Int. J. Human–Computer Studies 146 (2021) 102562

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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

Talking datasets – Understanding data sensemaking behaviours

Laura Koesten [a,1,*], Kathleen Gregory [b,1], Paul Groth [c], Elena Simperl [a]

[a] King's College London, England
[b] Data Archiving and Networked Services, Royal Netherlands Academy of Arts & Sciences, Netherlands
[c] University of Amsterdam, NL, Netherlands

ARTICLE INFO

Keywords:
Sensemaking
Human computer interaction
Human data interaction
Data reuse
Data sharing

ABSTRACT

The sharing and reuse of data are seen as critical to solving the most complex problems of today. Despite this potential, relatively little attention has been paid to a key step in data reuse: the behaviours involved in data-centric sensemaking. We aim to address this gap by presenting a mixed-methods study combining in-depth interviews, a think-aloud task and a screen recording analysis with 31 researchers from different disciplines as they summarised and interacted with both familiar and unfamiliar data. We use our findings to identify and detail common patterns of data-centric sensemaking across three clusters of activities that we present as a framework: *inspecting* data, *engaging* with content, and *placing* data within broader contexts. Additionally, we propose design recommendations for tools and documentation practices, which can be used to facilitate sensemaking and subsequent data reuse.

1. Introduction

Climate change; poverty; global hunger - all have been dubbed wicked problems (Peters, 2017) that have the best chance of being tackled by bringing together and using cross-disciplinary data in new ways (Walshe et al., 2020).[2,3] Although data reuse is increasingly encouraged (European Commission, 2018), it involves a host of challenges, such as providing rich, standardised metadata adequate for interoperability and reuse (Wilkinson et al., 2016). Fundamentally, reusing data also requires that data consumers make sense of data that others have created.

Even within their own disciplinary domains, understanding and making sense of data is a difficult and time–intensive process for researchers and data professionals (Kern and Mathiak, 2015; Muller et al., 2019) which is heightened by the demands of navigating an increasing amount of digital information (Eppler and Mengis, 2004). Also contributing to this difficulty is the fact that data do not speak for themselves, but require supporting structures – both social and technical – to convey the meaning necessary for reuse (Borgman, 2015). The effort and costs involved in sensemaking can potentially be reduced through the development of automated tools and systems (Russell et al., 1993). Designing such tools is contingent upon first understanding and describing the behaviours involved in data-centric sensemaking Rogers et al. (2012).

Here, we identify and detail *patterns of activities* involved in data exploration and sensemaking. In the context of this work, data can be thought of as collections of related observations organised and formatted for a particular purpose, reflecting the variety of concepts different actors have of data, see (Borgman, 2015). In order to identify these patterns, we draw on verbal summarizations as a method of uncovering sensemaking processes and build on the following ideas: (1) the act of summarizing is a form of sensemaking, (2) verbal summarization represents unique cognitive processes and (3) it is possible to identify common patterns in sensemaking activities when people describe data that they are familiar with and data that are unknown to them. We used these ideas to develop the following research questions:

- RQ1: What are common patterns of activities, both for known and for unknown data, in the initial phases of data-centric sensemaking?
- RQ2: How do patterns of data-centric sensemaking afford potential data reuse?

To explore these questions, we combined in-depth interviews with researchers, in which they performed a think-aloud task, and a screen

* Corresponding author.
[1] Both authors contributed equally to this research.
[2] See CODATA Decadal Program http://www.codata.org/strategic-initiatives/decadal-programme.
[3] See United Nations on Big Data and Development https://www.un.org/en/sections/issues-depth/big-data-sustainable-development.

https://doi.org/10.1016/j.ijhcs.2020.102562
Received 6 February 2020; Received in revised form 6 October 2020; Accepted 22 October 2020
Available online 26 October 2020

recording analysis. During the interviews, researchers interacted with and verbally summarised an example of their own research data and a dataset that was unknown to them. We present the results from this study and use our findings to identify activity patterns and data attributes which are important across three clusters of sensemaking activities: *inspecting* data, *engaging* with data content more deeply and *placing* data within broader contexts. Finally, we detail design recommendations for tools and documentation practices to facilitate sensemaking and subsequent data reuse.

The key contributions of this work are identifying: (i) patterns of data-centric sensemaking activities; (ii) a framework for these activity patterns and their related data attributes; (iii) user needs for data reuse; and (iv) a set of design recommendations to support data-centric sensemaking.

2. Background

Sensemaking has been studied across a range of disciplines, including psychology (e.g. Klein et al. (2006)), decision making (e.g. (Klein et al., 1993; Malakis and Kontogiannis, 2013)), organizational behaviour (e.g. see Maitlis and Christianson (2014) for a review), information seeking (Dervin, 1997; Marchionini and White, 2007), and human computer interaction (HCI) (e.g. (Russell et al., 1993)). In this work, we focus on sensemaking as discussed in information science and HCI. In these domains, sensemaking is defined as the process of constructing meaning from information (Blandford and Attfield, 2010), and is recognised as being an iterative process that involves linking different pieces of information into a single conceptual representation (Hearst, 2009; Russell, 2003).

2.1. Sensemaking and information seeking

Models of information seeking behaviour often present sensemaking as a key component. Traditional models detail the specific steps involved in sensemaking during information seeking as a sequential, yet evolving, process (Hearst, 2009; Kuhlthau, 2004; Sutcliffe and Ennis, 1998). While traditional models tend to be static, many of their authors emphasise that people's behaviour is complex and changes when being presented with new information. More recent, dynamic models acknowledge a variety of influencing factors in finding and making sense of information, e.g. skills, knowledge, preconceptions, culture or motivation (Kelly, 2009; Klein et al., 2007). Other work examines the cognitive mechanisms involved, framing sensemaking as a series of different information processing components taking data as input and producing conceptual changes as an output (Bechtel, 2008; Zhang and Soergel, 2020).

2.2. Data-centric sensemaking

While sensemaking of textual information has been well-explored, there is a relative gap in research that aims to understand the strategies involved in making sense of data. Compounding this is the fact that the very definition of "data", particularly "research data" has itself been the subject of much debate. An increasingly common conceptualisation of research data is that proposed by Borgman (2015): data are representations of observations, objects, or other entities that are used as evidence for the purposes of research or scholarship. This definition does not distinguish between data formats or qualitative or quantitative data, recognizing that what serves as data in one situation for one individual may not act as data in another situation for another individual (see also Pasquetto et al., 2017). Similary, in their data frame theory, Klein et al. (2007) emphasise how the perspective (or frame) of the data consumer shapes the data in terms of how they are perceived, interpreted and even acquired. Through engaging with data, preexisting frames either change or get reinforced, which can be seen as an aspect of sensemaking. Critical data studies also describe this as a collective process, due to interpretative layers built into the creation and use of data (Neff et al., 2017)

Studies in HCI tend to focus on quantitative data, addressing, e.g., the role that visualization plays in identifying patterns in data (Furnas and Russell, 2005; ah Kang and Stasko, 2012); this focus reflects the emergence of bespoke visual exploration environments (Marchionini et al., 2005; Yalçin et al., 2018). Other work proposes tools to aid in sensemaking activities, such as a visual analytics system tailored for particular groups of data analysts (Stasko et al., 2008) or agile display mechanisms for users accessing government statistics (Marchionini et al., 2005). Investigations of exploratory data analysis (EDA) strategies, where new data are explored with a set series of procedures until a high-level story emerges, are also of relevance. Common EDA techniques include performing rough statistical checks and analyses (e.g. calculating descriptive statistics) or looking for general trends or outliers in the data (Baker et al., 2009; Marchionini, 2006). Many EDA techniques are graphical in nature and are undertaken to help assess the quality of the data.

The first phase of getting to know data, which can involve exploratory data analysis techniques, has been shown to involve a high level of cognitive effort (Zhang and Soergel, 2014). Existing categories can prompt users to activate related memory content, resulting in converging categorization and verbalization processes; this influences how information is interpreted and potentially eases sensemaking efforts (Fiore et al., 2003; Ley and Seitlinger, 2015).

Engagement and sensemaking with data is also determined by the purpose of the engagement activity, usually connected to a task, which can range in specificity. The importance of quality indicators and uncertainty attached to data is task dependent (Boukhelifa et al., 2017; Koesten et al., 2017). While there are a variety of task classifications in the information seeking literature (e.g. (Freund, 2013; Li and Belkin, 2008), to this date there is no established taxonomy for data-centric work tasks, which might reflect rapidly changing work practices with data.

2.3. Evaluating data for reuse

There is a growing amount of literature, particularly within information science, that examines the reuse of research data. Key studies question and explore the definitions and types of data reuse within and across disciplines (Pasquetto et al., 2019; van de Sandt et al., 2019). Many studies characterize the contextual information required to make decisions about using (or not using) data within particular disciplinary fields (i.e. Faniel and Yakel, 2017; Kriesberg et al., 2013). Although a set definition of context remains challenging (Faniel et al., 2019), there is an overall agreement that data reuse without any contextual reference is almost impossible to do well (Birnholtz and Bietz, 2003; Borgman, 2015).

Building on studies of researchers in three disciplinary domains, Faniel et al. (2019) propose a typology of the information needed to support data evaluation, finding that information about data production, data repositories, and data usage are key in making decisions about reusing data. Similarly, Gregory et al. (2020) find that researchers across disciplines rely on information about data collection conditions, data processing, topic relevance and accessibility when evaluating data. This aligns to a large degree with Koesten et al.'s (2020) findings on dataset-specific selection criteria covering different aspects of relevance, quality and usability (Koesten et al., 2019b).

Other work looks specifically at how researchers develop trust in data. Yoon (2017), for instance, draws on interviews with quantitative social scientists to explore the social, multi-stage processes involved in trust development. She identifies data characteristics which can aid in building trust, such as the quality of documentation and the reputation of the data publisher. Passi and Jackson (2018) describe perceived trustworthiness of data or a data science system as a task dependent and collaborative accomplishment that involves assessing different types of

uncertainties. While the criteria used for both data evaluation and trust building likely play important roles in data-centric sensemaking, several authors highlight that much of the knowledge needed to make sense of data is tacit and not included in data documentation (Birnholtz and Bietz, 2003; Rolland and Lee, 2013).

2.4. Summarisation as a way to understand sensemaking

We adopt a study design that builds on work using summarisation as a way of exploring cognitive processes (Hidi and Anderson, 1986). Summarisation tasks, as studied in psychology, are described as involving three distinct cognitive activities: selection of which aspects should be included in the summary; condensation of source material to higher-level ideas or more specific lower-level concepts; and transformation by integrating and combining ideas from the source (Hidi and Anderson, 1986). As comprehension is viewed as a prerequisite for summarisation, text summarisation tasks have been used to assess recall and language abilities (Kintsch and Van Dijk, 1978). We build on these ideas and use summarization as a way of exploring the cognitive processes involved in comprehending data as an information source.

In a recent study, Koesten et al. (2019b) had participants produce written dataset summaries in order to better understand selection criteria for datasets. While these summaries provide insights into the conceptualisation of datasets, written summaries do not always capture the complex verbal sensemaking that precedes their creation (Mayernik, 2011). Verbal, or spoken, summarizations often reflect deeper, more spontaneous cognitive processes (Crestani and Du, 2006), but they have yet to be used to understand data sensemaking behaviours.

2.5. Summary of key points

We argue that in order to reuse data, data consumers must first be able to understand and make sense of those data. While we hypothesise that these sensemaking processes will include attributes similar to those identified in the above literature, we also postulate that data-centric sensemaking involves particular cognitive processes and social and technical interactions resulting in common patterns of sensemaking activities. Our work therefore takes into account not just attributes and categories related to engagement with data, but also considers the wider social, disciplinary and communication contexts existing in data work and their impact on consumer engagement with data.

Our argument builds on the assumption that sensemaking affords specific activities when engaging with data opposed to other information objects (e.g. textual sources), which is mirrored in the literature. However, the sensemaking and information seeking literature often either focuses on textual sources or does not clearly differentiate which source is addressed.

It is also worth noting that there is a significant amount of literature on dataset reuse that focuses on operational problems, machine readability and data interoperability (Koesten, et al., 2020). We do not review this literature in detail here, as our purpose is to focus on the less-studied practices and patterns involved in understanding data. We connect our work to these discussions regarding technical solutions for facilitating data reuse by providing empirical evidence of data-centric sensemaking and by identifying common patterns of sense-making activities to enable design efforts.

3. Methodology

Our past work in textual data summarization (Koesten et al., 2019b) and the reuse of research data (Gregory et al., 2019) informed the the creation of a semi-structured interview design examining how people verbally summarize and make sense of both familiar and unfamiliar data.

3.1. Study design

All participants were asked to bring data that they had used or were familiar with to share during the interview. We refer to this data as the *known data* in this paper. We left the decision about what constitutes "data" up to participants. The majority ($n = 27$) chose to bring data which they had created themselves. Most of the data brought by participants were spreadsheets ($n = 19$); other data included textual data (e.g. interview transcripts), images, videos or other artefacts. We did not ask for any documentation, supporting information or metadata from participants to see what they brought when not prompted.

We also prepared a dataset to share with participants; we refer to this data as the *unknown data* in this paper. This dataset was a modified version of a spreadsheet from a popular news source in the UK, the Guardian Data Blog, United Nations[4] which was used in a previous study (Koesten et al., 2019b) (see Figure 1; the entire spreadsheet is available on a GitHub repository[5] associated with this work). This dataset met specific selection criteria: it included numerical and textual data, missing values, inconsistencies in formatting and some ambiguous variables. At the same time, the data were understandable and not specific to a particular domain.

3.2. Data collection

The interview protocol (available on GitHub) consisted of two primary sections: questions about the *known* data, and questions about the *unknown data*. Interviews lasted 30 – 60 minutes and were held using the web-conferencing application Zoom. All interviews were audio-recorded; screen-recordings capturing participants' interactions with the data were obtained for 26 interviews. (Five recordings were not created due to technical problems).

Both sections of the interview began with the verbal summarization task. The task was for participants to provide a general summary description of the data to someone who is trying to decide whether to use the data, but who is unable to see it. We formulated the task in this fashion in order to elicit rich descriptions of the data, not aimed at a particular use case.

In the first section of the interview, after summarising the *known* data, participants were asked questions about data reuse and their data creation and documentation practices. In the second section of the interview, we shared our dataset and asked participants to perform the same verbal summarization task, this time on the *unknown* data. We then asked them to describe and discuss specific areas of the *unknown* dataset and posed follow-up questions about data reuse, data sharing and data search (examples can be seen in Table 1).

Two pilot interviews were conducted in October 2018. This allowed us to determine the interview duration, to fine-tune the interview questions and the set-up of the summarization task. The remaining interviews were held between November 2018 and January 2019 and were transcribed by a professional transcription firm.

Recruitment.

Our primary sample was drawn from a pool of individuals, past respondents to a large scale survey study conducted by Gregory et al. (2020), who had published at least one article indexed in Elsevier's Scopus literature database[6] in the last three years.

We sent recruitment emails ($n = 1000$) in November-December 2018 in two batches and received 47 positive responses. From those, we selected 27 participants who represented a range of disciplines and nationalities and were proficient in English. We recruited an additional four participants via convenience and purposive sampling, for a total of 31 participants. Participants for our pilot interviews were identified

[4] https://www.theguardian.com/.
[5] https://github.com/laurakoesten/talkingdatasets.
[6] https://www.scopus.com.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	Country	Deaths, confirmed swine flu	Deaths per million population	Confirmed cases (last updated 26/01/2010)	Infection rate per million people	LONG	LAT	POP (source: UN)	SOURCES (where no link, based on agency reports)	Second source, if needed		swineflu deaths
2	1	Afghanistan	17	0.6	837	29.73	65	33	28150000	http://www.emro.who.int/csr/h1n1/h1n1_update.htm		AF	yes
3	2	Albania	6	1.89	13	4.1	20	41	3169000	http://ecdc.europa.eu/en/activities/surveillance/EISN/Pages/EISN		AL	yes
4	3	Algeria	57	1.63	476	13.64	3	28	34895000	http://www.afro.who.int/ddc/influenza	http://www.ands.dz/grippe-p	DZ	yes
5	4	Andorra		0	1	11.63	1.5	42.5	86000	http://www.who.int/csr/don/2009_07_27/en/index.html		AD	yes
6	5	Angola	0		37	2	18.5	-12.5	18498000	http://www.afro.who.int/ddc/influenzaa/updates/		AO	no
7	6	Antigua and Barbuda	0	0	4	45.45	-61.8	17.05	88000	http://new.paho.org/hq/index.php?option=com_content&task=blog		AG	yes
8	7	Armenia	3	0.97					3090000	http://ecdc.europa.eu/en/activities/surveillance/EISN/Pages/EISN		AM	
9	8	Australia	191	8.97	37553	1,763.63	-27	133	21293000	http://ecdc.europa.eu/en/activities/surveillance/EISN/Pages/EISN		AU	yes
10	9	Austria *	0	0	361	43.16	47.3333	13.3333	8364000	http://ecdc.europa.eu/en/activities/surveillance/EISN/Pages/EISN		AT	yes
11	10	Argentina	626	15.54	10209	253.48	-34	-64	40276000	http://new.paho.org/hq/index.php?option=com_content&task=blog		AR	no
12	11	Azerbaijan	2	0.22	2	0.22	47.5	40.5	8934000	http://www.who.int/csr/don/2009_08_04/en/index.html		AZ	yes
13	12	Bahamas	4	11.7	24	70.18	-76	24	342000	http://new.paho.org/hq/index.php?option=com_content&task=blog		BS	yes

Fig. 1. Excerpt of the provided or *"unknown"* dataset describing the global occurrence and mortality rate of swine flu.

Table 1
Overview of interview schedule including the summarisation task and example topics for the different interview sections.

Section	EXAMPLES
0: Background	Demographics, job role, discipline or research area, experience of working with data
1: Summary of known data	*Verbal summarisation task*
1: Context of known data	Information structures needed for reuse Describe for colleagues vs for someone outside your domain
2: Summary of unknown data	*Verbal summarisation task*
2: Context of unknown data	Anything missing that you would like to know about the dataset
3: Specific areas of unknown data	Rows / columns with missing / ambiguous data, different variable types

Table 2
Description of participants (P) with their disciplinary domains and professional roles.

P	Domain	Role
1	Biological sciences	Project manager
2	Life sciences, Paleontology	Project acquisition manager
3	Biblical studies, Information Technology	Researcher
4	Musicology, Humanities	Project leader, project manager
5	Geophysics	Data curator
6	Physics, Chemistry	Post doctoral associate
7	Analytical Chemistry	Researcher
8	Material Science and Engineering	Professor emeritus, researcher
9	Social Science (Social Care, Social Work)	Senior research fellow
10	Social sciences, Computer science	Director of Research Services
11	Social justice, Socioeconomic Justice	Professor
12	Geology, Earth Sciences	Research scientist
13	Earth Sciences	PhD student
14	Fluid Mechanics	Researcher
15	Molecular Biology	Researcher
16	Tourism, Social Psychology	Senior lecturer
17	Mathematical education	Assistant professor
18	Telecommunications, Computer science	Associate professor
19	Biological anthropology	Postdoctoral research fellow
20	Medicine, Biomedicine	Researcher and teacher
21	Agriculture, Food science	PhD Student
22	Medicine	Surgeon, PhD student
23	Entomology (Biological Sciences)	Researcher, curator
24	Environmental sciences, agriculture	Lecturer
25	Biostatistics, Epidemiology	Associate professor, biostatistician
26	Material Science	Researcher
27	Psychology	Researcher, PhD student
28	Veterinarian, Obstetric Clinician	Assistant professor
29	Information science, Medicine	Associate director
30	Environmental sciences	Researcher
31	Medicine, Mental Health	Head of research group in a hospital

using purposive sampling. We did not offer incentives for participation in this study.

Participants.

Participants ranged from age 26 to age 73, with the majority being between 30 and 45 years old (Median 40.6). They reported 19 different countries of residence worldwide, with a skew towards the Netherlands ($n = 5$) and the UK/USA ($n = 3$). 13 out of the 19 countries are in Europe; 20 of our 31 participants live in European countries.

Although participants work in multiple countries, the majority were fluent in English; minor problems with language or internet connectivity were experienced in two of the interviews. Over half of the participants ($n = 18$) worked at a university or college at the time of the study, with six working in research institutions. Participants' disciplinary domains and roles are described in Table 2. All participants have previously published research papers. The majority were experienced with quantitative research; others categorised themselves as predominantly qualitative researchers, or used both quantitative and qualitative methods.

Ethics.

The study was approved by the University of Southampton's Ethical Advisory Committee under ERGO Number 45874. Informed written consent was given by the participants prior to the interview.

3.3. Data analysis

Coding strategy.

The coding strategy for thematic analysis was developed through a multi-step process of independent parallel coding (Thomas, 2006), using the the qualitative data analysis program NVivo. Two authors independently analyzed a sample of seven interview transcripts and developed an initial codebook with supporting examples, employing a combination of deductive and inductive thematic analysis (Robson and McCartan, 2016). Codes developed through deductive analysis were oriented on the different sections of the interview protocol and on existing literature in data summarization (Koesten et al., 2019b) and data reuse (Faniel et al., 2019; Faniel and Yakel, 2017).

Within these high-level themes, the authors iteratively developed codes based on a general inductive approach (Thomas, 2006) through sequential readings of the transcripts. The independently-developed codebooks were compared for similarities and differences and combined and modified to create a single unified codebook which was then used to re-code the sample transcripts, which were divided evenly between the two researchers. To further enhance the reliability of the coding scheme, two senior researchers checked and discussed the unified codebook for a sample of the data.

Based on this analysis, we made further modifications, resulting in a nested coding tree consisting of three primary codes with a total of 30 child codes (see Figure 2 for the most used codes). We consolidated these codes through axial coding (Straus and Corbin, 1990) drawing out those links which allowed us to answer our research questions. The themes identified through axial coding are used to structure the Findings section

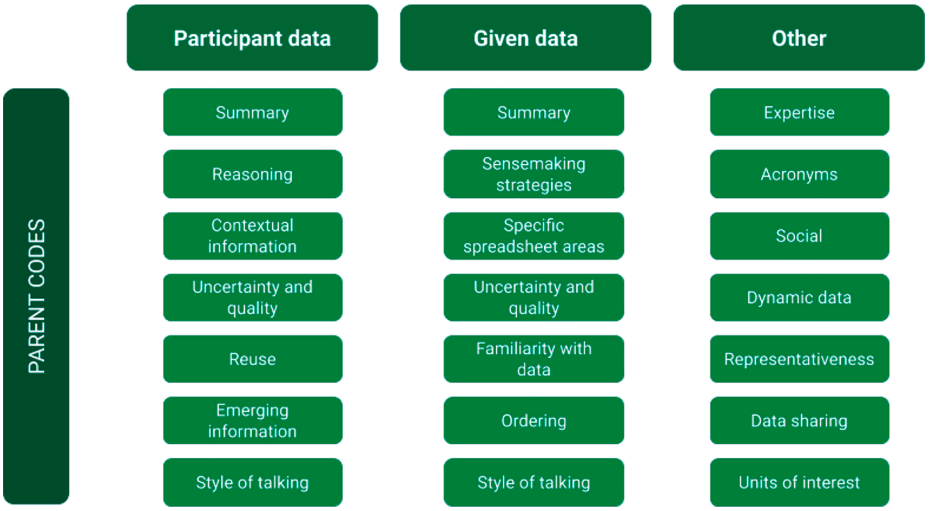

Fig. 2. Primary codes.

and form the basis of the synthesis presented in Figure 7.

Screen recording analysis.

We analyzed the 26 captured screen recordings to identify common interactions with the *unknown* dataset. We examined participants' actions during the general summarisation task. We did not analyse the screen recordings of the summarisation task for the *known* data, due to the diversity of participants data, participants privacy concerns and in order to encourage deeper descriptions of their data. Two of the authors independently viewed a sample of these screen recordings to identify common interactions. The authors then discussed the list of interactions and iteratively developed a list of 15 interactions to use in the video analysis. We used this list to identify the first occurrence of each possible action; we did not consider the duration of each action in our analysis. The coding of the screen recordings was done by one researcher only. Even then, a sample of the coded data was checked by the other researcher. Via the screen recordings, we could document how much of the spreadsheet participants could visibly access on their screens without scrolling. This allowed us to control for larger screens.

Visual analysis.

All plots were created using the statistical analysis program R. We used the color palette "viridis", as it has been shown to be more accessible than other comparable color schemes.[7]

4. Findings

We present our results along two dimensions: the research questions identified in the introduction section and the clusters of sensemaking activities which we identified via axial coding, namely *inspecting* the data, *engaging* with the content and *placing* data in broader contexts. Although we divide this section by research question to improve readability, the evidence we present often spans these divisions.

We pay special attention to both *activity patterns*, which we define as common physical and cognitive actions undertaken by participants when engaging with the data, and data *attributes*, or characteristics of the data with which participants interacted. We examine the findings in light of data reuse and synthesise them in the Discussion section to provide an overview of the patterns we identify.

4.1. RQ1: What are common patterns in sensemaking activities, both for known and for unknown data, in the initial phases of data-centric sensemaking?

4.1.1. Inspecting

When participants were first shown our dataset, we asked them to perform the verbal summarization task – to provide a general summary description, after taking a few minutes to explore the data silently. In this section, we examine both the order of how participants discussed attributes of the data (see Figure 3 and 4) and their actions in the spreadsheet during these verbal summarisations (Figure 5).

Order of verbal summarisation We observed two approaches when completing the verbal summarisation task: participants took either a linear or an interwoven approach. In linear summaries of the *unknown* data ($n = 24$), participants addressed the data attributes identified in Figure 3 (e.g. time, location, format) one-by-one before proceeding to the next attribute. In the interwoven summaries, participants interspersed descriptions of individual attributes with their analyses and comments ($n = 7$).

Figure 3 shows the attributes headers ($n = 64$), quality/uncertainty ($n = 42$), topic/title ($n = 33$), and analysis/dependencies ($n = 30$) were most frequently mentioned in the summarisation of the *unknown*dataset. The majority of participants mentioned the overall topic or title as one of the first two attributes ($n = 24$); roughly half of participants mentioned the format or shape of the data (e.g. the number of columns, rows or observations) either first or second ($n = 15$). The discussions of other attributes were likely influenced by the structure of the dataset itself. Location information was a prominent part of the dataset, e.g., as the data were ordered by country and the four columns containing geographic information were positioned on both the left and right sides of the spreadsheet. The data included only minimal temporal information. The majority of general summaries mentioned location ($n = 22$) toward either the beginning or end of the description, while temporal information was mentioned in just under half of the general summaries ($n = 13$).

In the linear general summaries, time and location were discussed or questioned at a general level:

> This communication shows us the deaths from swine flu in the countries around the world, Afghanistan, Albania, Columbia, Bolivia. (P31)
>
> The one thing that is not apparent immediately from the data is the time span. (P19)

Participants taking an interwoven approach to summarization

[7] https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html.

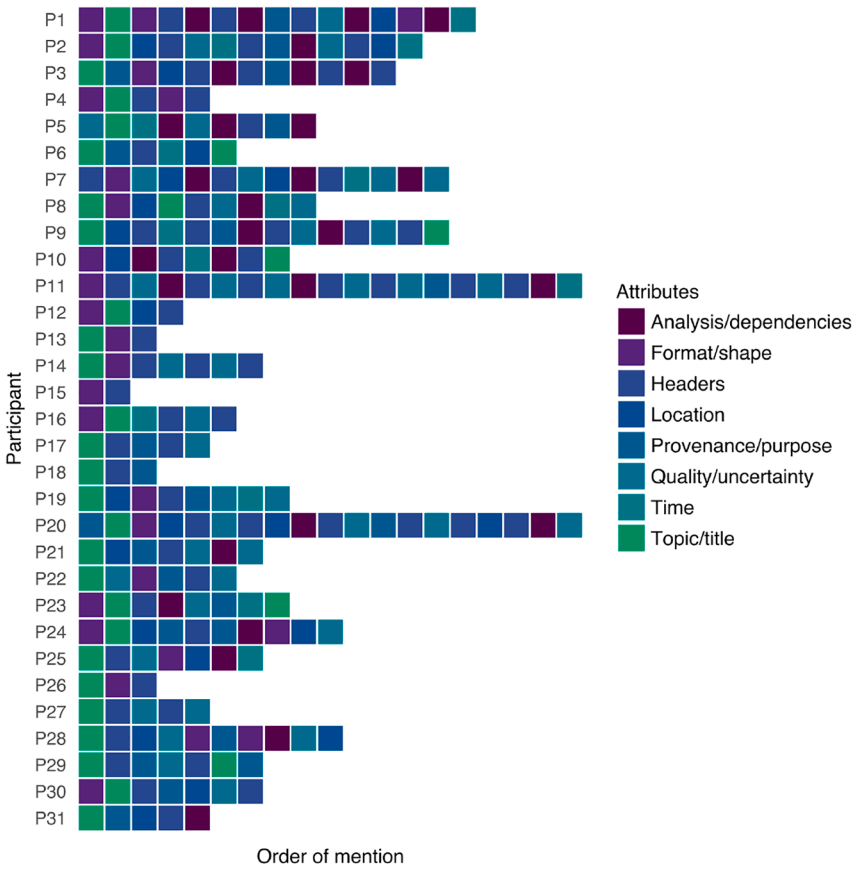

Fig. 3. The order in which participants discussed certain attributes in the *unknown* dataset.

engaged in more initial analysis, repeatedly seeking relationships and dependencies between the spreadsheet columns or expressing uncertainty about meaning or the quality of the dataset.

> I don't see any date or year, for purposes of comparison then it's a bit problematic, I can for example only do comparison charts for those with an asterisk for Austria and Bulgaria, for example, because they all have the data from 2009 but for number of deaths recorded in that country, then this data is useful, infection rate per population. (P5)

We observed similar attributes within the general summaries of participants' own data, but participants also mentioned additional attributes, i.e. details of their own field of research, methodology and details of the particular study, data availability and access restrictions, and the existence of additional information or documents needed to describe and understand their data (Figure 4).

Most of the general summaries of participants' data followed a linear pattern ($n = 24$). This could be because participants were not working to understand their own data, but were rather aiming to make their data understood. They were also to some extent better prepared for the requirements of the task, having already had experience working with and discussing their data. In interwoven summaries of their data ($n = 7$), participants mixed descriptions of study methodologies with descriptions of headers and data format; some relied on methodological descriptions to communicate the general topic of their data.

> These are experiments from a 50 metre long indoor set up that we have, where we ran gas and oil through the pipeline, through a 60 metre long pipeline, and we measured the average values - so pressure drop and build-up. And we did that for different gas and liquid velocities, and they also changed the type of oil, so we did this with one oil with a quite low viscosity and one with oil with a quite high viscosity. (P14)

Actions in the spreadsheet The actions captured in the screen recordings of the verbal summarization task for the *unknown* data support the attributes identified in Figure 3 and 4. Figure 5 shows the total number of actions observed, as well as the frequency of their order of occurrence. Scrolling right ($n = 24$) was the most frequently observed action, followed by scrolling down ($n = 23$). Participants also clicked on or indicated column headers and specific values. Clicking on both headers ($n = 18$) and particular cells ($n = 17$) occurred more often than other forms of indicating these areas of the spreadsheet. Four participants indicated headers in other ways, i.e. hovering over or circling them, yielding a total of 22 participants who either clicked on or indicated headers. Four participants also pointed out particular cell values using these alternative actions, resulting in a total of 21 participants who either clicked on or indicated cell values.

Analysing the order of these actions show that the majority of participants began by determining the length, breadth and general topic of the *unknown* dataset. Nine participants first scrolled down, while eight clicked on headers and seven initially scrolled to the far right of the spreadsheet. Once participants established the general shape of the data, more analysis-related actions were observed, most noticeably examining specific cell values by clicking or indicating and moving back and forth between different columns. One example of left and right scrolling was switching between the different types of geospatial columns which were not located in close proximity to one another.

Some participants prepared for analysis by reformatting the spreadsheet ($n = 10$), i.e. by adjusting the column width or freezing columns or rows. (For three recordings, the width of one of the columns was not optimised to allow reading one of the header names). Analysis features of the spreadsheet were only used in four instances, for actions such as sorting, filtering, or performing calculations. This reflects the nature of the think aloud task and the time limitations of our study.

We began examining screen recordings for participants' data after

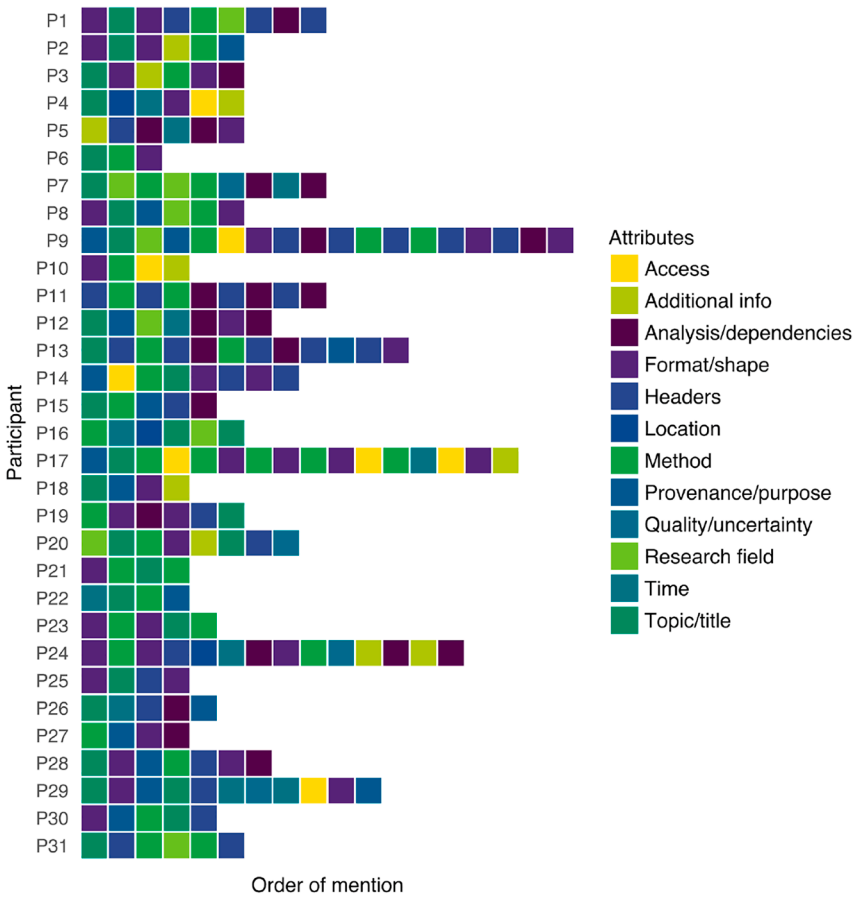

Fig. 4. The order in which participants discussed certain attributes in their own data (*the known data*).

Fig. 5. Total number of actions and order of actions observed in screen recordings of the *unknown* dataset. Size of circle represents number of participants engaging in activity. Figure is arranged according to which activity most frequently occurred first. Color represents purpose of action.

Table 3
Exemplary quotes illustrating participants' interactions with both their own and the *unknown* data.

	Known Data	Unknown Data
Encodings	(P23): Because of the way it's set up, while it may appear on screen in words, it's actually all in zero, one, two, three, four up to nine. But you can present it yourself in words, and that's really helpful if you're scoring, because you can actually click from one cell to the next, and down the base it will actually tell you what that character is in that cell and whether you just code a zero or a one. (P9): Age band and then I grouped the age bands into adult and older people, and that was one of the issues of each of our journeys, a different way of categorising people, so I ended up with a very broad age range really.	(P28): I don't know if it could be interesting to be banded in categories like, I don't know, continents... it depends. (P19): It looks like they started to code for if there are no deaths, then it's coded as a zero, but there are some instances where there are missing data.
Acronyms and abbreviations	(P22): That is a classic abbreviation in the field of hepatic surgery. AFP is alpha feto protein. It is a marker. It's very well known by everybody...the AFP score is a criterion for liver transplantation. (P28): So if there is strange code that people cannot understand, I make a legend. Normally the colleagues I'm working with, we use our terms, so I tend to use the most user terms like LD; like SEM is typical for...everyone in my field.	(P20): If I would make this assumption, I would say this is like geographical location of the countries, but I have no idea what is 'Long' and 'Lat'. In my work, I have never encountered these kind of acronyms, so it's currently hard for me to assume what would this mean in the context of swine flu. (P7): I'm not sure what 'long' means. I wonder if it's not something to do with longevity. On the other hand, no, it's got negative numbers. I can't make sense of this.
Identifiying "strange things"	(P7): Let's say from previous experiments and or runs, you know that repeating the experiment, you would get within an error of say 5% or 2%, whatever the case maybe. So obviously these three [indicating error bars] are huge, and it would mean that you will have to repeat. So either something is wrong in your system, or you get something wrong during the sample preparation, or the system's not stable, or something else is going on. Or that you're just not planning enough repetitions to get to the true value, so I think it is an important measure to determine if you've got reliable data.	(P20): If I would not go into those cases, like with these discrepancies, I would just assume that this column indicates only the optimised data about whether they're aware or they're not, that [deaths are] due to swine flu in these countries." (P14): That is simply a column saying if there are any deaths at all or not for a certain country related to the swine flu. I see there is a formula here, just simply checking if Column L is larger than zero. So exactly using this information...so then that means there is something wrong with the formula or I completely misunderstood what Column L is.

the general summarisation task. These screen recordings provide a different type of insight, revealing actions that participants took to ensure that the interviewer adequately understood their data. The actions that we observed ranged in complexity. Participants with spreadsheet data often clicked on each column header, as they provided more detail about each column. Others demonstrated how they would analyse the data, showing unique functions of their analysis software or creating sample spreadsheets and plots.

4.1.2. Engaging with the content

Participants engaged with data content in more depth as they worked to explain and understand the data. This stage of deeper interrogation sometimes began during the scanning phase; it occurred both when interacting with the *unknown* dataset and with the *known* data. Table 3, which is further discussed throughout this section, presents quotations demonstrating similarities and differences in how they engaged with both known and unknown data. This table is organised along three themes: *encodings*, or codes developed to understand and interact with data, *acronyms and abbreviations* and *identifying "strange things"* within the data.

In this phase, participants identified patterns and trends (e.g., via simple analyses or discovering relationships between columns) and discussed encodings, often related to categorisations, expressed within the data. They also explored uncertainties attached to the data and the data's overall integrity. In addition to further discussing these approaches, we point out two other particular instances observed in this level of engagement with the data: understanding strange things and collaborative sensemaking.

Data analysis, encodings and tools

When discussing their data, participants demonstrated how they seek patterns and relationships by creating plots, switching between layers on geospatial images, and developing scales and formulas. Participants also expressed a desire to create plots to visualize the unknown data to identify trends and sought anchor variables as they investigated individual columns and described sample rows. They further drew attention to columns with limited value ranges in their summary descriptions, e.g. columns with binary variables or those with only a few categorical variables; fewer analyzed the range of values in columns with continuous variables.

Participants "encode" their own data in ways that help them more easily identify trends and generate findings by, e.g., converting categorical variables to numerical values and vice versa. These encodings are often influenced by the specifications of the analysis tools and software which participants use, such as SPSS, R, or domain-specific programs; which can also influence how participants structure their data, at times increasing the data's machine readability.

> I use this data to create variables in SPSS. The one I'm looking at now has still got all the labels as words; I thought it would be easier to look at as a spreadsheet. There's another process I went through to translate the words into numbers. For SPSS, you really need numbers in the value labels. That was a whole process, to go through of coding the written, the categories, but just adapting those into numbers that I use. (P9)

> [We are] working in R and our supervisor wrote a package which can easily work on it, but the main aspect is that you have to have grouping variables and independent variables which are the sensor signs. Then you have to separate the data to these different types, so the grouping variables and the independent variables because the PC and the IDA in the R can work in this structure. (P21)

Other forms of encoding included developing broad categories or groupings to describe and analyze data, such as differentiating between raw data and derived data, or numerical and non-numerical data. Participants also created groups of certain columns according to their semantic meaning; demographic variables were mentioned together, as were descriptive attributes for the same instance, e.g. "columns with sources" or "socioeconomic measures". These types of encodings were observed when participants discussed both their own and the *unknown* data. When working with our data, participants also searched for how null values were encoded and represented (see Table 3).

While the majority of participants reported using spreadsheets or Microsoft Excel at some point in their data workflows, very few actually made use of the built-in analysis tools in our spreadsheet at any time during the task. This could be due to time limitations during the interviews or to the fact that participants were not familiar with the Google Sheets environment which we used. It could also be a result of the fact that some participants do not use spreadsheets to analyze data directly, but rather reported using them for other purposes, such as recording and organizing data or cleaning and preparing data for analysis. Spreadsheets are also used by participants to specifically enable sharing data in a way that is easily accessible or compatible with a variety of analysis programs, facilitating data reuse.

Expressing uncertainty, seeking quality and understanding strange things Both when discussing their own data as well as when engaging with our data, they expressed concerns about potential misinterpretations, focusing on questions that could arise due to misunderstandings about how data were cleaned and processed. For both quantitative and qualitative data, participants viewed the encodings and categories that they had constructed as major risk points for correctly interpreting their data. The encodings that facilitated their own use of the data (as a data producer) may not be helpful or be explained well enough to enable appropriate data reuse by potential consumers of their data.

> Although we've tried really hard, because we've put in a coding frame and how we manipulate all the data, I'm sure that there are things in there which we haven't recorded in terms of, well, what exactly does this mean? I hope we've covered it all but I'm sure we haven't. (P10)

They also questioned and critiqued the meaning of the *known* data, highlighting the lack of contextual information about how the dataset was created and the use of unexplained abbreviations in the dataset. When discussing their own data, however, participants often referred to unexplained acronyms or abbreviations common in their own disciplinary domains (see Table 3).

Participants combined their interpretations about the meaning of our dataset along with analyses of its completeness and how missing values are reported to make quality determinations. They also used missing values as checkpoints to identify relationships between columns and to identify potential errors or anomalies in the data.

> So the data is fairly complete with really limited missing values, so the quality of data looks good. (P29)

> It's got some blanks, which I presume means no data has been given. Although that's interesting...there's some missing data which shouldn't be missing. Because Armenia, for number seven say, it reports three deaths and yet the swine flu deaths is blank, so that's a bit of an anomaly, and there are quite a few blanks actually. (P9)

Participants looked for other unexpected values (e.g. outliers) or inconsistencies in formatting or standard ways of reporting to assess the precision and accuracy of the *unknown* dataset. Wrestling with these strange things often served as the entry point to a deeper engagement and understanding of the data, allowing participants to question their assumptions and initial understandings (see Table 3).

> Now that sounds quite high for the Falklands. I wouldn't have thought the population was all that great...and yet it's only one confirmed case. Okay [laughs]. So yes...one might need to actually examine that a little bit more carefully, because the population of the Falklands doesn't reach a million, so therefore you end up with this huge number of deaths per million population [laughs], but only one case and one death. (P23)

> Some of them have decimals, like a lot of decimals, and some don't have any decimals. So I don't know whether that means that those are supposed to be measured more precisely...or that there is an inconsistency of using the amount of decimals per cell. (P1)

Encountering the unexpected in their own data is a critical and normal part of participants' research processes (see Table 3). While anomalies can be indicative of possible mistakes or points for improvement in the study design, they can also reflect unexpected external changes to the study environment, e.g. people withdrawing from a study, or new technologies that have been adopted over the course of long-running studies. Participants repeatedly emphasized the need to communicate information about these changes or potential sources of error to possible data reusers:

> When I'm explaining the dataset by sharing a screen or showing them the file or to someone who would probably understand the data, from a dataset perspective, I would basically talk about the implausible values and the missing values and if I'm aware of the issues related to the data, I would like to point them out. (P29)

Sensemaking through "collaborations" Working with team members is key to making decisions about study design and analysis, i.e. deciding which data are important to record and analyse, how to develop scales, clarifying study details and making sense of mistakes or unexpected values in the data.

> I know roughly what it consists of, but I didn't know precisely, and I had to go back to the person who generated it and say "What does column D mean? And where is the location of the thermocouple whose temperature is measured in column E?". (P8)

> We have a table with...almost 30 columns with variables that were collected, including the names of the people who went into the field and collected each of the samples. So we are keeping track of who's responsible for each of the samples, then if we find any error, any mistake, then we can contact those people. (P24)

During the interviews, participants also collaborated with the interviewer to ensure that the interviewer correctly understood their data. Often, important details crucial to understanding the data emerged only when both the interviewer and the participant could see and interact with the data together. We saw this, e.g., in the case of learning about the importance of temporal information in coral reef imaging data or highlighting a key variable (inflammation) in a study about bipolar individuals. For some, it was nearly impossible to explain their data without being able to indicate specific areas of an image or demonstrating how error analysis was conducted.

4.1.3. Placing

As they engaged more deeply with data, participants placed data into existing contexts, practices and knowledges; this process of "placing" occurred at different scales (Figure 6). Data were placed within their immediate contexts of creation, e.g., when participants detailed study designs, experimental setups or the conditions surrounding data collection, including broader temporal or geographic details.

> And it describes, or rather it comprises the results of a laboratory experiment lasting about an hour in which the experimenter, [..], is

Fig. 6. Placing data in contexts.

> inducing the crystallisation of a salt in a porous rock sample. And as the crystallisation proceeds, two things happen. One is that the heat is evolved and so the temperature changes and the rock sample slightly increases in temperature, and we measure that temperature at three different positions. In addition there is a very slight expansion of the rock which we detect from the output of a very sensitive mechanical gauge. And then these four measurements, the three temperatures and the mechanical strain are measured at intervals of one second over a period of a few hours and so the dataset consists of this set of numbers. (P8)

These types of contextual details have the potential to impact the meaning which the values themselves carry.

> Error bars depend a lot on the experimental conditions and on the condition of the material. So, for example, if it was used on powder samples then the error bars would be bigger than the ones that were obtained on single crystal data. (P6)

At a broader level, participants conceptually placed data within the norms of their disciplinary domains, referencing discipline-specific methodologies and limitations, ways of analysing and verifying data or common data formats. They also recognised that broader social contexts can influence the sensemaking process.

Finally, participants attempted to place data within the world, gauging how representative data are of a particular phenomena. These judgements reflect assumptions about how much the data reflect reality, as data themselves are usually samples which are hardly ever complete, unbiased or without conflict or ambiguity.

> It's a pretty large sample size, again, 1,260. We have equal numbers of males and females. We have three ethnicities: Caucasian, African American and Hispanic, equal numbers of each of those. So it's a well-balanced data set and, because of that, if you were to be interested in how these different cultural values vary or not based on ethnicity, it would be an excellent dataset. (P19)

One simple example we observed in our dataset was a contention about the representativeness of the countries, which showed a range of interpretations and was expressed in a variety of ways. Participants questioned both the completeness of the list of countries and also whether the data represented the entirety of each country.

> P2: It's listing the countries for which data are available, not sure if this is truly all countries we know of...
> P8: It includes essentially every country in the world
> P29: Global data
> P30: I would like to know whether it's complete...it says 212 rows representing countries, whether I have data from all countries or only from 25% or something because then it's not really representative.
> P7: If it was the whole country that was affected or not, affecting the northern part, the western, eastern, southern parts
> P24: Was it sampled and then estimated for the whole country? Or is it the exact number of deaths that were got from hospitals and health agencies, for example? So is it a census or is it an estimate?

During placing activities, participants commonly reported the need to know the original purpose for which the data was created. Descriptions of their own data's original purpose were often complex, as they were intermingled with descriptions of the field of research. Participants floundered in their attempts to place our data, in part because the original study objective was unknown.

Although important across all dimensions of sensemaking, disciplinary and data expertise were key to placing data. Most participants felt that it was easier to describe their data than to summarise and try to understand our data.

> My data are much more easier, for sure, because I knew what I was talking about. I didn't have to go through, to understand, which was the quality of the data; I didn't have to understand what it was, the kind of information that this data was giving to me. If I have to go to a database that I've never seen normally and also that is not in my field, it is absolutely much more difficult. (P28)

4.2. RQ2: How do patterns of data-centric sensemaking afford potential data reuse?

Some participants believed that only experts from within their same discipline could reuse their data meaningfully, citing the specificity of their data or the need to analyze the data using specific programs. Others stated that appropriate reuse would require a deep understanding of evolving domain research practices; many had difficulty imagining alternative uses for their data outside of their area of research.

> I don't think it would be used for a radically different purpose, but I could imagine somebody taking the data and reanalysing it in relation to a different model of the underlying process, for example. Or confirming the interpretation that we've placed upon the data using our own model...But they would be people who'd be very close to the topic. (P8)

4.2.1. Structures needed for sensemaking and reuse

A few participants believed that the use of common data structures, terminologies and methodologies within their domains made it possible for their data to "speak for themselves" to others with similar expertise.

> I probably wouldn't have to describe it [the data]. Probably they would just get it. (P1)

We observed procedural reasons why data do not speak for themselves, but require additional structures to convey meaning. Participants did not always include column names in their data in order to make them more machine readable, e.g., or they divided datasets into various sub-sheets to ease processing. We also observed that additional information structures, i.e. documents and codebooks, are needed to support reuse for data consumers regardless of domain, as well as to support future (re)use by the original data producer.

> Ten years makes a big difference in my memory, too. So, even at the time when I was working in it, I didn't have to refer to the code book, I knew it all by heart. I would have to go back and look at the code book now myself, and that's why it's important to keep the notes on what you're doing with the variables and keep a copy of the survey that was used, the research instrument, those sorts of things. (P11)

Participants described a large variety of documentation and knowledge transfer practices surrounding datasets (Table 4). These practices and the formats used to provide additional information are shaped by journal restrictions, metadata schemas and repository requirements, and by the perceived usability of the information structures themselves. Sometimes this additional information is separate from the data; other times it is embedded within the data, i.e. in the case of annotations or descriptions of codes within a spreadsheet. Different data consumers may require different information structures for the same data.

> If they're using a different program, I can direct them to a character set, which you can get from this matrix, but the publication of that character set is quite separate but available online. (P23)

> So if you start with the README here, then we can take several directions. So, you can delve into the features, what they mean, and you can delve into the feature documentation. You can delve into ways to query it, and do that for yourself, and then you go to all kinds of programming documentation. And then, here I also pointed to

Table 4
Information structures supporting sensemaking.

INFORMATION STRUCTURES
Supplementary files (corresponding spreadsheets, text documents, README files)
Resource description document (including, e.g., explanations of columns)
Code
Documentation of the code
Emails / communication protocols
Figures / visualisations
Code book / sheet; can contain personal data
Repository
Presentations / slides
Technical reports
Publications
Maps
Audio folder
Slack channel
Annotations & interpretations (also on various levels of the data, e.g. on image layers)
Tutorials
Questionnaires / surveys (variables often created in order of the questions)

> tutorials, [..]. And you can read some papers about it and they're also cited...We also have a Slack community with 120 people, and if they have really hard questions, we invite them to Slack, and they are being answered by either me or people who know more about it. (P3)

The study also revealed attributes which should be present in information structures to avoid losing meaning and to enable data reuse. We present these attributes according to two perspectives which emerged in the interviews: the data consumer's distance to the data and the methodological approach of the original study in Table 5 and 6. We define "distance to the data" in terms of a data consumer's familiarity and expertise with particular data. Someone "close to the data" will have more knowledge of the data and how they were created; someone more distant from the data will not have this knowledge. In Table 6, we focus on two broad approaches to data collection: quantitative and qualitative

Table 5
PERSPECTIVE: Information needs related to distance to the data. Someone "close to the data" will have more knowledge of the data and how they were created. Someone "far from the data" will not have this knowledge. Certain attributes did not seem to be affected by distance to the data (Column 3).

Close to the data	Far from the data	No difference by distance to the data
More granular information about conditions, assumptions, errors, trends, possible questions the data can answer, variable types, analysis / programming details, sample creation details, study objective	More granular information about research explanation, explanation of all abbreviations / acronyms, how ratios / errors / columns derived	Supplemental materials
Benefits / problems of data	Less technical language	Study objective and expected outcome
Previous work that this data builds on, relation to standards in discipline, out-of-discipline abbreviation	Research explanation	Data collection details
Less granular information about field of research, common abbreviations, data format / structure	Tailor to field of interest of the data consumer	Sample details
	More general data presentation	Potential use of data
	Calculation of ratios/ standard deviation	Usage restrictions, confidentiality concerns
		Explanations of codes, categories, scores

Table 6
PERSPECTIVE: Methodological narrative *(characteristics are not necessarily unique to either approach)*.

Quantitative	Qualitative
Detailed experimental set-up: including time period, instrument settings, location, etc.; where the test conditions differ from real world settings or from standard procedures	*Detailed study set-up:* including time period, description of participant sample, sample size, mode of interaction (e.g., online or in person)
Who did which work (data collection, quality control, data cleaning, code, analysis)	Who did the research; researcher's relationship to participants and how this was mitigated (e.g., professional role of researcher / context of recruitment)
Are measurements individual measurements or multiple measurements of the same thing that were aggregated	Questions or schedule for surveys or interviews (including information about answer modes (e.g., predefined answers or free text)
Which section of the object was measured, on how much material a measurement was made	Analysis (e.g., type, coding strategy, groupings and narrative of categorisation)
Standard error, precision of measurements, uncertainties	How sample was chosen (inclusion / exclusion criteria), created, scope and characteristics of sample (e.g., age of participants)
Influencing factors (seasonal differences, external events, etc.)	How categories were chosen, how scores were created, variables of focus
Standard units of measurement in a field / setting of study (e.g., instruments – specifications, reliability, how calibrated, how they work and how they create the data output, software format used to capture or analyze data)	Social context
Number of repetitions of experiment	Description of labels / codebook / account of variables

methodologies. These tables do not aim to present a comprehensive list, but rather reflect the specific work scenarios of our participants.

We asked participants if they would describe their data differently to a colleague with similar expertise, i.e. someone close to the data. Rather than needing less information about the data due to prior knowledge, many participants believe that individuals with similar expertise need more granular information about data creation conditions, prior work which the data builds on, and the potential uses of individual variables. Some participants said that they would not describe their data differently to someone close to the data, emphasising instead common attributes that would be important, regardless of a data consumer's distance to the data (Table 5, Column 3).

> So if I'm talking to somebody who is data agnostic or who has not worked in a data science field, my description would be limited to the basic variables, the fields that are of interest to the person...If I'm talking to a data science person or a data scientist who's going to use the data, my description would be more granular. My description would be more helping the person understand the benefits as well as the problems associated with the data. (P29)

> I would maybe shorten up some things and focus on some others. For instance, I would expect that everyone I'm working with expects to code BMI in kilograms and to have birth weight in grams because it's a standard unit for those things in Danish health research...I would tell them more about the study design, because often people I work with are epidemiologists. So there one of the main things would be, where do these 2,000 women come from? Is it data from Denmark or from somewhere else? Is this from last year or from 30 years ago? Things like that, so more complex information so that they can decide if it's relevant for their interests. (P25)

Different methodological approaches also elicited particular details, although these details were not mutually exclusive of each other (Table 6). For quantitative data, participants reported needing extensive information about an experimental setup, including how experimental

designs differed from the real world environment.

> Well I would perhaps mention the size of the pipe diameter. That is something that they're often interested in, because in real pipelines, the pipe diameter is perhaps 12 inch and more, quite large, while in typical labs, you don't have this possibility. (P14)

Key findings from the qualitative perspective include the choice of categories, questions of representativeness and details of the study set-up that influence the data, such as whether participants are required to answer a survey question. Social context also influences how study participants communicate, e.g. in the case of interview participants in conflict areas who may not feel safe enough to respond truthfully to questions.

5. Discussion

We bring together different perspectives in this study, drawing together our findings about participants' summaries of familiar and unfamiliar data and our observations of how participants engaged with these data. We now synthesise our findings, identifying different *patterns of activities* and their related *data attributes* involved in data-centric sensemaking. The sensemaking efforts which we observed can be synthesized into three clusters of activities: *inspecting* the data, *engaging* with the data content more deeply and *placing* data within broader contexts (Figure 7). We also examine the relation of the clusters of sensemaking activities to information structures needed for reuse and discuss three emergent themes in the context of this synthesis. Here, we define:

- *Activity patterns* as the actions, both physical and cognitive, which people undertake when making sense of data
- *Data attributes* as characteristics of the data which people interact with as they perform a set of activities
- *Clusters* as the activity patterns, with their related attributes, which tend to occur together

C1, inspecting, contains activities and attributes that provide participants with a broad overview of the data, such as understanding the data's general topic, title, structure and format. In the *unknown* data, we observed that most participants scanned the spreadsheet first vertically to look at the number of rows and to get an idea of missing values and then horizontally to look at the headers.

C2 represents a deeper **engagement with the content of the data**, including activities such as establishing relationships between columns, performing simple analyses, picking out examples of particular values, conducting quality assessments and trying to understand uncertainties attached to the data, by questioning, e.g. the meaning of missing values or abbreviations and acronyms.

In **C3**, we observed participants **placing** data in relation to the world and different contexts. They worked to understand how the data were related to study designs, to disciplinary norms as well as to temporal and geographic considerations to understand the representativeness of the content. They questioned, e.g., the level of detail (granularity) presented in the data as well as the data's original purpose.

Our findings show that level one (C1) of Figure 7 was mostly done alone; level two and three (C2, C3) were often solved in collaboration. When participants described their own as well as our data, critical details emerged only after the initial summary description, when both the interviewer and the participant could see and interact with the data together. These conversations moved away from objective descriptions towards describing the complexity of qualitative judgements behind the (quantitative) variables, as well as to rich descriptions of factors influencing the origination of data. This echoes literature in critical data studies, conceptualising data as the product of socio-technical arrangements but also as a medium through which conversation and negotiation can occur (Neff et al., 2017).

When discussing their data, participants made use of the information structures identified in Table 4 and their related qualities (Table 5 and 6) across all dimensions as they worked to make their data understood. Many also referenced the lack of contextual information (e.g. purpose, collection methods) in the *unknown* data as being a stumbling block to understanding.

The importance of needing contextual information to support data reuse, at both the level of the data (Borgman, 2015; Faniel et al., 2013; 2012) and of digital collections (Baker and Yarmey, 2009; Chin and Lansing, 2004; Lee, 2011), has been extensively noted in the literature. Our work is in line with these findings, particularly in noting the

Fig. 7. Activity patterns and attributes in data-centric sensemaking.

importance of information describing data collection conditions and methodological details. Recent work (Faniel et al. (2019)) also draws attention to descriptions of what we term "encodings" in Table 3, to describe the codes that participants create and use when working with data.

Koesten et al. (2020) provide a summary of the literature examining particular data attributes for reuse, listing which papers document the importance of certain data and documentation characteristics. While our findings add to this literature, particularly by presenting important attributes along the lines of an individual s knowledge of the data and its creation process (Table 5) and the methodological narrative (Table 6), our primary aim is not to isolate data attributes needed for reuse, as much work has already investigated this problem. Rather, the work we present here takes the lens of analysing how those attributes are brought together by the activities they afford, which we group to patterns of sensemaking activities.

Translating these findings into interaction guidance and subsequently into tools supporting reuse presents a challenge, in part because of the dynamics and context-specific nature of working with data (Kross and Guo, 2019; Muller et al., 2019). The in-depth descriptions of study set-ups, purposes of data collection and domain specific knowledge brought by our participants underscores this problem. As a way to address this challenge, Figure 7 can be viewed in the context of work using design patterns in areas such as software engineering (Gamma, 1995), user interface design (Granlund et al., 2001), or ontology design (Gangemi and Presutti, 2009).

This approach identifies high-level patterns as a way to provide repeatable solutions to recurring design problems. This creates possibilities for a level of formalisation that enables the development of flexible designs and tools. Our results are in line with (Boukhelifa et al., 2017; Koesten et al., 2019a; Marchionini et al., 2005), who see flexibility as being key to supporting real-world data workflows. Figure 7 therefore represents a patterns-based approach to conceptualising the processes involved in the initial stages of data-centric sensemaking.

To further contextualise Figure 7 and to illustrate how our findings could spur design efforts, we discuss three specific themes that emerged in our research at the level of each identified cluster. For each theme, we present design recommendations. The recommendations we propose exist in parallel to research in information visualization (Shneiderman, 1996), which suggests visual support for a high-level taxonomy of data tasks and types. Our study brings a deeper perspective to understanding information needs focusing on structured data, suggesting a wider variety of data-related tasks undertaken by users, which may or may not be supported through visual exploration. Our recommendations build on what we have learned about researchers sensemaking activities and workflows; our aim is to disrupt these workflows as little as possible. We therefore propose functionalities and approaches to support sensemaking that could be integrated within analysis tools already used by researchers, such as common programming languages or libraries.

5.1. Understanding shape

When inspecting a dataset for the first time, see Cluster 1, participants either discussed the data in a linear fashion, addressing each attribute individually before moving to the next, or they took a more interwoven approach, mixing descriptions of dataset attributes with analyses and questions. This interwoven approach also has overlaps with activities in the second cluster of Figure 7.

As they engaged in inspecting acitvities, participants aimed to arrive at an overview, to create a high-level representation of the entire dataset in their head while engaging with it (see also Koesten et al., 2017). We observed different levels of focus in this process. Participants alternated between "zooming out" to describe the data at the level of the entire spreadsheet, e.g. the number of observations or format of the data, and "zooming in" to look at specific cell values or individual parts of the data. Participants adopting a more interwoven approach tended to engage in the process of zooming in and out more often than those using a linear approach.

This desire to understand the data as a whole has parallels in the information science literature, where the need to understand an entire information collection at a high level has been mentioned (Rieh et al., 2016). Discussing the visual aspects of sensemaking, Russell (2003) also mentions the need to understand what is in a whole collection. White and Roth (2009) recommends allowing users to filter, sort and explore different views of information on demand for complex search tasks. In our study, the information is distributed among the cells of the dataset, the structure and organization of the data, as well as any related information structures.

5.1.1. Recommendations

Understanding the shape of a dataset can be supported through interface design and functionalities in a number of ways. Our results show that data needs to be understood as a whole, on the **level of the entire dataset**. This suggests summarization methods, which can be of textual, visual or statistical nature, that provide a zoomed-out view of the data (e.g. Holland et al. (2018)). At the same time, participants also engaged with subsets of the data, particularly individual columns; these patterns could be supported through zooming in via **column level summaries**, including interactive plots and visualizations at the column level (e.g., this idea is partially realised by Kaggle in their dataset previews[8]). Future research could look at different ways of expressing a column-based notion of provenance, such as where the data in a column comes from, how it was created or from where it was derived. Given the importance of scanning and zooming in and out (as mentioned in literature such as Shneiderman (1996)), data search engines and displays should optimize this functionality to make these processes as fast as possible; including horizontal scrolling to accommodate spreadsheets with more columns.

Similarly, certain types of information structures attached to the dataset facilitate particular sensemaking patterns over others. A README file with a summary of the dataset's size and format may provide the information necessary for a zoomed out inspection of the data; an interactive map of the area where a specimen was collected may be more suitable to a zoomed in approach, as well as enabling the activities described in Cluster 2.

5.2. "Strange things" as an entrypoint, not an obstacle

Participants repeatedly encountered and dealt with "strange things" in both data sources, i.e. outliers, errors, missing data, and inconsistencies in formatting. As they wrestled with the unexpected in the data, they engaged in the patterns identified in Cluster 2, such as expressing uncertainty, seeking relationships or performing analyses.

Whereas (Zhang and Soergel, 2020) describe dealing with conflicts as a barrier to sensemaking, our findings suggest that conflict is a useful and accelerating moment in the exploration of data. The concept that real data is usually messy and complex was internalised by our participants. Participants were neither surprised nor alienated by conflicting data; in contrast, errors and uncertainties were expected and participants applied different analytical strategies to overcome them, a finding also in line with recent literature (Boukhelifa et al., 2017; Koesten et al., 2017; Neff et al., 2017). Participants repeatedly emphasized the need to communicate information about sources of error and possible uncertainties to potential data consumers, although there were a variety of communication methods used to do so, some of which are detailed in Table 4. Methods for communicating information about strange things in the data were sometimes chosen arbitrarily or convenience-based. Some of this information was embedded within data themselves, leading to potential problems in machine readability. Others were not linked

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

to the data in a sustainable way, making them unsuitable for long-term preservation of meaning.

5.2.1. Recommendations

Our findings suggest that errors can be seen as an entry point to sensemaking, as flags to investigate further. This provides an interesting direction to explore for sensemaking functionalities in tools. Rather than flattening out data by making it cleaner, tools could instead **flag and highlight strange things** to make users more aware of their presence. Column summaries, as mentioned in 5.1.1, could include explanations of abbreviations and missing values, metrics or links to other information structures necessary for understanding the column's content. Datasets should include **links to basic concepts** (used in the data or in the documentation) such as common practices in code documentation or "the web" (i.e. in Wikipedia / Wikidata) to provide context. Documentation about the narrative surrounding these strange things should also be more standardised and linked directly to these flags in a sustainable way.

Other sensemaking patterns identified in C2 can be supported by customised interactive visualisations. Displaying the entire data, as described in 5.1.1, but **highlighting relationships between columns or entities** could allow users to more easily pick up relationships between columns. Tools could also **display trends and patterns** extracted from the dataset and allow users to select those attributes of the data that are of interest. Following this idea, data producers could identify anchor variables, those which they consider most important in their dataset; this could further aid sensemaking activities by focusing summarisation efforts.

5.3. Perspectives in placing

Participants place data and their representativeness in a range of broader contexts (the world, disciplinary norms, methodological contexts of creation). While we present these placing activities separately in Cluster 3, they can in fact be closely related. We saw this particularly in how participants placed data in terms of a study's methods and their own disciplinary expertise.

Details about data creation are often implicit within a domain's epistemic norms (Leonelli, 2016). Even with the best documentation, this complicates cross-disciplinary data reuse. A data consumer from another domain may not have the experience necessary to understand or evaluate the appropriateness of a particular methodological approach. Additionally, our participants' concept of the details needed for reuse encompassed much more than just a step-by-step process of how a study was conducted. Rather, for both quantitative and qualitative data (see Table 6), participants needed details about the entire narrative surrounding data creation, i.e. why a certain method was chosen or the unique, local aspects about a study's set-up and their attached constraints. This need for expanded and robust methodological narratives mirrors recent calls for details beyond those provided by standardized metadata and reporting conventions, particularly for the reuse of qualitative data (Poth, 2019).

We also found that the granularity of these narratives is related to a potential data consumer's expertise or distance to the data, with experts needing more detailed information about study descriptions. Table 5 also shows common attributes, aside from methodology, that are important in facilitating understanding, independent of a data consumer's expertise with data, i.e. needing information about study objectives, usage restrictions, and explanations of categories and acronyms.

5.3.1. Recommendations

Our findings highlight the need for flexible designs to support placing activities across the three identified levels of placing: the world, disciplinary norms and the study-set-up. Rather than designing for a specific type of user, tools should be designed to embrace different levels of expertise, allowing a potential data consumer to **drill down to the desired level of detail**. Semantic technologies (Balog, 2018) also could be used to **link to standardized definitions** of disciplinary acronyms or terms, mirroring our recommendation in 5.1.1 to link to external knowledge bases. Geographic information could be linked to a map or country registry to allow judgements of representativeness; a similar approach could be taken for certain disciplinary standards and study set-ups, such as standard experiment conditions, expected result ranges or commonly used confidence levels. Data citations, in particular their associated metadata, can contain detailed provenance information needed for sensemaking, offering another emerging possibility for providing the necessary context for data reuse (Groth et al., 2020).

Our findings across all dimensions emphasize the collaborative nature of data-centric sensemaking and the omnipresent role of information structures throughout the sensemaking and reuse process. It has been suggested that the production and consumption of academic writing can be conceptualized as a form of dialogue (e.g. Lillis (2011)); the broader practice of reusing data could itself be seen as a form of collaboration or conversation between data producers and consumers. The data producer must communicate the many (often collaborative) decisions which influence the creation of a dataset (Mahyar and Tory, 2014; Neff et al., 2017) to potential data reusers.

A combination of focused documentation practices integrating different media types, together with prescribed interaction flows tailored to the sensemaking practices of both data producers and consumers, could facilitate the conversation implicit in reusing data. These could include solutions with adaptable data representations suited to varying levels of expertise and needs.

6. Study Limitations

Although they were working in a wide range of disciplinary domains and research related roles, our sample population consisted of a particular type of professional: researchers who have published an article indexed in the Scopus database[9].

Scopus comes with a skew towards certain research fields; the Arts & Humanities, for instance, are not as well-represented. Scopus has an extensive review process for the journals which it selects for inclusion; and there are roughly an equivalent number of journals from the broadly defined fields of social sciences, health sciences and physical sciences (Elsevier, 2020). While the limitations of Scopus could lead to a potential bias in our sample, the selection criteria we applied also ensured that the sample population met our study requirement of speaking with different types of researchers with data experience.

As the study was conducted with researchers, our findings may not be directly applicable to other individuals. Focusing on researchers met the goals of our study, particularly our aim of examining sensemaking in light of reuse. However, we believe the general sensemaking patterns emerging from this study are to some extent transferable between different skill sets; simply the execution of how these goals are achieved might look different for people with a higher or lower level of data literacy. Nonetheless, the study would need to be repeated with different populations in order to apply our findings more broadly.

Our participants work in a variety of countries; English was not the native language for all. To account for this, we selected our sample from those responding to our recruitment messages carefully to ensure that participants had a high degree of English fluency. While we see the global spread and disciplinary diversity of our sample as a strength of the study, we also recognise that data, research, and sensemaking practices are influenced by social, legal, and economic contexts unique to both country and disciplinary domains.

The sensemaking patterns which we identify could also be limited by the data themselves. Different data may have surfaced different data

[9] https://www.scopus.com.

attributes. By including participants' data, as well as the *unknown* dataset in the study, we attempted to balance this potential bias. Another potential limitation is that describing their data first might have primed participants for performing the verbal summarisation task with the *unknown* data, influencing the way that they performed the second task. Given that any data description will be based on a participants' prior experience, we believe this is a natural side effect of these types of studies.

Finally, it is important to note that we intentionally did not ask participants to bring metadata for their *known data* as we wanted to see what types of data, metadata and other contextual information they felt that they needed to bring without being prompted. Similarly, the study design allowed us to capture what participants felt was missing from the *unknown data* and to identify what additional information was needed. This is especially relevant as data does not always come with complete or accurate or meaningful metadata or enough information for reuse, as detailed in the background section.

7. Conclusions

In this study, we investigated common patterns in sensemaking activities in initial encounters with data, particularly in light of potential data reuse.

We identified three clusters of activities involved in initial data-centric sensemaking (*inspecting, engaging with content, and placing in context*), and detailed the observed activities and data attributes relevant in these clusters. This approach provides an avenue to bring focus to design efforts, narrowing down the many technologically feasible solutions to those specifically supporting the sensemaking needs of data consumers. To summarize, the contributions of the paper are:

- activity patterns for data-centric sensemaking;
- a framework of these patterns and their associated data attributes;
- user needs for data reuse;
- design recommendations to support the identified activity patterns.

Our work illustrates the large space for future research trajectories in this area to validate and apply insights within different contexts of data-centric work practices. This could include investigating the identified activity patterns with different data or with individuals working outside of research. Other work could focus on how to apply the detailed insights and recommendations to existing user workflows. Such work could focus on determining the best way to present and allow interaction with data to facilitate sensemaking. Similarly, such work could explore the integration of our findings into existing services and platforms, particularly with regard to multidisciplinary data.

Sensemaking allows individuals to create rational accounts of the world which enable action (Maitlis, 2005). In this work, data-centric sensemaking enables a particular type of action: the reuse of data in research. Understanding how people make sense of data, and exploring designs to support these practices, therefore, plays a key role in realizing the potential of data reuse.

CRediT authorship contribution statement

Laura Koesten: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Kathleen Gregory:** Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Paul Groth:** Conceptualization, Validation, Writing - review & editing. **Elena Simperl:** Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is partially supported by the Data Stories project, funded by EPSRC research grant No. EP/P025676/1 and by the NWO Grant 652.001.002 Re-search: Contextual search for scientific research data.

References

Baker, J., Jones, D.R., Burkman, J., 2009. Using visual representations of data to enhance sensemaking in data exploration tasks. J. AIS 10 (7), 2.
Baker, K.S., Yarmey, L., 2009. Data stewardship: Environmental data curation and a web-of-repositories. IJDC 4 (2), 12–27. https://doi.org/10.2218/ijdc.v4i2.90.
Balog, K., 2018. Entity-Oriented Search. In: The Information Retrieval Series, 39. Springer. https://doi.org/10.1007/978-3-319-93935-3.
Bechtel, W., 2008. Mechanisms in cognitive psychology: What are the operations? Philosophy of Science 75 (5), 983–994. https://doi.org/10.1086/594540.
Birnholtz, J.P., Bietz, M.J., 2003. Data at work: supporting sharing in science and engineering. In: Schmidt, K., Pendergast, M., Tremaine, M., Simone, C. (Eds.), Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2003, Sanibel Island, Florida, USA, November 9-12, 2003. ACM, pp. 339–348. https://doi.org/10.1145/958160.958215.
Blandford, A., Attfield, S., 2010. Interacting with Information. Morgan & Claypool Publishers. https://doi.org/10.2200/S00227ED1V01Y200911HCI006.
Borgman, C.L., 2015. Big data, little data, no data: Scholarship in the networked world. MIT press.
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.
Chin Jr., G., Lansing, C.S., 2004. Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work. ACM, New York, NY, USA. pp. 409–418. DOI: 10.1145/1031607.1031677
Crestani, F., Du, H., 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. JASIST 57 (7), 881–890. https://doi.org/10.1002/asi.20350.
Dervin, B., 1997. Given a context by any other name: Methodological tools for taming the unruly beast. Information seeking in context 13, 38.
Elsevier, 2020. Scopus content coverage guide. Amesterdam: Elsevier BV.
Eppler, M.J., Mengis, J., 2004. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. The Information Society 20 (5), 325–344. https://doi.org/10.1080/01972240490507974.
European Commission, 2018. Turning FAIR into reality. Final report and action plan from the European Commission expert group on FAIR data. European Commission. Directorate General for Research and Innovation. Directorate B Open Innovation and Open Science. Unit B2 Open Science.
Faniel, I.M., Frank, R.D., Yakel, E., 2019. Context from the data reusers point of view. Journal of Documentation 75 (6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133.
Faniel, I.M., Kansa, E., Kansa, S.W., Barrera-Gomez, J., Yakel, E., 2013. The challenges of digging data: a study of context in archaeological data reuse. 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, Indianapolis, IN, USA, July 22 - 26, 2013, pp. 295–304. https://doi.org/10.1145/2467696.2467712.
Faniel, I.M., Kriesberg, A., Yakel, E., 2012. Data reuse and sensemaking among novice social scientists. Information, Interaction, Innovation: Celebrating the Past, Constructing the Present and Creating the Future - Proceedings of the 75th ASIS&T Annual Meeting, ASIST 2012, Baltimore, MD, USA, October 26-30, 2012, pp. 1–10. https://doi.org/10.1002/meet.14504901068.
Faniel, I.M., Yakel, E., 2017. Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. Curating research data, volume one: Practical strategies for your digital repository 103–126.
Fiore, S.M., Cuevas, H.M., Oser, R.L., 2003. A picture is worth a thousand connections: the facilitative effects of diagrams on mental model development and task performance. Computers in Human Behavior 19 (2), 185–199. https://doi.org/10.1016/S0747-5632(02)00054-7.
Freund, L., 2013. A cross-domain analysis of task and genre effects on perceptions of usefulness. Inf. Process. Manage. 49 (5), 1108–1121. https://doi.org/10.1016/j.ipm.2012.08.007.
Furnas, G.W., Russell, D.M., 2005. Making sense of sensemaking. Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005, pp. 2115–2116. https://doi.org/10.1145/1056808.1057113.
Gamma, E., 1995. Design Patterns: Elements of Reusable Object-oriented Software. Pearson Education India.
Gangemi, A., Presutti, V., 2009. Ontology design patterns. Handbook on Ontologies. Springer, pp. 221–243. https://doi.org/10.1007/978-3-540-92673-3_10.
Granlund, Å., Lafrenière, D., Carr, D.A., 2001. A pattern-supported approach to the user interface design process. International Conference on Human-Computer Interaction, pp. 05/08/2001–10/08/2001.

Gregory, K., Groth, P., Scharnhorst, A., Wyatt, S., 2020. Lost or found? discovering data needed for research. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e38165eb.Https://hdsr.mitpress.mit.edu/pub/gw3r97ht

Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S., 2019. Understanding data search as a socio-technical practice. Journal of Information Science 0 (0). https://doi.org/10.1177/0165551519837182.0165551519837182

Groth, P., Cousijn, H., Clark, T., Goble, C., 2020. Fair data reuse–the path through data citation. Data Intelligence 78–86. https://doi.org/10.1162/dint_a_00030.

Hearst, M., 2009. Search user interfaces. Cambridge University Press.

Hidi, S., Anderson, V., 1986. Producing written summaries: Task demands, cognitive operations, and implications for instruction. Review of educational research 56 (4), 473–493.

Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K., 2018. The dataset nutrition label: A framework to drive higher data quality standards. CoRR abs/1805.03677.1805.03677

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.

Kelly, D., 2009. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval 3 (1-2), 1–224. https://doi.org/10.1561/1500000012.

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.

Kintsch, W., Van Dijk, T.A., 1978. Toward a model of text comprehension and production. Psychological review 85 (5), 363.

Klein, G., Moon, B.M., Hoffman, R.R., 2006. Making sense of sensemaking 1: Alternative perspectives. IEEE Intelligent Systems 21 (4), 70–73. https://doi.org/10.1109/MIS.2006.75.

Klein, G., Phillips, J.K., Rall, E.L., Peluso, D.A., 2007. A data–frame theory of sensemaking. Expertise out of context. Psychology Press, pp. 118–160.

Klein, G.A., Orasanu, J., Calderwood, R., Zsambok, C.E., et al., 1993. Decision making in action: Models and methods. Ablex Norwood, NJ.

Koesten, L., Kacprzak, E., Tennison, J., Simperl, E., 2019. Collaborative practices with structured data: Do tools support what users need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019, p. 100. https://doi.org/10.1145/3290605.3300330.

Koesten, L., Simperl, E., Blount, T., Kacprzak, E., Tennison, J., 2019. Everything you always wanted to know about a dataset: studies in data summarisation. International Journal of Human-Computer Studies.

Koesten, L., Vougiouklis, P., Simperl, E., Groth, P., 2020. Dataset reuse: Translating principles to practice (preprint). PATTERNS.

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, Denver, CO, USA, May 06-11, 2017., pp. 1277–1289. https://doi.org/10.1145/3025453.3025838.

Kriesberg, A., Frank, R.D., Faniel, I.M., Yakel, E., 2013. The role of data reuse in the apprenticeship process. Proceedings of the American Society for Information Science and Technology 50 (1), 1–10.

Kross, S., Guo, P.J., 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019, p. 263. https://doi.org/10.1145/3290605.3300493.

Kuhlthau, C., 2004. Seeking Meaning: A Process Approach to Library and Information Services. Libraries Unlimited.

Lee, C.A., 2011. A framework for contextual information in digital collections. Journal of Documentation 67 (1), 95–143. https://doi.org/10.1108/00220411111105470.

Leonelli, S., 2016. Data-centric biology: A philosophical study. University of Chicago Press.

Ley, T., Seitlinger, P., 2015. Dynamics of human categorization in a collaborative tagging system: How social processes of semantic stabilization shape individual sensemaking. Computers in Human Behavior 51, 140–151. https://doi.org/10.1016/j.chb.2015.04.053.

Li, Y., Belkin, N.J., 2008. A faceted approach to conceptualizing tasks in information seeking. Inf. Process. Manage. 44 (6), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005.

Lillis, T., 2011. Legitimizing dialogue as textual and ideological goal in academic writing for assessment and publication. Arts and Humanities in Higher Education 10 (4), 401–432. https://doi.org/10.1177/1474022211398106.

Mahyar, N., Tory, M., 2014. Supporting communication and coordination in collaborative sensemaking. IEEE Trans. Vis. Comput. Graph. 20 (12), 1633–1642. https://doi.org/10.1109/TVCG.2014.2346573.

Maitlis, S., 2005. The social processes of organizational sensemaking. Academy of Management Journal 48 (1), 21–49.

Maitlis, S., Christianson, M., 2014. Sensemaking in organizations: Taking stock and moving forward. Academy of Management Annals 8 (1), 57–125.

Malakis, S., Kontogiannis, T., 2013. A sensemaking perspective on framing the mental picture of air traffic controllers. Applied Ergonomics 44 (2), 327–339.

Marchionini, G., 2006. Exploratory search: from finding to understanding. Commun. ACM 49 (4), 41–46. https://doi.org/10.1145/1121949.1121979.

Marchionini, G., Haas, S.W., Zhang, J., Elsas, J.L., 2005. Accessing government statistical information. IEEE Computer 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. Interaction 23 (3), 205–237. https://doi.org/10.1080/10447310701702352.

Mayernik, M., 2011. Metadata realities for cyberinfrastructure: Data authors as metadata creators. Available at SSRN 2042653.

Muller, M.J., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q.V., Dugan, C., Erickson, T., 2019. How data science workers work with data: Discovery, capture, curation, design, creation. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019, p. 126. https://doi.org/10.1145/3290605.3300356.

Neff, G., Tanweer, A., Fiore-Gartland, B., Osburn, L., 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. Big Data 5 (2), 85–97. https://doi.org/10.1089/big.2016.0050.

Pasquetto, I.V., Borgman, C.L., Wofford, M.F., 2019. Uses and reuses of scientific data: The data creators advantage. Harvard Data Science Review 1 (2). https://doi.org/10.1162/99608f92.fc14bf2d.Https://hdsr.mitpress.mit.edu/pub/jduhd7og

Pasquetto, I.V., Randles, B.M., Borgman, C.L., 2017. On the reuse of scientific data. Data Science Journal 16. https://doi.org/10.5334/dsj-2017-008.

Passi, S., Jackson, S.J., 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. PACMHCI 2 (CSCW), 136:1–136:28. https://doi.org/10.1145/3274405.

Peters, B.G., 2017. What is so wicked about wicked problems? a conceptual analysis and a research program. Policy and Society 36 (3), 385–396.

Poth, C.N., 2019. Rigorous and ethical qualitative data reuse: Potential perils and promising practices. International Journal of Qualitative Methods 18. https://doi.org/10.1177/1609406919868870.1609406919868870

Rieh, S.Y., Collins-Thompson, K., Hansen, P., Lee, H., 2016. Towards searching as a learning process: A review of current perspectives and future directions. J. Information Science 42 (1), 19–34. https://doi.org/10.1177/0165551515615841.

Robson, C., McCartan, K., 2016. Real world research. John Wiley & Sons.

Rogers, Y., Sharp, H., Preece, J., 2012. Interaction Design - Beyond Human-Computer Interaction, 3rd Edition. Wiley.

Rolland, B., Lee, C.P., 2013. Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In: Bruckman, A., Counts, S., Lampe, C., Terveen, L.G. (Eds.), Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, February 23-27, 2013. ACM, pp. 435–444. https://doi.org/10.1145/2441776.2441826.

Russell, D.M., 2003. Learning to see, seeing to learn: visual aspects of sensemaking. Human Vision and Electronic Imaging VIII, Santa Clara, CA, USA, January 20, 2003, pp. 8–21. https://doi.org/10.1117/12.501132.

Russell, D.M., Stefik, M., Pirolli, P., Card, S.K., 1993. The cost structure of sensemaking. Human-Computer Interaction, INTERACT '93, IFIP TC13 International Conference on Human-Computer Interaction, 24-29 April 1993, Amsterdam, The Netherlands, jointly organised with ACM Conference on Human Aspects in Computing Systems CHI'93, pp. 269–276. https://doi.org/10.1145/169059.169209.

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., Petras, V., 2019. The Definition of Reuse. The Definition of Reuse. Data Science Journal 18, 22. https://doi.org/10.5334/dsj-2019-022.

Shneiderman, B., 1996. The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996, pp. 336–343. https://doi.org/10.1109/VL.1996.545307.

Stasko, J.T., Görg, C., Liu, Z., 2008. Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization 7 (2), 118–132. https://doi.org/10.1057/palgrave.ivs.9500180.

Straus, A., Corbin, J., 1990. Basics of qualitative research: Grounded theory procedures and techniques.

Sutcliffe, A.G., Ennis, M., 1998. Towards a cognitive theory of information retrieval. Interacting with Computers 10 (3), 321–351. https://doi.org/10.1016/S0953-5438(98)00013-7.

Thomas, D.R., 2006. A general inductive approach for analyzing qualitative evaluation data. American journal of evaluation 27 (2), 237–246.

Walshe, R., Casey, K., Kernan, J., Fitzpatrick, D., 2020. Introduction to the special issue on: Big data/ai standardization in the journal of ict standardization. Journal of ICT Standardization 8 (2).

White, R.W., Roth, R.A., 2009. Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool Publishers. https://doi.org/10.2200/S00174ED1V01Y200901ICR003.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al., 2016. The fair guiding principles for scientific data management and stewardship. Scientific data 3.

Yalçin, M.A., Elmqvist, N., Bederson, B.B., 2018. Keshif: Rapid and expressive tabular data exploration for novices. IEEE Trans. Vis. Comput. Graph. 24 (8), 2339–2352. https://doi.org/10.1109/TVCG.2017.2723393.

Yoon, A., 2017. Data reusers' trust development. J. Assoc. Inf. Sci. Technol. 68 (4), 946–956. https://doi.org/10.1002/asi.23730.

Zhang, P., Soergel, D., 2014. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. JASIST 65 (9), 1733–1756. https://doi.org/10.1002/asi.23125.

Zhang, P., Soergel, D., 2020. Cognitive mechanisms in sensemaking: A qualitative user study. Journal of the Association for Information Science and Technology 71 (2), 158–171. https://doi.org/10.1002/asi.24221.