The Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour

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

Emilia Kacprzak
The Open Data Institute, UK
Univ. of Southampton, UK
e.kacprzak@theodi.org

Jenifer F A Tennison
The Open Data Institute, UK
jeni@theodi.org

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

ABSTRACT
Structured data such as databases, spreadsheets and web tables is becoming critical in every domain and professional role. Yet we still do not know much about how people interact with it. Our research focuses on the information seeking behaviour of people looking for new sources of structured data online, including the task context in which the data will be used, data search, and the identification of relevant datasets from a set of possible candidates. We present a mixed-methods study covering in-depth interviews with 20 participants with various professional backgrounds, supported by the analysis of search logs of a large data portal. Based on this study, we propose a framework for human structured-data interaction and discuss challenges people encounter when trying to find and assess data that helps their daily work. We provide design recommendations for data publishers and developers of online data platforms such as data catalogs and marketplaces. These recommendations highlight important questions for HCI research to improve how people engage and make use of this incredibly useful online resource.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation (e.g. HCI): User-centered design

Author Keywords
Human Data Interaction; Data Search; Data Portal

INTRODUCTION
Structured data, which is data that is explicitly organised, for example in relational databases, spreadsheets and web tables, is becoming critical in every domain and professional role [46]. We use it in various ways - from consulting official statistics to running scientific experiments, finding travel routes, creating maps, predicting elections and designing better products. More and more of it can be accessed or purchased online - a 2011 study by Cafarella et al. found more than one billion structured data resources on the (deep) web (as HTML tables, lists, etc.) [11], while McKinsey estimated two years later that more than one million datasets have been made openly available by governments worldwide [43, 63]. At the same time, the demand for financial, economic and marketing data provided by vendors such as Bloomberg and others continues to increase [10]. And yet, despite its abundance and applications, we still know very little about how people search for data, understand it and put it to use. The tools that support someone's 'data journey' [3] - from specifying a goal and finding the data they need to exploring new resources and assessing their relevance - often do not offer the best user experience [61].

Various scientific communities, including information retrieval (IR), databases, Linked Data, data visualisation and HCI, have looked at such data journeys from different angles. They have proposed new interaction models to engage with a particular species of data, such as graphs [57] or time series [16]; studied information needs and how they are formulated [7]; and developed tools for specific data-related activities, for example statistical analysis [33], visualisation [19], personal information management [31] and teamwork [5].

The scenario we are targeting is slightly different: imagine a data journalist writing an article about the runway expansion at London's main airport 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, arguing for or against expansion plans at each airport location. A large share of the relevant data is already available online, published by governmental agencies, researchers and other journalists. However, finding and using it is not always straightforward. The journalist could use regular search engines in the same way she does when looking for less structured kinds of information (such as regular Web sites). She might also know of a particular

CHI 2017 May 06-11, 2017, Denver, CO, USA

ACM ISBN 978-1-4503-4655-9/17/05.
DOI: http://dx.doi.org/10.1145/3025453.3025838

data provider, which offers offers access to their data via an online data catalogue. Depending on the tools used to discover the data, this step might involve downloading data files; working with different formats (e.g., CSV, HTML, RDF, relational tables); choosing among 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.

In our research we study scenarios like this 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. On this 'web of data' [6], our data journalist would be able to discover hard facts for her story in the same way she googles the web today - she would tell the data search engine of her choice what she is looking for and the engine would return a ranked list of data sources, or a faceted search interface which would be designed to support her understanding of the data. For each source, the journalist would see a summary of the most important features of the dataset (attributes, descriptive statistics etc.); a sample of the data for easy exploration (perhaps similar to Google's rich snippets); individual data records matching the query; or a visualisation of the data (analogously to the image and video tabs in web search engines). We are still far away from this idea - the web of data currently resembles very much the early ages of the 'traditional' web, when people could access web pages only if they knew where they were located, or via manually crafted directories such as DMOZ [22]. Initiatives such as the Linked Open Data Cloud [6], schema.org [49] and Google's Knowledge Graph [59] show the way forward. They promote the use of interlinked data and vocabularies, which can be easily mixed and re-purposed to make structured data accessible to everyone as easily as any web page today. While this technology vision is slowly coming to life, we need to learn more about how people would engage with it, revisit theories and empirical findings from related literature, and identify data-centric activities that are still challenging to undertake. Just like previous researchers in Human Data Interaction (HDI) and other fields before us [1, 17, 23, 45, 67], our assumption is that, despite many parallels, searching for structured data - whether it is a budget spreadsheet, social network graph, weather measurement or train timetable - will most likely require different tools and interaction models than the ones available for web search.

To study the requirements of our scenario, we followed a mixed-methods approach, combining semi-structured in-depth interviews, thematic analysis and search log analysis from a large open government data portal. The interviews focused on the information seeking behaviour 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 across several domains. The responses showed that people with different skill sets and professional backgrounds follow common workflows when engaging with structured data. Capitalising on these commonalities, we then defined a *framework for human structured-data interaction* with five pillars: *tasks*, *search*, *evaluation*, *exploration* and *use*, extending models in related areas [40, 50]. The proposed framework can help portal providers and data publishers design a better experience for their users. One of the main findings of the interviews was that people often cannot locate the data they need. To understand their problems better, we then looked at one year of search logs of data.gov.uk [18]. We examined the characteristics of more than 100k user queries, which led to 577k dataset downloads, to offer a quantitative view on the common themes of the framework. Finally, we derived design recommendations for data publishers and developers of data catalogs and marketplaces, and identified routes for future research to improve how people engage with and make use of structured data online.

BACKGROUND

Our work is in the broader context of **Human Data Interaction (HDI)**. The term is used by Crabtree and Mortier to refer to how people engage with their personal data and tackle privacy issues [17]. Elmqvist proposed a broader definition that includes the manipulation, analysis and sensemaking of data [23]. In our work, we consider the whole interaction process and its context, covering both people looking for answers to questions, and those interested in a particular dataset or combinations of datasets. For example, someone trying to find the number of schools in a given post code area would need to extract the answer from a larger dataset containing all entries for all schools in a country in 2016. Someone studying how the number of schools across different regions has changed over time would need to process and aggregate several versions of the same data, published year after year. Finally, schools data could be mixed with house prices statistics to understand how one aspect influences the other. All three settings have elements of dataset search, sensemaking and use, which consider different data properties, and our study considers them equally. In the following we explain how our work has been informed by three related fields of inquiry: (1) *data search*; (2) *data sensemaking*; and finally (3) *information seeking*.

Data search presents many challenges, as ideas and tools from web search cannot be (yet) directly applied [53]. Web search engines are designed for documents, not for data [11]. Putting aside the question of links between datasets, 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 fulfill the specific needs that people have when looking for data [67]. Keyword-based matching works less well on structured and semi-structured data [42] and a majority of data catalogues online rely on metadata descriptors to discover datasets instead of assessing their content. This metadata often varies in quality and availability, further limiting the user's ability what she needs [2]. Initiatives such as Open Data Monitor [47] and the European Data Portal [51], which build meta-portals with integrated access to multiple data repositories require manually curated mappings between the attributes that describe datasets on each site. Finally, contextual or personalised results are practically non-existent for data search.

In addition to these technology limitations, there is very little research examining data search from a user perspective. Most existing studies 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) and exploratory search (addressing more complex information needs) [1, 8, 21, 44, 55]. Other models distinguish between question answering [64] and document retrieval [29] 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 user's query. 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, the de facto standard in existing portals, cannot support this type of query. Closer to our area is the HCI literature on interaction with databases [14, 32, 56]. The focus there has been mostly on how users compose queries and the degree to which these queries can be translated into SQL or similar [14]. Our scenario is much more open in the range of data sources it targets - some of them could be relational databases, but many others will use other formats or will need to be discovered first.

Search log analysis is routinely used to understand the behaviour of users and evaluate search on the web and elsewhere [30, 37, 38, 69], using metrics such as time spent on a page and clickthrough streams, alongside specific activities such as tagging, printing and purchasing [25]. However, their findings are not straightforward to apply to new contexts such as data portals, which, our interviews showed, practitioners routinely used to search for data. Some work has been done in understanding structured queries against online databases such as the Linked Open Data Cloud [4, 48]. However, their aim was to study the popularity of certain parts of the database and query constructs, and not the interaction between users and structured data. In this paper, we take first steps in this direction. To supplement the in-depth interviews, we analyse the logs of a large open government data portal.

Making sense of structured data has mostly been studied in connection to information visualisation, which can help find patterns in data [26, 45]. The visual analytics system built by Stasko et al. [58] was tailored for sensemaking tasks of particular groups of data analysts. In their work on accessing government statistics, the system by Marchionini et al. [45] allowed users to explore data from different perspectives and understand relationships within the data via agile display mechanisms. As the web of data enters mainstream, this type of ideas need to be better integrated into productive environments and data work practice.

Both search and sensemaking are related to the broader area of *information seeking*. Previous research in this space has proposed a series of user-centred models that describe multistage and iterative processes of seeking information [7, 24, 40, 66, 68]. They are analysed and compared in [13, 20, 68]. Our study does not put forward a new model of information seeking. Instead, our aim is to investigate how people find, evaluate and explore structured sources of data that meet their information needs and to inform designers and improve the user experience. The information seeking scenario discussed in the introduction is real and increasingly important for various domains and professional roles, but it is not yet well understood and supported by tools and technologies. Existing models are a reasonable starting point, but to be truly useful, they would need to consider the specific search and sensemaking activities people carry out when working with structured data. For example, in Wilson's problem solving model [68], this would mean that at the 'problem resolution' stage the user decides that structured data is needed to solve the problem. We believe this decision affects how the rest of the information seeking process is carried out and the type of support the user requires. Another well-established model in this space is Kulthau's [40]. She defines six stages: task initiation, topic selection, pre-focus exploration, focus formation, information collection and search closure. In contrast to it, we are using the term 'exploration' as a means to understand whether a search result, in our case a dataset, matches an information need. This interpretation is more aligned with data science frameworks such as [50], which is centered around activities to collect relevant data, explore it to make sense of it, and build an analysis model to draw conclusions from it.

Further on, Fidel proposes an ecological approach to information seeking [24], 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 draws upon, including types and formats of sources, how reliable they are, how often they change, and whether one or more datasets need to be brought together to complete the user task. Current IR technology focuses on one of these dimensions, - the type of data - and are best suited for unstructured data, such as text or images [1, 17, 23, 45, 67]. Data-centric activities are likely to be dependent on the task people intend to use the data for [34]. Li et al. propose a faceted classification of information seeking tasks, in which one of the facets is the 'product of the task', which can be 'factual information' such as data, facts or similar [41]. We believe that a classification for just data-centric tasks could be useful as a preliminary step to improve the user experience in scenarios such as those discussed in the introduction.

METHODOLOGY

We used a mixed-methods approach, informed by [9], with a focus on qualitative methods in order to improve our understanding of how people work with data. To *expand* our findings [9] 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 the data.gov.uk 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 [36], of which our interviewees are a subset of; and (3) we used the quantitative insights only as a way to add more breadth to our enquiry, without making any claims about any direct links between the two samples.

In-depth interviews

User-system interactions are influenced by factors that are not easily observable or measurable [39]. For this reasons, and given the investigative nature of our 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.

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 UK's governmental Open Data User Group (around 15 people). For social media recruitment we used the ODIHQ account (at the time of the study over 31k followers, 5.3k impressions, 90 interactions, 20 retweets). For the interviews we chose only people who use data in their day-to-day work. We tried to cover various domains and professional backgrounds to gain a broad overview and avoid any unintended biases. The scoping survey helped us select the participants in a more targeted way (see additional material). It 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 respondents identified as relevant at this stage were contacted via email to arrange an interview.

Participants The sample consisted of $n = 20$ data workers (17 male and 3 female), based in the UK ($n = 16$), Germany ($n = 1$), USA ($n = 1$), France ($n = 1$) and globally ($n = 1$). Their roles, as reported by the participants, are shown in Table 1. They used both public (open) and proprietary data in different areas: environmental research, criminal intelligence, retail, social media, transport, education, geospatial information and smart cities. 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. The majority ($n = 17$) reported acquiring the skills to work with data on the job, from colleagues or self taught, by *doing it or experimenting with data (P5)*, though some people also mentioned formal education (in particular on core pre-requisites such as the fundamentals of statistics) and professional training.

Data collection and analysis We used 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. They were analysed using thematic analysis [54] using Nvivo, a qualitative data analysis package for coding. The coding was done by one researcher, but to enhance reliability two senior researchers checked the analysis for a sample of the data. 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: what people use data for; what information they need about the data to decide what to select and whether the search results were useful; how and where to search; and how to go about exploring and understanding datasets. For each of these themes we applied a second layer of coding, in which we used an inductive approach [12] to draw out further details. The resulting themes were then used to further categorise tasks and user needs, as we will see in the findings.

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 quantitative 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 1. Description of participants (P) with gender (G), their profession (Role) and sector they are working in (Sector)

Search logs

The interviews showed that finding data is a major issue for data practitioners. To understand these challenges better, we analysed the search logs of a governmental open data portal in the UK, data.gov.uk [18]. The portal has its own internal search capability and, as of September 2016, contains 39,983 dataset packages, which are collections of datasets (in different formats such as PDF, CSV, RDF etc.), grouped by topics. The portals SERP (search engine results page) shows a ranked list of dataset packages, which resembles the commonly used ten blue-links paradigm. When clicking on a result, the user is taken to a web page that displays dataset metadata, including a textual description and an option for downloading it.

We had access to the search logs generated via Google Analytics (GA) between 01 May 2015 and 30 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 [52], as GA clusters only identical queries). 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 sort 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 [60]. 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% [65]. 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 is 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; and (6) *Search depth* - the number of pages visitors viewed.

Ethics

The interview study was approved by our institution's Ethical Advisory Committees. 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 data.gov.uk for research purposes. No personal information was used for the search logs analysis.

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.

Data-centric tasks

Taxonomy When asked about their activities with data, participants reported a wide range of tasks, 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; and (2) *Goal-oriented tasks* - people think of data as a means to an end. Most participants reported to have engaged in both process- and goal-oriented tasks at some point. Process-oriented tasks include: building tools with data; integrating data in a database; defining predictive statistical models; producing data; publishing data; visualising it, etc. Goal-oriented ones include: seeking the answer to a question; comparing datasets or data items; finding patterns; allocating or managing resources in a data-informed way, etc. While the boundaries between the two categories are somewhat fluid, the primary difference between them lies in what we call the 'user information needs' - that is, the details people need to know about the data in order to interact with it effectively (see also section on information needs). For process-oriented 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 played a bigger role.

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) *time series analysis*; (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. 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 a *Time series analysis* ($n = 10$), data is ordered by time; the aim is to identify trends, or detect and predict events. [16] 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 it 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.

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. This is in line with the web of data idea discussed the introduction, which exploits links between datasets to allow people to find and browse structured data in the same way they engage with the web today. For example: 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.

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 and 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 the 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 different approaches to data quality (from assessing a given quality dimension automatically, to reviews and other annotations).

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.

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 prefering using fewer keywords to make sure that the search engine returns a broad range of (perhaps less accurate) results, which they then filter manually. We discuss below what happens after a query is issued to Google or similar search engines and the user chooses to click on the link of a data portal such as data.gov.uk.

Searching on portals

This section includes insights gained from interviews and analysis of 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 our methodology.

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 labeled *Search for data*. However, the format of the queries had other notable features, which hint at the qualities of data that count for open data portal users. 555 (0.55%) queries mentioned a 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 he 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 [62]. The participants in the interview 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

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, in particular 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 [27].

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

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 seem 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 2: (1) *relevance* to the task; (2) *usability*; and (3) *quality*. 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

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

Table 2. Information needs when selecting datasets

The first look We compiled a list of actions participants reported in Table 3. 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 / 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 schema / format
understanding semantics of the data
looking at documentation / metadata
checking the publication date and provenance
understand the purpose of original dataset

Table 3. What people do to get a first look at a new dataset

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:

(*P*17) 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.

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:

(*P*14) 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

(*P*5) 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

DISCUSSION

In this section we elaborate on the implications of our 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.

Framework for Human Structured-Data Interaction

Based on the models described in the background and on insights gained during our 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' [15] 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).

Similar to Yi et al. [70], 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 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*; *time series analysis*; *summarising*; *presenting*; and *exporting*. This taxonomy, and the broader framework

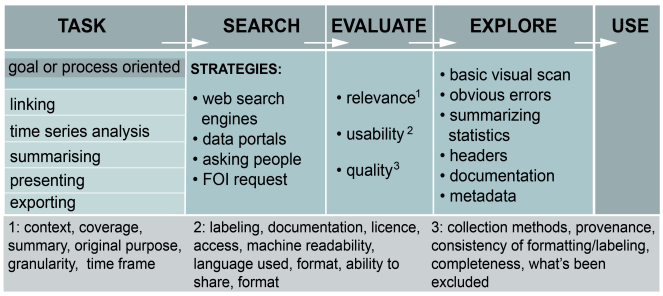

Figure 1. Framework for interacting with structured data

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 1 can be read from left to right, taking into account that this can be 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 a 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 pillar 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 pillar 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. 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 [40].

In the following section, we will 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, our 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.

Design recommendations

Data portals 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 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 [35]. 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. For instance, *linking* require 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* (*P*5). One approach for data sensemaking was proposed by Marchionini et al. in their work on relation browsers for statistical information [45]. 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 cleansing. 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.

Data publishers 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 the metadata and other 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.

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

This paper 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 activites. Thus our research can be seen as a step towards understanding what usability of data interaction tools means.

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 [15]. 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 [28].

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 researcher 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.

There is a large space for future studies extending the current understanding of how people interact with data. Additional research is needed to deepen our understanding of data-centric activites. This could include a direct observation of activities with data, developing an ontology of tasks and investigate the validity of the task taxonomies resulting from this study. An in-depth log analysis, in which a connection between queries and resulting downloads can be made, would support the understanding of user behaviour in dataset search. This can be a step towards modeling the information seeking behaviour of data users. Further studies are needed to develop better methods for automatic validation of structured data and creating additional metadata fields as proposed in our discussion. Advancing state-of-the-art interfaces, tailored for structured data, is not yet fully explored. Determining where in the search process the additional information we specified in our recommendations should be displayed needs to be established based on theoretical and empirical insight. We believe that these contributions can inform the design of data discovery tools, support exploratory assessment of datasets and make the exploration of structured data easier for a wider range of users.

ACKNOWLEDGEMENTS

This project is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 642795. We thank our participants and data.gov.uk for giving us insights and log data.

ADDITIONS

Interview schedule, scoping survey

REFERENCES

1. Michael J Albers. 2015. Human–Information Interaction with Complex Information for Decision-Making. In *Informatics*, Vol. 2. Multidisciplinary Digital Publishing Institute, 4–19.
2. Ulrich Atz. 2014. The tau of data: A new metric to assess the timeliness of data in catalogues. In *Conference for E-Democracy and Open Governement*. 257.
3. Jo Bates, Yu-Wei Lin, and Paula Goodale. 2016. Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society* 3, 2 (2016).
4. Wouter Beek, Laurens Rietveld, Hamid R Bazoobandi, Jan Wielemaker, and Stefan Schlobach. 2014. LOD laundromat: a uniform way of publishing other people's dirty data. In *International Semantic Web Conference*. Springer, 213–228.
5. Anant Bhardwaj, Amol Deshpande, Aaron J Elmore, David Karger, Sam Madden, Aditya Parameswaran, Harihar Subramanyam, Eugene Wu, and Rebecca Zhang. 2015. Collaborative data analytics with DataHub. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1916–1919.
6. Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (2009), 205–227.
7. Ann Blandford and Simon Attfield. 2010. Interacting with information. *Synthesis Lectures on Human-Centered Informatics* 3, 1 (2010), 1–99.
8. Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.
9. Alan Bryman. 2006. Integrating quantitative and qualitative research: how is it done? *Qualitative research* 6, 1 (2006), 97–113.
10. Burton-Taylor. 2015. Demand for financial market data and news. (24 March 2015). `https://burton-taylor.com/demand-for-financial-market-data-news-up-4-07-in-2014-highest-since-2011-2/`.
11. Michael J Cafarella, Alon Halevy, and Jayant Madhavan. 2011. Structured data on the web. *Commun. ACM* 54, 2 (2011), 72–79.
12. John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* (2013).
13. Donald Owen Case. 2012. *Looking for information: A survey of research on information seeking, needs and behavior*. Emerald Group Publishing.
14. Tiziana Catarci. 2000. What happened when database researchers met usability. *Information Systems* 25, 3 (2000), 177–212.
15. Gabriella Cattaneo, Mike Glennon, Rosanna Lifonti, Giorgio Micheletti, Alys Woodward, Marianne Kolding, Angela Vacca, Carla La Croce, and David Osimo. 2015. European Data Market SMART 2013/0063, D6 - First Interim Report. (October 2015). `https://idc-emea.app.box.com/s/k7xv0u3gl6xfvq1rl667xqmw69pzk790`.
16. Chris Chatfield. 2016. *The analysis of time series: an introduction*. CRC press.
17. Andy Crabtree and Richard Mortier. September 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*. Springer, 3–21.
18. Data.gov.uk. 2016. data.gov.uk, Opening up Government. (2016). `https://data.gov.uk/`
19. William Dilla, Diane J Janvrin, and Robyn Raschke. 2010. Interactive data visualization: New directions for accounting information systems research. *Journal of Information Systems* 24, 2 (2010), 1–37.
20. Jérôme Dinet, Aline Chevalier, and André Tricot. 2012. Information search activity: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology* 62, 2 (2012), 49–62.
21. Abdigani Diriye, Max L Wilson, Ann Blandford, and Anastasios Tombros. 2010. Revisiting exploratory search from the HCI perspective. *HCIR 2010* (2010), 99.
22. DMOX. 2016. Welcome to DMOZ! It's the Web, Organized. (2016). `https://www.dmoz.org/`
23. N Elmqvist. 2011. Embodied human-data interaction. In *ACM CHI Workshop "Embodied Interaction: Theory and Practice in HCI"*. 104–107.
24. Raya Fidel. 2012. *Human information interaction: an ecological approach to information behavior*. MIT Press.
25. Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
26. George W Furnas and Daniel M Russell. 2005. Making sense of sensemaking. In *CHI'05 extended abstracts on Human factors in computing systems*. ACM, 2115–2116.
27. GOV.UK. 2014. How to make a freedom of information (FOI) request. (November 2014). Retrieved September 05, 2016 from `https://www.gov.uk/make-a-freedom-of-information-request/the-freedom-of-information-act`.
28. Judith Green and Nicki Thorogood. 2013. *Qualitative methods for health research*. Sage.

29. Shweta Gupta, Sunita Yadav, and Rajesh Prasad. 2016. Document Retrieval using Efficient Indexing Techniques: A Review. *International Journal of Business Analytics (IJBAN)* 3, 4 (2016), 64–82.

30. Ido Guy, Sigalit Ur, Inbal Ronen, Sara Weber, and Tolga Oral. 2012. Best Faces Forward: A Large-scale Study of People Search in the Enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 1775–1784.

31. Hamed Haddadi, Richard Mortier, Derek McAuley, and Jon Crowcroft. 2013. Human-data interaction. *University of Cambridge* (2013).

32. Jozef Hvoreckỳ, Martin Drlík, and Michal Munk. 2010. Enhancing database querying skills by choosing a more appropriate interface. In *IEEE EDUCON 2010 Conference*. IEEE, 1897–1905.

33. Ross Ihaka and Robert Gentleman. 1996. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* 5, 3 (1996), 299–314.

34. Peter Emil Rerup Ingwersen. 1992. *Information Retrieval Interaction*. Taylor Graham.

35. Open Knowledge International. 2016. Good Tables. (2016). `http://goodtables.okfnlabs.org/`

36. Bernard J Jansen. 2006. Search log analysis: What it is, what's been done, how to do it. *Library & information science research* 28, 3 (2006), 407–432.

37. Daxin Jiang, Jian Pei, and Hang Li. 2013. Mining Search and Browse Logs for Web Search: A Survey. *ACM Trans. Intell. Syst. Technol.* 4, 4, Article 57 (October 2013), 57:1–57:37 pages.

38. Steve Jones, Sally Jo Cunningham, Rodger McNab, and Stefan Boddie. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3, 2 (2000), 152–169.

39. Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224.

40. Carol Collier Kuhlthau. 2004. *Seeking meaning: A process approach to library and information services*. Libraries Unltd Incorporated.

41. Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management* 44, 6 (2008), 1822–1837.

42. Jaime I Lopez-Veyna, Victor J Sosa-Sosa, and Ivan Lopez-Arevalo. 2012. KESOSD: keyword search over structured data. In *Proceedings of the Third International Workshop on Keyword Search on Structured Data*. ACM, 23–31.

43. McKinsey Global Institute: James Manyika, Michael Chui, Peter Groves, Diana Farrell, Steve Van Kuiken, and Elizabeth Almasi Doshi. 2013. Open data: Unlocking innovation and performance with liquid information. (2013). `http://www.mckinsey.com/business-functions/business-technology/our-nsights/open-data-unlocking-innovation-and-performance-with-liquid-information`.

44. Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

45. Gary Marchionini, Stephanie W Haas, Junliang Zhang, and Jonathan Elsas. 2005. Accessing government statistical information. *Computer* 38, 12 (2005), 52–61.

46. Viktor Mayer-Schönberger and Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

47. Open Data Monitor. 2016. Open Data Monitor. (2016). `http://project.opendatamonitor.eu/`

48. Mohamed Morsey, Jens Lehmann, Sören Auer, and Axel-Cyrille Ngonga Ngomo. 2011. DBpedia SPARQL benchmark–performance assessment with real queries on real data. In *International Semantic Web Conference*. Springer, 454–469.

49. Peter F Patel-Schneider. 2014. Analyzing schema. org. In *International Semantic Web Conference*. Springer, 261–276.

50. Hanspeter Pfister and Joe Blitzstein. 2015. cs109/2015, Lectures 01-Introduction. `https://github.com/cs109/2015/tree/master/Lectures`. (2015).

51. European Data Portal. 2016. (2016). `https://www.europeandataportal.eu/`

52. Jean-Baptiste Pressac. 2016. Open Refine Documentation. (2016). `https://github.com/OpenRefine/OpenRefine/wiki`

53. Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

54. Colin Robson and Kieran McCartan. 2016. *Real world research*. John Wiley & Sons.

55. Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM, 13–19.

56. Stefano Spaccapietra and Ramesh Jain. 2013. *Visual Database Systems 3: Visual Information Management*. Springer.

57. Andre Suslik Spritzer and Carla Maria Dal Sasso Freitas. 2008. Navigation and interaction in graph visualizations. *Revista de informática teórica e aplicada* 15, 1 (2008), 111–136.

58. John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

59. Thomas Steiner, Ruben Verborgh, Raphaël Troncy, Joaquim Gabarro, and Rik Van de Walle. 2012. Adding realtime coverage to the Google knowledge graph. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914*. 65–68.

60. Greg Sterling. 2015. It's Official: Google Says More Searches Now On Mobile Than On Desktop Company officially confirms what many have been anticipating for years. (May 2015). `http://searchengineland.com/its-official-google-says-more-searches-now-on-mobile-than-on-desktop-220369`.

61. Monica Swamiraj and Luanne Freund. 2015. Facilitating the discovery of open government datasets through an exploratory data search interface. (2015).

62. Mona Taghavi, Ahmed Patel, Nikita Schmidt, Christopher Wills, and Yiqi Tew. 2012. An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces* 34, 1 (2012), 162–170.

63. Barbara Ubaldi. 2013. Open Government Data. (2013).

64. Ellen M Voorhees and others. 1999. The TREC-8 Question Answering Track Report.. In *Trec*, Vol. 99. 77–82.

65. W3Schools.com. 2016. Browser Statistics. `http://www.w3schools.com/browsers/default.asp`. (2016).

66. Ryen W White. 2016. *Interactions with search systems*. Cambridge University Press.

67. Max L Wilson, Bill Kules, Ben Shneiderman, and others. 2010. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science* 2, 1 (2010), 1–97.

68. Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation* 55, 3 (1999), 249–270.

69. Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. 2011. Query log analysis of an electronic health record search engine. In *AMIA annual symposium proceedings*, Vol. 2011. 915–924.

70. Ji Soo Yi, Youn ah Kang, John Stasko, and Julie Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1224–1231.