DOI:
10.1016/j.ijhcs.2019.10.004

Document Version
Publisher's PDF, also known as Version of record

International Journal of Human-Computer Studies 135 (2020) 102367

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Everything you always wanted to know about a dataset: Studies in data summarisation

Laura Koesten[*,a,b], Elena Simperl[a], Tom Blount[a], Emilia Kacprzak[a,b], Jeni Tennison[b]

[a] University of Southampton, UK
[b] The Open Data Institute, UK

ARTICLE INFO

Keywords:
Data summarisation
Data search
Data sensemaking
Human data interaction

ABSTRACT

Summarising data as text helps people make sense of it. It also improves data discovery, as search algorithms can match this text against keyword queries. In this paper, we explore the characteristics of text summaries of data in order to understand how meaningful summaries look like. We present two complementary studies: a data-search diary study with 69 students, which offers insight into the information needs of people searching for data; and a summarisation study, with a lab and a crowdsourcing component with overall 80 data-literate 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. The results helped us design a template to create more meaningful textual representations of data, alongside guidelines for improving data-search experience overall.

1. Introduction

As digital technology has advanced over the past years, there has been a huge surge in the availability of data. 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 and professional roles (Manyika et al., 2013). With the rise of data science, millions of datasets have been published, sometimes under an open license, in institutional repositories, online marketplaces, and on social networks, in sectors from science and finance, to marketing and government (Gregory et al., 2018; Verhulst and Young, 2016).[1] People use such data to improve services, design public policies, generate business value, advance science, and make more informed decisions (Verhulst and Young, 2016). Recently, Google released an initiative to use its schema.org markup language[2] to index datasets alongside text documents, images and products in a vertical search engine for datasets (Noy et al., 2019).

Previous research has shown that, despite increased availability, this data cannot be easily reused, as people still experience many difficulties in finding, accessing and assessing it (Koesten et al., 2017). In Koesten et al. (2017) we discussed three major aspects that matter to data practitioners when selecting a dataset to work with: *relevance, usability* and *quality*. 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. While this applies to all information seeking activities this process demonstrates unique interaction characteristics, which have been subject to several human data interaction studies (Boukhelifa et al., 2017; Gregory et al., 2017; Kern and Mathiak, 2015; Koesten et al., 2017).

Data search often starts on a data portal with an interface as depicted in Fig. 1. 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. Fig. 2 shows an example. Metadata is often limited and might not provide enough content to decide whether a dataset is useful for a task (Noy et al., 2019). 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). It helps 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 also improves data discovery, as search algorithms can

* The corresponding author at: University of Southampton, University Rd, Southampton SO17 1BJ, UK
E-mail address: L.M.Koesten@soton.ac.uk (L. Koesten).

[1] In this paper, 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 artifacts returned by a search algorithm in response to a user query.

[2] http://schema.org/Dataset

https://doi.org/10.1016/j.ijhcs.2019.10.004
Received 24 October 2018; Received in revised form 12 August 2019; Accepted 14 October 2019
Available online 14 October 2019

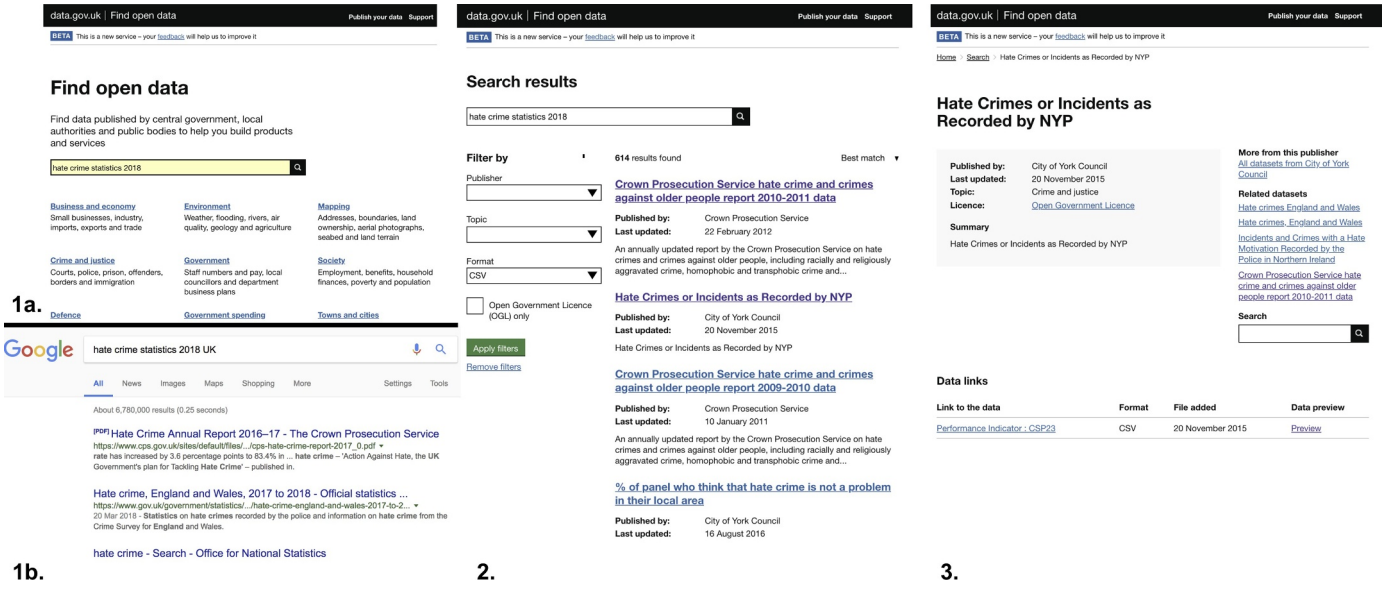

Fig. 1. Data search on a data portal. 1: search box of (a) a data portal, (b) Google; 2: SERP of a data portal; 3: dataset preview page.

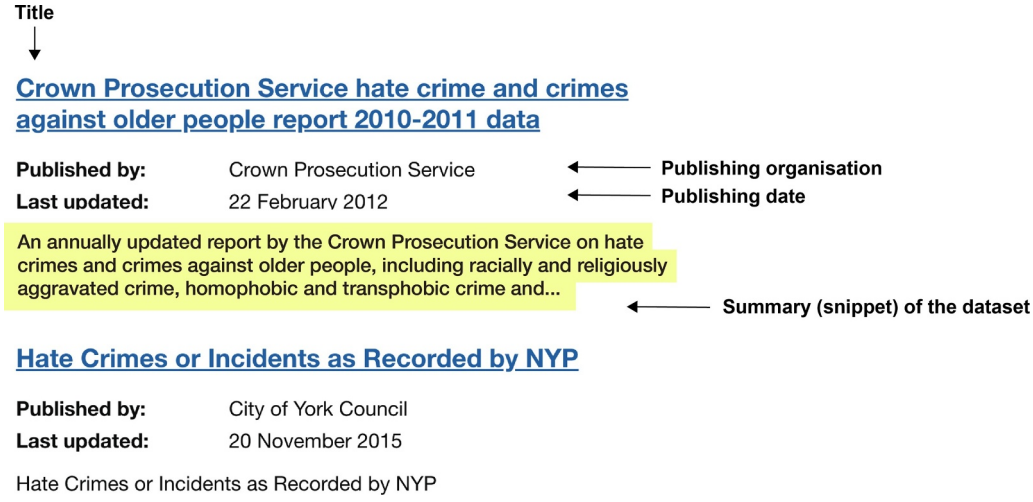

Fig. 2. Example of data search results on one of the most popular open government data portals. Next to title, publisher, domain and format, we see a textual description of the dataset.

match the text against keyword queries (Thomas et al., 2015).

In general, a good summary must be able to represent the core idea, and effectively convey the meaning of the source (Zhuge, 2015). In this paper, 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 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,[3] schema.org,[4] and SharePSI)[5]). 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 (Koesten et al., 2017; Neumaier et al., 2016).

We have undertaken two complementary studies in dataset selection and summarisation. Both studies build on previous research of ours from Koesten et al. (2017). In that work, we have reported on the results of a series of interviews with 20 data practitioners, which have helped us define a general *framework for dataset search*, built around the three themes mentioned earlier: *relevance, usability* and *quality*. The first study presented in this paper takes the next step: 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 qualitatively starting from the framework from Koesten et al. (2017). This resulted in a *list of dataset selection attributes*. In the second study, we carried out lab 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*

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

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.

Our 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 (Marienfeld et al., 2013; Reiche and Höfig, 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. The findings advance our current understanding of dataset summaries that are tailored to the needs of general purpose data consumers.

Summary of research questions and contributions.

This paper explores the following research questions:

RQ1 Study 1: What data attributes do people consider when determining the relevance, usability and quality of a dataset?

RQ2 Study 2: What data attributes do people choose to mention when summarising a dataset to others?

Our paper contributes to the emerging field of human 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. The work expands our understanding of how people interact with and communicate about data, and can further inform the design of data publishing platforms, metadata standards, and algorithms for natural language generation and snippet generation for dataset retrieval.

2. Motivating scenario

Before describing the two studies and their context, we will expand on the data search example introduced earlier to give an overview of the state of the art, and of the challenges that motivate our work. 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. An overview of the process is depicted in Fig. 1.

You enter a search query such as *"hate crime statistics2018"* in the search box of a UK data portal (see Fig. 1 (step 1)). The search results may look like in Fig. 1 (step 2). On the left hand side, you can find a classification of the results based on metadata attributes such as license and format, as well as the number of datasets that fall into each category. 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, main domain and available formats. In most cases, the dataset is accompanied by a short text summary, as can be seen in Fig. 2. The results can be sorted according to different criteria, including relevance. You select one of the results to explore further, based on what else is on the list and on the information displayed in the snippet. This commonly takes you to a new page (see Fig. 1 (step 3)), where you can also download the dataset to examine it on your own computer.

Data search results on general-purpose web search engine have a similar look and feel although the hits are a mix of datasets and other types of sources, as can be seen in Fig. 1 (step 1 - *Google*). 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.

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.

3. Related work

In this section we outline current practices for selecting and describing data on the web and related work on summarisation. We draw on metadata standards and community guidelines for data publishing; literature in the fields of data search and sensemaking, as well as natural language generation; and related HCI and HDI (human data interaction) studies.

3.1. Selecting and making sense of datasets

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). We also 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., 2010; Yu et al., 2005), people (Weerkamp et al., 2011) or products (Gysel et al., 2016; Rowley, 2000).

Previous works have highlighted the distinct characteristics of dataset compared to document retrieval. Data requires context to create meaning and make sense of it. While this applies to information seeking in general (Klein et al., 2006; Marchionini and White, 2007; Russell et al., 1993), choosing a dataset greatly depends on the information provided alongside it. Making sense of structured data has mostly been studied in connection to information visualisation, which can help to see patterns in data (e.g. (ah Kang and Stasko, 2012; Furnas and Russell, 2005)) and in contexts of exploratory data analysis (Marchionini et al., 2005; Pirolli and Card, 2005). However, visualisations are often not available; especially when searching data on the web, users mostly rely on metadata. Only few studies have looked at sensemaking with structured data specifically in information seeking scenarios. For example, Wynholds et al. (2012) show that, in the context of digital libraries, seeking and using documents and data for research purposes are different in terms of information needs, processes and required level of support. In user studies with social scientists, Kern and Mathiak (2015) found that the quantity and quality of metadata is more critical in dataset search than in literature search, where convenience prevails. Empirical social scientists in that study were willing to put more effort into the retrieval of research data than in literature retrieval. In our prior mixed-methods study mentioned earlier (Koesten et al., 2017) 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, missing values, the granularity of the captured information, as well as the ability to understand the schema used to organise a dataset and to process it in the form it was published.

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.

As a starting point for the studies presented in this paper, we thus made 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. In addition, across the two studies reported here we were able to identify common themes and attributes of summaries that match the information needs of potential readers.

For the remainder of this section, we will elaborate on existing practices and techniques to create text about structured data.

3.2. Practices around text summaries for datasets

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 returned documents (Bando et al., 2010). Snippets adjust their content based on the user query to make selection more effective, but these capabilities have evolved over time (Baeza-Yates and Ribeiro-Neto, 2011; 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. To name some of the reasons: We lack a taxonomy of information seeking tasks for data that such approaches could be modelled on. As we describe in Section 3.4 we also do not currently have the technical capabilities to provide useful snippets automatically and dataset summaries are often manually written by data publishers. Based on prior research we see that many information seeking tasks for data are currently exploratory (Gregory et al., 2017; Koesten et al., 2017), and queries for datasets are currently not very expressive (Kacprzak et al., 2019), which is why we focus this work on static general purpose dataset summaries. However, we believe there is a large space for future research on creating query-biased, personalised dataset summaries for specific information seeking scenarios and their respective tasks.

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, such as the W3C's Data on the Web Best Practices[6] 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, a vocabulary to describe datasets in catalogues, 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)*.

On collaborative platforms that are used for sharing and working with data, such as Kaggle[7] or Github[8], we also see users describing datasets in textual format. Dataset summaries take different shapes and forms, but there is a lack of clear and consistent guidelines to summarise datasets for the purpose of reuse. For instance on Kaggle instructions are formulated as:

> The description should explain what the dataset is about in long-form text. A great description is extremely useful to Kaggle community members looking to get started with your data.

Based on the general lack of guidance, we focus the current paper on understanding the composition of meaningful summaries rather than exploring whether people find the resulting summaries useful, which we believe is a necessary next step. 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, in this paper we refer to textual descriptions of datasets as *summaries*.

3.3. 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 paper we consider the latter.

Our research, as much of the related work in human data interaction, is based on the assumption that, in order to offer the best user experience, we cannot simply reuse or re-purpose principles and models that have been proposed for less structured sources of information (Marchionini et al., 2005; Wilson et al., 2010). 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 cognitive work in order to contextualise it in relation to other information, and context to make it meaningful (Albers, 2015); arguably more than when

[6] https://www.w3.org/TR/dwbp/

[7] https://www.kaggle.com/

[8] https://github.com/

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). In our studies we follow a similar view - we compare themes and datasets attributes derived from the summaries created in our lab and crowdsourcing experiments with analogue attributes prevalent to the task context in which such summaries would be most likely used, which is dataset selection and sensemaking.

3.4. 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; Gupta et al., 2019; Mosa et al., 2019; Reiter and Dale, 1997). 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 (Boukhelifa et al., 2017; Gregory et al., 2017; Kern and Mathiak, 2015; Koesten 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., 2016). 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.

Our work helps define commonly used strategies for abstracting the content of a dataset in a summarisation context; as such, it can inform the design of abstractive approaches by pointing to types of information that an algorithm should aim to include in a summary to make it more useful for its readers.

Another closely related area of research is data profiling, which 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 papers 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). There are significant efforts in the database community to develop automatic techniques to extract insights from large amounts of data (e.g. (Liu and Jagadish, 2009; Saint-Paul et al., 2005; Tang et al., 2017)) which could be interesting to explore in the context of data-to-text generation. However, we believe that an in-depth understanding of what a meaningful summary should contain is a necessary first step.

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 is not available - most studies use criteria such as accuracy, readability, coverage,. 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. Study design

To answer **RQ1** we conducted a thematic analysis of data-search diaries which resulted in a *list of data selection attributes*. For **RQ2** we applied a mixed-methods approach (Bryman, 2006) combining a task-based lab experiment and a crowdsourcing experiment, in which participants summarised datasets in a writing task. This led to a *list of data summary attributes*. Fig. 3 gives an overview of the research carried out.

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.

We report on the two studies, in Section 5 and Section 6

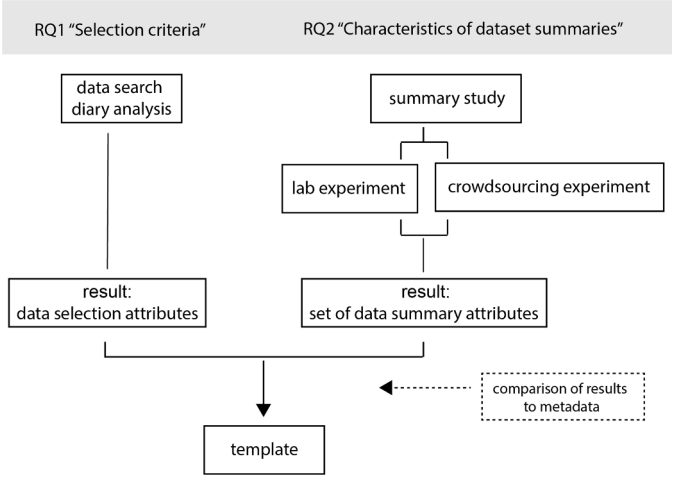

Fig. 3. Overview of research methods and outcomes.

respectively.

The first study (Study 1) used 269 data-search diaries by 69 students to understand what data attributes are relevant in dataset selection, and, hence, how data summaries should look like to be useful to their readers.

The second study (Study 2) analysed what attributes people choose to describe previously unknown datasets, based on a total of 360 data summaries for 25 datasets created by 80 participants. We compare both sets of attributes, discuss differences in summary creation across datasets and summary authors, and highlight common themes and characteristics.

5. Study 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.

5.1. Methods: Data-search diaries

Process.

We conducted a thematic analysis of 269 data-search diaries that were completed by 69 students[9] 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[10]). 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 yea

Example 2: Details on the molecular composition of Titan's atmosphere.

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 of the authors 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 inductively coded attributes emerging in each of these areas (Thomas, 2006). 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, 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. Participants consented to the data being used for research when joining the course. No personal data was analysed or reported.

5.2. 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 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 Koesten et al. (2017): *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 1, as we see these as the main contribution of the diary analysis. 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.

5.2.1. 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

[9] 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 ($n = 1$)

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

Table 1
Findings on selection criteria for datasets, based on thematic analysis of the data-search diaries. Prevalence can be seen as indicative for importance, but needs further validation.

THEME	ATTRIBUTE	%
Relevance	Scope (e.g., topical, geographical, temporal)	36
	Granularity (e.g., number of traffic incidents per hour, day, week) Comparability	19
	Context (e.g., original purpose of the data)	14
	Documentation (e.g., understandability of variables, samples)	11
		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

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 (Erete et al., 2016; Koesten et al., 2017). 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.

5.2.2. 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 basic statistics in relevance. This is probably due to the fact that students were mindfull 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.

5.2.3. Quality

Participants mentioned unique attributes such as *provenance* – in a broad sense of the term – that would allow judgements around the authoritativeness and trustworthiness of the publisher and so the 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 (Simmhan et al., 2008; Stamatogiannakis et al., 2014).

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 an earlier study of ours (Koesten et al., 2017), which 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 8 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.

6. Study 2: Dataset summaries

The second study explored the attributes that people chose to describe a dataset, both in a lab-based experiment with 30 participants and in a crowdsourcing experiment with 50 crowdworkers. This provided us with insights into the features of a dataset people consider important to include in a summary.

6.1. DATASETS: The Set − 5 and Set − 20 corpora

We used openly published datasets available as CSV files from three different news sources: *FiveThirtyEight*[11], *The Guardian*[12], and *Buzzfeed*[13]. 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 *D*1 to *D*5, which are described in more detail in Table 2. 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 *E*1 to *E*20 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 the supplementary material connected to this paper). All datasets are available on GitHub[14].

[11] http://fivethirtyeight.com/
[12] https://www.theguardian.com/
[13] https://www.buzzfeed.com/news

Table 2
Datasets in *Set*-5.

Dataset	Topic	Example characteristics
*D*1	Earthquakes	>10k rows, 10 columns, dates inconsistently formatted, ambiguous headers, granular geospatial information
*D*2	Marvel comic characters	>16.000 rows, 13 columns, no geospatial information, many string values, limited value ranges, missing values, yearly and monthly values
*D*3	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
*D*4	Refugees	192 rows, 17 columns, mostly text values, formatting inconsistencies, ambiguous headers, identifiers, geospatial information (continent/region/country), no temporal information
*D*5	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

6.1.1. Labbased experimen

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. Each participant was asked to summarise the datasets from *Set*-5 as explained below. 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 the 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 some 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 one of 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.[15] Our sample consisted of n = 30 participants (19 male and 11 female), all based in the UK at the time of the study. Two 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 2017.

Process.

Respondents to the study call were contacted via email to receive an information sheet. We arranged a time for the experiment at the author's organisation 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, 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 Fig. 4).

To get a sense of the surface form of the summaries, we counted how many of them used full sentences, bullet points or a mixture of the two. 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 6.2.1). 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 Fig. 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 6. 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.

Ethics.

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

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

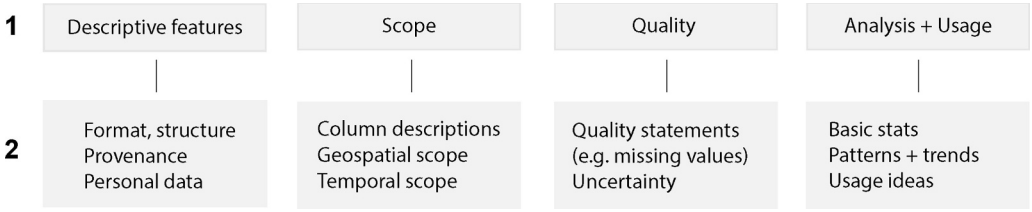

Fig. 4. Information types (1) and emerging **summary attributes (2)** from the thematic analysis of the lab summaries, reflecting our coding process.

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

6.1.2. Crowdsourcing experiment

Following the lab experiment, we undertook a data summaries experiment on the crowdsourcing platform CrowdFlower (now Figure Eight)[16]. 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[17].

Process.

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 (Fig. 5).

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

Overview

In this task, you will be required to write 5 short (50-100 word) descriptions of 5 datasets. Datasets are spreadsheets or tables containing data about a topic (which can be numbers or words).

The descriptions should summarise the content of the dataset they describe. You will get a link to each dataset which you can look at while you are writing the description. **We ask you to describe the dataset in a way that other people, who cannot see the data, can understand what it is about.**

A good quality description summarises the content of the dataset in an understandable way. You will see examples of different good quality descriptions before the task. You will not be able to submit your description if the length is outside the range of 50-100 words. You will be asked qualification questions that require basic reading, reasoning and data skills. You'll have the opportunity to comment on our task at the end.

The datasets vary a lot and so we would expect the descriptions of smaller datasets to be closer to the minimum of 50 words, whereas descriptions of larger datasets would be expected to be comparatively longer. (Smaller datasets have about 10 columns and all rows are visible without scrolling down)

Steps

- Answer the 4 demographic questions
- Answer the qualification questions
- Click on the link provided to view the first dataset (if you can't view the dataset, try using a different browser)
- Take a look at the WHOLE dataset (scroll until you know how many rows and columns it has, look at all headers before you start writing your description)
- Provide an accurate description of the dataset in the box below. This description has to be between 50 and 100 words long.
- Repeat this for all 5 datasets
- You can optionally leave any comments you have regarding the task
- Submit the task

Fig. 5. CrowdFlower task instructions in the crowdsourcing experiment.

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

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

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 on the GitHub repository created for this work[18].

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 6.1.1). 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.

Ethics.

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

6.2. 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: (i) summary *form* and *length*; (ii) *information types*; and (iii) detailed *summary attributes*.

6.2.1. 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: the majority (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 on their potential information gain (Maxwell et al., 2017), as well as in the context of the summary template we propose in Section 8.

6.2.2. 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 (detailed in Fig. 4). 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 110,172 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 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

Table 3
Percentages of information types per dataset in *Set*-5, based on 150 lab summaries.

	Total	*D*1	*D*2	*D*3	*D*4	*D*5
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

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

150 lab summaries analysed and in Section 9 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.

6.2.3. 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:

- Subtitle
- Headers
- Geographical scope

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. Quotes in these sections are from the lab-based experiment. Across the 360 summaries created in the two experiments we have identified the following attributes (Table 4):

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, lon, magnitude and location name.

Percentages of these attributes over all summaries can be seen in Table 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 *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 just under half of the *Set*-20 summaries.

Table 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

Table 4
Most frequent summary attributes, based on 360 summaries of datasets from *Set*-5 and *Set*-20.

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 5
Comparison of percentage of summaries created in the lab (*L*) and via crowdsourcing (*C*) that mention summary respective 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 5.2.2).

	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 6
Percentage of lab summaries containing respective attributes, per dataset from *Set*-5 (n = 150). Darker fields have higher percentages.

	*D*1	*D*2	*D*3	*D*4	*D*5
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 7
Percentage of crowdsourced summaries from *Set*-5 (n = 120) containing respective attributes, per dataset. Darker fields have higher percentages.

	*D*1	*D*2	*D*3	*D*4	*D*5
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

consistently in more than half of the cases (55%). *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.

Table 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 setting: *subtitle, format* and *headers* remain important, but *basic statistics* are more consistently mentioned than in the other experiment.

At the same time, the crowd focused on *groupings* of headers as well, much more so than the data practitioners who participated in the lab experiment - overall 70% of the crowd-generated summaries of *Set*-5 mentioned this attribute, compared to just under half of the lab summaries (see Table 5); the scores for the individual datasets varied more in the lab than in the summaries created by the crowd.

The 20 datasets from *Set*-20 the summaries created by the crowd reinforce some of these trends. In *Set*-20, *subtitle, geographical* and *temporal* scope and *headers* are mentioned in the majority of summaries, just as with the summaries of *Set*-5. *Groupings* seem to be popular among crowd workers across all datasets (for instance 69% of the *Set*-20 summaries mention them, as Table 5 shows) but less frequent in just under half of the lab summaries. By contrast, *Set*-20 summaries showed a lower prevalence to the crowd-produced *Set*-5 summaries for the attributes: *format* and *basic statistics* and to some extent *geographical* scope.

The popularity of *groupings* aside, a second surprising result was the popularity of *patterns/trends* – this attribute was mentioned in less than a third of the lab and the crowd summaries for the same *Set*-5 datasets, but in more than half of the *Set*-20 summaries (see Table 5). 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).

Synopsis:

- The top-5 attributes from the lab experiment *Set*-5 are: subtitle, geographical scope, headers, uncertainty and format
- The top-5 attributes from the crowdsourcing experiment *Set*-5 are: subtitle, headers, basic statistics, geographical scope and groupings.

We compare the distributions of attributes in the two experiments shown in Table 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 6), geospatial attributes, as well as provenance appear to have the highest dependency on the dataset. *D*5 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. *D*2 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. *D*2 was set in a fictional universe and may have therefore not 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).

Synopsis:

- The top-5 attributes from the crowdsourcing experiment *Set*-5 are: subtitle, headers, basic statistics, geographical scope and groupings.
- The top-5 attributes from the crowdsourcing experiment *Set*-20 are: subtitle, headers, groupings, geographical scope, temporal scope.

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 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, *E*10 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, *E*11 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 that get highlighted in a summary. For example in the example of a dataset (*E*10) 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 - e.g. *E*4 had mainly numerical values describing the elderly population worldwide - or in connection to uncertainty about the provenance - e.g. *E*12 was about US weather data, but did not make any reference to the source of the data.

6.2.4. 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 a fifth of the participants and mostly mentioned in connection to *D*3 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 *D*3), 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 (just under half of the lab summaries and 35% of the crowdsourced summaries), or to the temporal scope of the dataset (33% of the lab summaries, 17% *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 15,000 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 *D*2, 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 *D*1 and *D*3 (87%) and less often in connection to *D*2 and *D*4 (< 40%). *D*3 had three date columns separating day, month and year from each other which might prompt including this information in the summaries. *D*1 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. *D*1 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 6.

*D*4 on the other hand did not contain temporal information explicitly which explains the significantly lower percentage. This was reflected in the crowdsourced summaries for *D*4. *D*2 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.

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.

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?

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. We present an initial prototype of a template based summary writing tool to exemplify how it could be used[19].

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.

8. Discussion

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.

8.1. 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 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 Fig. 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 suggest 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 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 (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.

In 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

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

Table 8
Comparison of summary attributes to data-search diary and metadata standards. Summary = results from this study; Diary = Analysis of selection criteria in a data-search diary; Schema (S) = http://schema.org/Dataset; DCAT (D) = https://www.w3.org/TR/vocab-dcat/ (as per 05/2019) - Attributes "description" excluded.

Category	Summaries	Diary	Schema and DCAT
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
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 coverage, release date, update and modification date, frequency of publishing
Quality	*quality dimensions:*		D: refers to Data Quality Vocabulary (focus on metrics)
	completeness consistency in formatting understandability (headers, acronyms, abbreviations) representativeness, coverage	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

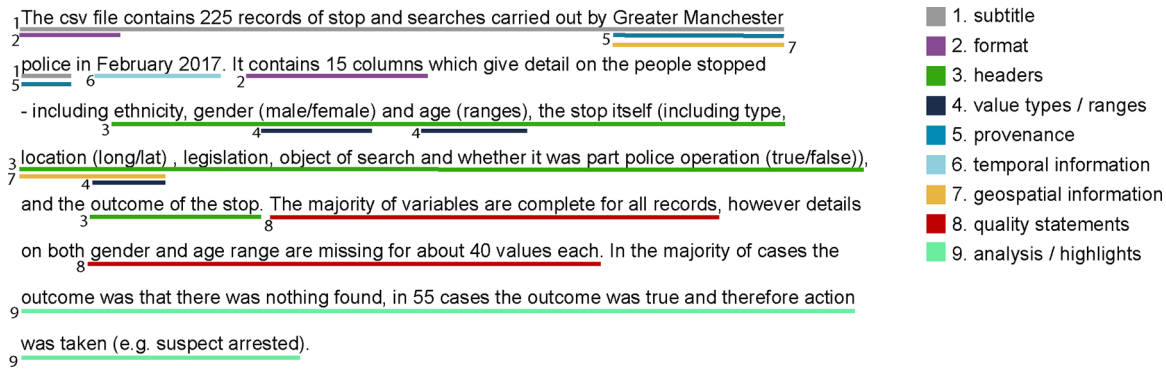

Fig. 6. Example of an annotated dataset summary.

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.

8.2. Comparison to metadata standards and data search diaries

Table 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 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.

8.3. Making better summaries

Prior work has 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[20]. 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[21]. 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 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); Koesten 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. 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.

8.4. 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 relevance 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

[20] http://schema.org/Dataset

[21] https://www.w3.org/TR/dwbp/

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.

9. 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 (Koesten et al., 2017). 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 data presented in a spreadsheet; 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 or visualisation (such as a graph presentation) might vary. Future work could investigate how the composition of summaries changes for different presentations of data.

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.

10. Conclusion and future work

With the overabundance of structured data, techniques to represent it in a condensed form are becoming more important. Text summaries serve this function and they have the potential to make data on the web more user friendly and accessible. We 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.

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.

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.

This work could be extended in a number of directions. We aim to evaluate the perceived usefulness of summaries created according to the proposed template as a next step. Follow-up studies could include crowdworkers iterating on the summaries created by the template, which has been proven useful for image descriptions and text shortening (Bernstein et al., 2015; Little et al., 2010). Additional work could be carried out on refining a semi-automatic approach to generating summaries, using the template by prompting crowd workers 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. 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. Similarly, approaches to generate query-biased summaries, such as those shown by Au et al. (2016) to generate task dependent summaries, are an interesting area for further research that could significantly improve user experience in dataset search.

Declaration of Competing Interest

The authors declare that they do not have any financial or non-financial conflict of interests.

Acknowledgements

This project was supported by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642795,and No 780247 and by the EPSRC Data Stories project EP/P025676/1. We thank our participants for taking part in this study.

References

ah Kang, Y., 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. https://doi.org/10.1109/TVCG.2012.224.

Albers, M.J., 2015. Human–information interaction with complex information for decision–making. Informatics 2 (2), 4–19. https://doi.org/10.3390/informatics2020004. URL http://www.mdpi.com/2227-9709/2/2/4

Au, V., Thomas, P., Jayasinghe, G.K., 2016. Query-biased summaries for tabular data. Proceedings of the 21st Australasian Document Computing Symposium, ADCS 2016, Caulfield, VIC, Australia, December 5–7, 2016. pp. 69–72. URL http://dl.acm.org/citation.cfm?id=3015027

Baeza-Yates, R.A., Ribeiro-Neto, B.A., 2011. Modern Information Retrieval - the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England. URL http://www.mir2ed.org/

Balatsoukas, P., Morris, A., O'Brien, A., 2009. An evaluation framework of user interaction with metadata surrogates. J. Inf. Sci. 35 (3), 321–339. https://doi.org/10.1177/0165551508099090.

Bando, L.L., Scholer, F., Turpin, A., 2010. Constructing query-biased summaries: A comparison of human and system generated snippets. Proceedings of the Third Symposium on Information Interaction in Context. ACM, New York, NY, USA, pp. 195–204. https://doi.org/10.1145/1840784.1840813.

Bargmeyer, B.E., Gillman, D.W., 2000. Metadata standards and metadata registries: An overview. 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. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<149::AID-ASI5>3.0.CO;2-J.

Bechchi, M., Raschia, G., Mouaddib, N., 2007. Merging distributed database summaries. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. ACM, New York, NY, USA, pp. 419–428. https://doi.org/10.1145/1321440.1321500.

Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K., 2015. Soylent: a word processor with a crowd inside. Commun. ACM 58 (8), 85–94. https://doi.org/10.1145/2791285.

Borromeo, R.M., Alsaysneh, M., Amer-Yahia, S., Leroy, V., 2017. Crowdsourcing strategies for text creation tasks. EDBT. pp. 450–453.

Boukhelifa, N., Perrin, M.-E., Huron, S., Eagan, J., 2017. How data workers cope with uncertainty: A task characterisation study. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3645–3656. https://doi.org/10.1145/3025453.3025738.

Boydell, O., Smyth, B., 2007. From social bookmarking to social summarization: an experiment in community-based summary generation. Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, Honolulu, Hawaii, USA, January 28–31, 2007. pp. 42–51. https://doi.org/10.1145/1216295.1216311.

Bryman, A., 2006. Integrating quantitative and qualitative research: how is it done? Qual. Res. 6 (1), 97–113. https://doi.org/10.1177/1468794106058877.

Ceolin, D., Groth, P.T., Maccatrozzo, V., Fokkink, W., van Hage, W.R., Nottamkandath, A., 2016. Combining user reputation and provenance analysis for trust assessment. J. Data Inf. Qual. 7 (1–2), 6:1–6:28. https://doi.org/10.1145/2818382.

Convertino, G., Echenique, A., 2017. Self-service data preparation and analysis by business users: New needs, skills, and tools. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1075–1083. https://doi.org/10.1145/3027063.3053359.

Cormode, G., 2015. Compact summaries over large datasets. Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. ACM, New York, NY, USA, pp. 157–158. https://doi.org/10.1145/2745754.2745781.

Cutrell, E., Guan, Z., 2007. What are you looking for?: an eye-tracking study of information usage in web search. Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007. pp. 407–416. https://doi.org/10.1145/1240624.1240690.

Erete, S., Ryou, E., Smith, G., Fassett, K.M., Duda, S., 2016. Storytelling with data: Examining the use of data by non-profit organizations. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. ACM, New York, NY, USA, pp. 1273–1283. https://doi.org/10.1145/2818048.2820068.

Ermilov, I., Ngomo, A.N., 2016. TAIPAN: automatic property mapping for tabular data. Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19–23, 2016, Proceedings. pp. 163–179. https://doi.org/10.1007/978-3-319-49004-5_11.

Ferreira, R., de Souza Cabral, L., Lins, R.D., e Silva, G.P., Freitas, F., Cavalcanti, G.D.C., Lima, R., Simske, S.J., Favaro, L., 2013. Assessing sentence scoring techniques for extractive text summarization. Expert Syst. Appl. 40 (14), 5755–5764. https://doi.org/10.1016/j.eswa.2013.04.023.

Fetahu, B., Dietze, S., Nunes, B.P., Casanova, M.A., Taibi, D., Nejdl, W., 2014. A scalable approach for efficiently generating structured dataset topic profiles. The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25–29, 2014. Proceedings. pp. 519–534. https://doi.org/10.1007/978-3-319-07443-6_35.

Furnas, G.W., Russell, D.M., 2005. Making sense of sensemaking. CHI'05 extended abstracts on Human factors in computing systems. ACM, pp. 2115–2116.

Gambhir, M., Gupta, V., 2017. Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. 47 (1), 1–66. https://doi.org/10.1007/s10462-016-9475-9.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., 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. https://doi.org/10.3233/AIC-2009-0453.

Gkatzia, D., 2016. Content selection in data-to-text systems: a survey. CoRR abs/1610.08375.

Gkatzia, D., Lemon, O., Rieser, V., 2016. Natural language generation enhances human decision-making with uncertain information. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 2: Short Papers. URL http://aclweb.org/anthology/P/P16/P16-2043.pdf

Greenberg, J., 2010. Metadata and digital information. Encyclopedia of Library and Information Sciences, 3610–3623.

Gregory, K., Cousijn, H., Groth, P.T., Scharnhorst, A., Wyatt, S., 2018. Understanding data retrieval practices: a social informatics perspective. CoRR abs/1801.04971.

Gregory, K., Groth, P.T., Cousijn, H., Scharnhorst, A., Wyatt, S., 2017. Searching data: a review of observational data retrieval practices. CoRR abs/1707.06937.

Gupta, V., Bansal, N., Sharma, A., 2019. Text summarization for big data: A comprehensive survey. In: Bhattacharyya, S., Hassanien, A.E., Gupta, D., Khanna, A., Pan, I. (Eds.), International Conference on Innovative Computing and Communications. Springer Singapore, Singapore, pp. 503–516.

Gysel, C.V., de Rijke, M., Kanoulas, E., 2016. Learning latent vector spaces for product search. Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24–28, 2016. pp. 165–174. https://doi.org/10.1145/2983323.2983702.

He, J., Duboue, P., Nie, J., 2012. Bridging the gap between intrinsic and perceived relevance in snippet generation. COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8–15 December 2012, Mumbai, India. pp. 1129–1146. URL http://aclweb.org/anthology/C/C12/C12-1069.pdf

Hidi, S., Anderson, V., 1986. Producing written summaries: task demands, cognitive operations, and implications for instruction. Rev. Edu. Res. 56 (4), 473–493. https://doi.org/10.3102/00346543056004473.

Joslyn, S., LeClerc, J., 2013. Decisions with uncertainty: the glass half full. Curr.

Directions Psychol. Sci. 22 (4), 308–315. https://doi.org/10.1177/0963721413481473.
Kacprzak, E., Koesten, L., Ibáñez, L.-D., Blount, T., Tennison, J., Simperl, E., 2019. Characterising dataset searchâ an analysis of search logs and data requests. J. Web Semant. 55, 37–55. https://doi.org/10.1016/j.websem.2018.11.003. URL http://www.sciencedirect.com/science/article/pii/S1570826818300556
Kay, M., Kola, T., Hullman, J.R., Munson, S.A., 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 5092–5103. https://doi.org/10.1145/2858036.2858558.
Kay, M., Morris, D., schraefel, m., Kientz, J.A., 2013. There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, New York, NY, USA, pp. 401–410. https://doi.org/10.1145/2493432.2493456.
Kern, D., Mathiak, B., 2015. Are there any differences in data set retrieval compared to well-known literature retrieval? 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. pp. 197–208. https://doi.org/10.1007/978-3-319-24592-8_15.
Kim, J., Monroy-Hernandez, A., 2016. Storia: Summarizing social media content based on narrative theory using crowdsourcing. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. ACM, New York, NY, USA, pp. 1018–1027. https://doi.org/10.1145/2818048.2820072.
Klein, G., Moon, B.M., Hoffman, R.R., 2006. Making sense of sensemaking 2: a macrocognitive model. IEEE Intell. Syst. 21 (5), 88–92. https://doi.org/10.1109/MIS.2006.100.
Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E., 2017. The trials and tribulations of working with structured data: -a study on information seeking behaviour. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1277–1289. https://doi.org/10.1145/3025453.3025838.
Kukich, K., 1983. Design of a knowledge-based report generator. Proceedings of the 21st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 145–150. https://doi.org/10.3115/981311.981340.
Law, A.S., Freer, Y., Hunter, J., Logie, R.H., Mcintosh, N., 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. J. Clin. Monit. Comput. 19 (3), 183–194. https://doi.org/10.1007/s10877-005-0879-3.
Lehmberg, O., Ritze, D., Meusel, R., Bizer, C., 2016. A large public corpus of web tables containing time and context metadata. Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016, Companion Volume. pp. 75–76. https://doi.org/10.1145/2872518.2889386.
Li, X., Liu, B., Yu, P.S., 2010. Time Sensitive Ranking with Application to Publication Search. Springer, New York, pp. 187–209.
Lin, S., Fortuna, J., Kulkarni, C., Stone, M.C., Heer, J., 2013. Selecting semantically-resonant colors for data visualization. Comput. Graph. Forum 32 (3), 401–410. https://doi.org/10.1111/cgf.12127.
Little, G., Chilton, L.B., Goldman, M., Miller, R.C., 2010. Exploring iterative and parallel human computation processes. Proceedings of the ACM SIGKDD Workshop on Human Computation. ACM, New York, NY, USA, pp. 68–76. https://doi.org/10.1145/1837885.1837907.
Liu, B., Jagadish, H.V., 2009. Datalens: making a good first impression. Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29, - July 2, 2009. pp. 1115–1118. https://doi.org/10.1145/1559845.1559997.
Liu, X., Tian, Y., He, Q., Lee, W.-C., McPherson, J., 2014. Distributed graph summarization. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, New York, NY, USA, pp. 799–808. https://doi.org/10.1145/2661829.2661862.
Malaverri, J.E.G., Mota, M.S., Medeiros, C.B., 2013. Estimating the quality of data using provenance: a case study in escience. 19th Americas Conference on Information Systems, AMCIS 2013, Chicago, Illinois, USA, August 15–17, 2013. URL http://aisel.aisnet.org/amcis2013/DataQuality/GeneralPresentations/4
Manyika, M. G. I. J., Chui, M., Groves, P., Farrell, D., Kuiken, S. V., Doshi, E. A., 2013. Open data: Unlocking innovation and performance with liquid information. http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information-and-performance-with-liquid-information.
Marchionini, G., Haas, S.W., Zhang, J., Elsas, J.L., 2005. Accessing government statistical information. IEEE Comput. 38 (12), 52–61. https://doi.org/10.1109/MC.2005.393.
Marchionini, G., White, R., 2007. Find what you need, understand what you find. Int. J. Hum. Comput. Interact. 23 (3), 205–237. https://doi.org/10.1080/10447310701702352.
Marcu, D., 2000. The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, MA, USA.
Marienfeld, F., Schieferdecker, I., Lapi, E., Tcholtchev, N., 2013. Metadata aggregation at govdata.de: an experience report. Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, August 05 - 07, 2013. pp. 21:1–21:5. https://doi.org/10.1145/2491055.2491077.
Maxwell, D., Azzopardi, L., Moshfeghi, Y., 2017. A study of snippet length and informativeness: Behaviour, performance and user experience. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017. pp. 135–144. https://doi.org/10.1145/3077136.3080824.
Mei, H., Bansal, M., Walter, M.R., 2016. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016. pp. 720–730. http://aclweb.org/anthology/N/N16/N16-1086.pdf.
van der Meulen, M., Logie, R.H., Freer, Y., Sykes, C., McIntosh, N., 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. https://doi.org/10.1002/acp.1545.
Mizzaro, S., 1997. Relevance: the whole history. JASIS 48 (9), 810–832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U.
Mosa, M.A., Anwar, A.S., Hamouda, A., 2019. A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. Knowl.-Based Syst. 163, 518–532. https://doi.org/10.1016/j.knosys.2018.09.008.
Naumann, F., 2014. Data profiling revisited. SIGMOD Rec. 42 (4), 40–49. https://doi.org/10.1145/2590989.2590995.
Neumaier, S., Umbrich, J., Polleres, A., 2016. Automated quality assessment of metadata across open data portals. J. Data Inf. Qual. 8 (1), 2:1–2:29. https://doi.org/10.1145/2964909.
Noy, N., Burgess, M., Brickley, D., 2019. Google dataset search: Building a search engine for datasets in an open web ecosystem. 28th Web Conference (WebConf 2019).
Owczarzak, K., Dang, H.T., 2009. Evaluation of automatic summaries: Metrics under varying data conditions. Proceedings of the 2009 Workshop on Language Generation and Summarisation. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 23–30. http://dl.acm.org/citation.cfm?id=1708155.1708161.
Park, T.K., 1993. The nature of relevance in information retrieval: an empirical study. The Lib. Q. 63 (3), 318–351. https://doi.org/10.1086/602592.
Pham, M., Alse, S., Knoblock, C.A., Szekely, P., 2016. Semantic Labeling: A Domain-Independent Approach. Springer International Publishing, Cham, pp. 446–462.
Pirolli, P., Card, S., 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of international conference on intelligence analysis. 5. McLean, VA, USA, pp. 2–4. URL https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf
Pirolli, P., Rao, R., 1996. Table lens as a tool for making sense of data. Proceedings of the workshop on Advanced visual interfaces 1996, Gubbio, Italy, May 27–29, 1996. pp. 67–80. https://doi.org/10.1145/948449.948460.
Reiche, K.J., Höfig, E., 2013. Implementation of metadata quality metrics and application on public government data. IEEE 37th Annual Computer Software and Applications Conference, COMPSAC Workshops 2013, Kyoto, Japan, July 22–26, 2013. pp. 236–241. https://doi.org/10.1109/COMPSACW.2013.32.
Reiter, E., Dale, R., 1997. Building applied natural language generation systems. Natural Lang. Eng. 3 (1), 57–87. https://doi.org/10.1017/S1351324997001502.
Robson, C., McCartan, K., 2016. Real world research, 4th Edition. John Wiley & Sons.
Roddick, J.F., Mohania, M.K., Madria, S.K., 1999. Methods and Interpretation of Database Summarisation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 604–615.
Rowley, J., 2000. Product search in eâ shopping: a review and research propositions. J. Consum. Marketing 17 (1), 20–35. https://doi.org/10.1108/07363760010309528.
Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K., 1993. The cost structure of sensemaking. Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 269–276. https://doi.org/10.1145/169059.169209.
Saint-Paul, R., Raschia, G., Mouaddib, N., 2005. General purpose database summarization. Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30, - September 2, 2005. pp. 733–744. URL http://www.vldb.org/archives/website/2005/program/paper/thu/p733-saint-paul.pdf
Schamber, L., Eisenberg, M.B., Nilan, M.S., 1990. A re-examination of relevance: toward a dynamic, situational definition. Inf. Process. Manage. 26 (6), 755–776. https://doi.org/10.1016/0306-4573(90)90050-C.
Scott, D., Hallett, C., Fettiplace, R., 2013. Data-to-text summarisation of patient records: using computer-generated summaries to access patient histories. Patient Edu. Couns. 92 (2), 153–159. URL http://www.sciencedirect.com/science/article/pii/S0738399113001778
Shekhar, S., Celik, M., George, B., Mohan, P., Levine, N., Wilson, R.E., Mohanty, P., 2010. Spatial analysis of crime report datasets. National Science Foundation (NSF), Washington, DC., USA.
Simianu, V.V., Grounds, M.A., Joslyn, S.L., LeClerc, J.E., Ehlers, A.P., Agrawal, N., Alfonso-Cristancho, R., Flaxman, A.D., Flum, D.R., 2016. Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey. BMC Med. Inf. Decis. Making 16 (1), 153. https://doi.org/10.1186/s12911-016-0391-3.
Simmhan, Y.L., Plale, B., Gannon, D., 2008. Karma2: provenance management for data-driven workflows. Int. J. Web Service Res. 5 (2), 1–22. https://doi.org/10.4018/jwsr.2008040101.
Sripada, S.G., Reiter, E., Davy, I., Nilssen, K., 2004. Lessons from deploying nlg technology for marine weather forecast text generation. Proceedings of the 16th European Conference on Artificial Intelligence. IOS Press, Amsterdam, The Netherlands, The Netherlands, pp. 760–764. URL http://dl.acm.org/citation.cfm?id=3000001.3000161
Sripada, S.G., Reiter, E., Hunter, J., Yu, J., 2003. Summarizing neonatal time series data. Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 167–170. https://doi.org/10.3115/1067737.1067775.
Stamatogiannakis, M., Groth, P.T., Bos, H., 2014. Looking inside the black-box: Capturing data provenance using dynamic instrumentation. Provenance and Annotation of Data and Processes - 5th International Provenance and Annotation Workshop, IPAW 2014,

Cologne, Germany, June 9–13, 2014. Revised Selected Papers. pp. 155–167. https://doi.org/10.1007/978-3-319-16462-5_12.

Sultanum, N., Brudno, M., Wigdor, D., Chevalier, F., 2018. More text please! understanding and supporting the use of visualization for clinical text overview. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21–26, 2018. pp. 422. https://doi.org/10.1145/3173574.3173996.

Tam, N.T., Hung, N.Q.V., Weidlich, M., Aberer, K., 2015. Result selection and summarization for web table search. 31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13–17, 2015. pp. 231–242. https://doi.org/10.1109/ICDE.2015.7113287.

Tang, B., Han, S., Yiu, M.L., Ding, R., Zhang, D., 2017. Extracting top-k insights from multi-dimensional data. Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14–19, 2017. pp. 1509–1524. https://doi.org/10.1145/3035918.3035922.

Thomas, D.R., 2006. A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. 27 (2), 237–246.

Thomas, P., Omari, R.M., Rowlands, T., 2015. Towards searching amongst tables. Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8–9, 2015. pp. 8:1–8:4. https://doi.org/10.1145/2838931.2838941.

Tombros, A., Sanderson, M., Gray, P., 1998. Advantages of query biased summaries in information retrieval. SIGIR. 98. pp. 2–10.

Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G., Wu, C., 2011. Recovering semantics of tables on the web. Proc. VLDB Endow. 4 (9), 528–538. https://doi.org/10.14778/2002938.2002939.

Verhulst, S., Young, A., 2016. Open data impact when demand and supply meet. Technical Report. GOVLAB. URL https://ssrn.com/abstract=3141474

Wand, Y., Wang, R.Y., 1996. Anchoring data quality dimensions in ontological foundations. Commun. ACM 39 (11), 86–95. https://doi.org/10.1145/240455.240479.

Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 12 (4), 5–33. URL http://www.jmis-web.org/articles/1002

Weerkamp, W., Berendsen, R., Kovachev, B., Meij, E., Balog, K., de Rijke, M., 2011. People searching for people: Analysis of a people search engine log. New York, NY, USA. https://doi.org/10.1145/2009916.2009927.

Willett, W., Ginosar, S., Steinitz, A., Hartmann, B., Agrawala, M., 2013. Identifying redundancy and exposing provenance in crowdsourced data analysis. IEEE Trans. Vis. Comput. Graph. 19 (12), 2198–2206. https://doi.org/10.1109/TVCG.2013.164.

Wilson, M.L., Kules, B., m. c. schraefel, Shneiderman, B., 2010. From keyword search to exploration: designing future search interfaces for the web. Found. Trends Web Sci. 2 (1), 1–97. https://doi.org/10.1561/1800000003.

Wiseman, S., Shieber, S.M., Rush, A.M., 2017. Challenges in data-to-document generation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017. pp. 2253–2263. URL https://aclanthology.info/papers/D17-1239/d17-1239

Wynholds, L., Wallis, J.C., Borgman, C.L., Sands, A.E., Traweek, S., 2012. Data, data use, and scientific inquiry: two case studies of data practices. Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10–14, 2012. pp. 19–22. https://doi.org/10.1145/2232817.2232822.

Yu, G., 2009. The shifting sands in the effects of source text summarizability on summary writing. Assessing Writing 14 (2), 116–137. URL http://www.sciencedirect.com/science/article/pii/S1075293509000142

Yu, J., Reiter, E., Hunter, J., Mellish, C., 2007. Choosing the content of textual summaries of large time-series data sets. Nat. Lang. Eng. 13 (1), 25–49. https://doi.org/10.1017/S1351324905004031.

Yu, P.S., Li, X., Liu, B., 2005. Adding the temporal dimension to search–a case study in publication search. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, Washington, DC, USA, pp. 543–549. https://doi.org/10.1109/WI.2005.21.

Zhuge, H., 2015. Dimensionality on summarization. CoRR URL http://arxiv.org/abs/1507.00209