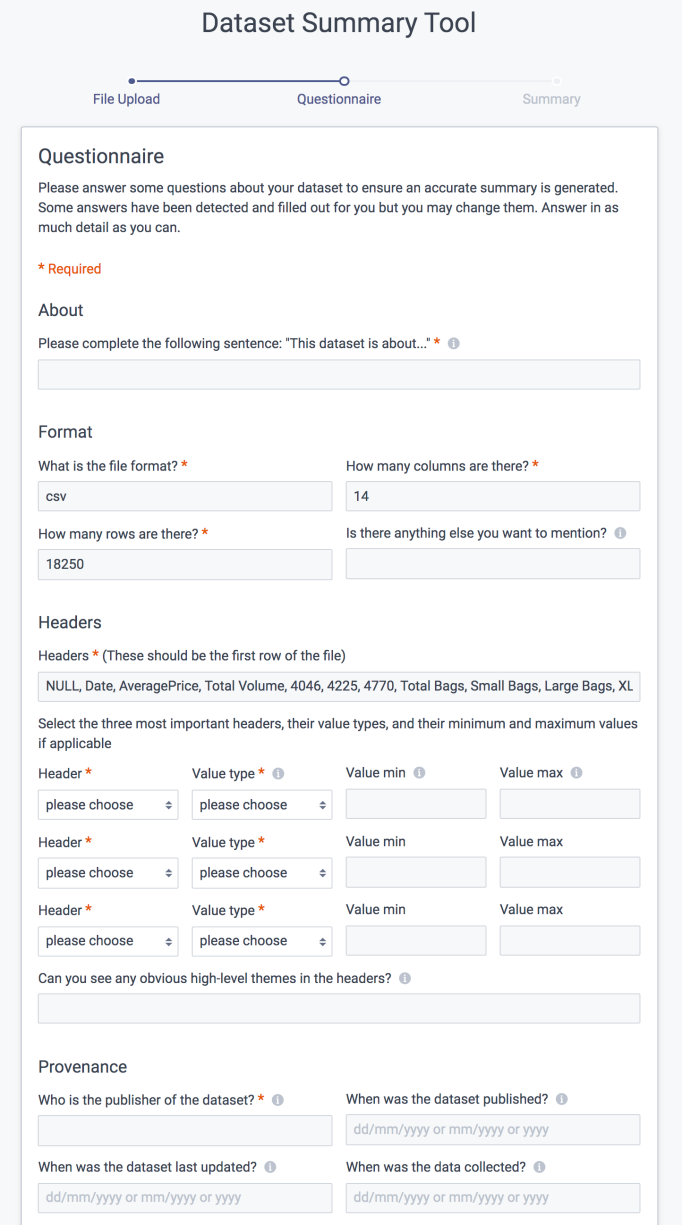


FIGURE 4.8: Dataset summary tool: Questionnaire (1/2)

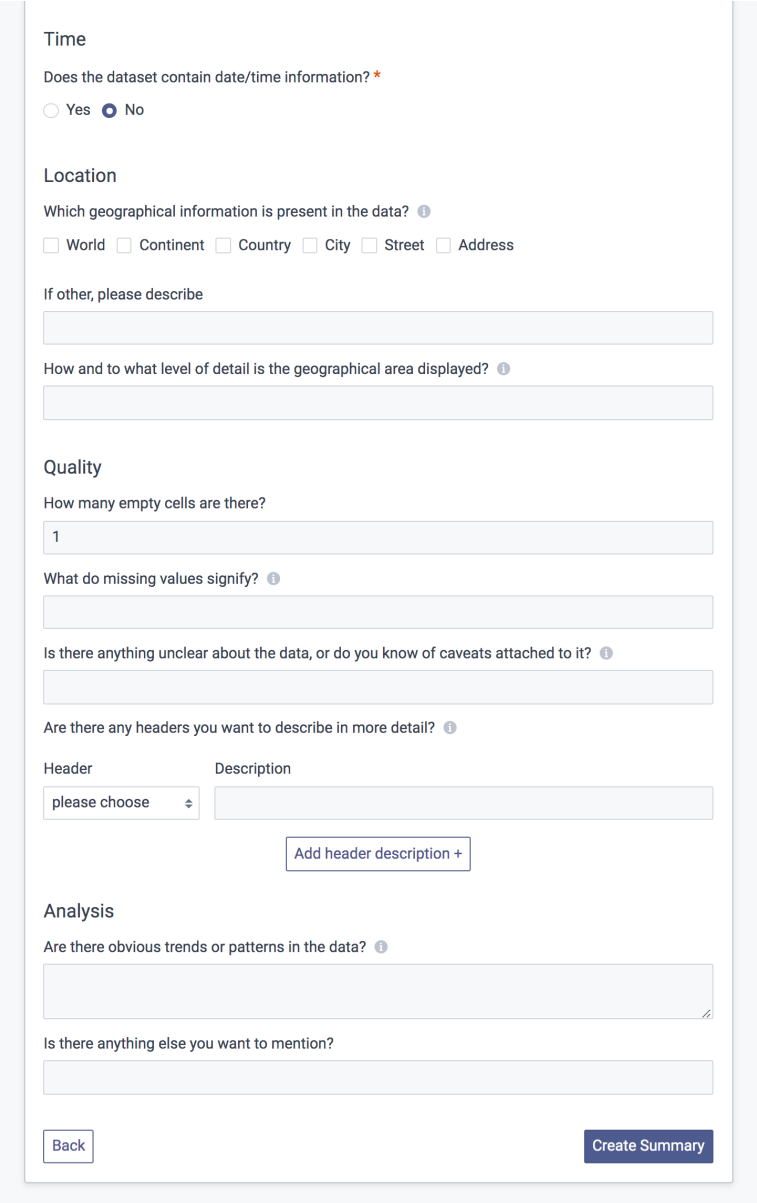


FIGURE 4.9: Dataset summary tool: Questionnaire (2/2)

4.8 Limitations

The dataset searching diaries consisted of set questions that asked the person writing the response to think about the partially subconscious selection process of datasets in an abstract way and requires them to articulate their information needs. Although that is a potentially complex task our findings suggest that participants expressed real information needs and the results generally overlapped with those in Chapter 3. However, observational studies could be done in future work to confirm or complement these findings through different methods. This could include a controlled lab study, with several means to log user behaviour such as search queries and refinements, session length,

eye-tracking and voice-recording. The diaries described in this study were conducted over a period of two weeks, with a submitted diary record only after a data search session. As the results are self-reported we cannot verify whether the diaries contained all search session students conducted during that time. However, we believe the diary entries contained valuable insights in dataset selection criteria in this context and have the benefit of being conducted in a more natural setting than a lab experiment, without being intrusive.

There are several confounding factors in the task of summary generation, due to the complexity of the task, which was also discussed in previous research on textual summary creation (Bernstein et al., 2015). The overarching aim of this study was to gain an understanding of peoples' conceptualisation of data, within the boundaries of this task (instructions, environment, time constraints). We did not specify the desired output in our lab experiment in terms of structure, style, choice of features and type of language, as we wanted to see what type of summaries people produce without guidance.

The study was carried out using files in CSV format; while we assume that elements of the summaries, particularly the high-level information extraction, would likely remain the same for all structured or semi-structured data, the description of the structure and representation (such as the number of rows, headings, etc.) of a dataset using a different format might vary. We found the particular datasets influenced the composition of the summaries in some instances, such as quality statements, geospatial attributes and provenance. However, despite these differences, we believe that there were sufficient commonalities in the summaries, both between datasets and the methods used, to derive recommendations and identify directions for further improvement.

Our datasets were relatively small, all in the same format, openly available and (as they came from different news sources) represent topics of potential public interest. While we do not believe that the resulting summaries are exhaustive for all data search needs, we believe they are applicable to the majority of Open Governmental and research datasets published on the web, in that they can give an initial insight into the dataset and have the potential to significantly improve the data search experience. We acknowledge that in any attempt to develop a more standardised way of documentation in a domain as open as data search, guidelines will not fit every scenario to the same extent. This is why we chose a variety of dimensions in the dataset sample, aiming for a template that covers different types of data and could potentially be extended for more specific requirements. Due to the explorative nature of our research question we believe there is a large space for further research investigating the applicability and comprehensiveness of these summaries to other types of data and for them to be tailored to domain specific contexts.

Participants in the lab experiment were data literate and used data in their work, but did not necessarily classify themselves as data professionals. As a result, they may

not have been aware of additional needs of data professionals, such as information on licensing or formatting that might have been mentioned, had they more specialised knowledge. Furthermore, we suspect personal data would in reality play a bigger role in a different study, for relevant datasets. More research would be needed to understand how summaries would change when sensitive information is present.

We used publicly available datasets that are not known to be popular, though we cannot be certain that none of our participants were familiar with the datasets. However, literature on text summarisation found that prior knowledge did not have significant effects on written summarisation performances (Yu, 2009). While we believe that there is intrinsic value in textual summaries of datasets - as they cannot only be used to inform selection by users, but could also be useful in search - we do not test the best representation of summary content in this work. Further studies are needed to determine optimal presentation modes of summary content for user interaction in a dataset selection activity.

4.9 Summary

This work contributes to the emerging field of human structured data interaction by presenting, to the best of our knowledge, the first in-depth characterisation of human-generated dataset summaries. The two studies helped us identify, on the one hand, dataset attributes which people find useful to make sense of a dataset, and, on the other hand, attributes they choose to describe a dataset to others. Both informed the design of the summary template. Our aim was to create practical, user-centric guidelines for data summarisation that reflect the needs and expectations of data consumers rather than what data publishers consider important.

With the overabundance of structured data, techniques to represent it in a condensed form are becoming increasingly important. Text summaries serve this function and they have the potential to make data on the web more user friendly and accessible. The studies described in this chapter contribute to a better understanding of human-data interaction, identifying attributes that people consider important when they are describing a dataset. We have shown that text summaries in our study are laid out according to common structures; contain four main information types; and cover a set of dataset features *(RQ1)*. This enables us to better define evaluation criteria for textual summaries of datasets; gives insights into selection criteria in dataset search; and can potentially inform metadata standards.

We conclude that our results are consistent enough between different participants and between different types of datasets to assume their generalisability for our scenario *(RQ2)*. We found general overlap between the information needs expressed in the data-search diaries *(RQ1)* and in the summaries created as a result of this study. Based on a subset

of attributes, we found that summaries of data practitioners have a higher prevalence of provenance, quality statements and usage ideas as well as a slightly more geospatial information. We also found that a number of attributes depend more on the dataset than others and which could influence the application of the dataset summary template.

The summary template is primarily meant as a tool for data publishers, but also for data scientists and engineers. It could be integrated into data publication forms alongside common metadata fields. It could also help build data-to-text algorithms that do a better job at reflecting the information needs and expectations of summary readers, and improve dataset indexing strategies, which are currently relying on metadata (Reiche and Höfig, 2013; Marienfeld et al., 2013). The findings of the two studies also suggest much needed extensions to existing metadata standards in order to cover aspects such as the numbers of rows and columns in a dataset; the levels of granularity of temporal and geospatial information; quality assessments; and meaningful groupings of headers.

Our results further suggest that crowdsourcing could be applied for large-scale dataset summarisation, however the validity would need to be studied in more depth. This study gives first insights into the feasibility of such an approach. Furthermore, when indexing dataset content to support search, we need to make a selection of important attributes based on what people search for and choose to summarise about a dataset. These attributes might vary in domain specific contexts, or might require extension to be more conclusive in specific data search scenarios. In that context, it would be interesting to investigate summaries created, for instance by researchers from different fields as well as by statisticians or professional data scientists and investigate commonalities and differences.

Chapter 5

Summary evaluation

While we discussed the content of dataset summaries in detail in Chapter 4, in this chapter we focus on a study to evaluate such summaries in a data search scenario.

5.1 Motivation

As discussed in Chapter 2, in general web search we are used to being presented with a snippet, which is the short summarising text that is returned by search engines. This helps users to make a decision about the relevance of a search result when searching for documents (Bando et al., 2010). The snippets are automatically generated, usually based on content from the web pages and displayed on a Search Engines Result Page (SERP) (Marchionini and White, 2007). Search results on data portals are displayed in a similar way to web search, with a title and short snippet following the ten blue links paradigm. Clicking on a result commonly takes the user to a page that contains metadata and a description of the dataset.

Little research has to date looked specifically at dataset surrogates for SERPs and how to support dataset-specific selection scenarios. As discussed in Chapter 4, we currently have no established way of representing dataset search. While there is research on automated snippet generation from text (Au et al., 2016) it exists mainly in research settings and does not provide anything close to the same user experience that we know from web search. One of the open questions in this space is how to present datasets as search results to users, both from a content, as well as a presentation point of view.

The study in this chapter is, to the best of our knowledge, the first to investigate the effect of dataset summaries on interaction with the SERP. We examine how presenting dataset summaries (as described in Chapter 4) as 'snippets' affects user interaction with the SERP and their ability to select relevant search results. We are interested in how the content of result summaries for datasets affects the ability to select relevant over

non-relevant items. We investigate search behaviour and performance by varying the SERP between four conditions.

We created four SERPs that vary according to their content (and to some extent their layout). All four presentation modes of the SERPs follow a traditional ten blue links paradigm, but they vary in snippet content and layout. Two conditions include the dataset summaries from Chapter 4 as snippets: one in textual form, the other as a structured table. The two control conditions include a preview of the first few rows of the dataset, or only the title of the dataset. All conditions further display the file format and the publisher of the dataset and are described in more detail in Section 5.3 below.

We conducted a between-subjects online experiment with the following research questions:

RQ1: How does the dataset presentation mode (text summary, table summary, preview, title) affect search time, performance, user behaviour and experience?

RQ2: How does the presentation mode of a dataset summary (as text or as a table) affect search time, performance, user behaviour and experience?

We hypothesise that SERPs displaying dataset summaries will result in better performance (i.e. selection of more relevant results), more confidence in the selected results and a better search experience. We measure several metrics for performance, interaction and experience. The findings contribute to a better understanding of user behaviour in dataset selection scenarios, and inform the design of data-search SERPs and the development of automatic snippet generation.

5.2 Related work

5.2.1 Snippets

The search engine results page is core to a user's success and experience in interactive information retrieval (Maxwell et al., 2017). People spend most of their search time examining results returned by the search system (Marchionini and White, 2007). SERPs for general web search have been studied from numerous angles (layout, number of results, snippet length and content and features, ranking of results, etc. (Maxwell et al., 2017; Kelly and Azzopardi, 2015; Kammerer and Gerjets, 2010; Marcos et al., 2015)). However, we still know little about how SERPs for structured data search should look like. We see a large design space to explore how SERPs for datasets should be presented to users. Within this space, this work focuses on the usefulness of dataset summaries as a component of SERPs.

Data requires context to create meaning (Dervin, 1997), to *make sense of it*. Thomas et al. (2015) show that dataset repositories have poor search over and inside datasets. It is difficult for a user to tell from a repository whether a useful dataset is available, and this problem is only likely to get worse the more datasets are available. The display of a search result should provide the user with sufficient information to judge the relevance, quality and usability of the results of and often fails to do so (Koesten et al., 2017). In user studies with social scientists, Kern and Mathiak (2015) found that the quantity and quality of metadata are far more critical in dataset search than in literature search, where convenience prevails.

Currently, web search engines construct snippets from two or three sentences extracted from the document which have a close relationship with query terms (Bando et al., 2010). The ranking position has shown to be the most influential factor to determine relevance for web search (Craswell et al., 2008), however studies on snippets show that user performance, especially for informational tasks can increase significantly with informative snippets (Cutrell and Guan, 2007).

Several studies focus on automatically generating text summaries for data, however the focus is on the technicalities of the algorithm, rather than how useful they are in a search process. Most proposals are domain specific, for instance for medical data (Scott et al., 2013), sports data (Wiseman et al., 2017) or financial data (Kukich, 1983). None of these have been evaluated from a user experience point of view in a search context. Other work has looked at creating summaries of databases, or query-based subsets of databases, in which they try to generalise content to higher-level categories to produce a more condensed version of a column, or a table (Saint-Paul et al., 2005). Other approaches focus on summarising particular data types, such as time series (Yu et al., 2007; Sripada et al., 2003) or graphs (Liu et al., 2014). Au et al. (2016) describe a method to produce query biased summaries from tabular data which are presented as a graph or as a table in a data search scenario. In an initial small scale user study, these tabular summaries performed slightly better than textual summaries. However the automatically-created textual summaries used in this study are not coherent, or easy-to-read, and might therefore not represent the potential benefits of textual datatset summaries in a data search scenario.

5.2.2 SERP design

Interface design plays a key role in representing context (Greenberg, 2001). There is a large body of research exploring how SERP design and composition influences user behaviour. For instance, studies such as by Kelly and Azzopardi (2015) have explored how the number of results per page influences search behaviour and user experience. Resnick et al. (2001) looked at how a conventional list SERP compared against a tabular one. The columns of the table represented the different elements of the search results (title,

excerpt, URL, metadata). The results suggest that the tabular interface supported a wider variety of search strategies than the conventional list. While this is not comparable to our scenario, it gives insight into the context of comparing our two summary versions, which are natural language text and a tabular format (however, in our case on the level of an individual search result). Several studies compare list and grid interfaces. For instance, Kammerer and Gerjets (2010) noted differences in user behaviour and perceived trustworthiness of search results depending on their ranked position in a list. The effects are less pronounced than their position in a grid interface. For other search verticals, such as for images (Xie et al., 2017), videos (Schoeffmann et al., 2015), or in e-commerce (Rowley, 2000), it is common to present results differently to best support users in their specific information needs for the source. Some work has explored snippets for aggregated search results (Marcos et al., 2015; Arguello and Capra, 2014). We see a parallel with our scenario in the aim of condensing larger amounts of information into one snippet, similar to when trying to summarise large datasets.

5.3 Methodology

We conducted a between-subjects experiment in which participants were randomly assigned to one of four interface conditions, which are described in Table 5.1 and in Appendix C.1 to C.7. Two conditions (type 2 and 3) acted as baselines. The other two conditions were designed based on prior work in Chapter 4 to create general-purpose dataset summaries to aid search.

Viewtype	Description
`0 (text)`	dataset summary as text snippet
`1 (table)`	dataset summary as table
`2 (title)`	low baseline
`3 (preview)`	higher baseline

TABLE 5.1: Interface conditions

As shown in Figure 5.3, all tasks had a search box resembling general web search and a search button. All SERPs presented in the task followed a ten blue links paradigm with 10 results per page. Each participant completed three search tasks and was presented with one interface condition, which remained unchanged across the three tasks.

The two conditions including a dataset summary can be seen in Figure 5.1 and 5.2), and each presentation mode can be seen in Appendix C.4.

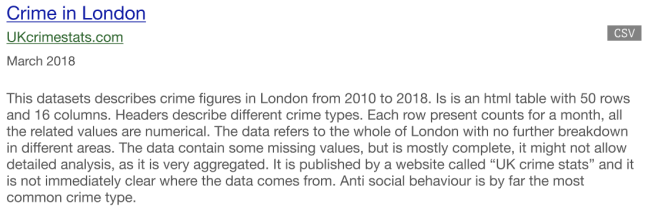


FIGURE 5.1: Viewtype 0 - text

FIGURE 5.2: Viewtype 1 - table

The independent variable is one of four conditions of the presentation mode of the SERP page:

- Textual summary of the dataset according to the template presented in Chapter 4, plus metadata (file format, publishing organisation, publishing date).
- Tabular summary of the dataset according to the template presented in Chapter 4, plus metadata (file format, publishing organisation, publishing date)
- Title of the dataset plus metadata (file format, publishing organisation, publishing date)
- Preview of the first few rows of the dataset plus metadata (file format, publishing organisation, publishing date)

Recruitment

The 174 participants were recruited using social media and mailing lists. These included lists targeted at university students, researchers and open data users. The recruitment email and a description of participant demographics can be seen in the Appendix (C.2 and C.6). Participants were incentivised with the option of taking part in a prize draw to win one of two £50 Amazon vouchers.

5.3.1 Tasks

The survey consisted of a consent form, a pre-task questionnaire, a set of tasks (described below) and three post-task questions. The tasks were chosen and constructed based on tasks described in the information seeking study from Chapter 3, and informed by Borlund (2003)'s notion of simulated work-tasks. These are particular types of tasks, in which a pre-defined goal can be reached by an assigned unit of work. The goal is situated in an 'embedding task' which describes the cognitive environment in which the search is conducted and establishes the criteria for evaluating the search results. In our study we focused on the problem that needs to be solved, as it makes the participant understand the objective of the search (Kelly, 2009). All tasks can be understood as requiring informational search activities, as defined by Broder (2002), who distinguishes between navigational and transactional search activities. The tasks were rotated in the experiment to prevent learning effects and fatigue (Kelly, 2009).

Table 5.2 shows the tasks used in the study. All three tasks were structured in the same way, including a temporal and a geospatial component.

	Information seeking tasks for data
1 Pay Gap	Imagine you want to write a report on the gender pay gap in the UK and how it has changed over the last five years across industries. You need to find data on people's salaries, split by gender and industry, over the last five years so you can understand in which industry the gender pay gap has decreased the most.
2 Crime	Imagine you want to write an article about crime rates in London over the last ten years. You need to find data on crime rates, split by area, over the last ten years so you can understand in which part of London crime rates have changed the most.
3 Obesity	Imagine you want to write an article about obesity trends over the last thirty years in the UK. You need to find data on obesity, split by age, over the last thirty years so you can understand which age ranges have the highest growth in obesity.

TABLE 5.2: Information seeking task for data

5.3.2 Process

The survey was built using SurveyJs[1] and hosted on Heroku. No identifying data was collected about participants, with the exception of email addresses if participants wanted

[1] https://surveyjs.io/Overview/Library/

to take part in the prize draw we used for incentivisation - in this case the data was anonymised for analysis.

Participants were approached via a call on social media or email, in which the study goal and process were described with a link to the survey. The first page included information about the study, data protection and an estimation of the time it would take to participate. This page served as the informed consent form to which the participants needed to click to agree. They completed a pre-task survey, including demographic questions on age range, gender, country of residence and job role. They were further asked three questions, each of which captured on a Likert scale, (i) frequency of search for information online; (ii) frequency of search for data online; (iii) whether they use data for their work and if yes, what for. Data was defined as '*information (can be numbers or words) in tables, spreadsheets or databases*'.

Participants were then presented with a paragraph describing a task with an information need for which they need data as the information source, as can be seen in Table 5.2. The interface displayed a search search bar resembling web search, highlighting that the search is for data (Figure 5.3). They typed a query and were subsequently presented with ten search results, all of which datasets.

To make the search activity as realistic as possible the interface displayed a search box and subsequently search results. Participants could always scroll up to the task, which was indicated by a scroll bar. They were allowed to freely formulate search queries in an interface that resembles traditional web search (Figure 5.3).

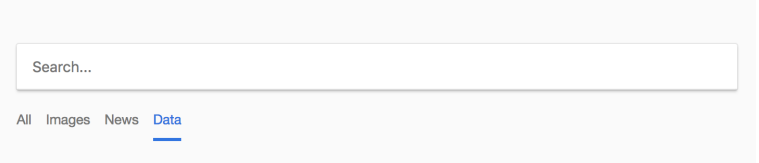
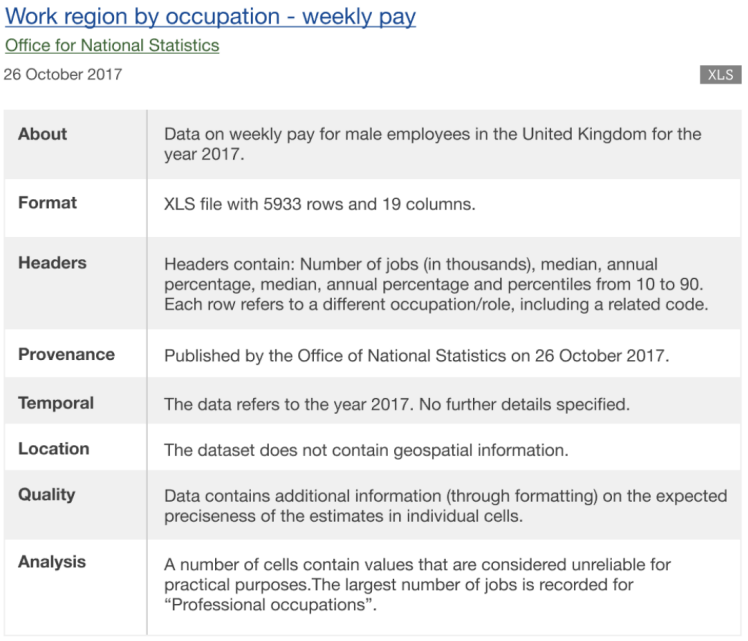


FIGURE 5.3: Searchbox

All participants were presented with the same search results, in different order, displayed in different ways according to the experimental condition. An example of a search result for viewtype 1 (table) can be seen in Figure 5.4:



Work region by occupation - weekly pay

Office for National Statistics

26 October 2017 XLS

About	Data on weekly pay for male employees in the United Kingdom for the year 2017.
Format	XLS file with 5933 rows and 19 columns.
Headers	Headers contain: Number of jobs (in thousands), median, annual percentage, median, annual percentage and percentiles from 10 to 90. Each row refers to a different occupation/role, including a related code.
Provenance	Published by the Office of National Statistics on 26 October 2017.
Temporal	The data refers to the year 2017. No further details specified.
Location	The dataset does not contain geospatial information.
Quality	Data contains additional information (through formatting) on the expected preciseness of the estimates in individual cells.
Analysis	A number of cells contain values that are considered unreliable for practical purposes.The largest number of jobs is recorded for "Professional occupations".

FIGURE 5.4: Viewtype 1 (table) of a relevant dataset for Task 1

Participants were asked to select those datasets that they perceived to be relevant for the described information need. There was no facility to download the data; as we were only interested in the decision processes based on the SERP. Subsequently they were asked to answer three post-task questions:

- *How difficult did you find the task?*
- *Do you think you had enough information to judge whether a dataset was useful for the task?*
- *Was there anything missing that would have helped you to decide if the dataset was useful for the task?*

After selecting search results (by clicking on them) participants were asked to review their selection of datasets and determine their confidence in whether each selected dataset was useful for the given task; as shown in Figure 5.5:

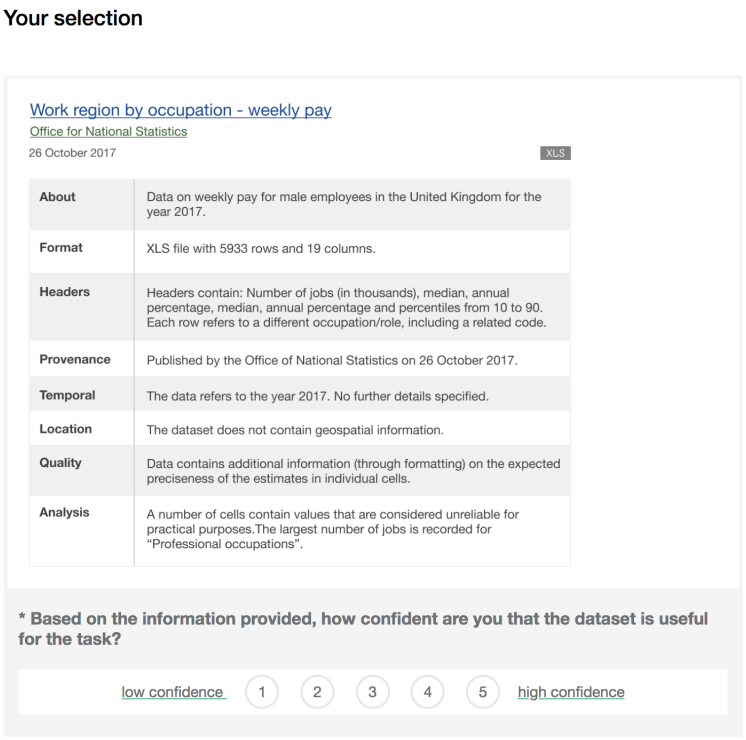


FIGURE 5.5: Confidence rating

5.3.3 Corpus

Datasets were collected by performing Google searches and searches on two UK data portals for the given tasks. All datasets were among the top-20 search results. Each SERP also contained one control dataset, which was clearly off-topic, and was therefore not within the top-20. Usefulness judgements for each dataset against the respective tasks were performed by three independent experts, all of whom work with data in their jobs. The judgements were made on a 5-point Likert scale (not useful to very useful). The results were aggregated to be binary: all ratings above 3 were reported as relevant and the final relevance judgement was established using a majority vote. The ratings reached moderate agreement (Fleiss' Kappa ($\kappa = 0.52$)) amongst the three raters (Landis and Koch, 1977). In total, each SERP contained 10 results.

To avoid position bias (Craswell et al., 2008) the result list was randomly reordered for each participant and each task. For task 1 (pay gap) there were four relevant results in the result list, for task 2 (crime) three and for task 3 (obesity) there were five. Screenshots of the presentation mode can be seen in Appendix C.4.

Summaries

For the condition 0 (viewtype text) and 1 (viewtype table) summaries of the corresponding datasets were presented as a textual snippet or a table. These were manually created based on the template provided in Chapter 4. The content of the textual summary and

the table are the same (using the attributes of the template), the difference is in the presentation and the table contained less natural language text.

5.3.4 Metrics

In this section we describe the choices in our experimental design. In our setup we have four conditions, two of which are displaying the dataset summaries in different ways and two types of baselines to compare with: a preview as it is commonly used on dataportals (e.g. data.gov.uk [2]) and, as an alternative, merely the title of the dataset. This design is inline with literature: in IIR evaluations baselines are often introduced as one of the experimental conditions, instead of measuring them in advance as described by Kelly (2009) in her extensive work on evaluating IIR systems with users. We chose a between-subjects design to not expose subjects to all conditions of the experiment and so avoid contamination (Kelly, 2009). This allows us to focus on the effect of the interface rather than requiring a conscious choice from participants between interfaces. The primary objective of the design is to compare the four presentation modes and not the tasks. The tasks function as control variables and to discuss potential learning effects; we expect minimal differences between them. The order of tasks is controlled through rotation (Kelly, 2009).

We collect self-reported user characteristics in the form of a pre-task questionnaire to describe our sample and potentially explain differences in performance (Boyce et al., 1994).

Time-based measures describe the time participants interact with the search results, which is a common measure in IIR to understand the effectiveness of a SERP as well as potentially the depth of engagement with the search results (Kelly, 2009).

We recorded the time from issuing a query (through pressing enter or the *search data* button) until selection of all datasets considered useful and pressing *next* to leave the page. We then measured differences between the time it took participants to select datasets as search results between viewtypes using a Mann-Whitney U test, as the data was not normally distributed.

Performance and interaction-based measures are related to the outcome of the interaction, such as the number of relevant documents selected and the mean average precision (the mean of the average precision scores for each query). These are commonly used in IIR experiments and typically computed from log data (Fenichel, 1981). We report on: the total number of selected search results, the number of relevant (true-positives) and non-relevant selected (false-positives) search results, and the self-reported confidence of participants of whether each selected result is in fact relevant for the task.

[2] https://data.gov.uk/

As self-reported *experience measures* we chose 'usefulness' and 'perceived amount of information', measured on a 5-point Likert scale, as can be seen in Appendix C.5. We elaborate in Chapter 2 on the difference between relevance and usefulness in the context of IIR evaluations. We test differences between viewtypes using a Kruskal Wallis test, as we compare four groups and the data is considered to be ordinal.

We further present a *qualitative analysis* of free text comments participants could add after completing the task to *complement* the results and add context by presenting insights into the experience during the task (Bryman, 2006).

5.3.5 Ethics

The study was approved by the University of Southampton's Ethical Advisory Committee[3]. Consent was given by participants through a tick box to initiate the online survey. The consent form can be seen in Appendix C.3. It was possible to take part in the study anonymously.

5.4 Findings

We report the results of our study according to time-based, performance-based and interaction-based metrics, looking at differences between the four viewtypes.

5.4.1 Participants

The majority of our 174 participants were from the UK ($n = 119$), some from the US ($n = 9$) or Austria ($n = 7$), Germany($n = 6$), Italy ($n = 4$) and France ($n = 4$). A detailed breakdown of countries can be seen in Appendix C.6. Their age range was between 18-25 ($n = 61$) and 26-35 ($n = 62$). The rest of the participants were older. 61% of participants were female ($n = 107$) and 35% male ($n = 62$), the rest either classified themselves as other or preferred not to say. While this might not be a representative distribution of gender amongst data users, we do not assume that gender influences the result and report this number merely to describe our sample in more detail. Most participants ($n = 162$) reported that they very often search for information online. 81% of all participants stated to use data for their work ($n = 142$), so the assumption is that the sample can be described as at minimum basic data literate. When asked how often they specifically search for data online 34% of the participants reported to very often search for data online and 35% at least once a week. 28% rarely search for data online, but only 1% of the participants stated to never search for data online. (We report results

[3]ERGO Number 41384

by viewtype, the number of participants was for viewtype 0 (text) (n=42), for viewtype 1 (table) (n=37), for viewtype 2 (title) (n=42) and for viewtype 3 (preview) (n=53).)

5.4.2 Time-based results

For time-based results we recorded the time from issuing a query until selection of all datasets considered useful to pressing *next* to leave the page. The dataset selection task took participants on average between 2.25 to 5.5 minutes to complete, as can be seen in Table 5.3.

Viewtype	Total mean time	med	std	min	max
`summary as text(0)`	257.3	199.6	208	8.7	766.6
`summary as table(1)`	333.2	273.8	305	24.6	1633.9
`title(2)`	135	123.9	72.8	24.2	396.7
`preview(3)`	215.4	179.7	159.2	39.6	967.1

TABLE 5.3: Total time across all tasks per viewtype in seconds

We tested the data for normal distribution using the Shaprio-Wilk test and plotting of the data and found that it was not normally distributed.

Our hypothesis was as follows:

$H0$ = the distribution of total task time between viewtypes is equal
$H1$ = the distribution of total task time between viewtypes is not equal

A two sided Mann-Whitney U test has shown that viewtype 2 (title) was statistically significantly different from all other viewtypes ($p < 0.0001$) with a Bonferroni corrected significance level alpha $\bar{\alpha} = 0.0083$. $\bar{\alpha}$ is the significance level alpha = 0.05 divided by the total number of comparisons (n=6). We know that for viewtype 2 (title) the participant had less information on the SERP to read, as can be seen in Appendix C.4, so it is unsurprising they took less time to select datasets.

Non significant observations
However, we further observed that on average participants took longest for viewtype 1 (table) (mean = 333 s , std = 305 s), followed by viewtype 0 (text) (mean = 257s , std = 208 s) and were slightly quicker for viewtype 3 (preview) (mean = 215s , std = 159 s). The mean and standard deviation for the total time per viewtype can be seen in Figure 5.6. We hypothesise that the longer overall task time with viewtype 1 (table) could point to a more in-depth engagement with the table representation. The structured layout of the table potentially prompts a more thorough investigation by participants, which takes more time.

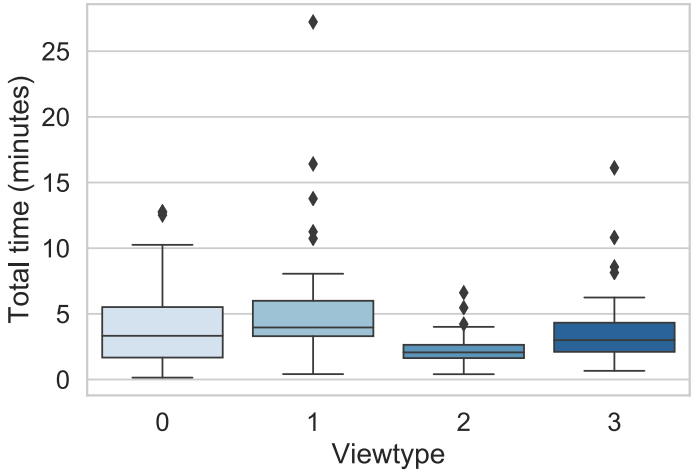


FIGURE 5.6: Total time grouped per viewtype

5.4.3 Interaction and performance based results

Total number of selected results

Each participant could choose a maximum of 10 results for each of the three tasks. The number of relevant results were three for task 'crime', four for task 'gender' and five for task 'obesity'. Hence we expected a total of 12 correct results over all tasks. We observed that the average of selected results over all tasks was lower (mean of 10.8 for viewtype 0, 10.1 for viewtype 1, 10.6 for viewtype 2 and 9.6 for viewtype 3; compared to the total of 12 correct results across all tasks). A two sided Mann-Whitney U test showed no statistically significant differences between the total number of selected results.

Non significant observations

However, we observed differences in the average number of selected results per viewtype. For viewtype 1 (table) and viewtype 3 (preview) participants selected slightly fewer results overall. This could point to a more targeted selection process, potentially because the table and the preview condition displayed content that was easily accessible and matched the participants selection criteria. However, this assumption would need to be validated in future work.

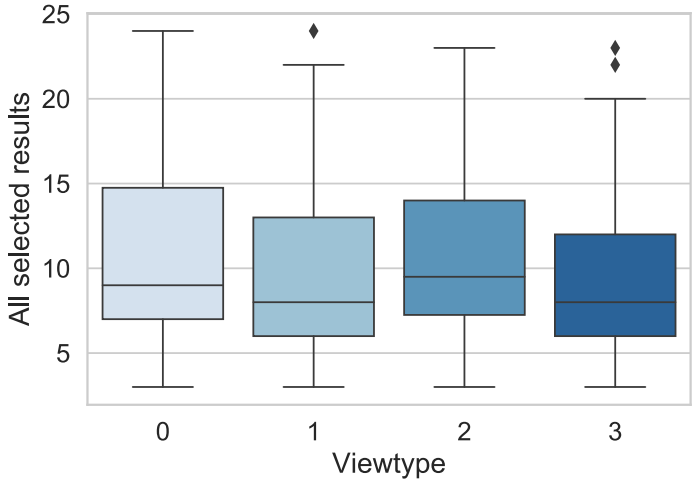


FIGURE 5.7: Total number of datasets selected per viewtype

Viewtype	T1	T2	T3	Av. selected	Av. relevant selected (out of 12)	min	max
`text(0)`	3.6	3.5	3.7	10.9	3.9	3	24
`table(1)`	3.2	3.1	3.8	10.1	3.6	3	24
`title(2)`	3.5	3.4	3.8	10.7	3.6	3	23
`preview(3)`	2.8	3.2	3.5	9.6	3.45	3	23

TABLE 5.4: Selection of search results (mean). T1 = Task 1'gender', T2 = Task 2'crime', T3 = Task 3 'obesity'

True positives

We describe the number of correctly selected relevant results across the three tasks. A two sided Mann-Whitney U test showed no statistically significant differences in the number of selected and relevant results between viewtypes.

Non significant observations
However, we observed that across all tasks viewtype 0 (text) and viewtype 2 (title) resulted in a higher median of relevant and selected results. Viewtype 2 (title) showed a larger range in comparison to viewtype 0 (text). Viewtype 1 (table) and 3 (preview) performed similar to each other.

Viewtype	mean	med	std	min	max
`text(0)`	3.88	4.0	2.03	1	8
`table(1)`	3.65	3.0	2.06	1	9
`title(2)`	3.62	4.0	1.75	0	8
`preview(3)`	3.49	3.0	1.74	1	7

TABLE 5.5: True positives - selected relevant results

The average number of correct choices over all tasks was four. The table and the preview viewtype both led to fewer wrongly selected results in total. This mirrors the tendency

to select fewer search results for viewtype 1 (table) and 3 (preview), as can be seen in Figure 5.8.

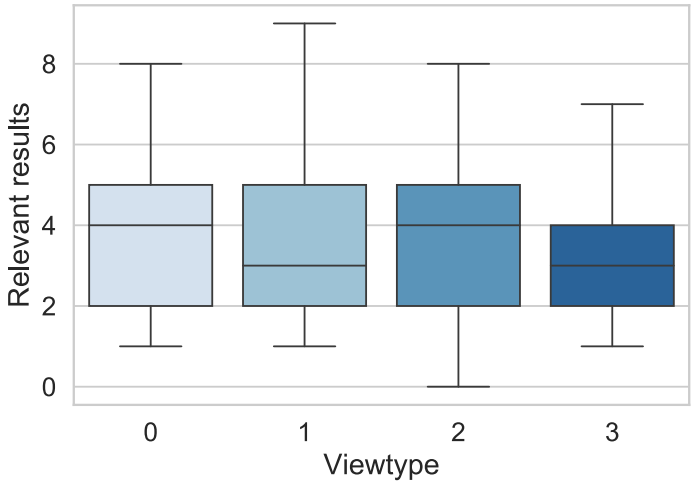


FIGURE 5.8: True positives selected per viewtype

False positives

We describe the number of non-relevant selected results across the three tasks. A two sided Mann-Whitney U test showed no statistically significant differences in the number of selected and non-relevant results between viewtypes.

Non significant observations
The maximum number of non-relevant results a user could choose was $n = 18$. We observed that across all tasks the average number of selected non-relevant results was lower for viewtype 1 (table) (mean = 6.46, median = 5, std = 4.17) and viewtype 3 (preview) (mean = 6.13 , median = 5, std = 3.69) and higher for viewtype 2 (text) (mean = 7.05, median = 6.5, std = 3.85).

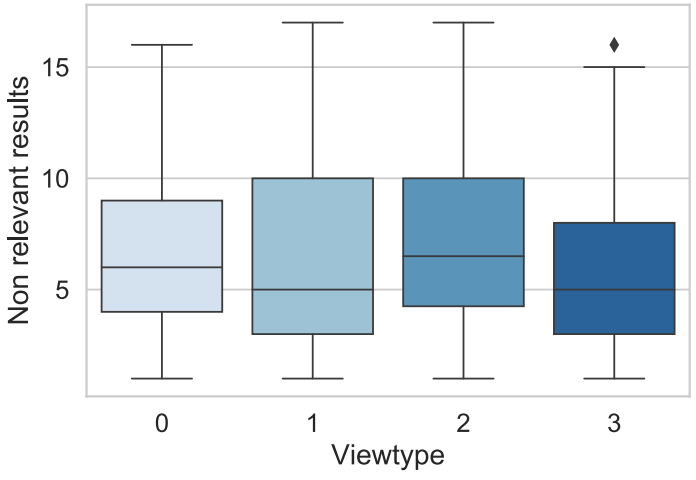


FIGURE 5.9: False positives selected per viewtype

Viewtype	mean	med	std	min	max
`text(0)`	6.98	6.0	3.80	1	16
`table(1)`	6.46	5.0	4.17	1	17
`title(2)`	7.08	6.5	3.85	1	17
`preview(3)`	6.13	5.0	3.70	1	16

TABLE 5.6: False positives - selected, but not relevant results

5.4.4 Interaction based results

We collected interaction based measures on a 5-point Likert scale. To understand differences between groups we used a Mann-Whitney U test, as Likert scale responses are considered ordinal (Kuzon et al., 1996).

Confidence

For each selected search result participants were asked to rate their confidence that the dataset is useful for the task on a 5-point Likert scale (1 - low confidence to 5 - high confidence).

Our hypothesis was as follows:

$H0$ = Confidence in selecting a useful dataset for the task is rated equally between all viewtypes
$H1$ = Confidence in selecting a useful dataset is not rated equally between all viewtypes

A Kruskal Wallis test showed significant differences of confidence scores between viewtypes (H(2) = 8.34, p = 0.039). For pairwise comparison, a two-sided Mann-Whitney U test has shown statisitically signifcantly different mean confidence scores between viewtype 0 (text) and viewtype 1 (table) with $p < 0.01$ as well as between viewtype 1 (table) and viewtype 3 (preview) ($p < 0.05$). However, follow-up Bonferroni tests with a corrected significance level alpha of $\bar{\alpha} = 0.0083$ showed that the significant difference existed only between viewtype 0 (text) and viewtype 1 (table). $\bar{\alpha}$ is the significance level alpha = 0.05, divided by the total number of comparisons (n=6).

As can be seen in Table 5.7, the lowest average confidence score was 3.5, which considering the 5-point Likert scale is more confident than neutral. We observed that the confidence scores across all tasks were highest for viewtype 1 (table) ($mean$ = 3.85, $median$ = 4, std = 0.72). We hypothesise that the structured layout of the table enabled participants to easily target those attributes that matter for dataset selection in the context of the task.

Non significant observations

There were no significant differences between tasks, however task 3 showed a higher variability of confidence scores for viewtype 1 (table) ($mean = 3.91$, $median = 3.95$ $std = 0.91$).

Figures 5.10 and 5.11 show the average confidence rating per viewtype across all tasks. Figure 5.12 to 5.17 show the same information split by task to illustrate the slight variations in confidence for each of the tasks. The tasks varied in the number of relevant results that were represented in the result list, as well as in their topic. We see a slightly higher variation for viewtype 1 (table) in task 2, and to some extent in task 3. Task 1 had the lowest number of as relevant results, which might have an influence on this result. Overall, as expected, the differences between tasks were not significant.

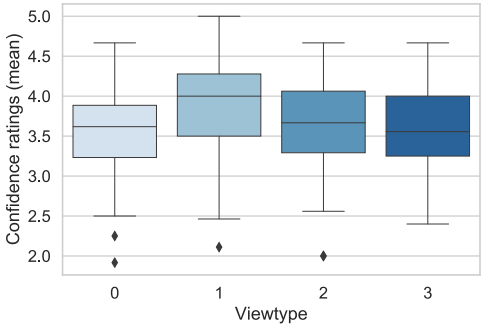


FIGURE 5.10: Average confidence ratings per viewtype across tasks

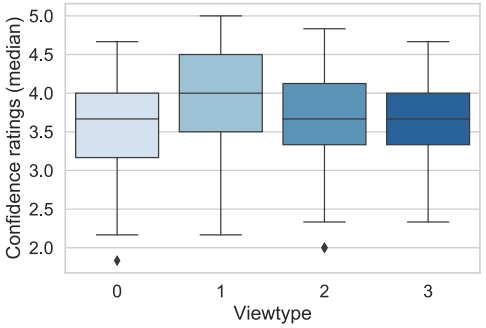


FIGURE 5.11: Median confidence ratings per viewtype across tasks

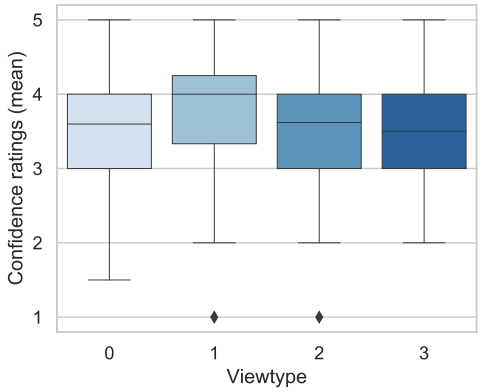


FIGURE 5.12: Average confidence ratings per viewtype (task 1)

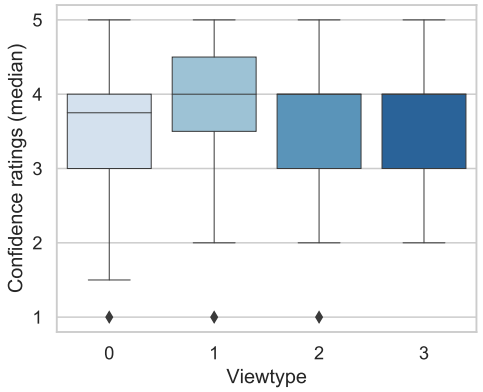


FIGURE 5.13: Median confidence ratings per viewtype (task 1)

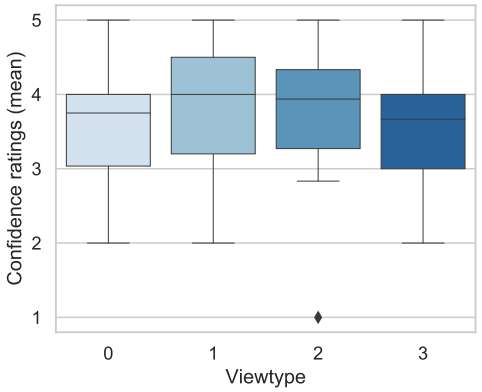


FIGURE 5.14: Average confidence ratings per viewtype (task 2)

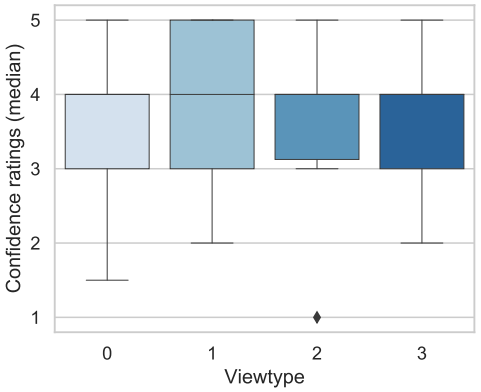


FIGURE 5.15: Median confidence ratings per viewtype (task 2)

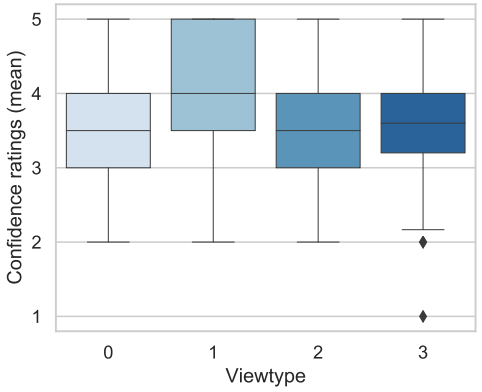


FIGURE 5.16: Average confidence ratings per viewtype (task 3)

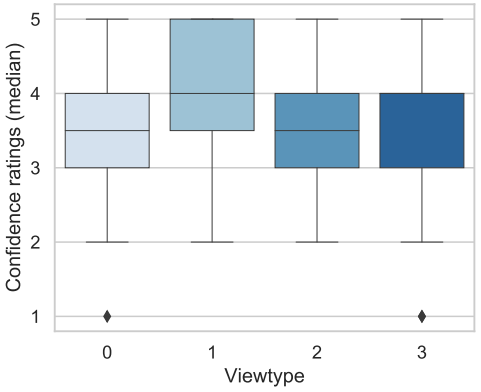


FIGURE 5.17: Median confidence ratings per viewtype (task 3)

Viewtype	mean	med	std	min	max
`text(0)`	3.5	3.67	0.56	1.9	4.7
`table(1)`	3.85	4.00	0.72	2.1	5.0
`title(2)`	3.62	3.67	0.62	2.0	4.7
`preview(3)`	3.60	3.67	0.57	2.4	4.7

TABLE 5.7: Confidence ratings across tasks based on a 5-point Likert scale

5.4.5 Experience based results

We collected experience based measures on a 5-point Likert scale.

To capture user experiences, we asked subjects to complete post-task questions, as described in Section 5.3.2. These questions asked about the perceived difficulty of the tasks (1 - not difficult to 5 - very difficult) and about whether the participants felt they

had enough information to judge whether the datasets were useful for the tasks (1 - not useful to 5 - very useful).

We used a Kruskal-Wallis test as a non-parametric method to analyse variance, which uses the median for comparison between more than two groups of ordinal or continuous variables.

Perceived difficulty

Perceived difficulty was captured in the post-task questionnaire on a 5-point Likert scale.

Our hypothesis was as follows:
$H0$ = The perceived difficulty is rated equally between all viewtypes
$H1$ = The perceived difficulty is not rated equally between all viewtypes

A Kruskal Wallis test showed no significant differences between viewtypes for perceived difficulty.

Non significant observations

However, Figure 5.18 shows a larger variability for viewtype 2 (title) with a tendency to rate perceived difficulty higher for this viewtype (mean = 3.6, std = 1.17, median = 4). Viewtype 0 (text) and viewtype 3 (preview) were perceived to be slightly less difficult (viewtype 0, mean = 3.3, std = 0.98; viewtype 3, mean = 3.25, std = 1.11).

Viewtype	mean	med	std	min	max
`text(0)`	3.3	3.0	0.1	1	5
`table(1)`	3.57	4.0	1.07	1	5
`title(2)`	3.56	4.0	1.17	1	5
`preview(3)`	3.25	4.0	1.12	1	5

TABLE 5.8: Perceived difficulty based on a 5-point Likert scale

Perceived amount of information

Perceived amount of information was captured in the post task questionnaire on a 5-point Likert scale.

Our hypothesis was as follows:
$H0$ = The perceived amount of information is rated equally between all viewtypes
$H1$ = The perceived amount of information is not rated equally between all viewtypes

A Kruskal Wallis test showed significant differences for the perceived amount of information per viewtype. ($H(2) = 14.80$, $p < 0.001$), with a Bonferroni corrected significance level alpha $\bar{\alpha} = 0.0083$. $\bar{\alpha}$ is the significance level alpha = 0.05 divided by the total number of comparisons (n=6). A two sided, equally Bonferroni corrected Mann-Whitney U test showed that the differences exist between viewtype 1 and viewtype 2 ($p < 0.001$), as well as viewtype 1 and viewtype 3 ($p < 0.01$). Without the correction for multiple comparisons there was a also a difference between viewtype 0 and viewtype 2 ($p < 0.05$).

Figure 5.19 shows that for viewtype 2 (title) the participants rated the amount of information shown to decide whether the dataset is useful for the task as less satisfactory than for the other viewtypes (viewtype 2, mean= 2.36, std = 1.12, median = 2).

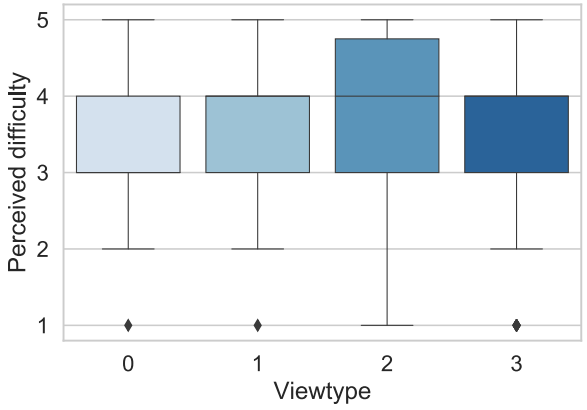


FIGURE 5.18: Perceived difficulty, captured on a 5-point Likert scale

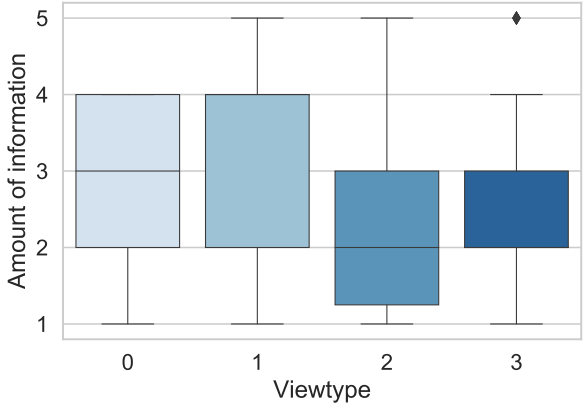


FIGURE 5.19: Amount of information, captured on a 5-point Likert scale

Viewtype	mean	med	std	min	max
`text(0)`	2.85	3.0	1.04	1	4
`table(1)`	3.38	4.0	1.32	1	5
`title(2)`	2.36	2.0	1.12	1	5
`preview(3)`	2.57	2.0	1.23	1	5

TABLE 5.9: Amount of information across the four conditions based on a 5-point Likert scale

5.4.6 Qualitative findings

After completion of all three tasks participants were asked whether they believed anything was missing in the presentation of the search results. In total we collected 102 comments. Qualitative results can add context to quantitative results and indicate where to focus research questions for further research (Bryman, 2006). The comments varied according to viewtype and are presented on an exemplary basis. Viewtype 0 (text) had 24 comments, viewtype 1 (table) 15 comments, viewtype 2 (title) 31 comments and viewtype 3 (preview) with 32 comments. Note that, due to the between-subjects design, participants' comments do not reflect a comparison of different viewtypes, as they were only presented with one viewtype each. Hence, when they suggest different presentation types they were not prompted to do so, and were not influenced by other viewtypes in the study, which may indicate a stronger statement.

Comments for viewtype 0 (text)

Comments for viewtype 0 (text) mentioned that it would be useful to be presented with an excerpt or preview of the data, mostly to get an understanding of the column headers:

> *(P) Data set column headings*
>
> *(P) Small sample from dataset to quickly evaluate included fields*
>
> *(P) I would like to see a sample of the data (i.e. a couple of rows). More structured information (e.g. in bullet points or a table) would have been better, as it would have made the datasets easier to compare.*

Some participants also mentioned preferring a more structured presentation than the text snippet and suggested a table presentation of the summary:

> *(P) Quick view of the data. Make it easier to scan the descriptions – table rather than paragraphs.*

When presented with only a textual summary some participants expressed the need to get a better idea of the data itself and the usefulness of being able to quickly see the included variables or columns.

(P) Easier way to filter by years included in data.

Comments for viewtype 1 (table)

On the other hand when presented with the table summary a number of comments also mentioned the desire to see a preview of the data:

(P) Example of the first 2 rows of the dataset.
(P) Seeing the actual data set.
(P) A preview of the dataset.

One participant mentioned that summaries need to be trusted as they present curated content:

(P) No everything is there... just need to believe it.

Comments for viewtype 2 (title)

For viewtype 2 (title) participants expressed the need for additional information about the dataset:

(P) For a lot of the items that I marked as less confident, I would like to know what the dataset covered.
(P) A short summary of the 'document'.
(P) The data content to decide whether it can use or not.

Some participants described particular attributes they would like to know about, such as temporal information, provenance or collection methods:

(P) Additional metadata e.g. date ranges.
(P) It would be nice to know how the data was derived.
(P) Maybe knowing at a glance the different variables included – like whether analysis is by age, or time etc. Also time spans covered rather than just year of publication.

Some explicitly mentioned previews of the data, to better understand what the dataset contains:

(P) A preview of eg. the header lines of a csv would be very useful, so that you can assess what information you might be downloading.

Some participants specifically mentioned that a summary or snippet of the data would help them decide:

(P) It would be helpful to offer a quick 1 to 2 sentence summary about the data within the dataset so that better judgement could be made about a dataset during the searching process.
(P) Having a snippet of the dataset would have been useful.

Others stated that given the information they were presented with (the title) they would need to open the dataset to make a selection:

(P) I would have looked at the chosen datasets before deciding on their usefulness. This might mean having to go through the ones not chosen at times.

In summary, the comments suggest that the viewtype 2 (title) was missing information that participants need to select a dataset out of a pool of search results.

Comments for viewtype 3 (preview)

For viewtype 3 (preview) comments discussed the need of additional information about the dataset, such as about provenance or temporal scope:

(P) More detailed metadata of when, how and by whom data were collected and the dataset was generated.
(P) Source of Data, more information on provider of data.
(P) More records e.g.: to see whether more than one year is covered [..].
(P) Sample size and date of when data was collected.

Other comments explicitly mentioned a summary:

(P) More detailed summary of the data.
(P) More data and a summary of what the data shows, if possible.

Others mentioned the need for further documentation and metadata, such as

(P) Description of the dataset, source with confidence, visualizations would also be useful.
(P) Once the table was selected still to understand them was a bit hard because the legend of the table was missing. I at least would give hint to understand the information of the table (legend or what the acronyms means).

The comments for viewtype 3 (preview) suggest that additional information, in addition to displaying the first few rows of a dataset would be perceived as helpful in the selection process of datasets in a search scenario.

5.5 Discussion of findings

In this study, we investigated the influence of different search results presentation types on search performance, interaction and experience. The analysis was focused around addressing the two research questions, which explored (RQ1) *How does the dataset presentation mode (text summary, table summary, preview, title) affect search performance, user behaviour and experience?* and (RQ2) *How does the presentation mode of a dataset summary (as text or as a table) affect search performance, user behaviour and experience?*

In this section we discuss the differences we observed regarding interaction and performance-based metrics, as well as experience-based metrics and how these potentially relate to one another. We hypothesise that with a higher number of participants we would have observed more statistically significant differences between viewtypes.

Intuitively, viewtype 2 (title) takes less time to look at as there is less information displayed, which was reflected in our results. We further found that on average viewtype 1 (table) took participants the longest time to select datasets. On average both versions of the summary conditions took participants longer to select datasets compared to the baselines without summaries. This potentially suggests a deeper engagement with the search results when being presented with a dataset summary on the SERP. Some comments point towards participants' interest for even deeper engagement, such as to understand comparability between different datasets already on the SERP:

> *(P) Would be good to know if the locations in one data group could be compared and weighted against other data sets. Ie. is location and date in one data set comparable to the same location data with ages accompanied to get more complete data?*

> *(P) Consistent and clearly presented definitions of the dimensions of each dataset*

In the study in Chapter 3 participants reported that when searching for data 'in the wild' they lack sufficient information to make an informed selection of a dataset, which confirms the findings of this study.

Viewtype 1 (table) also showed the highest confidence ratings on average, we hypothesise that there is a connection between the time participants examined the search results and their confidence that they had selected relevant search results. In IIR studies investigating snippet performance those which take a shorter time can be considered to be more effective (Kelly, 2009), due to the associated lower cost and effort. In our scenario however, we hypothesise that participants engaging with the SERP for a longer time is a sign of higher interest and deeper exploration of the search result, similar to Maxwell et al. (2017) who investigated snippet length in web search. Thus, longer search time does not need to be seen as an indicator of negative performance in this context.

Looking at the total number of selected search results we found that participants selected on average fewer results for viewtype 1 (table) and viewtype 3 (preview). While this difference was not found to be statistically significant, it is consistently reflected for both the number of relevant select results, as well as the non-relevant ones that were selected by participants. We assume this points towards a more targeted selection of search results. For viewtype 1 (table) this is confirmed by higher confidence scores on average in contrast to the other viewtypes. We hypothesise that the table representation allows users to easier target those attributes that make sense in the context of the data search task, which then make them more confident in their selection. Interestingly, the highest average number of correctly selected results could be observed for viewtype 0 (text) but this viewtype also showed the lowest average confidence scores. This could mean that the textual summary might lead to more precision in the selection of search results, but without making users more confident in their selection. For a successful interaction with a SERP users should both be able to select relevant results, but also be confident in their own selection (e.g. trust their selection is correct). However, these results have to be interpreted with caution due to their limited statistical significance.

We observed that viewtype 2 (title) had the highest number of selected but not-relevant results, which supports our assumption that merely displaying the title might lead to less targeted selection of search results. Viewtype 2 (title) was also on average perceived to be more difficult than the others and was considered the least useful, but that was not reflected in the performance measures. We hypothesise that a larger sample and a completely naturalistic setting in which participants follow their own information needs might display this trend also in the ability to select relevant search results, as the qualitative comments suggested a clear perception of not having enough information to make an informed selection. The nature of our result pool and of our tasks might afford people to select datasets they might not normally select. In a natural setting they might exit the search tasks even prior to selection, however in our study they were forced to select a minimum of one dataset.

Looking at the qualitative results, the need for summarisation was expressed both by those commenting on the title as well as on the preview viewtype. However, when presented with a summary (either as text or as a table) participants also mentioned the need to get an idea of what the data looks like. They described that seeing a preview of the data or an easy to digest summary of all headers or columns of the dataset would aid to properly evaluate the usefulness of a search result, in addition to the summary. These comments were more prevalent with the viewtype 0, the text view, than for the presentation as a table. We assume this is due to the higher structuredness of the table summary which might makes a quick assessment of the dataset easier. When comparing traditional lists of search results to a tabular interface, Resnick et al. (2001) found that the tabular structure supported faster consumption of the SERP and was preferred by users. Here they changed the design of the entire SERP into a tabular structure, however

each search result included additional structured facets, such as data, size or keyword count. While this presentation mode is not directly comparable to our table viewtype, as we presented each of the dataset summaries as an individual table, but the SERP design was still a list, we believe that it indicates that search results as tables might be an interesting direction for further research.

We hypothesise that with a larger sample size the differences between participants confidence in their selected results as well as. Potentially the difference in the metrics we measured were too small to be picked up by such a small scale study. One of the reasons we chose an between-subjects design was to avoid participants being biased by seeing different versions of the SERP. Ideally, such a set-up would allow a more objective measure of the effect of SERP design on performance and the other metrics than the subjective choice of a within-subjects design. It would be interesting to explore this in future research.

For viewtype 2 (title) the comments clearly showed that participants did not perceive to have been given enough information to make an informed search result selection. Some comments just expressed the need for more information about the covered content, others explicitly mentioned the need for a summary or snippet and a high level content description of the dataset. Interestingly, both the time span covered by the dataset as well as information about the datasets creation were mentioned by the participants. This confirms findings from our previous two studies presented in Chaper 3 and 4.

Overall, we found that participants did not perceive viewtype 2 (title) as displaying enough information to judge whether a dataset is useful for the data search tasks we explored in this study. This was confirmed both by the results of the post task questionnaire, as well as our qualitative results.

A small scale user study by Au et al. (2016) tested the presentation mode for automatically created query-biased summaries from structured data and suggests a preference for non-textual summaries. However, their textual summaries are limited in scope and fluency and are so not comparable to what we refer to as meaningful summaries in the context of this work.

The table summary performed best in the interaction and experience based metrics that we measured. Based on the findings of this work, including insights from Chapter 3, we hypothesise that a table summary together with a dataset preview would perform best in a dataset search scenario. However, this would need to be validated in future research.

Our study was conducted using a between-subjects design, in contrast to many user studies comparing interfaces. This was a choice in experimental set-up to avoid influences of the different presentations modes during the study. We believe that asking

participants for their preference between different viewtypes might not create valid results as personal preference might not indicate the viewtype with the best performance and confidence. Being able to compare between viewtypes allows participants to draw conclusions about the expected results of the study and can so lead to contamination (Kelly, 2009). Therefore we chose a between-subjects design, despite the expected difficulties with this approach. The disadvantage of this set-up is that a larger number of participants is needed to create reliable results. Although our overall participation was > 170, when divided by four groups a larger sample size could further improve the statistical significance of the results. However, we believe that our results are potentially more meaningful as they represent user performance and interaction completely focused on the viewtype under concern and so indicate important directions for further research.

We believe it is worth considering observational methods for studying the effect of different dataset SERP's on user interaction in more detail as the performance focused nature of log data potentially misses the rich, in-depth interaction that users engage in when selecting a dataset out of a pool of search results. This could include eye tracking studies as well as think aloud protocols. These methods could allow a better understanding whether people look at the summaries and use them in their decision making process of selecting a dataset.

5.6 Limitations

As described in the methods, our tasks did not include participants' real information needs. However we designed the tasks carefully, based on real information needs that require datasets as the information source from our previous studies, amongst other considerations.

The presentation of the summaries could be done in many different ways. Although we displayed tables for one of the viewtypes the general structure of our SERP was still following a ten blue link paradigm. There are approaches investigating other presentation layouts for data, as mentioned in Chaper 2, but these are outside the scope of this study.

Due to the nature of the study we could not control for confounding variables in the participants' environments. They are conducting the study online and unsupervised. Therefore the study cannot be treated as a controlled experiment.

We further cannot control for a self-selection bias of participants. Our results showed that that more than 80% use data for their work, so we can assume that our sample was fairly data literate and UK centric (just under 70%). We believe that similar studies with participants who are less data literate are of importance to design inclusive data search interfaces.

5.7 Summary

Overall, we found that participants did not perceive viewtype 2 (title) as displaying enough information to judge whether a dataset is useful for the data search tasks we explored in this study. This was confirmed both by the results of the post-task questionnaire, as well as participants free text comments; and is in line with the qualitative findings from Chapter 3. Generally, the results somewhat suggest that viewtype 1 (table) and viewtype 3 (preview) performed best, and we believe a combination of the two presentation modes would be an interesting direction to investigate. Viewtype 1 (table) presents a structured version of the summaries discussed in 4, in which each row represents one attribute of the proposed template and the results suggest that this presentation mode leads to higher confidence in participants selection of search results. The experiments in 4 have shown that both our lab participants and crowdworkers can produce dataset summaries of similar quality. This leads to the assumption that the task of describing individual attributes (from the summary template) might be easier to perform, as constructing a free text paragraph without guidance is know to be a complex task. Furthermore, a table version of a dataset summary would not need grammatically correct sentences, but focuses more on a bare version of the content of the attributes which might be easier to produce automatically. A further advantage of a table view is its clear structure and the potential to extract particular rows where possible from the dataset and rely on human input for those attributes that we are currently not able to produce automatically.

This study can be used as an example of how IIR experiments for dataset search can be conducted and there is a large number of possible variations that would be interesting to investigate, including different task types; a more diverse set of participants both in terms or location and data literacy; a higher number of participants; or different presentation modes of the summary content – potentially interactive views that allow a more in-depth exploration of the datasets already on the SERP.

Chapter 6

Discussion

In this chapter we discuss the contributions of this work and synthesise the insights gained through the different studies. We then discuss future research directions based on the lessons learned.

Due to the exploratory nature of the research space, which is still emerging, we first conducted a broad scoping study about the information seeking process for structured data. Subsequently, we identified the themes as the focus of this work: selection criteria and summarisation of datasets for the purpose of selecting a dataset in a search scenario. We used a mixed-methods approach combining semi-structured in-depth interviews with data professionals, supported by a search log analysis of data search queries from the UK open governmental data portal (Chapter 3). The research questions were directed at people's *search strategies, their work tasks with data, the selection criteria they apply when choosing a dataset, as well as their exploration strategies for new data.* We found that data search is perceived to be a major issue for data professionals and that they often rely on human recommendation. Participants reported often not having enough information about the content and context of a dataset to make an informed decision without downloading and looking at the data. We found that specific relevance, usability and quality aspects were perceived to be different for data than for documents – for example, the methodology used to collect and clean the data, the values missing, the granularity of the captured information, as well as the ability to understand the schema used to organise a dataset.

In Chapter 4 we presented two complementary studies in which we looked at selection criteria for datasets to validate the findings from Chapter 3 and we explore the composition of dataset summaries for the purpose of selecting suitable datasets in a search scenario. With the overabundance of structured data, techniques to represent it in a condensed form are becoming more important. Through a lab experiment and a crowdsourcing experiment we investigated what data attributes people choose to mention when summarising a dataset to others. We identified common attributes that people consider

important when they are describing a dataset and contribute so to a better understanding of Human Data Interaction. We showed that text summaries are laid out according to common structures, contain four main information types, and cover multiple dataset attributes. This enabled us to better define criteria for textual summaries of datasets for human consumption and give additional insights into the selection criteria that apply to dataset search.

In Chapter 5 we described a follow-up study that investigated how presenting dataset summaries (as suggested in Chapter 4) as snippets, affects user interaction with the SERP and the ability to select relevant search results. We found indications that presenting dataset summaries as a table on a SERP (potentially together with a preview of the dataset) is an interesting direction to explore further.

6.1 Contributions

The main contributions of this work lie in advancing our knowledge of user behaviour in dataset search regarding the following topics: *categorisation of data-centric work tasks, search strategies in data search, selection criteria in data search, exploration activities for datasets, composition of dataset summaries and insights into SERP design and IIR experimentation for data search.* We further present insights into information seeking for structured data in a framework for Human Structured Data Interaction, synthesising findings from this work.

6.1.1 Framework for Human Structured Data Interaction

We presented the results of the study in Chapter 3 as a framework for Human Structured Data Interaction. This provides a novel perspective on the conceptualisation of data-centric tasks, a better understanding of people's search strategies, as well as of the selection and exploration strategies in data search.

The five pillars of this framework are: *tasks, search, evaluation, exploration* and dataset *use*, as can be seen in Figure 6.1.

An initial version of the framework was suggested based on findings from the study described in Chapter 3, and the further work described in this thesis confirmed the individual pillars as described below. The adjustments resulting from further insights are explained, per individual pillar, below.

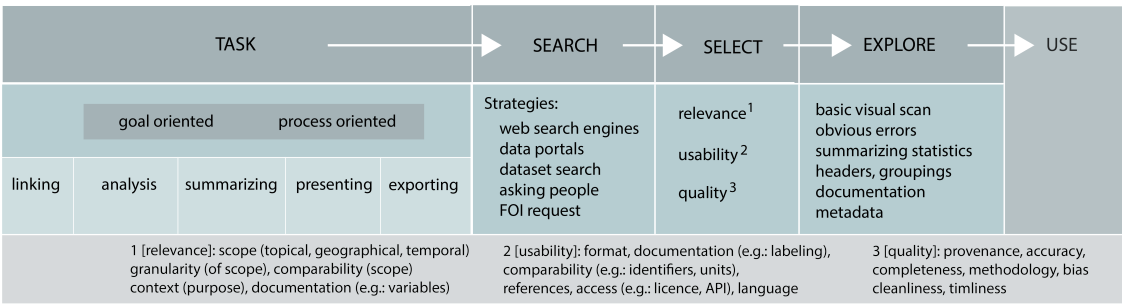


FIGURE 6.1: Framework for interacting with structured data

6.1.1.1 Taxonomy of data-centric work tasks

Based on the qualitative study in Chapter 3 we proposed a taxonomy of data-centric tasks oriented on the requirements expressed by our participants for selection criteria with two dimensions - *process* or *goal-oriented*. Process-oriented tasks describe tasks in which data is used for something transformative, such as using data in machine learning processes; and goal-oriented tasks in which data is, e.g., used to answer a question. 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 attached information needs (the details a user needs to know about the data in order to select it). 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.

We proposed another taxonomy, taking a different perspective. Focusing on the specific activities a user intends to accomplish with data, we differentiate between five types of data-centric tasks: *linking*; *analysing*; *summarising*; *presenting*; and *exporting*. This can inform system designers and publishers of data to cater functionalities to different interaction requirements of each task type.

In the initial framework we only discussed time series analysis as an example for analysis tasks. We adjusted the tasks described in framework to the higher-level term *analysis* as this better describes the diversity of potential data analysis activities. The students data searching diaries described in Chapter 4 also reflected this wider definition of data analysis, for instance by mentioning comparison or modelling tasks.

These taxonomies, and the broader framework they are part of, are aimed to help system designers and publishers of data understand what people do when searching for and engaging with datasets.

6.1.1.2 Better understanding of search strategies in data search

The study in Chapter 3 gave insights into search strategies, both from in-depth interviews with data professionals, as well as by analysing search logs of a data portal. We found that dataset search is often exploratory, involving iterations and complex cognitive processing. Data professionals in our study reported often struggling to find the data that they search for. Queries for datasets are short and under-specified Kacprzak et al. (2019) and we found that participants were not confident that the search functionality will be able to provide relevant data for longer and more specific queries. It appears that the search box of a data portal is currently treated more as a starting point for further exploration, rather than as a direct access route to a suitable dataset. Hence, human recommendations were reported as a widely used search strategy by the majority of the participants. We included dataset search in the final framework to reflect recent developments for dataset specific vertical search engines.

6.1.1.3 In-depth knowledge on selection criteria in data search

We learned that people experience difficulties finding datasets, also due to the fact that the information they need to evaluate a datasets fitness for use is not always available or easy to interpret out of context. The findings in Chapter 3 showed that participants evaluate three higher level categories when selecting a dataset's usefulness for a task: *relevance, usability and quality.* While these are common categories in selection criteria for information objects in general (Bales and Wang, 2005), our findings suggest that the practical details of what this means are specific to datasets.

The analysis of data search diaries described in Chapter 4 showed similar results and contributes to the robustness of our findings. This allowed us to refine and extend the insights into selection criteria by including a wider variety of attributes with more in-depth examples.

We adjusted the terminology of the initial framework from *evaluation* criteria to *selection* criteria, as we wanted to focus on the activity of selecting a dataset in a dataset search scenario.We outline how the attributes included in each of these criteria evolved in the course of this work:

Relevance describes whether a dataset content is considered applicable to a particular task; e.g. it is on the correct topic. Temporal and geographical scope were also explicitly mentioned to assess relevance. In the original framework we referred to scope as 'coverage'. We further detailed the concept of context by pointing to the original purpose of the data as well as external requirements or norms influencing data creation.

Usability covers how suitable the dataset is considering practical implications of, e.g., format or license. In order to judge a datasets 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).

Quality refers to anything that participants use to judge a datasets condition or standard for a task, such as e.g. completeness or accuracy. Building on the original framework (Chapter 3) we extended the framework to include timeliness, more in-depth info about methodological set-up, as well as a datasets perceived accuracy and information about missing values.

Combining insights from the interview study, the analysis of the data search diaries, as well as of over 300 dataset summaries and comparing them to existing metadata guidelines we arrived at four information types and nine attributes that, if available in a search scenario, are likely to allow people to make an informed selection of a dataset.

The value of selection criteria is manifold. Understanding user needs is the first step towards aiding decision making in dataset search, and can inform interface design and information architecture for data discovery systems and focus research efforts on how to best communicate these information needs to users. This is of great importance, given the exploratory nature of data search tasks, as described in Chapter 3. While we were focusing on general data search tasks it is likely that there are domain specific selection criteria of more granular nature, which would be an interesting direction for future work.

6.1.1.4 Analysis of exploration activities for datasets

Both the study in Chapter 3 and in Chapter 3 gave insights into exploration activities for datasets. The interviews provide in-depth reports of participants exploration strategies and allowed us to identify common struggles and barriers. We found a number of key activities that participants engaged in to assess a datasets suitability for a given task, such as such as basic visual scans or investigating summarising statistics as well as metadata.

The study in Chapter 4 allowed us a different perspective on exploration strategies as we had 30 people describing data that is new to them in a lab experiment and further collected summaries through a crowdsourcing experiment. The analysis of these summaries, as described in 6.1.2 showed that participants arrived at common information types and attributes that they described after engaging with a dataset that is new to them for a short amount of time. We extended the framework to include groupings of headers. Observational studies would enable us to validate exploration strategies in practice and to investigate differences between users of different skill sets.

6.1.2 Composition of dataset summaries created by different types of participants (data literate lab participants versus crowd workers)

We presented an in-depth characterisation of the composition of dataset summaries created by different types of participants (data literate lab participants versus crowd workers). We consolidated our findings together with existing guidelines on metadata creation and proposed recommendations for the creation of textual summaries of datasets for human consumption.

We identify key themes and commonly mentioned dataset attributes, which people consider when searching and making sense of data. The most prevalent ones over all the experiments were subtitle (a high-level one-phrase summary describing the topic of the dataset), headers (explicit references to dataset headers) and geographical scope (geospatial scope of the data at different levels of granularity). Our findings further suggest that dataset summaries can be created using crowdsourcing. We presented these attributes as a template to guide the summary writing process.

This template could be used to build tool that semi-automatically supports data publishers in the summary writing process. We presented a detailed specification that could be used to guide the development of a dataset summarisation tool incorporating publisher input and automated features, described in more detail as well as with a preliminary prototype, in Chapter 4, Section 6.1.2.

The attributes mentioned in the summaries could also indicate those that are useful in search, which, if validated in future work, could increase the discoverability of data on the web. Web search functionalities are tailored to textual sources, therefore having a textual summary containing meaningful content on the dataset could potentially allow general web search engines to index data sources in a similar way as web pages.

6.1.3 Insights into SERP design and interactive IR experimentation for data search

The experiment described in Chapter 5 presented initial results on how to represent dataset summaries (based on results from Chapter 4) in a data search scenario. Some participants who were presented with just the title or a preview of the data explicitly expressed the need for dataset summarisation in their comments. This was in line with findings from Chapter 3 in which data professionals mentioned summaries as a solution that could allow them to easily assess a datasets fitness for use.

While there were few statistically significant differences regarding participants search performance, the table summary performed best in the interaction and experience based metrics that we measured. Both, when asked about whether the SERP provided enough information to judge whether a dataset was useful for the task, and in the confidence

scores which were recorded for each selected dataset, the table summary showed promising results. In contrast, merely displaying the title performed poorly, both in terms of perceived difficulty and usefulness, as well as for false positive selected results – an outcome that was supported by participants comments. On average both versions of the summary conditions took participants longer to select datasets compared to those without summaries. This potentially suggests a deeper engagement with the search results when being presented with a dataset summary on the SERP.

Overall we learned that summaries presented as a table together with a preview of the data would be an interesting direction to investigate further. This reflects the findings from Chapter 3 were participants mentioned the need to access the data, potentially to create a mental model of what it looks like, as well as expressed the need for a summary of the data that guides and aids interpretation of the data to understand it's suitability for a task.

We believe the study set-up in itself provides valuable insights for user studies in dataset search. We drew from interactive IR experiments and highlight the need for open-source test corpora, task collections and benchmarks for dataset search as described in Section 6.2. We discuss our choice of experimental design and metrics in Chapter 5. We believe that there is scope to think about sensitive metrics that allow to differentiate participants performance also in small scale studies which points to future work as described in 6.2.

The findings of this thesis revealed unique interaction characteristics in information seeking for structured data. They confirm previous work in Human Data Interaction, such as in Gregory et al. (2017); Kern and Mathiak (2015); Boukhelifa et al. (2017). This not only helps us build our understanding of how people interact with and communicate about data, but should inform metadata standards as well as automatic summary generation for datasets. Overall, our contributions can inform the design of data discovery tools, support the assessment of datasets and help make the exploration of structured data easier for a wider range of users, including less data literate audiences.

We now discuss where we see the most promising direction for future work.

6.2 Future work

This work could be extended in several directions. We grouped them into several themes: *data-centric tasks, summary creation, interfaces for dataset search, benchmarks and conversational data search.*

We discuss where we see the gaps in literature for the following topics: *data-centric tasks, summary creation, interfaces for dataset search, benchmarks and conversational data search* as these emerged as promising directions to extend this work and illustrate the breadth of potential research in dataset search.

6.2.1 Data-centric tasks

Additional research is needed to deepen our understanding of activities people carry out with structured data. An accepted taxonomy/ontology of data-centric work tasks in general, based on a large scale study across domains and skill sets, would be valuable to advance our knowledge of how people interact with structured data. Similarly, we lack a detailed classification of information seeking tasks (which can be seen as sub-tasks of work tasks (Byström and Hansen, 2005)) for structured data that interactive components of data discovery tools could be modelled on. We proposed two initial classifications of data-centric tasks in the study in Chapter 3, however further investigation with a larger number of participants per role type and domain would be needed to characterise them in depth and to examine how they differ between domains and between users with different skill sets.

Such classifications can be valuable in different ways, including incorporating this knowledge in the back-end of systems, for instance in ranking algorithms in dataset search (to improve ranking efficiency), as well as for task-specific or query biased snippet generation. At the same time, this knowledge could be integrated in the front-end through task-specific interface conditions. For instance, through tailored interfaces displaying maps for geospatial attributes, enabling comparisons between datasets, or advanced filtering options. Literature suggests that task-enabled search systems have the potential to increase the users' chances to access useful content when interacting with large collections (Freund, 2013). Similar to Li and Belkin (2008), faceted classification of information seeking tasks could be of value to examine data-centric information seeking tasks in a similar manner. Considering that data-centric tasks might increasingly be performed in collaboration with others, a more granular understanding of such tasks could inform the design of tools to support user interaction at different stages of the workflow with data, grounded in current work practices with data.

Furthermore, a more in-depth log analysis of data search logs would be interesting, in which connections between queries and resulting downloads can be made, as well as investigating query refinements within sessions. This could advance our understanding of task specific user behaviour in dataset search. The search log data that was kindly made available to us by the UK governmental open data portal described in Chapter 3 did not allow us to make a link between the search query and a resulting dataset download, nor information about query refinements, which is a limitation that comes with the data. Availability of such data could also allow us to get a sense of granular user activities.

6.2.2 Summary creation

Further studies are needed to develop methods to semi-automatically create dataset summaries suitable for human consumption, as well as to investigate the usefulness of summaries as snippets in data-search scenarios. Chapter 5 showed first steps in the direction of such an evaluation. Given the lack of guidance for dataset summary creation detailed in Chapter 4, we believe that having a template and recommendations for what to include in a summary is a necessary first step. The crowdsourcing experiments showed that summaries can be created by people with basic data literacy skills who are not familiar with a dataset. Follow-up studies could include crowdworkers iterating on the summaries created by the template, a process that has proven useful for image descriptions and text shortening (Bernstein et al., 2015; Little et al., 2010). Users of a dataset could potentially iteratively improve summaries, based on their in-depth engagement with the data. However, as the production of a dataset summary is a complex and potentially time-intensive task we believe there is also a large scope for developing semi-automatic summary creation tools that guide a data publisher through the summary writing process. We presented a preliminary prototype of what such a tool could look like in Chapter 4.

Future work could focus on how to improve both the interaction flow and the pre-processing capabilities of similar tools, and develop smart solutions of how to minimise human input while at the same time creating meaningful dataset summaries for data search.

An interesting direction could also be to refine semi-automatic approaches to generate summaries, using the template to prompt crowdworkers to extract these elements from datasets. This may also have the side-effect of producing higher quality descriptions overall, simply by providing more structure to the task and clearer examples and guidance to the crowd workers, as well as validation and training.

We further see a large scope for data-to-text generation, using approaches that learn from manually written summaries, which abstract the content of a dataset in a meaningful way, as detailed in Chapter 4. While we believe that this process should be automated as much as possible in order to scale, it is likely the best solution currently could involve a mix of machine learning and rule-based approaches, in combination with involvement from the publisher. As mentioned earlier, we currently lack both training data, as well as a clear understanding of people's interactions with dataset summaries in a search process.

Being able to generate dataset summaries at scale would allow search systems to use them and thus to investigate their usefulness in data search scenarios in the wild. This could include a range of (interactive) information retrieval experiments to evaluate if indexing such summaries results in better search performance. Web search is tailored

to textual sources, therefore having a textual summary containing meaningful content on the dataset could potentially allow general web search engines to index data sources in a similar way as web pages.

As mentioned in Section 6.2.1 there is still a large space to understand the generalisability of the summaries we proposed, and to explore task and domain specific summaries. Query-biased summaries, as proposed by (Au et al., 2016), are an interesting step in that direction. There is a large body of research aiming to understand visual representations of data for different contexts. Similarly, we believe that we need to further examine textual representations for data in much more detail to understand how to tailor them to specific users and their contexts, as discussed in Section 6.2.1. Another potential direction are dataset summaries tailored to different skill sets. However, we believe that in the context of universal design (Lidwell et al., 2010) the likelihood that basic, easily understandable summaries, which do not require in-depth knowledge of certain technologies, will be beneficial for everyone.

Indicators for dataset quality are an interesting area for further work and could be included in dataset summaries. The summaries produced by our participants included qualitative statements about completeness or comprehensiveness of the dataset and the proposed summary tool displays the percentage of missing values. However, our findings from Chapter 3 suggest that quality is inherently task-specific and would therefore need to be tailored to task contexts. If quality is seen as task-dependent the importance of capturing accurate, understandable and useful metadata attributes becomes even more evident. We could imagine search systems that allow users to select quality metrics that make sense to them in the context of their task, eventually influencing the ranking algorithm.

Not least, other aspects for future work include exploring the optimal length of dataset summaries displayed on a SERP, as has been done for snippets for web pages; and the potential benefits of interactive summaries, as explained in Section 6.2.3. At the same time, our experiment in Chapter 5 suggests that users might benefit from a summary as well as a preview of the datatset in a search scenario. We have yet to explore the best surface representation and placement for them.

6.2.3 Interfaces for dataset search

We still know little about what advanced state-of-the-art search interfaces tailored to structured data should look like. While the visual aspects of SERP design for data search were outside the scope of this thesis, our findings can provide valuable insights on where to focus future research efforts for data search result displays.

Interfaces tailored to user needs in data search should better support users in evaluating the relevance, quality and usability of a dataset result, getting a suitable overview of

the dataset; and finding related datasets; amongst other factors. Systems could benefit from understanding whether a user is searching for a dataset (the main focus of this work), for a data point within a dataset, or for an aggregation of data points. For each of these scenarios the supporting information that enables users to evaluate the search result could be slightly different.

Determining where in the search process the additional information that satisfies a user's selection criteria should be displayed needs to be established based on theoretical and empirical insight. For search-result displays we believe that presenting a summary of the dataset's content would support users in assessing whether a result is suited for their task. Similarly, our results from Chapter 5 suggest that providing a preview alongside a textual summary could be beneficial in dataset selection scenarios. Future work could be dedicated to generating dataset previews based on user studies investigating the optimal number of rows, choice of columns and interactive features. Similarly, future work could explore whether specific individual elements of the dataset are more useful to display alongside search results than others, such as for instance *headers, entities and/or summarising statistics* of a relevant column or field. We further envision visual or textual indicators of data quality on the interface, backed up by automatically computed metrics, user-generated reviews and annotations or reuse statistics.

As data search is often exploratory and datasets are often used in combination with other datasets or information sources we believe that future data search interfaces will go beyond the traditional 10-blue-link paradigm. As a dataset is often accompanied by a dataset preview page there are many options to access and display metadata during a search. For instance, Google's dataset search currently displays a split interface showing a large list of search results for scrolling on the left side and a reduced version of a dataset preview page with links to one (or multiple) dataset repositories that hold the selected dataset, on the right side of the screen (Figure 6.3).

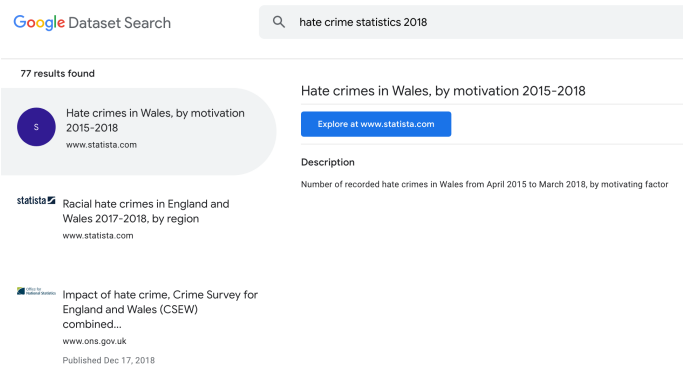


FIGURE 6.2: Google dataset search: SERP

Elsevier's DataSearch[1] for instance, provides several filtering options on the left side and

[1] https://datasearch.elsevier.com/

allows the expansion of detailed information about a dataset within the SERP at the same time. One the one hand this caters to user needs such as specifying a time frame (date) and displaying the dataset within it's context (in this case by showing connected files). On the other hand this results in a cluttered interface, potentially connected to a high cognitive load for the user.

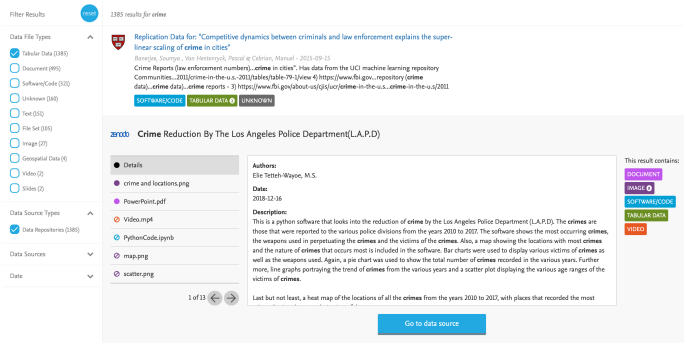


FIGURE 6.3: Elsevier DataSearch: SERP + expanded preview for one dataset

Thinking about ways to allow exploration of data through the SERP there are also opportunities to enrich dataset preview pages with interactive visualisations. This could allow users to choose their area of interest within a larger dataset, as well as providing a comprehensive overview of the content. Allowing users to filter, sort and explore different views of the data on demand is recommended for complex search tasks (White and Roth, 2009). Similarly, the action of zooming in and out of levels of data to support exploration, either already on a SERP or on a dataset preview page, could be explored further.

Our findings, as well as literature in this space, suggest that approaches that allow an overview over the potential result space are a promising direction for further work. Liu and Jagadish (2009a) discuss how in the context of database queries user struggle to issue precise queries without understanding the whole possible result set. Discussing the visual aspects of sensemaking, Russell (2003) also mention the need to understand what is in a whole collection. Short, underspecfied queries are common in dataset search (Kacprzak et al., 2019) and our participants in Chapter 3 reported to submit high-level queries as they do not expect the search functionality to work well. Another reason they reported for underspecified queries was the feeling to lack an overview of what data is 'out there'. Hence, allowing users to to interact with the entire result set and adjust ranking based on their personal exploration of the result set (Klouche et al., 2017) would be a promising direction to explore for dataset search. This could include re-ranking based on different facets that have been shown to be important for dataset search, such as temporal or geospatial scope of the data, provenance, or others. Users could first get an overview of the available data and then tailor the results to their specific needs, based on a learning experience during the search activity. In general web search users

refine their query, however currently in dataset search over-specified or longer queries often do not produce results. Hence, this would be a strategy to overcome this problem and still personalise the result set where possible.

We further see a large space to explore interfaces that allow more complex interaction with datasets such as advanced querying (Jagadish et al., 2007). This involves mechanisms beyond keyword queries, such as taking a dataset as input and searching for similar ones or being able to follow links between entities in datasets. In tabular search, the query consists of an existing table and the aim is either to add data to the table or search for similar tables (based on their schema) within a result set (e.g. Zhang (2018); Cafarella et al. (2008). At the same time, being able to query datasets on a column, row and cell level (e.g. Pimplikar and Sarawagi (2012) to match specific selection criteria at the time of the search activity on the search result would be an interesting direction to explore further.

We believe useful data discovery tools should support users in making relationships between pieces of information visible. In complex search tasks, users commonly need to link multiple datasets together. The discovery and exploration of links could be supported by 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 based on content or structure of the dataset. Techniques used in semantic web technologies to provide graph based visualisations that display connections between search results and other datasets could be a promising direction to explore further. This includes, for instance, Cluster Maps (e.g. Fluit et al. (2006)) or trying to understand relationships between datasets based on entity similarity between the datasets (e.g. Zhang and Balog (2018)).

The development of new approaches and interaction paradigms needs to be tested and iteratively improved. This requires benchmarks to evaluate the success both from a systems' and a user perspective. For instance, even if we could create well performing dataset recommendation approaches, we still lack an in-depth understanding of the signals users are interested in dataset recommender systems. While there is some work on dataset recommendation, there is to this date little awareness of user needs. To give an example, Ellefi et al. (2016) propose an approach to recommend datasets based on schema similarity, however we would argue this assumption needs to be tested with real users and in dataset search scenarios. The next section describes the need for benchmarks and testing within the data search space in greater detail.

6.2.4 Benchmarks

Research in dataset search would hugely benefit from benchmark and test collections for (interactive) IR experiments. For traditional text retrieval there is a large space of established test collections (Voorhees et al., 2005) that facilitate experimentation. Given that structured datasets have unique interaction characteristics we believe that there is a need to establish similar collections for dataset search. This includes corpora of information seeking tasks, connected queries and results annotated with relevance rankings. At the same time, there is a need to improve our understanding of how dataset search success should be measured, both from a systems and a user perspective. Not all traditional metrics used in evaluation of web search results might be applicable to dataset search and additional metrics might be interesting for datasets, such as e.g. completeness, openness, license, cost, or indications of reuse.

To the best of our knowledge there is currently no large-scale open-source log data for dataset search that we can learn from, and anecdotal evidence suggests that many small organisations that capture data-search logs do not necessarily record in-depth interaction data, from which inferences on user behaviour could be made. A collection of standard dataset search tasks, together with available corpora, could accelerate research in this area by facilitating easier and more comparable experimentation.

6.2.5 Conversational data search and future data interaction paradigms

With advances in speech recognition and text to speech synthesis, it has become possible to search for information in natural language queries and for an information retrieval system to verbally respond to queries with systems such as Google Home or Alexa. This means that eventually we will be able to search for datasets using voice, which might imply different information needs for navigating data in this way. We currently do not know what a useful dataset snippet looks like in a spoken conversational search setting. This is interesting as even if we think about Question Answering where a system responds with a datapoint rather than a dataset, a user might want to understand where the answer comes from in order to evaluate it and could then be interested in a spoken dataset summary.

It has been shown that simply transforming textual content into audio does not provide the same user experience that we know from textual/visual systems (Sahib et al., 2012; Trippas et al., 2017). One assumption is that to use dataset summaries in an audio setting these need to be shorter as it has been shown that the amount of information that is possible to convey to users via audio is limited by the nature of the medium (Yankelovich and Lai, 1999).

Therefore we believe that studies exploring how to shorten the summaries proposed in Chapter 4 can be valuable. Short dataset summaries can be useful as dataset snippets

in conversational interfaces, but also potentially for smaller text interfaces, such as on smartphones and wearable devices.

This could be done by exploring a potential ordering preference of the individual summary attributes. Moreover, investigating whether some attributes are considered more useful than others in a dataset selection scenario could be an interesting direction for future work. This could be investigated for both text and voice representations of dataset summaries. Understanding the content categories that users find most useful in judging a dataset's fitness for their task can further provide us with valuable insight in how future question answering systems based on structured data should present the underlying dataset in a meaningful way to users.

This might involve not just presenting the data but pre-empting analysis based on our task context. We imagine task sensitive data search systems that warn us if the data is not suited for our task, or if there is risk (such as bias) or uncertainty attached to using a dataset. Data search systems could make users aware of the representativeness of the data they are looking at, or scrutinise datasets in other ways. We imagine the tools used to search and to subsequently use data to be less fragmented and to support people's entire workflow with data; from formulating an information need, to searching and exploring a datasets to using it. Such systems might automatically update the datasets we create and makes us aware of changes in connected datasets that are relevant for us. We will get better at automatically creating synergies between heterogeneous data, which might make it easier to attach context to datasets in a way that allows us to make sense of them.

Our findings confirmed the general notion of data to be dependent on context and situated in the culture of its creation process. Future systems supporting data-centric work will hopefully have a less deterministic approach to numbers embedded deep in their design; and potentially treat data more like images, as they capture a version of reality, but can also be manipulated. Research advancing dataset search in the direction of more customisable and responsible interactions with data can contribute to increase data literacy by supporting users in searching, finding and understanding data.

Chapter 7

Conclusions

Here we summarise the contributions of this work and conclude with were we see dataset search as a research area.

7.1 Summary of contributions

The primary goal of this work was to understand how we can support people to select data that is useful for their task. We approached this by advancing our understanding of user behaviour in structured data discovery. The contributions of this work include:

- Categorisations of data-centric work tasks
- A better understanding of search strategies in data search
- In-depth knowledge on selection criteria in data search
- An analysis of exploration activities for datasets
- Insights into the composition of dataset summaries created by different types of participants
- Insights into SERP design and interactive IR experimentation for data search

We translated the contributions (detailed in Section 6.1) to actionable insights into how people could be supported in selecting datasets from a pool of search results and how to further advance Human Data Interaction research for dataset search:

- A framework for Human Structured Data Interaction that can be used as guidance for data publishers, data portals and tool designers of data discovery tools

- A proposed template to create user centred dataset summaries
- Initial suggestions how to present these summaries in a data search scenario
- Specifications for a data summarisation tool, together with an initial prototype
- A detailed discussion of potential directions for future research in data search

Data search has received increasing attention in the last years: Google launched dataset search in beta in 2018. At the same time we see an increasing interest in the scientific community to discover and reuse research datasets, as well as recent efforts to establish data search as a research area in its own right, for instance through international data search workshops. Given the large amount of available data, and the increasing need for intelligent data discovery (be that on the web or in an enterprise setting), we believe that we have only just started to understand the unique characteristics of structured data as an information source, and the interaction challenges system designers are facing.

We believe there is scope to extend research on user centred dataset discovery in numerous directions, including, but not limited to: *data-centric work tasks, summary creation for dataset, interfaces for dataset search, benchmark creation and conversational dataset search* (as discussed in Chapter 6).

It will be interesting to see how the broader area of data search (including its technical and infrastructure related challenges) will be approached by researchers and practitioners, and to what extent techniques from other fields (including e.g. databases, information retrieval, and semantic web search) can be applied towards the problem of dataset search.

There is large scope to investigate data search behaviour amongst different user groups, according to common skill sets or domain expertise. Only with a thorough understanding of user needs in data search can we start to translate these insights into best practices and functionalities that enable data providers to facilitate data search and selection.

Moving from data search to data use there is an even wider space to explore the current and constantly evolving work practices with structured data and the tools available to do so. Many of the data-centric tasks described by our participants in Chapter 3 are in fact performed in collaboration. Hence, investigating how existing methods, which facilitate collaboration for documents or code, could be applied efficiently to collaboration with structured data, taking the interaction characteristics discussed in this work into account, would be another interesting direction to explore further.

We hope that this work could contribute to increase our understanding of data discovery from a user perspective, and that many others will continue to work to make data easier to find, understand and use, for as many people as possible.

Appendix A

Appendix Chapter 3

A.1 Risk assessment form

Researchers name: Laura M Koesten

Part 1 Dissertation/project activities

What do you intend to do? (Please provide a brief description of your project and details of your proposed methods.)

I intend to do an interview study with people who work with data in their jobs. The aim is to understand how people understand data and the processes of how data is used in a work context. It will include a brief online survey, sent to people or organisations identified as relevant, with the aim to find suitable participants. In the survey the people will be asked if theyd be interested in participating in an interview study concerning their engagement with data in their work. The interviews will be one off, will be around 45 minutes long, audio recorded and analysed using semantic analysis. Results will be presented at conferences and/or published in conference proceedings or scientific journals without revealing peoples identities, all data will be stored password secured.

Will this involve collection of information from other people? (In the case of projects involving fieldwork, please provide a description of your proposed sample/case study site.)

As describes above, people will be interviewed and these interview will be audio recorded. Interviews will take place either via telephone or in a place where other people are present (Office of the Open Data Institute in London or the participants work place, to their convenience). To recruit participants a short online survey will be sent out in which participants can enter their email addresses in case they are interested to be contacted for an interview. Participants will be selected to represent a wide range of different user

groups. Informed consent will be received beforehand - for both the survey as well as for the interviews. The optimal number of participants would be between 10 and 20 people.

If relevant, what location/s is/are involved?

Possibly the office of the Open Data Institute (65 Clifton Street, London, EC2A 4JE) or participants workplaces if suitable. Participants will be offered to choose from skype or face-toface interviews to their convenience.

Will you be working alone or with others?

Two supervisors are involved (Dr. Elena Simperl and Dr. Jeni Tennsion).

Part 2 Potential safety issues / risk assessment.

Potential safety issues arising from proposed activity?

No potential safety issues expected.

Person/s likely to be affected?

n/a

Likelihood of risk?

No risk anticipated.

Part 3 Precautions / risk reduction

Existing precautions:

Participants will be assured that they can withdraw at any point during the interviews and can withdraw their participation in the study up to one week after the interview was conducted. Participants will not be harassed or coerced into participating. If interviews should be conducted face to face it will be in a place where either other people are present or a room with glass walls with other people present outside and being able to look into the room. Confidentiality will be ensured by securing collected data with a password and ensuring anonymity in storage and presentation of data. Any personal information collected will be stored separately from the interview data and will be deleted after the research activity.

Proposed risk reduction strategies if existing precautions are not adequate:

Participants will be reminded that they can withdraw at any time during the interview without have to give any reason.

Part 4 International Travel

If you intend to travel overseas to carry out fieldwork then you must carry out a risk assessment for each trip you make and attach a copy of the International Travel form to

this document Download the Risk Assessment for International Travel Form Guidelines on risk assessment for international travel at can be located at: www.southampton.ac.uk/socscinet/safety (risk assessment section). Before undertaking international travel and overseas visits all students must: Ensure a risk assessment has been undertaken for all journeys including to conferences and visits to other Universities and organisations. This is University policy and is not optional. Consult the University Finance/Insurance website for information on travel and insurance. Ensure that you take a copy of the University travel insurance information with you and know what to do if you should need medical assistance. Obtain from Occupational Health Service advice on any medical requirements for travel to areas to be visited. Ensure next of kin are aware of itinerary, contact person and telephone number at the University. Where possible arrange to be met by your host on arrival. If you are unsure if you are covered by the University insurance scheme for the trip you are undertaking and for the country/-countries you intend visiting, then you should contact the University's Insurance Office at insure@soton.ac.uk and check the Foreign and Commonwealth Office website.

Risk Assessment Form for International Travel attached? NO

A.2 Participant information sheet

Project title: Interview study with data workers

You have been invited to take part in a research study that aims at understanding how people engage with data in their job. We want to understand how people gather data, how they understand and assess data and which tools they use for theit respective tasks. We are interviewing people from a variety of backgrounds, domains and with different levels of expertise.

About us
We are researchers working at the Open Data Institute in London, UK and at the Web and Internet Science group, in the Faculty of Physical Sciences and Engineering of the University of Southampton. This project has been approved by the Facultys Ethics Committee.

How this study is conducted
We recruited interviewees via email and social media and we have done a scoping survey to find the best spread of participants for our study. You have been asked to participate in one interview, which will be conducted face-to-face, via video call or phone, according to the participants preference. Each interview will last approximately 45 minutes and will be recorded for further analysis.

Data usage policy
Responses will be treated with the strictest confidentiality. Your contact information

will not be shared with anyone else without your permission. The results of the data gathered will be analysed and can be presented at scientific conferences, and/or published in conference proceedings or journals. You can decide to leave the study at any time during data collection and without explanation. You will have the right to ask that any data you have supplied to that point be withdrawn or destroyed. You can find further information on data usage in the informed consent sheet you have been given.

Questions and the right to withdraw
If any question appears unclear, far-fetched or difficult, you are welcome to interrupt and ask for clarification, omit or refuse to answer without giving a reason. Moreover, you have the right to have questions about the procedure answered (unless answering these questions would interfere with the studys outcome). If you have any questions regarding this information sheet, please ask the researcher before the beginning of the study. No risks or benefits are known for participants in this study. Your participation is voluntary.

Further information
Laura Koesten PhD researcher at The Open Data Institute and the University of Southampton, UK Email: laura.koesten@theodi.org You are welcome to contact us with any questions about the study.

A.3 Scoping survey

Online Survey

DO YOU WORK WITH DATA?

We want to improve how people engage with data. We are putting together a study on how different people use data in their work and are looking for interviewees for that study. If you are interested in helping us, please complete this short survey which will help us to identify suitable participants.

It should not take longer than 3 minutes to complete. There are no right or wrong answers, so please respond freely and honestly.

Data Usage Policy

The data of this survey will be used to decide whether you would be a relevant interview participant for our study. You will be asked for your contact information for the case a follow-on interview would be useful. Your contact information will not be shared with anyone else without your permission. You are also giving consent for your information to be reported in aggregated statistics, which can be published as open data. These may be reported in academic work and in conferences, but no individual responses that you may choose to provide will be published.

For further information about the survey or the study please contact:

Laura Koesten PhD researcher at The Open Data Institute and the University of Southampton, UK Email: laura.koesten@theodi.org

Survey questions

1. I confirm that I have read the Data Usage Policy above.
Yes / No

2. Do you collect, store, manage, analyse, interpret or visualise data in your work?
Yes / No / I dont know

3. Do you use data to help make decisions in your work?
Yes / No / I dont know

4. How often do you use data that is new to you?
Daily / weekly / monthly / less frequently than monthly / never

5. Do you ever search for data for your work?
Daily / weekly / monthly / less frequently than monthly / never

6. What were you looking for when you last searched for data? Textfield

7. What tool or website did you use to search for data? Textfield

8. When you use data which of these activities do you typically do?
Look at visualisations / Create visualisations / Perform data analysis / Create applications or tools / Other, such as: Textfield

9. Which applications or tools do you use when working with data? Textfield

We are asking the following questions to make sure we have a variety of participants in the study:

10. What is your job title? Textfield

11. Which sector do you work in? Dropdown list of sectors, including the option other

12. Which country do you work in? Textfield

13. What is your gender? Male / Female / Other- textfield

14. Are you 18 years or older? Yes/No

15. Please provide your email address so we can contact you for the follow up interview study: Textfield

Thank you very much for taking the time to complete this scoping survey!

A.4 Interview schedule

Interview Schedule
Thank you for agreeing to be interviewed
Thank you for the information you provided in the scoping survey

Just going through a few points before we start:
Approximately 45 minutes
Informed consent
You don t need to answer any questions you don t like to
No right or wrong answers
You can stop at any point without giving a reason
Will be recorded but data kept confidentially, secured and deleted after analysis
If quotes will be used, they will be redacted to not reveal your identity

Outlook and purpose
About your experiences of using data in your work
Interviews will be analysed and the results can help to understand the processes of how data is
Help improve how people engage with data

Content
We are interested in the whole process of how you are using data in work to help make decisions
That includes: how you search for data, how you evaluate data to be relevant for you, what you do with the data.
We are interested in your experiences and thoughts when using data

Interview questions

- Whats your job title, please describe your role?
- How do you work with data: What types of data, how does the data look like (e.g. spreadsheets, graphs, CSV,..)
- What do you typically do with the data? (high level task) A typical question in your work that you are trying to solve by using data?
- How often do you make decisions within your work that are fully or partially based on data? Think back to a particular decision you made recently. Talk me through how you made that decision? What sources did you consult?

- With how many different datasets (sources) are you interacting in a day / in a week?
- Do you search for new datasets?
- If so, where and how do you search for datasets? What are your experiences when searching for datasets?
- Can you think of a search query you made?
- Exploring a dataset: When you see data on something which is new for you what do you look at first? When you open a dataset youve never seen before, what do you look at first?
- How do you evaluate data / a new data set to whether it is useful for you?
- What are your requirements on a data set - which kind of information are you looking for initially?
- Which tools do you use to work with the data? What do you like about the tool you are using? What do you dislike?
- Where would you say have you learned the skills to search and use the data? In your formal education, on the job, colleagues, self-taught?
- What are your personal challenges in working with data? Things that you find tricky, things you complain about
- How would your ideal process of using data to make a decision (personal example given in the beginning of the interview) look like? Describe the steps and what would need to work for that ideal process What would make it easier?

Debrief

Is there anything else you would like to say before we end the interview?
Thank you for taking the time. We very much appreciate it! If you like a follow-up information to get to know more about the results let us know. You have our email address.

Appendix B

Appendix Chapter 4

B.1 Risk assessment form - lab experiment

ERGO/FPSE/28636, Risk Assessment Form

Researchers name: Laura M Koesten, Emilia Kacprzak

Part 1 Dissertation/project activities *What do you intend to do? (Please provide a brief description of your project and details of your proposed methods.)* We intend to do a task based labs study, with participants who have basic data literacy skills. The task consists of looking at a dataset in a spreadsheet and describing is in a written form. After that participants will be asked to summarise their own description in max 100 words. For some participants that will involve 1 dataset and approximately 20 minutes. For others this involves 3 datasets and a maximum of 60 minutes. Participants will be asked to describe the dataset as they were describing it to another person who cannot see the data but has to decide whether or not to use it for their data project. Potential participants will have responded to a call via social media or email to participate in the study. After expressing interest to participate they will be asked a few questions assessing their data literacy and collecting basic demographic information (such as age, gender, education level). If considered suitable they will be contacted to agree on the time of the study. If considered unsuitable they will receive a polite response that we either have enough participants or that inclusion criteria regarding demographics and data literacy didnt match with theirs. Participants will be given a written information sheet and informed consent will be given before the study. The tasks will be conducted face-to-face. After the task has been conducted, the investigator will thank the participant for their time, and ask whether they need any further information. Results of the study will be presented at conferences and/or published in conference proceedings or scientific journals, without revealing peoples identities, all data will be stored password secured. Will this involve

collection of information from other people? (In the case of projects involving fieldwork, please provide a description of your proposed sample/case study site.) As described above, people will be taking part in a one-off task based study in which they produce a written output. The study will place where other people are present (Office of the Open Data Institute in London), in a room with see through walls. Informed consent will be received beforehand. The optimal number of participants would be between 25 and 30 people.

If relevant, what location/s is/are involved? The office of the Open Data Institute (65 Clifton Street, London, EC2A 4JE). *Will you be working alone or with others?* Two supervisors are involved (Prof. Elena Simperl and Dr. Jeni Tennsion).

Part 2 Potential safety issues / risk assessment. *Potential safety issues arising from proposed activity?* No potential safety issues expected. *Person/s likely to be affected?* n/a Likelihood of risk? No risk anticipated.

Part 3 Precautions / risk reduction *Existing precautions:* Participants will be assured that they can withdraw at any point during the task. Participants will not be harassed or coerced into participating. The task based study will be in a place where either other people are present or a room with glass walls with other people present outside and being able to look into the room. Confidentiality will be ensured by securing collected data with a password and ensuring anonymity in storage and presentation of data. Any personal information will be deleted after the research activity. *Proposed risk reduction strategies if existing precautions are not adequate:* Participants will be reminded that they can withdraw at any time during the task without having to give any reason.

Part 4 International Travel No

B.2 Risk assessment form - crowdsourcing experiment

Researchers name: Laura Koesten
Part 1 Dissertation/project activities

What do you intend to do? (Please provide a brief description of your project and details of your proposed methods.)

I intend to do a crowdsourcing experiment research to understand how people describe and summarise data. We will do this in 3 rounds using the crowdsourcing platform CrowdFlower. In the first round people will be asked to write 5 short (50-100 words) description of 5 datasets (spreadsheets). In the second round, we will validate these descriptions by showing them 5 descriptions of the same dataset and ask them to rank

them according to their quality and discard wrong descriptions. In the 3rd round we will ask participants to identify is there is anything missing in the description of a dataset and to edit the description to improve its quality. Results of the analysis of these descriptions will be presented at conferences and/or published in conference proceedings or scientific journals. Participants will be asked to self-report basic demographic information, namely their age, gender, nationality and their highest level of education obtained. Responses will be recorded through the CrowdFlower platform. No sensitive information will be collected. Users will have an anonymous identifier and will not be contacted directly by the researchers. All data will be stored password secured. Participants will be paid and will voluntarily decide to take part in the experiment.

Will this involve collection of information from other people? (In the case of projects involving fieldwork, please provide a description of your proposed sample/case study site.)

As describes above, responses will be recorded through the CrowdFlower platform. This platform provides an environment to set up the tasks and works as a mediator between its customer, the researchers in this case, and the participants. These must register to the platform, or to one of the user channels connected to that. Users will have an anonymous identifier and will not be contacted directly by the researchers and no sensitive information will be collected.

If relevant, what location/s is/are involved?

CrowdFlowers channels include people from all over the world and in the case of our experiment we will limit participation to people from countries in which the majority of the population is native-English speaking [1]. Participants will carry out the tasks from their computers.

Will you be working alone or with others?

Two other researchers (Emilia Kacprzak and Tom Blount) and two supervisors are involved (Dr Elena Simperl and Dr Jeni Tennsion).

Part 2 Potential safety issues / risk assessment.

Potential safety issues arising from proposed activity? No potential safety issues expected.

Person/s likely to be affected? n/a
Likelihood of risk?
No risk anticipated.

Part 3 Precautions / risk reduction

[1] ://www.southampton.ac.uk/studentadmin/admissions/admissions-policies/language.page

Existing precautions:

All data will be stored password secured. Participants have an anonymous identifier provided by the CrowdFlower platform.

All participants are managed by Crowdflower, the study never gets to identify or interact with the participants. Payments are handled by the CrowdFlower platform.

The process of the task will be explained thoroughly before participants decide to do the task to avoid them spending time on a task they do not want to finish. Proposed risk reduction strategies if existing precautions are not adequate:

Participants will be reminded that they can withdraw at any time during the interview without have to give any reason.

Part 4 International Travel
No

B.3 Participant information sheet - lab experiment

Project title: Describing data

You have been invited to take part in a research study that aims at understanding how people understand and describe data. For this study we are asking people from a variety of domains and with different levels of expertise to describe datasets, that they are not familiar with, in their own words.

About us
We are researchers working at the Open Data Institute in London, UK and at the Web and Internet Science group, in the Faculty of Physical Sciences and Engineering of the University of Southampton. This project has been approved by the Facultys Ethics Committee. How this study is conducted We recruited our participants via email and social media. You have been asked to participate in an exercise, which will be conducted face-to-face or via video according to your preference. We will give you a task for which you need to use data and we will ask you to look at a dataset and write a description of this dataset for a maximum of 12 minutes per dataset. The purpose of this description is that another person, who is not able to see the dataset, can decide whether or not to use that dataset for their data project. We will then ask you to summarise your own description in max 100 words. We will repeat this for 5 datasets, which means the session will last about 60 minutes. We will ask you a few questions before you start describing the data for the purpose of collecting demographics and to make sure you havent worked with the datasets before. After the task you will have an opportunity to

give us verbal feedback on your experiences if you want to.

Data usage policy

Responses will be treated with the strictest confidentiality. Your contact information will not be shared with anyone else without your permission. The results of the data gathered will be analysed and can be presented at scientific conferences, and/or published in conference proceedings or journals. You can decide to leave the study at any time and without explanation. You can find further information on data usage in the informed consent sheet you have been given.

Questions and the right to withdraw

If the task appears unclear, far-fetched or difficult, you are welcome to interrupt and ask for clarification, omit or refuse to do the task without giving a reason. Moreover, you have the right to have questions about the procedure answered (unless answering these questions would interfere with the studys outcome). If you have any questions regarding this information sheet, please ask the researcher before the beginning of the study. No risks or benefits are known for participants in this study. Your participation is voluntary.

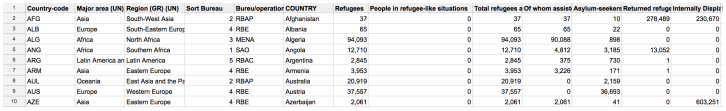
Further information

Laura Koesten + Emilia Kacprzak
PhD researcher at The Open Data Institute and the University of Southampton, UK
Email: laura.koesten@theodi.org / emilia.kacprzak@theodi.org

You are welcome to contact us with any questions about the study.

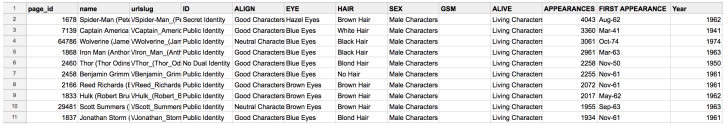
B.4 Datasets in *Set–5*

The full datasets can be seen on a Github repository created for the study [2] as described in Chapter 4.



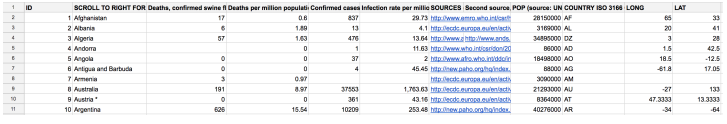
Country-code	Major area (UN)	Region (GR) (UN)	Sort Bureau	Bureu/operatior	COUNTRY	Refugees	People in refugee-like situations	Total refugees a	Of whom assist	Asylum-seekers	Returned refuge	Internally Displ
AFG	Asia	South-West Asia	2	RBAP	Afghanistan	37	0	37	37	10	278,489	230,670
ALB	Europe	South-Eastern Europ	4	RBE	Albania	65	0	65	65	22	0	0
ALG	Africa	North Africa	3	MENA	Algeria	94,093	0	94,093	90,088	898	0	0
ANG	Africa	Southern Africa	1	SAO	Angola	12,710	0	12,710	4,812	3,185	13,052	0
ARG	Latin America an	Latin America	5	RBAC	Argentina	2,845	0	2,845	375	730	1	0
ARM	Asia	Eastern Europe	4	RBE	Armenia	3,953	0	3,953	3,226	171	1	0
AUL	Oceania	East Asia and the Pa	2	RBAP	Australia	20,919	0	20,919	0	2,159	0	0
AUS	Europe	Western Europe	4	RBE	Austria	37,557	0	37,557	0	36,693	0	0
AZE	Asia	Eastern Europe	4	RBE	Azerbaijan	2,061	0	2,061	2,061	41	0	603,251

FIGURE B.1: Refugee movements



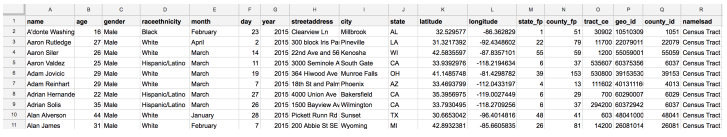
page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES	FIRST APPEARANCE	Year
1678	Spider-Man (Pete	\/Spider-Man_(Pe	Secret Identity	Good Characters	Hazel Eyes	Brown Hair	Male Characters		Living Characters	4043	Aug-62	1962
7139	Captain America	\/Captain_Americ	Public Identity	Good Characters	Blue Eyes	White Hair	Male Characters		Living Characters	3360	Mar-41	1941
64786	Wolverine (Jame	\/Wolverine_(Jam	Public Identity	Neutral Characte	Blue Eyes	Black Hair	Male Characters		Living Characters	3061	Oct-74	1974
1868	Iron Man (Anthor	\/Iron_Man_(Anth	Public Identity	Good Characters	Blue Eyes	Black Hair	Male Characters		Living Characters	2961	Mar-63	1963
2460	Thor (Thor Odins	\/Thor_(Thor_Od	No Dual Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters		Living Characters	2258	Nov-50	1950
2458	Benjamin Grimm	\/Benjamin_Grim	Public Identity	Good Characters	Blue Eyes	No Hair	Male Characters		Living Characters	2255	Nov-61	1961
2166	Reed Richards (I	\/Reed_Richards	Public Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters		Living Characters	2072	Nov-61	1961
1833	Hulk (Robert Bru	\/Hulk_(Robert_B	Public Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters		Living Characters	2017	May-62	1962
29481	Scott Summers (	\/Scott_Summers	Public Identity	Neutral Characte	Brown Eyes	Brown Hair	Male Characters		Living Characters	1955	Sep-63	1963
1837	Jonathan Storm (	\/Jonathan_Storm	Public Identity	Good Characters	Blue Eyes	Blond Hair	Male Characters		Living Characters	1934	Nov-61	1961

FIGURE B.2: Marvel comic characters



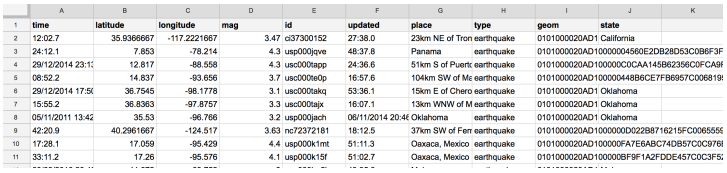
ID	SCROLL TO RIGHT FOR	Deaths, confirmed swine fl	Deaths per million populati	Confirmed cases	Infection rate per millio	SOURCES	Second source	POP (source: UN	COUNTRY ISO 3166	LONG	LAT
1	Afghanistan	17	0.6	837	29.73	http://www.emro.who.int/asr/		28150000	AF	65	33
2	Albania	6	1.89	13	4.1	http://ecdc.europa.eu/en/acti		3169000	AL	20	41
3	Algeria	57	1.63	476	13.64	http://www.p http://www.ands.		34895000	DZ	3	28
4	Andorra		0	1	11.63	http://www.who.int/csr/don/20		86000	AD	1.5	42.5
5	Angola	0	0	37	2	http://www.afro.who.int/dsc/dr		18498000	AO	18.5	-12.5
6	Antigua and Barbuda	0	0	4	45.45	http://new.paho.org/hq/index.		88000	AG	-61.8	17.05
7	Armenia	3	0.97			http://ecdc.europa.eu/en/acti		3090000	AM		
8	Australia	191	8.97	37553	1,763.63	http://ecdc.europa.eu/en/acti		21293000	AU	-27	133
9	Austria *	0	0	361	43.16	http://ecdc.europa.eu/en/acti		8364000	AT	47.3333	13.3333
10	Argentina	626	15.54	10209	253.48	http://new.paho.org/hq/index.		40276000	AR	-34	-64

FIGURE B.3: Swineflu deaths



name	age	gender	raceethnicity	month	day	year	streetaddress	city	state	latitude	longitude	state_fp	county_fp	tract_ce	geo_id	county_id	namelsad
A'donte Washing	16	Male	Black	February	23	2015	Clearview Ln	Millbrook	AL	32.529577	-86.362829	1	51	30902	10510309	1051	Census Tract
Aaron Rutledge	27	Male	White	April	2	2015	300 block Iris Pa	Pineville	LA	31.3217392	-92.4348602	22	79	11700	22079011	22079	Census Tract
Aaron Siler	26	Male	White	March	14	2015	22nd Ave and 56	Kenosha	WI	42.5835597	-87.8357101	55	59	1200	55059001	55059	Census Tract
Aaron Valdez	25	Male	Hispanic/Latino	March	11	2015	3000 Seminole A	South Gate	CA	33.9392976	-118.2194634	6	37	535607	60375356	6037	Census Tract
Adam Jovicic	29	Male	White	March	19	2015	364 Hwood Ave	Munroe Falls	OH	41.1485748	-81.4298782	39	153	530800	39153530	39153	Census Tract
Adam Reinhart	29	Male	White	March	7	2015	18th St and Palm	Phoenix	AZ	33.4693799	-112.0433197	4	13	111602	40131119	4013	Census Tract
Adrian Hernande	22	Male	Hispanic/Latino	March	27	2015	4000 Union Ave	Bakersfield	CA	35.3956975	-119.0027449	6	29	700	60290007	6029	Census Tract
Adrian Solis	35	Male	Hispanic/Latino	March	26	2015	1500 Bayview Av	Wilmington	CA	33.7930495	-118.2709256	6	37	294200	60372942	6037	Census Tract
Alan Alverson	44	Male	White	January	28	2015	Pickett Runn Rd	Sunset	TX	30.8653042	-96.4014816	48	41	603	48041060	48041	Census Tract
Alan James	31	Male	White	February	7	2015	200 Abbie St SE	Wyoming	MI	42.8932381	-85.6605835	26	81	14200	26081014	26081	Census Tract

FIGURE B.4: Police killings



time	latitude	longitude	mag	id	updated	place	type	geom	state
12:02.7	35.9366667	-117.2221667	3.47	ci37300152	27:38.0	23km NE of Tron	earthquake	0101000020AD1	California
24:12.1	7.853	-78.214	4.3	usp000jqvw	48:37.8	Panama	earthquake	0101000020AD10000004560E2DB28D53C0B6F3F	
29/12/2014 23:1	12.817	-88.558	4.3	usc000tapp	24:36.6	51km S of Puerto	earthquake	0101000020AD1000000C0CAA145B62356C0FCA9F	
08:52.2	14.837	-93.656	3.7	usc000te0p	16:57.6	104km SW of Mc	earthquake	0101000020AD10000044BB6CE7FB6957C006819	
29/12/2014 17:5	36.7545	-98.1778	3.1	usc000takq	53:36.1	15km E of Chero	earthquake	0101000020AD1	Oklahoma
15:55.2	36.8363	-97.8757	3.3	usc000tajx	16:07.1	13km WNW of M	earthquake	0101000020AD1	Oklahoma
05/11/2011 13:42	35.53	-96.766	3.2	usp000jach	08/11/2014 20:4	Oklahoma	earthquake	0101000020AD1	Oklahoma
42:20.9	40.2961667	-124.517	3.63	nc72372181	18:12.5	37km SW of Fern	earthquake	0101000020AD1000000D022B8716215FC0655559	
17:28.1	17.059	-95.429	4.4	usp000k1mt	51:11.3	Oaxaca, Mexico	earthquake	0101000020AD100000FA7E6ABC74DB57C0C976	
33:11.2	17.26	-95.576	4.1	usp000k15f	51:02.7	Oaxaca, Mexico	earthquake	0101000020AD100000BF9F1A2FDDE457C0C3F52	

FIGURE B.5: Earthquakes

[2] https://github.com/describemydataset/DatasetSummaryData2018

B.5 Crowdsourcing Task

Figure A.6 and A.5 shows the crowdsourcing task described in Chapter 4 in which we asked crowdworkers to write summaries of datasets.

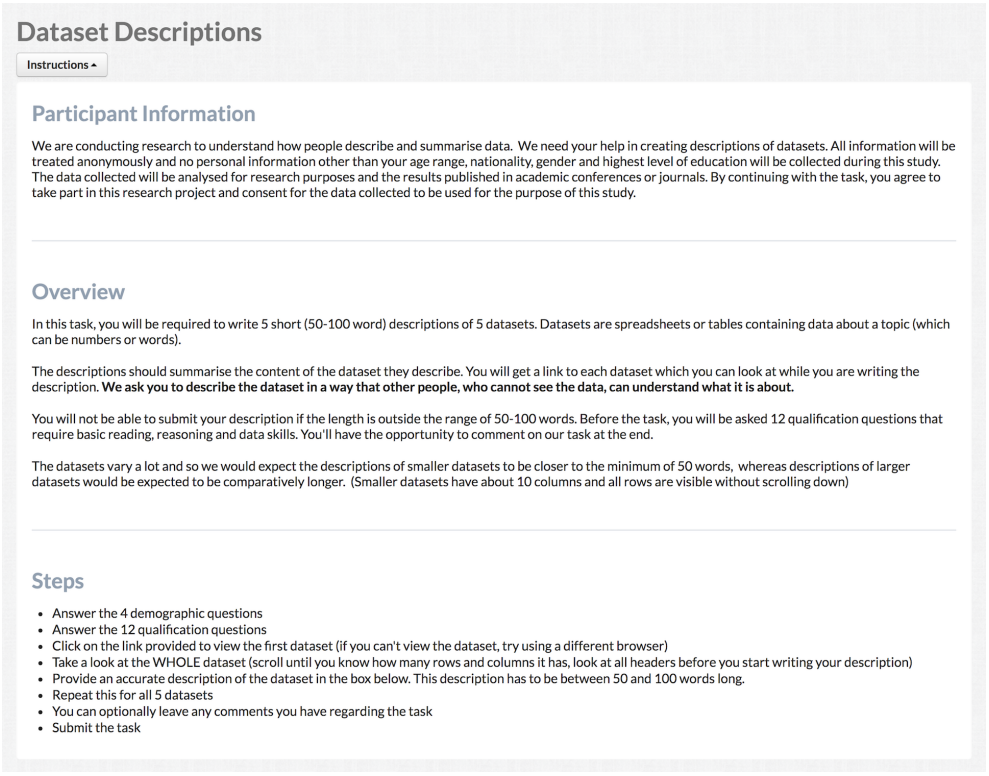


FIGURE B.6: Task 1 (1/2)

The task

We would like you now to write good quality descriptions of 5 datasets. A good quality description summarises the content of the dataset in an understandable way. Have a look at examples of different good quality descriptions below.

Examples

Here is a screenshot of part of a dataset:

Type	Date	Part of a policing operation	Policing operation	Latitude	Longitude	Gender	Age range	Self-defined ethnicity	Officer-defined ethnicity	Legislation
Person search	2017-02-01	False		53.569438	-2.424162	Male	over 34	White - White British (W1)	White	Police and Criminal Evide
Person and Vehicle search	2017-02-01	False		53.60951	-2.156453	Male	25-34	Asian or Asian British - Pakistani (A2)	Asian	Misuse of Drugs Act 1971
Person and Vehicle search	2017-02-01	False		53.609402	-2.156801	Male	18-24	Asian or Asian British - Pakistani (A2)	Asian	Misuse of Drugs Act 1971
Person search	2017-02-01	False		53.509444	-2.251033			White - Any other White ethnic backgrou	White	Police and Criminal Evide
Person search	2017-02-01	False		53.509444	-2.251033	Male	25-34	White - Any other White ethnic backgrou	White	Police and Criminal Evide
Person and Vehicle search	2017-02-01	False		53.495022	-2.253088			Asian or Asian British - Any other Asian	Asian	Misuse of Drugs Act 1971
Person and Vehicle search	2017-02-01	False		53.494647	-2.251715			Asian or Asian British - Any other Asian	Asian	Misuse of Drugs Act 1971
Person and Vehicle search	2017-02-01	False		53.496563	-2.251304			Asian or Asian British - Any other Asian	Asian	Misuse of Drugs Act 1971
Person search	2017-02-02	False		53.591656	-2.45605	Male	over 34	White - White British (W1)	White	Police and Criminal Evide

The following 3 example descriptions are all considered good quality:

Example description 1

Police stop and search activities in the Greater Manchester area from February 2017. Contains:
- Type of search: person or person plus vehicle search
- Date, time, location (in longitude and latitude) of the search
- Whether the search was part of a policing operation
- Information about the searched person: gender, age range, ethnicity
- Information about the stop and search activity: legislation, object of search, the outcome (not clear), whether more than outer clothing was removed

CSV file, 225 rows and 15 columns. Never was more than outer clothing removed. In majority there was nothing found. Majority of the searched persons were male.

Example description 2

The csv file contains 225 records of stop and searches carried out by Greater Manchester police in February 2017. It contains 15 variables which give detail on the people stopped (including ethnicity, gender and age), the stop itself (including type, location, legislation, object of search and police operation), and the outcome of the stop. The majority of variables, where they are not dependent on the result of another variable, are complete for all records. Several are missing only 3-5 records, however details on both gender and age range are missing for about 40 values each.

Example description 3

Details police stop and search data from the Greater Manchester area. There were 225 cases. Headings: Type, Date, Part of policing operation, Policing operation, Latitude, Longitude, Gender, Age range, Self-defined ethnicity, Officer defined ethnicity, Legislation, Object of search, Outcome, Outcome linked to object of search and Removal of more than just outer clothing.

Necessary to detail whether it was part of a police operation or not?
Titles unclear unless given context
Gaps in the data; needs cleaning
Gender terminology unclear - officer determined or self determined? Sex a better term
Necessary to define difference between officer and self perceived ethnicity?

Tips

Imagine you are describing the dataset to another person who does not have access to it.
Try to summarise the whole dataset in your description.
If the dataset doesn't open try doing the task in a different browser (e.g. Google Chrome or Safari) or open the link in Google Sheets.

Click here to get to dataset 1:

Word count: 0

Enter your description of dataset 1 here (required)

FIGURE B.7: Task 1 (2/2)

B.6 Datasets in $Set-20$

Dataset	Topic	Example characteristics
$E1$	Unemployement rates worldwide	8 rows, 15 columns, temporal information (years), no geospatial info, no empty cells (1 NA)
$E2$	Election polls	7978 rows, 23 columns, temporal information (dates), geospatial info (location codes), string + numeric values, empty cells,
$E3$	Death penalty US	52 rows, 12 columns, temporal information (in headers only, referring to historical data), geospatial information (US States), mostly numeric values (integer + float), empty cells, headers clearly understandable
$E4$	Shootings US	337 rows, 12 columns, temporal information (dates), geospatial info (US county level), string + numeric values, 1 empty column, (otherwise complete), several columns with links to sources, personal data
$E5$	Religion survey	922 rows, 47 columns, temporal information (as prevalence reports, age range), geospatial info (region), string values only, all headers are survey questions,
$E6$	Earthquake survey	1014 rows, 11 columns, geospatial information (US region), string values + currency, many columns appear like answers to fixed categories columns few empty cells, headers are survey questions
$E7$	Health & well-being	21 rows, 17 columns, geospatial information (country), mostly numeric values (integer + float), percentages, empty cells, some domain specific language, inconsistencies in formatting: link
$E8$	House of commons speaker	563 rows, 22 columns, temporal information only in headers, string + numeric values, currencies, empty cells, personal data, links
$E9$	Visa decisions	7470 rows, 15 columns, geospatial information, temporal information (date, time), not all headers understandable
$E10$	Love actually appearances	72 rows, 15 columns, many empty cells, only string values (true/false)
$E11$	Police incidents	¿22,000 rows, 10 columns, geospatial information (state codes), temporal information (different date formats, year), complete, string + numeric values, personal data, not all headers easily understandable,
$E12$	Government websites	49 rows, 11 columns, mostly numeric values (integers + floats), currency, few empty cells, many null values,
$E13$	College grads	174 rows, 21 columns, string + numeric (integers + floats) values, complete

$E14$	Government spending	1923 rows, 11 columns, temporal information (dates), string + numeric values, complete (some ## values), undefined currency
$E15$	Facebook fact checking	2283 rows, 12 columns, temporal information (date), string + numeric values, links, columns with limited categories, empty cells
$E16$	Population statistics	53 rows, 18 columns, geospatial information (country, region), mostly numeric values (integers + floats), complete headers not easily understandable
$E17$	IRS scams	8892 rows, 10 columns, geospatial information (county), temporal information (date), string + numeric values, empty cells, personal data
$E18$	Weather	366 rows, 13 columns, temporal information, (date, year), numeric values only (integer + float), complete
$E19$	Human Development Index	188 rows, 15 columns, geospatial information (country + ISO code), string + numeric values, empty cells
$E20$	Life expectancy UK	435 rows, 10 columns, geospatial information (UK area, area codes), string + numeric values, empty cells, redundant information

B.7 Dataset summary tool

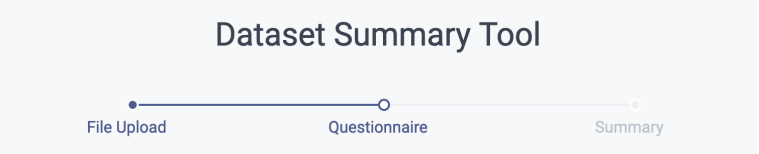


FIGURE B.8: Title

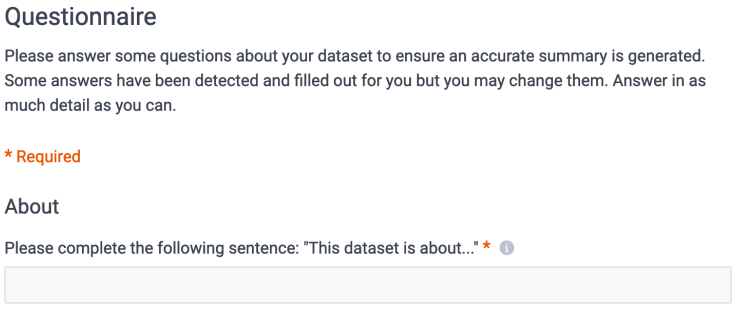


FIGURE B.9: About

FIGURE B.10: Format

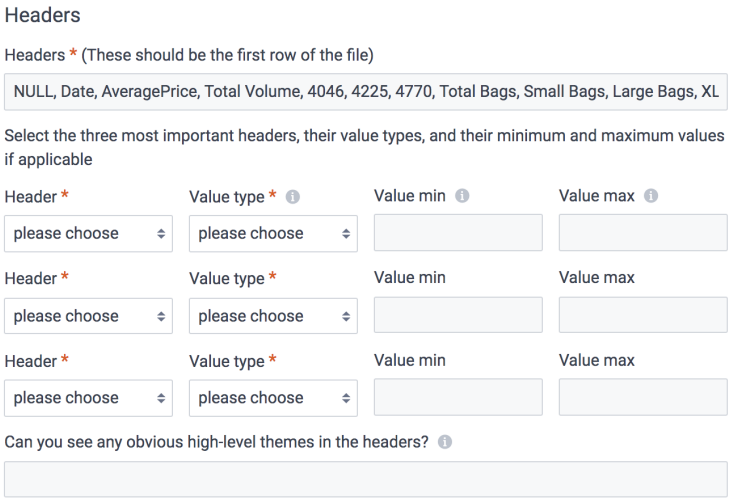


FIGURE B.11: Header

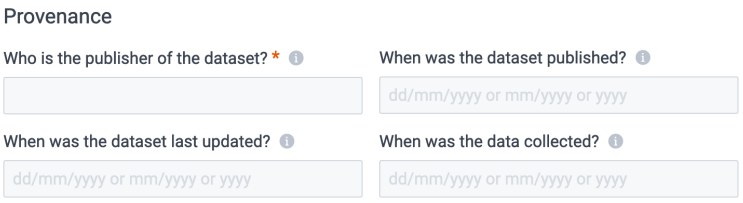


FIGURE B.12: Provenance

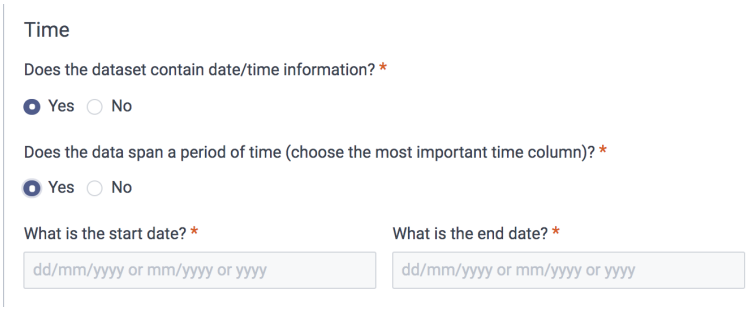


FIGURE B.13: Time

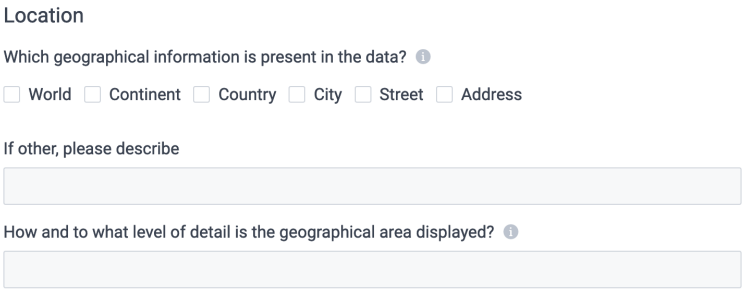


FIGURE B.14: Location

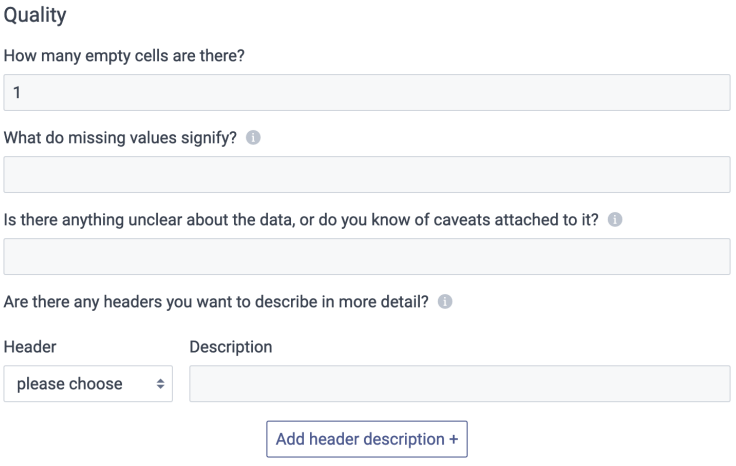


FIGURE B.15: Quality

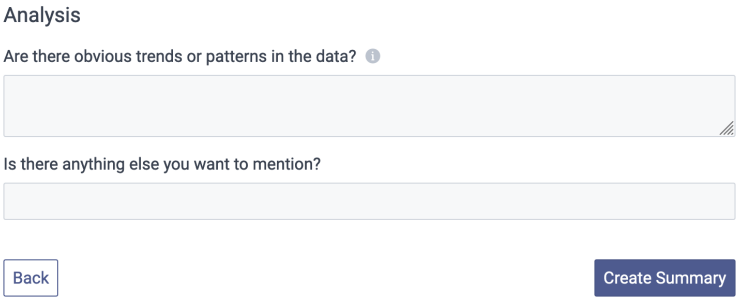


FIGURE B.16: Analysis

B.7.1 Requirements summary tool

	Questions	Sub questions	Answer type	Restrictions
1	About	How would you describe the dataset in one sentence?	Text field	not longer than 20 words / requires at least 3 words.
2	What does the dataset look like? *machine readability*	2a. File format? 2b. Number of rows? 2c. Number of columns 2d. Anything else?	2a. single choice and one other option with a textfield 2b. integer input 2c. integer input 2d .text field	2a. required 2b. required 2c. required 2d. not required
3i	What are the headers in the csv file? *(the tool should display all the headers, and allow to choose the most relevant ones)* ADD: ability to overwrite prefilled headers	3i.a. What are the most important / meaningful columns (up to three)? *(Explanation in text underneath - tbc Laura)* 3i.b. Can you group the headers in a sensible way?	3i.a. selectable 3i.b. text field	3i.a. required 3i.b. not required
3ii	What are the value types and value ranges for the most important headers? *(selected in 3i.a)* ADD: ability to overwrite prefilled header values	3ii.a. Take the 3 columns from 3b and ask for value types 3ii.b. Value ranges (what is the lowest value and what is the highest value in this column?).	3ii.a. value type for up to 3 columns (depending on 3b): dropdown/multiple choice with the options: words, numbers, dates 3ii.b.. values ranges up to 3 text fields (depending on 3b)	3ii.a. required 3ii.b. not required
4	Where is the data from?	5a. Who is the publisher? 5b. When was it published? 5c. When was it last updated? 5d. When was the data collected?	5a. Textfield 5b. Textfield /date input 5c. Textfield /date input 5d. Textfield /date input	5a. required 5b. required 5c. required 5d. required
5	In what way does the dataset mention time?	6a.What is the timeframe covered by / mentioned in the data? 6b. To what level of detail is time displayed? (E.g. Year /month /week / day/ hour…etc)	6a. Give a timeframe based on the choice - smarter display (timeline?) 6b. Choose level of detail: Year/Months/Weeks/Hours/Minutes/Seconds/other (buttons, only other is a textfield)	6a. required 6b. not required
6	In what way does the dataset mention location?	7a. What is the geographical area covered by the data? 7b. How and to what level of detail is the geographical area displayed? (E.g. latitude/longitude, street name, city, county, country etc.)	7a. Choose level of detail: World/Continent/Country/City/Street/Address/Other/ No geographical Info in this dataset (*Map display*?) 7b. Textfield	7a. required 7b. not required
7	Is there anything unclear about the data, or do you have reason to doubt the quality?	8a. How complete is the data (are there missing values)? 8b. Are all column names self explanatory? 8c. What do missing values mean?	8a. Textfield 8b. Textfield 8c. Textfield	8a. required 8b. not required 8c. not required
8	Is there anything that you would like to point out, or analyse in more detail?	9a. Do you see obvious trends in the data? 9a. Anything else you want to mention?	9a. Textfield 9b. Textfield	9a. not required 9b. not required

FIGURE B.17: Requirements for summary tool based on the template in Chapter 4

Appendix C

Appendix Chapter 5

C.1 Risk assessment form

There is no direct risk posed to participants as the study is done online, on a voluntary basis and there will be an option to participate anonymously.

There is a very low likelihood of the following risks: Breach of the database Breach of the individual investigators computers

We mitigate these risks by storing data in an encrypted database, through a securely generated password. The event of a database breach would have a low impact as no sensitive data is collected for the purpose of the study.

Data analysis will be done on the investigators computers, which are password protected and the data will not be accessible to anyone other than the named investigators.

The event of a breach would have a very low impact as no sensitive data is collected for the purpose of the study and participants names and email addresses will not be stored locally; as they are only used for communication purposes in the case of 2 participants who win in the random prize draw for Amazon vouchers.

For further information related to data protection and storage refer to the DPA plan.

C.2 Recruitment email

Hi, I am part of a team of researchers at the Open Data Institute and at the University of Southampton. We want to understand how we could improve the way people find data online. We are conducting a study on how people search for data to get a better idea of how we should present datasets in search results.

If you would be interested in helping us, please complete this survey:
https://wdaqua-survey.herokuapp.com/
It should take no more than 10-15 minutes to complete and there are no right or wrong answers. You will be able to withdraw from the task at any time.

If you complete the survey you can take part in a prize draw for one of two 50 Amazon vouchers! Many thanks for participating - we really appreciate it!

Best regards,
Laura
(Please use Chrome, Firefox, or IE as a browser).

C.3 Participant information and consent

Searching data on the web

We want to improve how people find data online, therefore we are conducting a study on how people search for data. If you are interested in helping us, please complete this survey. It will help us understand how we should present datasets as search results.

The task takes 10-15 minutes to complete. There are no right or wrong answers and you can withdraw from the task at any time.

Data Usage Policy

You can take part in the study anonymously (without giving your contact information).

The study consists of demographic questions and you will be asked about your experience of working with data. Then you will complete a data searching task, and afterwards you will be asked questions about this searching experience.

By agreeing to participate, you are giving consent for this information to be reported in aggregated statistics, which can be published as open data. These may be reported in academic work and in conferences, but no individual responses that you may choose to provide will be published.

If you want to take part in a prize draw for one of two 50 Amazon vouchers you will be asked to enter your email address and name at the end of the survey. This data will only be used for the purpose of informing the winners of the prize draw . You can separately opt in to being contacted for future studies.

Your contact information will not be shared with anyone else other than the study team, and you have the right to request the deletion of your response. However, if you decide

not to provide contact information you will not be eligible for the prize draw and we cannot remove your response at a later time, as we will not be able to identify it.

For further information about the study please contact:

Laura Koesten
PhD researcher at The Open Data Institute and the University of Southampton, UK
Email: laura.koesten@theodi.org

(This study has been approved by the University of Southampton Ethical Advisory Committee (ERGO number 41384))

C.4 Interface conditions

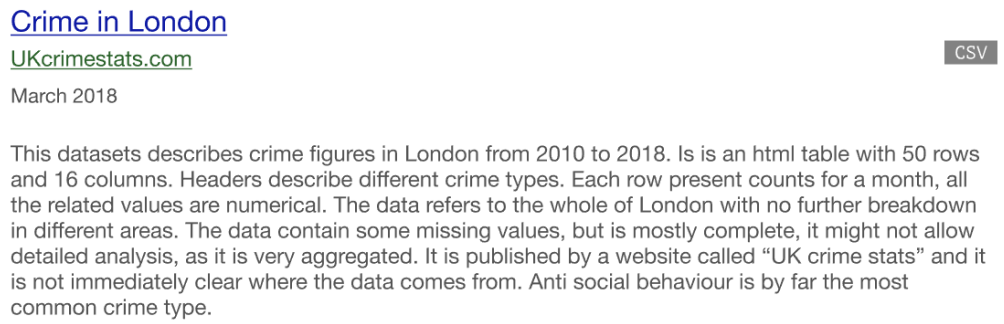


FIGURE C.1: Viewtype 0 - text

FIGURE C.2: Viewtype 1 - table

FIGURE C.3: Viewtype 2 - title

Crime in London
UKcrimestats.com
March 2018
HTML

	ASB	Burglary	Robbery	Vehicle	Violent	Shoplifting	CD&A	Other Theft	Drugs	Bike Theft	Theft From the Person	Weapons	Public Order	Other	Total
Mar 2018	16,061	6,373	2,565	8,914	18,345	3,809	4,572	8,929	2,568	1,034	3,624	567	3,933	890	82,184
Feb 2018	14,009	6,582	2,425	8,712	15,776	3,620	4,219	8,219	2,592	1,015	3,283	496	3,286	765	74,999
Jan	15,920	7,601	2,811	9,399	17,406	3,887	4,866	8,876	2,854	1,132	3,509	577	3,455	889	83,182

FIGURE C.4: Viewtype 3 - preview

C.5 Post-task questions

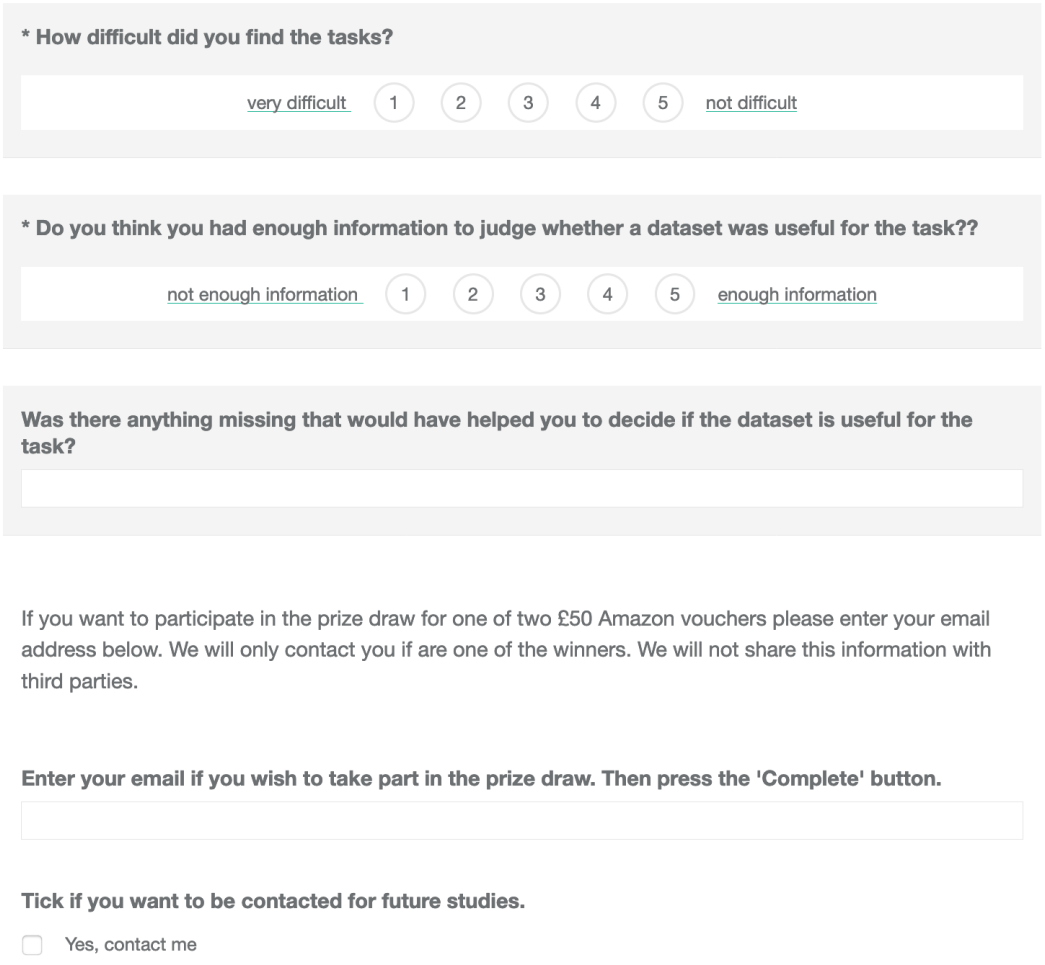


FIGURE C.5: Post-task questions

C.6 Participant demographics per country

Country	Number of participants
UK	119
US	9
AT	7
DE	6
IT, FR	4
IN, AU	3
NL, MX , RU, MY	2
MK, RO, AG, CH,SA, FI, Bouvet Island, PK, MD JP, DZ	1

TABLE C.1: Participants countries of residence, self reported (Countries abbreviated using ISO 3166-1, Alpha-2 code)

C.7 Searching for information

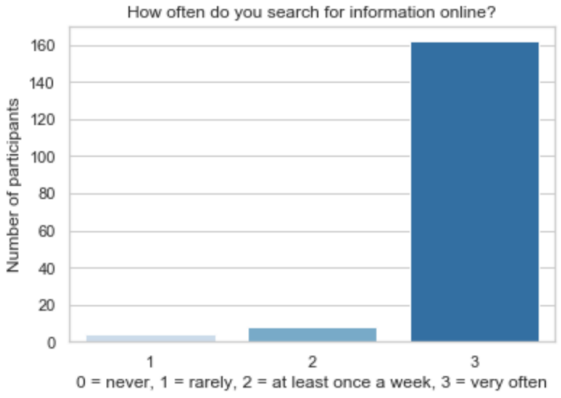


FIGURE C.6: Frequency of information search by participants

C.8 Searching for data

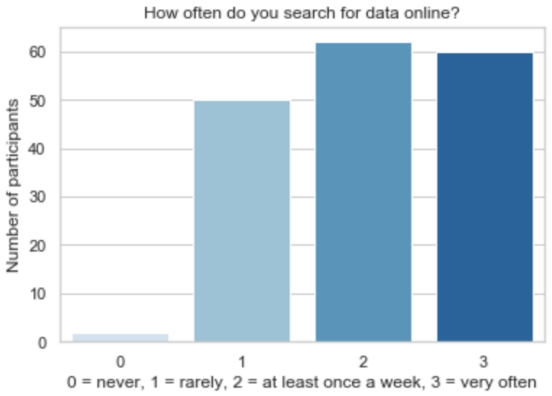


FIGURE C.7: Frequency of data search by participants

Bibliography

Adams, A. and Blandford, A. (2005). Digital libraries' support for the user's 'information journey'. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CO, USA, June 7-11, 2005, Proceedings*, pages 160–169.

Agrawal, S., Chaudhuri, S., and Das, G. (2002). Dbxplorer: A system for keyword-based search over relational databases. In *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002*, pages 5–16.

ah Kang, Y. and Stasko, J. T. (2012). Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2869–2878.

Albers, M. J. (2015). Human–information interaction with complex information for decision–making. *Informatics*, 2(2):4–19.

Altman, M., Castro, E., Crosas, M., Durbin, P., Garnett, A., and Whitney, J. (2015). Open journal systems and dataverse integration– helping journals to upgrade data publication for reusable research. *The Code4Lib Journal*, 30.

Arguello, J. and Capra, R. G. (2014). The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 539–548.

Assaf, A., Troncy, R., and Senart, A. (2015). What's up lod cloud? In Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., and Zimmermann, A., editors, *The Semantic Web: ESWC 2015 Satellite Events*, pages 247–254, Cham. Springer International Publishing.

Atz, U. (2014). The tau of data: A new metric to assess the timeliness of data in catalogues. In *Conference for E-Democracy and Open Governement*, page 257.

Au, V., Thomas, P., and Jayasinghe, G. K. (2016). Query-biased summaries for tabular data. In *Proceedings of the 21st Australasian Document Computing Symposium, ADCS 2016, Caulfield, VIC, Australia, December 5-7, 2016*, pages 69–72.

Baeza-Yates, R. A. (2018). Bias on the web. *Commun. ACM*, 61(6):54–61.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England.

Baker, J., Jones, D. R., and Burkman, J. (2009). Using visual representations of data to enhance sensemaking in data exploration tasks. *J. AIS*, 10(7):2.

Balakrishnan, A. D., Fussell, S. R., and Kiesler, S. (2008). Do visualizations improve synchronous remote collaboration? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1227–1236, New York, NY, USA. ACM.

Balatsoukas, P., Morris, A., and O'Brien, A. (2009). An evaluation framework of user interaction with metadata surrogates. *J. Information Science*, 35(3):321–339.

Balazinska, M., Howe, B., Koutris, P., Suciu, D., and Upadhyaya, P. (2013). *A Discussion on Pricing Relational Data*, pages 167–173. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bales, S. and Wang, P. (2005). Consolidating user relevance criteria: A meta-ethnography of empirical studies. *Proceedings of the American Society for Information Science and Technology*, 42(1).

Balog, K., Meij, E., and de Rijke, M. (2010). Entity search: Building bridges between two worlds. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, pages 9:1–9:5, New York, NY, USA. ACM.

Bando, L. L., Scholer, F., and Turpin, A. (2010). Constructing query-biased summaries: A comparison of human and system generated snippets. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 195–204, New York, NY, USA. ACM.

Bargmeyer, B. E. and Gillman, D. W. (2000). Metadata standards and metadata registries: An overview. In *International Conference on Establishment Surveys II, Buffalo, New York*.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *JASIS*, 45(3):149–159.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52.

Bechchi, M., Raschia, G., and Mouaddib, N. (2007). Merging distributed database summaries. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 419–428, New York, NY, USA. ACM.

Beek, W., Rietveld, L., Bazoobandi, H. R., Wielemaker, J., and Schlobach, S. (2014). Lod laundromat: a uniform way of publishing other people's dirty data. In *International Semantic Web Conference*, pages 213–228. Springer.

Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, 4(2-3):111–129.

Belkin, N. J., Cole, M., and Liu, J. (2009). A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 7–8.

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2015). Soylent: A word processor with a crowd inside. *Commun. ACM*, 58(8):85–94.

Bertin, J. (2010). *Semiology of Graphics - Diagrams, Networks, Maps*. ESRI.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

Blandford, A. and Attfield, S. (2010). *Interacting with Information*. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool Publishers.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf. Res.*, 8(3).

Borromeo, R. M., Alsaysneh, M., Amer-Yahia, S., and Leroy, V. (2017). Crowdsourcing strategies for text creation tasks. In *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017.*, pages 450–453.

Boukhelifa, N., Perrin, M.-E., Huron, S., and Eagan, J. (2017). 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, pages 3645–3656, New York, NY, USA. ACM.

Boyce, B. R., Meadow, C. T., and Kraft, D. H. (1994). *Measurement in information science*. Academic Pr.

Boydell, O. and Smyth, B. (2007). From social bookmarking to social summarization: an experiment in community-based summary generation. In *Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, Honolulu, Hawaii, USA, January 28-31, 2007*, pages 42–51.

Broder, A. Z. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative Research*, 6(1):97–113.

Buchanan, G., Blandford, A., Thimbleby, H. W., and Jones, M. (2004). Integrating information seeking and structuring: exploring the role of spatial hypertext in a digital library. In *HYPERTEXT 2004, Proceedings of the 15th ACM Conference on Hypertext and Hypermedia, August 9-13, 2004, Santa Cruz, California, USA*, pages 225–234.

Burton-Taylor (2015). Demand for financial market data and news - Report. *Burton-Taylor International Consulting LLC*.

Byström, K. and Hansen, P. (2005). Conceptual framework for tasks in information studies. *JASIST*, 56(10):1050–1061.

Cafarella, M. J., Halevy, A., and Madhavan, J. (2011). Structured data on the web. *Commun. ACM*, 54(2):72–79.

Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., and Zhang, Y. (2008). Webtables: Exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549.

Calzada Prado, J. and Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2):123–134.

Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding in-depth semistructured interviews problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization - using vision to think*. Academic Press.

Catarci, T. (2000). What happened when database researchers met usability. *Inf. Syst.*, 25(3):177–212.

Cattaneo, G., Glennon, M., Lifonti, R., Micheletti, G., Woodward, A., Kolding, M., Vacca, A., Croce, C. L., and Osimo, D. (2015). European data market SMART 2013/0063, D6 - First Interim Report.

Ceolin, D., Groth, P. T., Maccatrozzo, V., Fokkink, W., van Hage, W. R., and Nottamkandath, A. (2016). Combining user reputation and provenance analysis for trust assessment. *J. Data and Information Quality*, 7(1-2):6:1–6:28.

Chatfield, C. (2016). *The analysis of time series: an introduction*. CRC press.

Choi, J. and Tausczik, Y. R. (2017). Characteristics of collaboration in the emerging practice of open data analysis. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 835–846.

Churchill, E. F. (2012). From data divination to data-aware design. *Interactions*, 19(5):10–13.

Cleverdon, C., Mills, J., and Keen, M. (1966). Factors determining the performance of indexing systems. volume 1. design.

Cole, M., Liu, J., Belkin, N., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., and Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR*, pages 1–4.

Convertino, G. and Echenique, A. (2017). Self-service data preparation and analysis by business users: New needs, skills, and tools. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 1075–1083, New York, NY, USA. ACM.

Cormode, G. (2015). Compact summaries over large datasets. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '15, pages 157–158, New York, NY, USA. ACM.

Crabtree, A. and Mortier, R. (2015). Human data interaction: Historical lessons from social studies and CSCW. In *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway*, pages 3–21.

Craswell, N., Zoeter, O., Taylor, M. J., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 87–94.

Cutrell, E. and Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 407–416.

Daniels, P. J. (1986). Cognitive models in information retrieval - an evaluative review. *Journal of Documentation*, 42(4):272–304.

DataGovUk (2018). Uk open data portal. https://data.gov.uk/.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60.

Davies, T. and Frank, M. (2013). 'there's no such thing as raw data': exploring the socio-technical life of a government dataset. In *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*, pages 75–78.

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

Diriye, A., Wilson, M. L., Blandford, A., and Tombros, A. (2010). Revisiting exploratory search from the hci perspective. *HCIR 2010*, page 99.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., and Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5):667–690.

Ellefi, M. B., Bellahsene, Z., Dietze, S., and Todorov, K. (2016). Dataset recommendation for data linking: An intensional approach. In *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, pages 36–51.

Ellis, S. E. and Groth, D. P. (2004). A collaborative annotation system for data visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 411–414, New York, NY, USA. ACM.

Elmqvist, N. (2011). Embodied human-data interaction. In *ACM CHI Workshop "Embodied Interaction: Theory and Practice in HCI"*, pages 104–107.

Elsevier (2018). Elsevier scientific repository. https://datasearch.elsevier.com/.

Erete, S., Ryou, E., Smith, G., Fassett, K. M., and Duda, S. (2016). Storytelling with data: Examining the use of data by non-profit organizations. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1273–1283, New York, NY, USA. ACM.

Ermilov, I. and Ngomo, A.-C. N. (2016). TAIPAN: Automatic Property Mapping for Tabular Data. In *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management*.

Faisal, S., Cairns, P. A., and Blandford, A. (2007). Building for users not for experts: Designing a visualization of the literature domain. In *11th International Conference on Information Visualisation, IV 2007, 2-6 July 2007, Zürich, Switzerland*, pages 707–712.

Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experiences. *JASIS*, 32(1):23–32.

Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., Lima, R., Simske, S. J., and Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, 40(14):5755 – 5764.

Fetahu, B., Dietze, S., Nunes, B. P., Casanova, M. A., Taibi, D., and Nejdl, W. (2014). A scalable approach for efficiently generating structured dataset topic profiles. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 519–534.

Fidel, R. (2012). *Human Information Interaction: An Ecological Approach to Information Behavior*. Mit Press.

Fluit, C., Sabou, M., and Van Harmelen, F. (2006). Ontology-based information visualization: toward semantic web applications. In *Visualizing the semantic web*, pages 45–58. Springer.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168.

Freund, L. (2013). A cross-domain analysis of task and genre effects on perceptions of usefulness. *Inf. Process. Manage.*, 49(5):1108–1121.

Furnas, G. W. and Russell, D. M. (2005). Making sense of sensemaking. In *Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005*, pages 2115–2116.

Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.*, 47(1):1–66.

Ganoe, C. H., Somervell, J. P., Neale, D. C., Isenhour, P. L., Carroll, J. M., Rosson, M. B., and McCrickard, D. S. (2003). Classroom bridge: Using collaborative public and desktop timelines to support activity awareness. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, UIST '03, pages 21–30, New York, NY, USA. ACM.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., and Sripada, S. (2009). From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Commun.*, 22(3):153–186.

GESIS (2018). Gesis data search.

Gkatzia, D. (2016). Content selection in data-to-text systems: A survey. *CoRR*, abs/1610.08375.

Gkatzia, D., Lemon, O., and Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. *CoRR*, abs/1606.03254.

Gonzalez, H., Halevy, A., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., and Shen, W. (2010a). Google fusion tables: Data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, pages 175–180, New York, NY, USA. ACM.

Gonzalez, H., Halevy, A. Y., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., Shen, W., and Goldberg-Kidon, J. (2010b). Google fusion tables: web-centered data management and collaboration. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, pages 1061–1066.

GOV.UK (2014). How to make a freedom of information (foi) request. Retrieved September 05, 2016 from https://www.gov.uk/make-a-freedom-of-information-request/the-freedom-of-information-act.

Goyal, N. and Fussell, S. R. (2016). Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 288–302, New York, NY, USA. ACM.

Green, J. and Thorogood, N. (2013). *Qualitative methods for health research*. Sage.

Greenberg, J. (2010). Metadata and digital information. *Encyclopedia of Library and Information Sciences,*, pages 3610–3623.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction*, 16(2-4):257–268.

Gregory, K., Cousijn, H., Groth, P. T., Scharnhorst, A., and Wyatt, S. (2018). Understanding data retrieval practices: A social informatics perspective. *CoRR*, abs/1801.04971.

Gregory, K., Groth, P., Scharnhorst, A., and Wyatt, S. (2019a). Lost or found? discovering data needed for research.

Gregory, K., Groth, P. T., Cousijn, H., Scharnhorst, A., and Wyatt, S. (2017). Searching data: A review of observational data retrieval practices. *CoRR*, abs/1707.06937.

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., and Wyatt, S. (2019b). Understanding data search as a socio-technical practice. *Journal of Information Science*, 0(0):0165551519837182.

Greis, M., Joshi, A., Singer, K., Schmidt, A., and Machulla, T. (2018). Uncertainty visualization influences how humans aggregate discrepant information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 505:1–505:12, New York, NY, USA. ACM.

Grubenmann, T., Bernstein, A., Moor, D., and Seuken, S. (2018). Financing the web of data with delayed-answer auctions. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1033–1042, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema.org: evolution of structured data on the web. *Commun. ACM*, 59(2):44–51.

Gupta, S., Yadav, S., and Prasad, R. (2016). Document retrieval using efficient indexing techniques: A review. *International Journal of Business Analytics (IJBAN)*, 3(4):64–82.

Gysel, C. V., de Rijke, M., and Kanoulas, E. (2016). Learning latent vector spaces for product search. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 165–174.

Hajizadeh, A. H., Tory, M., and Leung, R. (2013). Supporting awareness through collaborative brushing and linking of tabular data. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2189–2197.

Halevy, A. Y., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. (2016). Goods: Organizing google's datasets. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 795–806.

He, J., Duboue, P., and Nie, J. (2012). Bridging the gap between intrinsic and perceived relevance in snippet generation. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1129–1146.

Hearst, M. (2009). *Search user interfaces*. Cambridge University Press.

Heer, J. and Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Commun. ACM*, 55(4):45–54.

Heer, J., Viégas, F. B., and Wattenberg, M. (2007). Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1029–1038, New York, NY, USA. ACM.

Hendler, J., Holm, J., Musialek, C., and Thomas, G. (2012). Us government linked open data: Semantic.data.gov. *IEEE Intelligent Systems*, 27(3):25–31.

Hidi, S. and Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, 56(4):473–493.

Ingwersen, P. and Järvelin, K. (2005). *The Turn - Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Kluwer International Series on Information Retrieval*. Kluwer.

Ingwersen, P. E. R. (1992). *Information Retrieval Interaction*. Taylor Graham.

International, O. K. (2016). Good tables.

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., and Yu, C. (2007). Making database systems usable. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007*, pages 13–24.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3):407–432.

Joslyn, S. and LeClerc, J. (2013). Decisions with uncertainty: the glass half full. *Current Directions in Psychological Science*, 22(4):308–315.

Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., and Simperl, E. (2019). Characterising dataset search - an analysis of search logs and data requests. *J. Web Semant.*, 55:37–55.

Kammerer, Y. and Gerjets, P. (2010). How the interface design influences users' spontaneous trustworthiness evaluations of web search results: Comparing a list and a grid interface. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 299–306, New York, NY, USA. ACM.

Kassen, M. (2013). A promising phenomenon of open data: A case study of the chicago open data project. *Government Information Quarterly*, 30(4):508 – 513.

Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. (2016). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5092–5103, New York, NY, USA. ACM.

Kay, M., Morris, D., schraefel, m., and Kientz, J. A. (2013). There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 401–410, New York, NY, USA. ACM.

Keim, D. A., Andrienko, G. L., Fekete, J., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Information Visualization - Human-Centered Issues and Perspectives*, pages 154–175.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224.

Kelly, D. and Azzopardi, L. (2015). How many results per page?: A study of SERP size, search behavior and user experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 183–192.

Kern, D. and Mathiak, B. (2015). Are there any differences in data set retrieval compared to well-known literature retrieval? In *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings*, pages 197–208.

Kim, J. and Monroy-Hernandez, A. (2016). Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1018–1027, New York, NY, USA. ACM.

Kindling, M., van de Sandt, S., Rücknagel, J., Schirmbacher, P., Pampel, H., Vierkant, P., Bertelmann, R., Kloska, G., Scholze, F., and Witt, M. (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, 23(3/4).

Klouche, K., Ruotsalo, T., Micallef, L., Andolina, S., and Jacucci, G. (2017). Visual re-ranking for multi-aspect information retrieval. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 57–66.

Knight, S. and Burn, J. M. (2005). Developing a framework for assessing information quality on the world wide web. *InformingSciJ*, 8:159–172.

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., and Simperl, E. (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, Denver, CO, USA, May 06-11, 2017.*, pages 1277–1289.

Kuhlthau, C. (2004). *Seeking Meaning: A Process Approach to Library and Information Services*. Information management, policy, and services. Libraries Unlimited.

Kukich, K. (1983). Design of a knowledge-based report generator. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, pages 145–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kuzon, W., Urbanchek, M., and McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of plastic surgery*, 37:265–272.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 52(2).

Law, A. S., Freer, Y., Hunter, J., Logie, R. H., McIntosh, N., and Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of clinical monitoring and computing*, 19(3):183–194.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *IJDC*, 6(2):4–37.

Lehmberg, O., Ritze, D., Meusel, R., and Bizer, C. (2016). A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 75–76.

Li, X., Liu, B., and Yu, P. S. (2008). Time sensitive ranking with application to publication search. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 893–898.

Li, Y. and Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837.

Lidwell, W., Holden, K., and Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.

Lin, S., Fortuna, J., Kulkarni, C., Stone, M., and Heer, J. (2013). Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, volume 32, pages 401–410. Wiley Online Library.

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010). Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 68–76, New York, NY, USA. ACM.

Liu, B. and Jagadish, H. V. (2009a). 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*, pages 1115–1118.

Liu, B. and Jagadish, H. V. (2009b). A spreadsheet algebra for a direct data manipulation query interface. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29 2009 - April 2 2009, Shanghai, China*, pages 417–428.

Liu, X., Tian, Y., He, Q., Lee, W., and McPherson, J. (2014). Distributed graph summarization. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 799–808.

Lopez-Veyna, J. I., Sosa, V. J. S., and López-Arévalo, I. (2012). KESOSD: keyword search over structured data. In *Proceedings of the Third International Workshop on Keyword Search on Structured Data, KEYS 2012, Scottsdale, AZ, USA, May 20, 2012*, pages 23–31.

Madhavan, J., Halevy, A. Y., Cohen, S., Dong, X. L., Jeffery, S. R., Ko, D., and Yu, C. (2006). Structured data meets the web: A few observations. *IEEE Data Eng. Bull.*, 29(4):19–26.

Maguire, D. J. and Longley, P. A. (2005). The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, 29(1):3–14.

Malaverri, J. E. G., Mota, M. S., and Medeiros, C. B. (2013). Estimating the quality of data using provenance: a case study in escience. In *19th Americas Conference on Information Systems, AMCIS 2013, Chicago, Illinois, USA, August 15-17, 2013*.

Manyika, M. G. I. J., Chui, M., Groves, P., Farrell, D., Kuiken, S. V., and Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information.

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., and Luo, H. (2016). When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472, New York, NY, USA. ACM.

Marchionini, G. (1995). Information seeking in electronic environments. In *Cambridge Series on Human-Computer Interaction*.

Marchionini, G. (2006a). Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46.

Marchionini, G. (2006b). Toward human-computer information retrieval. *Bulletin of the American Society for Information Science and Technology*, 32(5):20–22.

Marchionini, G. (2008). Human-information interaction research and development. *Library & Information Science Research*, 30(3):165–174.

Marchionini, G., Haas, S. W., Zhang, J., and Elsas, J. L. (2005). Accessing government statistical information. *IEEE Computer*, 38(12):52–61.

Marchionini, G. and White, R. (2007). Find what you need, understand what you find. *Int. J. Hum. Comput. Interaction*, 23(3):205–237.

Marcos, M., Gavin, F., and Arapakis, I. (2015). Effect of snippets on user experience in web search. In *Proceedings of the XVI International Conference on Human Computer Interaction, Interacción 2015, Vilanova i la Geltrú, Spain, September 7-9, 2015*, pages 47:1–47:8.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Marienfeld, F., Schieferdecker, I., Lapi, E., and Tcholtchev, N. (2013). Metadata aggregation at govdata. de: An experience report. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 21. ACM.

Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 135–144.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mayernik, M. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators.

Mei, H., Bansal, M., and Walter, M. R. (2015). What to talk about and how? selective generation using lstms with coarse-to-fine alignment. *CoRR*, abs/1509.00838.

Mizzaro, S. (1997). Relevance: The whole history. *JASIS*, 48(9):810–832.

Morsey, M., Lehmann, J., Auer, S., and Ngomo, A.-C. N. (2011). Dbpedia sparql benchmark–performance assessment with real queries on real data. In *International Semantic Web Conference*, pages 454–469. Springer.

Morton, K., Balazinska, M., Grossman, D., and Mackinlay, J. D. (2014). Support the data enthusiast: Challenges for next-generation data-analysis systems. *PVLDB*, 7(6):453–456.

Muller, M. J., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., and Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 126.

Naumann, F. (2014). Data profiling revisited. *SIGMOD Rec.*, 42(4):40–49.

Naumer, C., Fisher, K., and Dervin, B. (2008). Sense-making: a methodological perspective. In *Sensemaking Workshop, CHI'08*.

Neumaier, S. and Polleres, A. (2018). Enabling spatio-temporal search in open data. Technical report, Department für Informationsverarbeitung und Prozessmanagement, WU Vienna University of Economics and Business.

Neumaier, S., Umbrich, J., and Polleres, A. (2016). Automated quality assessment of metadata across open data portals. *J. Data and Information Quality*, 8(1):2:1–2:29.

Nguyen, T. T., Nguyen, Q. V. H., Weidlich, M., and Aberer, K. (2015). Result selection and summarization for web table search. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 231–242. IEEE.

Noy, N., Burgess, M., and Brickley, D. (2019). Google dataset search: Building a search engine for datasets in an open web ecosystem.

Owczarzak, K. and Dang, H. T. (2009). Evaluation of automatic summaries: Metrics under varying data conditions. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, UCNLG+Sum '09, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: bringing order to the web.

Park, T. K. (1993). The nature of relevance in information retrieval: An empirical study. *The library quarterly*, 63(3):318–351.

Pasquetto, I. V., Randles, B. M., and Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, 16.

Pfister, H. and Blitzstein, J. (2015). cs109/2015, lectures 01-introduction. https://github.com/cs109/2015/tree/master/Lectures.

Pham, M., Alse, S., Knoblock, C. A., and Szekely, P. A. (2016). Semantic labeling: A domain-independent approach. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, pages 446–462.

Pimplikar, R. and Sarawagi, S. (2012). Answering table queries on the web using column keywords. *Proceedings of the VLDB Endowment*, 5(10):908–919.

Pirolli, P. and Rao, R. (1996). 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*, pages 67–80.

Pressac, J.-B. (2016). Open refine documentation.

Rao, R. and Card, S. K. (1995). Exploring large tables with the table lens. In *Conference Companion on Human Factors in Computing Systems*, CHI '95, pages 403–404, New York, NY, USA. ACM.

Rasmussen, J. (1985). Trends in human reliability analysis. *Ergonomics*, 28(8):1185–1195.

Reiche, K. J. and Höfig, E. (2013). Implementation of metadata quality metrics and application on public government data. In *IEEE 37th Annual Computer Software and Applications Conference, COMPSAC Workshops 2013, Kyoto, Japan, July 22-26, 2013*, pages 236–241.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Resnick, M. L., Maldonado, C., Santos, J. M., and Lergier, R. (2001). Modeling on-line search behavior using alternative output structures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 45, pages 1166–1170. SAGE Publications Sage CA: Los Angeles, CA.

Rieh, S. Y., Collins-Thompson, K., Hansen, P., and Lee, H. (2016). Towards searching as a learning process: A review of current perspectives and future directions. *J. Information Science*, 42(1):19–34.

Robertson, S. (2008). On the history of evaluation in IR. *J. Information Science*, 34(4):439–456.

Robson, C. and McCartan, K. (2016). *Real world research*. John Wiley & Sons.

Roddick, J. F., Mohania, M. K., and Madria, S. K. (1999). Methods and interpretation of database summarisation. In *Database and Expert Systems Applications, 10th International Conference, DEXA '99, Florence, Italy, August 30 - September 3, 1999, Proceedings*, pages 604–615.

Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 13–19.

Roth, E. M., Woods, D. D., and Pople Jr, H. E. (1992). Cognitive simulation as a tool for cognitive task analysis. *Ergonomics*, 35(10):1163–1198.

Rowley, J. (2000). Product search in eshopping: a review and research propositions. *Journal of Consumer Marketing*, 17(1):20–35.

Russell, D. M. (2003). Learning to see, seeing to learn: visual aspects of sensemaking. In *Human Vision and Electronic Imaging VIII, Santa Clara, CA, USA, January 20, 2003*, pages 8–21.

Sahib, N. G., Tombros, A., and Stockman, T. (2012). A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *JASIST*, 63(2):377–391.

Saint-Paul, R., Raschia, G., and Mouaddib, N. (2005). General purpose database summarization. In *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005*, pages 733–744.

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, pages 201–218.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the Annual Meeting-American Society for Information Science*, volume 34, pages 313–327. LEARNED INFORMATION (EUROPE) LTD.

Schamber, L., Eisenberg, M. B., and Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Inf. Process. Manage.*, 26(6):755–776.

Schoeffmann, K., Hudelist, M. A., and Huber, J. (2015). Video interaction tools: A survey of recent work. *ACM Comput. Surv.*, 48(1):14:1–14:34.

Scott, D., Hallett, C., and Fettiplace, R. (2013). Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories. *Patient Education and Counseling*, 92(2):153 – 159.

Shekhar, S., Celik, M., George, B., Mohan, P., Levine, N., Wilson, R. E., and Mohanty, P. (2010). Spatial analysis of crime report datasets. *National Science Foundation (NSF), Washington, DC., USA*.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996*, pages 336–343.

Sieg, A., Mobasher, B., and Burke, R. D. (2007). Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 525–534.

Simianu, V. V., Grounds, M. A., Joslyn, S. L., LeClerc, J. E., Ehlers, A. P., Agrawal, N., Alfonso-Cristancho, R., Flaxman, A. D., and Flum, D. R. (2016). Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey. *BMC Medical Informatics and Decision Making*, 16(1):153.

Simmhan, Y. L., Plale, B., and Gannon, D. (2008). Karma2: Provenance management for data-driven workflows. *Int. J. Web Service Res.*, 5(2):1–22.

Simon, H. A. (1973). The structure of ill structured problems. *Artif. Intell.*, 4(3):181–201.

Sripada, S. G., Reiter, E., Davy, I., and Nilssen, K. (2004). Lessons from deploying nlg technology for marine weather forecast text generation. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pages 760–764, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Sripada, S. G., Reiter, E., Hunter, J., and Yu, J. (2003). Summarizing neonatal time series data. In *Proceedings of the Tenth Conference on European Chapter of the*

Association for Computational Linguistics - Volume 2, EACL '03, pages 167–170, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stamatogiannakis, M., Groth, P. T., and Bos, H. (2014). Looking inside the black-box: Capturing data provenance using dynamic instrumentation. In *Provenance and Annotation of Data and Processes - 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*, pages 155–167.

Stasko, J. T., Görg, C., and Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132.

Sterling, G. (2015). It's Official: Google Says More Searches Now On Mobile Than On Desktop. Company officially confirms what many have been anticipating for years. *Search Engine Land*.

Sultanum, N., Brudno, M., Wigdor, D., and Chevalier, F. (2018). More text please! understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 422.

Sutcliffe, A. G. and Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3):321–351.

Swamiraj, M. and Freund, L. (2015). Facilitating the discovery of open government datasets through an exploratory data search interface. Open Data Research Symposium.

Taghavi, M., Patel, A., Schmidt, N., Wills, C., and Tew, Y. (2012). An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces*, 34(1):162–170.

Tam, N. T., Hung, N. Q. V., Weidlich, M., and Aberer, K. (2015). Result selection and summarization for web table search. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 231–242.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246.

Thomas, P., Omari, R. M., and Rowlands, T. (2015). Towards searching amongst tables. In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*, pages 8:1–8:4.

Tombros, A., Sanderson, M., and Gray, P. (1998). Advantages of query biased summaries in information retrieval. In *SIGIR*, volume 98, pages 2–10.

Trippas, J. R., Spina, D., Cavedon, L., and Sanderson, M. (2017). How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 325–328.

Tukey, J. W. (1970). *Exploratory Data Analysis: Limited Preliminary Ed.* Addison-Wesley Publishing Company.

Ubaldi, B. (2013). *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*. OECD Publishing.

Umbrich, J., Neumaier, S., and Polleres, A. (2015). Quality assessment and evolution of open data portals. pages 404–411.

van der Meulen, M., Logie, R. H., Freer, Y., Sykes, C., McIntosh, N., and Hunter, J. (2010). When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1):77–89.

Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G., and Wu, C. (2011). Recovering semantics of tables on the web. *Proc. VLDB Endow.*, 4(9):528–538.

Verhulst, S. and Young, A. (2016). Open data impact when demand and supply meet. Technical report, GOVLAB.

Viégas, F. B., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. M. (2007). Manyeyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1121–1128.

Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*.

Voorhees, E. M., Harman, D. K., et al. (2005). *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge.

W3Schools.com (2016). Browser statistics. http://www.w3schools.com/browsers/default.asp.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, 39(11):86–95.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33.

Weerkamp, W., Berendsen, R., Kovachev, B., Meij, E., Balog, K., and de Rijke, M. (2011). 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*, pages 45–54.

Wen, Y., Zhu, X., Roy, S., and Yang, J. (2018). Interactive summarization and exploration of top aggregate query answers. *PVLDB*, 11(13):2196–2208.

White, R. W., Bailey, P., and Chen, L. (2009). 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*, pages 363–370.

White, R. W. and Roth, R. A. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

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., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Willett, W., Ginosar, S., Steinitz, A., Hartmann, B., and Agrawala, M. (2013). Identifying redundancy and exposing provenance in crowdsourced data analysis. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2198–2206.

Wilson, M. L. (2011). *Search User Interface Design*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Wilson, M. L., Kules, B., m. c. schraefel, and Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97.

Wilson, M. L., m. c. schraefel, and White, R. W. (2009). Evaluating advanced search interfaces using established information-seeking models. *JASIST*, 60(7):1407–1422.

Wiseman, S., Shieber, S. M., and Rush, A. M. (2017). Challenges in data-to-document generation. *CoRR*, abs/1707.08052.

Woods, D. D. (1991). The cognitive engineering of problem representations. *Human-computer interaction and complex systems*, pages 169–188.

Woods, D. D., Patterson, E. S., and Roth, E. M. (2002). Can we ever escape from data overload? a cognitive systems diagnosis. *Cognition, Technology & Work*, 4(1):22–36.

Wynholds, L., Jr., D. S. F., Borgman, C. L., and Traweek, S. (2011). When use cases are not useful: data practices, astronomy, and digital libraries. In *Proceedings of the*

2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011, pages 383–386.

Xie, X., Liu, Y., Wang, X., Wang, M., Wu, Z., Wu, Y., Zhang, M., and Ma, S. (2017). Investigating examination behavior of image search users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 275–284, New York, NY, USA. ACM.

Yankelovich, N. and Lai, J. (1999). Designing speech user interfaces. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems, CHI Extended Abstracts '99, Pittsburgh, Pennsylvania, USA, May 15-20, 1999*, pages 124–125.

Yi, J. S., ah Kang, Y., Stasko, J. T., and Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1224–1231.

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing*, 14(2):116–137.

Yu, J., Reiter, E., Hunter, J., and Mellish, C. (2007). Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49.

Yu, P. S., Li, X., and Liu, B. (2005). 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*, pages 543–549.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., and Keim, D. (2012). Visual analytics for the big data eraa comparative review of state-of-the-art commercial systems. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 173–182. IEEE.

Zhang, S. (2018). Smarttable: Equipping spreadsheets with intelligent assistancefunctionalities. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 1447–1447, New York, NY, USA. ACM.

Zhang, S. and Balog, K. (2018). 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*, pages 1553–1562.

Zhuge, H. (2016). *Multi-dimensional summarization in cyber-physical society*. Morgan Kaufmann.