Collaborative Practices with Structured Data: Do Tools Support What Users Need?

Laura Koesten
University of Southampton, ECS
Southampton, UK
laura.koesten@gmail.com

Emilia Kacprzak
University of Southampton, ECS
Southampton, UK
e.kacprzak@theodi.org

Jeni Tennison
The Open Data Institute
London, UK
jeni@theodi.org

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

ABSTRACT

Collaborative work with data is increasingly common and spans a broad range of activities - from creating or analysing data in a team, to sharing it with others, to reusing someone else's data in a new context. In this paper, we explore collaboration practices around structured data and how they are supported by current technology. We present the results of an interview study with twenty data practitioners, from which we derive four high-level user needs for tool support. We compare them against the capabilities of twenty systems that are commonly associated with data activities, including data publishing software, wikis, web-based collaboration tools, and online community platforms. Our findings suggest that data-centric collaborative work would benefit from: structured documentation of data and its lifecycle; advanced affordances for conversations among collaborators; better change control; and custom data access. The findings help us formalise practices around data teamwork, and build a better understanding how people's motivations and barriers when working with structured data.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Information systems** → *Collaborative and social computing systems and tools*; Data exchange.

ACM ISBN 978-1-4503-5970-2/19/05.
https://doi.org/10.1145/3290605.3300330

KEYWORDS

Collaborative tools, structured data, Human Data Interaction

ACM Reference Format:
Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300330

1 INTRODUCTION

Working with structured data - that is, data that is organised as spreadsheets, tables or databases - has become a critical part of many professions [35, 60]. More and more of it is available online: for instance, in 2015 the Web Data Commons project extracted 233 million data tables from the Common Crawl [39], while the site Data Planet lists to date no less than 6.2 billion statistical datasets, many of which are public.

However, having data available does not always mean it can be used purposefully [22]. In this paper, we focus on the challenges that people face when working together on a data task, and explore how current information systems and tools support them in their work. Making sense of data requires cognitive effort to put it in context and relate it to other information sources, arguably more than in the case of text documents [1, 42, 67]. Data projects are often carried out in teams, drawing upon skills from several areas, including domain knowledge, statistics, data engineering and interaction design [8, 14]. In these projects, participants make decisions about how and how much of their work they document for the rest of the team, to allow them to reproduce what happened even when they do not share the same set of skills or experiences [14].

The digitisation of the workplace has led to remote work and teams fragmented across multiple locations [54], both in data-related contexts and beyond. More often than not, we collaborate with others without extensive direct contact. As

data becomes ubiquitous, we seamlessly switch between data producer and consumer roles, sometimes in the frame of the same project. However, while being in the same place is not a requirement anymore [3], building a common information space and understanding among collaborators is even more important for the success of a project [4, 16].

In this paper we define 'collaboration with data' on a **spectrum** that spans across a wide range of scenarios - from creating, processing or analysing data in an interdisciplinary team to sharing data with others to reusing someone else's data in a new context, with little to no interaction with the producer. Collaboration may involve anything from a small group of people sharing an office to large distributed teams to open communities where people join all the time. It includes core data science activities such as data sampling, exploratory data analysis, curation, interlinking and machine learning [34], as well as supporting ones such as data discovery, archiving or open data publishing.

We aim to understand collaborative practices along this spectrum in greater detail and compile a set of guidelines for designers of collaborative data technologies. We present an interview study with 20 participants in different professional roles and domains. Based on the interviews and related literature, we establish four high-level user needs for collaborative data tools: *documentation*, *conversation*, *change control* and *custom data access*. We then survey 20 systems that are commonly associated with data activities and compare their capabilities against the user needs. To manage the scope of the study, we consider only systems that support work on the actual data, as opposed to 'derivatives' such as descriptive statistics or charts built from a dataset. Our findings suggest that data-centric collaborative work would benefit from: a thorough, structured documentation of data and its lifecycle; advanced affordances for conversations among collaborators; better change control to help people understand the history of a dataset; and custom data access to overcome differences in technical skills in data projects.

2 BACKGROUND

In this section we define the spectrum of collaborative activities with data in more detail and discuss how collaborative sensemaking relates to activities across the entire spectrum.

Spectrum of Collaboration

Our notion of 'collaboration with data' is based on scenarios that have been considered in previous studies [24] or that have been mentioned by our interviewees. Scientists reuse the data of their peers to reproduce previous experiments. Developers jointly define benchmark datasets and gold standards that everyone can use to establish to compare related algorithms and approaches. We also consider data sharing and reuse activities as part of this spectrum. In those cases, collaboration then takes a less direct form, and sometimes involves more or less formal interactions between data producers and consumers via comments, questions and feedback [17]. In that context a dataset can be seen in some instances as a boundary object in the sense of [36], an artifact that can be created in a shared way and exists in multiple contexts and communities of practice simultaneously; and thus needs to satisfy potentially different informational requirements [36, 58].

Existing research on data and collaboration discusses specific types of data activities, such as collaborative data creation [19, 29, 63]; maintenance [27]; analysis [28, 41]; and visualisation [46]. In particular, the benefits of collaborative data analysis have been widely discussed in [22, 28, 46, 66] and there are several tools that support it, including Many Eyes [62] and Tableau Public[1]. There is also some work on data sharing as a form of collaboration. For example, [17] studies how users of a government dataset released in the public domain engage with the data to make sense of it.

Collaboration takes place in teams of different sizes and compositions [3]. It may be explicit or implicit [33] and include tighter or looser interactions, synchronous and asynchronous [50]. Data-centric collaboration takes all these forms and concerns activities that span from *co-creation* to *feedback* and *discussion* to data *reuse*. In the remainder of this section, we introduce each of them in greater detail before discussing the collaborative sensemaking aspects they involve.

Co-creation Direct collaboration includes co-creating, as well as jointly maintaining and analysing a dataset, such as when a team is working on the same Google spreadsheet. Other scenarios may rely on less explicit forms of collaboration such as in community-curated projects like OpenStreetMap,[2] or the Gene Ontology,[3] where contributions to the shared dataset are crowdsourced [27]. Participants take various roles. For instance, in Wikidata[4] only a specific group of editors is allowed to define or change critical parts of the data to minimise potential disruptions [49]. In other cases, data users are sought to add values to datasets which are already published online, collaborating with the data producers. This is supported in various ways, including pull requests on GitHub.

Feedback This involves making producers or publishers of data aware of errors or other limitations of their data, using a range of communication channels including social media. People comment upon the data, flag mistakes or missing values, or enquire about the way the data was collected and processed prior to publication [37]. For instance, a journalist might leave comments on a site publishing official statistics,

[1] https://public.tableau.com/s/
[2] https://www.openstreetmap.org
[3] http://www.geneontology.org/
[4] https://www.wikidata.org/

aiming to engage with the data owner. Data users commonly report on their experience with a dataset for a particular task in discussion forums, for example on Kaggle.[5]
Discussion We use the term to refer more generally to conversations within user communities that are not explicitly addressed at the data producer or publisher. The aim is rather to discuss the data and help each other understand and handle it. For example, statistical datasets released by public authorities are often shared and discussed in different communities, including journalists, researchers and the public. Some data portals have discussion forums connected to key datasets. In wiki-based data projects such as Wikidata each data entry is accompanied by a talk page. Conversations about popular datasets can also be found on general-purpose question answering sites such as Quora or on reddit.
Reuse The most implicit form of collaboration happens when data is reused by other people remotely and asynchronously, with a minimum of interaction between collaborators. This is a common occurrences in many areas, including science, data markets, and open government data portals [34]. Data can be reused in different contexts - for instance, historical data is used to assess trends over time or to measure the impact of interventions. There is extensive work on guidance and technologies to make data easier to reuse, possibly in combination with other sources [6]. However, the focus is mostly on helping the individual data producer or consumer rather than viewing reuse as an interaction between the two.

Independently of the activities it involves, collaboration requires some form of information sharing between participants to develop a common understanding of the task. Even when we consider data reuse, data producers and consumers are united in the act of engaging with the data artifact and its meaning [24]. Most challenges in publishing open data are concerned with finding formats and capabilities to make the data as useful as possible in as many contexts as possible [56]. A shared information space between producer and consumer cannot be defined a priori as the aim is to increase government accountability and foster innovation. The same is true about scientific datasets, which are shared and reused to enable reproducibility and advance science, often across domains [20, 23, 64]. Open collaboration is on the rise - participants work towards a common goal, but the cohesiveness of the group and the level of participation vary [3, 24]. This makes efforts to build a shared understanding of data between collaborators, such as this paper, even more timely and important.

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

Collaborative Sensemaking

Our collaboration spectrum is related with the concept of 'collaboration information behaviour' [31], which covers aspects such as identifying an information need together; retrieving; seeking; sharing; evaluating; synthesising; making sense of; and utilising information as potential collaborative activities. Collaborative sensemaking affects all of them, as multiple actors bring in different perspectives, but share a common sense of purpose [47].

However, literature on collaborative sensemaking focuses on work with documents or textual representations of information, or does not make the distinction between documents and data. As discussed in multiple studies [14, 17, 22, 29, 30], engaging with data involves complex processes and interaction patterns, which are not yet well understood. There is added complexity due to the fact that data on its own is difficult to interpret and needs context to create meaning from it, [8, 17, 18] and because of the skills involved, from knowing how to handle technical formats to understanding licences and terms of use. Tasks with data are often exploratory and iterative, with feedback loops at several stages in the workflow [14]. Sensemaking in secondary data use has been subject of studies in several domains, e.g. science [20, 69] and open government data [16, 43].

3 METHODOLOGY

We conducted in-depth interviews with data practitioners and supplemented the analysis with a review of selected tools currently used to collaborate with data.

In-depth Interviews

Recruitment The purpose of our sampling strategy was to include a spread of sectors and a wide range of skill sets and roles. Participants were recruited via targeted emails and social media and asked to fill out an online scoping survey. The survey was taken by 65 people. We chose only people who self-reported using data in their day-to-day work. We tried to include various domains and professional backgrounds to gain a broad overview and avoid unintended biases. The respondents identified as relevant at this stage were contacted via email to arrange an interview. The sample consisted of $n = 20$ data professionals, 17 male and 3 female (see Table1 of participants, their roles and countries of residence). They used both public (open) and proprietary data in different areas: environmental research, criminal intelligence, retail, social media, transport, education, geospatial information and smart cities. Most interviewees stated that their tasks with data vary greatly, but all participants described collaborative activities when asked about their tasks with data.

Ethics The study was approved by our institution's Ethical Advisory Committees. Informed consent was given by the participants.

Data collection and analysis We used semi-structured, in-depth interviews of circa 45 minutes, which were audio-recorded and subsequently transcribed. They were carried out via Skype or face-to-face. The interviews were organised around the participants' data-centric tasks; the search for new data; and the evaluation and exploration of potentially relevant data sources. Data types mentioned in the interviews included a range of structured and semi-structured data, such as spreadsheets, JSON, RDF graphs, geospatial data, survey data or sensor streams.

The responses were analysed using thematic analysis [51] with NVivo, a qualitative data analysis package. This paper focuses on those questions in the interviews which gave insights on collaborative tasks with data, including motivations and barriers for data sharing and reuse. The coding was done by one researcher, but to enhance reliability two senior researchers checked and discussed the analysis for a sample of the data. We applied two layers of coding to be able to look into the data at different levels of generality and from different viewpoints. For this work, we used deductive categories based on the data science process model [48] as a primary layer: *how people collaborate with data; tasks; issues and barriers; tools used, and how people make sense of data through interaction.* All codes potentially related to these categories were recorded.

For each of the categories we applied an inductive approach [59] to draw out emerging themes, which were then consolidated into four core high-level user needs: *documentation*, *conversation*, *change control* and *custom data access*. These were linked to a set of 11 capabilities, which we used to review the tools.

Tools Review

Our aim was to investigate the range of functionalities which support collaboration centred around data. We chose a cross-section of tools from responses to a one question Twitter survey and from the interviews. The tweet read as follows: *What are the tools and platforms you use to *collaborate* with data? Include tools for working with data, maintaining it, updating it, asking for feedback on and discussing data, publishing and reusing data* and was posted on two Twitter accounts.[6]

Due to the nature of the network of the Twitter accounts, the question was likely to be answered by data professionals.

[6]The Twitter accounts had at the time of the tweet more than 7*k* / 45*k* followers. The post reached 24.261 / 5179 impressions, 234 / 36 engagements, 64 / 7 retweets and 25 / 4 responses, see also https://support.twitter.com/articles/20171990.

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

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

The objective was to review a range of popular tools that people use to work with data; to either co-create, edit, publish, share or reuse it. This includes direct and indirect interactions, such as feedback and discussion, between multiple actors (e.g., data producers, data users etc.), with varying levels of input into the collaboration process. In addition to the responses collected via social media, we included the tools reported by our interview participants. We disregarded tools mentioned only once, and excluded programming languages; general-purpose tools handling documents rather than data; cloud storage; and tools which are focused on creating new data representations such as data exports to new formats, descriptive statistics and charts.

The remaining 20 tools (Table 2) can be roughly grouped into four categories: data portals, which offer access to datasets held by one or more parties; online tools that support collaborative forms of specific data activities such as shared spreadsheets or notebooks; wikis; and versioning tools. This is a general high-level categorisation only used for narrative purposes in this work. The tools were analysed individually. Detailed information about them and the number of responses we received can be found in the GitHub repository that accompanies this paper.[7]

This is by no means an exhaustive list of tools, but it was chosen to represent the diversity of the space and capabilities on an exemplary basis. To streamline the review, we derived 11 functional categories (Table 3) related to the four core user needs that came out of the thematic analysis of the interviews.

The review of the tools was performed by two researchers, during the first half of 2018, assessing how each of the tools

[7]https://github.com/confsubmission/chi2019

REVIEWED TOOLS
• *Data portals*: CKAN, Socrata (open data management systems); Mendeley data, Figshare (scientific data portals); ONS, Data.world, Kaggle (*Data portals*)
• *Online tools*: Google sheets, Excel Online, Airtables, Jupyter Notebooks, EtherCalc, Zoho Sheet, Floobits
• *Wiki-based platforms*: Wikidata, OpenStreetMaps
• *Versioning tools*: GitHub, Git, Dat project, OSF

Table 2: Tools reviewed

supports collaboration with data from the perspective of each of the 11 categories. The objective was to identify existing functionalities to provide an overview of the means by which tools support collaboration with data currently, and not to assess their performance or compare their usability. Where possible, the structure of the review was as follows: we set up a collaborative test project and each researcher individually went through the categories, noting the relevant functionalities available. Following this, both researchers consulted with one another and compared the two lists, discussing points of disagreement and referring to the tools, in order to produce a common, unified table, which is available in our GitHub repository.

Exemplary Discussion Analysis

To add depth to our findings about *conversations* we carried out an exemplary analysis of public discussions of datasets on two popular data-science sites: Kaggle and Data.world.[8] Both target public communication as a key functionality and discussions are encouraged and common between users. While there are other platforms that facilitate discussions, they often take place in private between direct collaborators. Our aim with this additional analysis was to build an understanding of the topics people raise in data-centric conversations, as these can point to requirements for collaborative data tools.

For Kaggle we took the 20 most trending datasets on 30/08/2018 that had a discussion thread attached to them, based on the category *hotness*, which, according to the Kaggle documentation, is a metric for how interesting and recent a dataset is. For Data.world we chose the 20 most recently updated datasets on 31/09/2018. We searched the site using the query *updated:2018-08-01 TO 2018-08-31 resourcetype:dataset* and excluded datasets with no discussion threads or non-English discussions.[9] The resulting sample of 20 datasets per platform contained 755 Kaggle comments, with an average of 18.5 unique users commenting per dataset (with a median of 10.5, min of 2 and max of 128); and 219 Data.world comments, with an average 5.5 unique users commenting per dataset (with a median of 3, min of 1 and max of 44). We analysed the discussion threads for each dataset thematically, using an inductive approach [59]. Results were interpreted in the context of the research goals to draw out user needs that could be supported by functionalities on collaborative tools for data.

[8] https://data.world/

[9] For Kaggle we used the first five discussion threads of each chosen dataset. For Data.world we analysed each thread as there was a lower number of separate discussions for each dataset.

4 FINDINGS AND DISCUSSION

We present the findings of our study and discuss their implications. The section is structured according to the four high-level user needs identified from the interviews: *documentation*, *conversation*, *change control* and *custom data access*. A comprehensive description of the functionalities that support these needs in the 20 tools we reviewed is available in our GitHub repository.

User Need: Documentation of Data & Its Lifecycle

Scope, forms and granularity of documentation Data work may involve reusing someone else's data without having access to additional information about the meaning of the data, its purpose or the way it was collected and processed prior to publication [50, 54]. Capturing this process can take many forms and includes text descriptions, annotations, metadata, previews and categories. Our interviewees commonly mentioned a perceived lack of documentation, referring to aspects as diverse as ambiguous variable names, unclear data provenance or under-specified methodologies. Participants discussed the difficulties of interpreting data with limited documentation and context:

> P10: What we do need is good quality data with some idea where it's come from, with some indication of what the data is saying. A good example of that would be when columns are numbered 1, 2, 3, 4, it doesn't really mean that much!

> P6: I think documentation is most frustrating, there's often data without documentation and fishing for this information is the hardest bit.

From the tools reviewed, only a few (e.g. Socrata, a data publishing software), provide a description of headers with their data in a structured way.

Documentation on granularity of the data emerged to be of key importance. Being able to understand minimum and maximum ranges and coverage were mentioned especially for geospatial and temporal data:

> P8: I'm looking at is what I would call the coverage of the data, so does it cover the geographic area I'm interested in? Or the time period that I'm interested in? Does it do that to the level of detail I need [..]?

Metadata and text descriptions Most online tools we reviewed collect and present metadata, as well as text descriptions of

DOCUMENTATION OF DATA & CREATION PROCESS	CUSTOM DATA ACCESS
• *Description and context* - everything that gives more information about the data, including its underlying methodology, e.g. textual description, structured metadata, README files, previews, etc. • *Annotation* - documentation, notes, tags • *Connections to other datasets* - datasets that the data was based on or are related	• *Access to subsets of the data* - creation of customisable slices of the data • *Data support* - type of data that is supported, such as tabular, geospatial data • *Visualisation* - is it possible to create custom visualisations • *Alternative formats* - is the data provided in alternative formats, does the tool support transformation between formats
CHANGE CONTROL	**CONVERSATION**
• *Subscription and notifications of updates* - including notification through emails, or via the tool to specific users or to a project • *Version control* - including a history of changes, ability of reverting back to previous version and tracking changes done by individual users	• *Conversation* - all functionalities that can enable users to communicate with each other, or with the creator or publisher of the data • *Feedback* - push requests, notifications, instruction of what to do when an issue with the data was found, contact information

Table 3: Categories used to review the tools

datasets, as it can be typically seen in most data portals. However, our findings suggest that this metadata is of limited use for human consumption and point to specific elements that are required to facilitate collaboration, such as the context of a dataset and with it the importance to document methodology.

Metadata often uses domain-specific vocabularies and technical formats. Its creation and understanding is challenging [20, 44]. Collaborators share specific aspects of it informally [69], hence the need for functionalities to capture and make use of it more effectively. While most tools provide capabilities to enter metadata, the review suggests a lack of structured mechanisms to formulate and produce detailed, useful metadata. For instance README files or descriptions allow for free-text input, but do not offer further support such as automatic extraction of specific metadata fields.

Datasets are commonly accompanies by a text description. Our interviewees reported that these descriptions do not necessarily increase understanding. There is no standardised way of collecting and presenting this information to users. Our findings suggest that collaborative data tools could benefit from supporting the creation of dataset summaries, which could be reviewed and revised by multiple participants to ensure they are as useful as possible for a variety of audiences.

Looking at other areas, the closest parallels can be found in code where README files are an established practice. GitHub for instance supports line-oriented descriptions. While not all data formats have canonical representations to allow for line-oriented documentation, our analysis of public data discussions point to several data aspects that users often ask for: variables, columns, cells etc.

Data lifecycle Many interview participants showed a strong interest in understanding methodological choices that were made in the creation of the dataset. This may involve the original purpose of the data, sample size, collection method, supporting tools and technologies etc. Understanding why and how the data was created helps assess potential risks connected to reuse:

> P1: There are always loads of limitations with every dataset that you use.
>
> P15: So although the data is a good quality, it's not really designed for my purpose so actually there's quite a lot of uncertainty and risk in that, it's still the best data we have but it's that knowledge of how it was, why it was created, against how I want to use it.
>
> P5: There may be lots of hidden implications about the values, the meaning of the records or the fields and if there is not very clear documentation about it, then it may be misleading [...].

Reporting on methodology is essential to the interpretation of data. For data reuse interviewees reported difficulties understanding data in its context, and emphasised that knowing the original purpose of the dataset is key to understanding what it could be used for. Users engage in communication with each other, or with publishers, to understand more about the decisions made during the creation of the dataset. Bannon has emphasised issues with reconstructing the intended meaning of information produced by others when the context of its creation is not documented [4].

Capturing methodological choices and the reasoning behind them is widespread in science, but does not yet appear to be a common practice for all types of data and data science activities. Insights in the thought processes that went into the creation of a dataset can be helpful for collaborators, which is supported by findings of [5, 44]. Some systems allow users to publish the code that was used to create or process

the data alongside the data itself (e.g. GitHub, Kaggle) or record the entire history of a dataset for each data entry (e.g. Wikidata).

Annotations are used to convey information about a resource or about relations between resources [10, 12]. From the tools we reviewed annotations were often supported at a dataset level; and sometimes at a row (e.g. GitHub, Airtables) or cell level (e.g. Google Sheets and EtherCalc). However, none of the tools provides a flexible and easy way of annotating the exact subset of the data a user might want to refer to. Predefined themes were provided in the form of tags in some tools (e.g. ONS, CKAN). Tags may also indicate the file type, such as *csv, html* (e.g. CKAN), technology used *python, java* (e.g. GitHub) or dataset content (Data.world). Annotations are supported in various forms in other areas - for web-based content [12, 53] e.g. via annotation servers. Web annotations build a layer of interactivity on top of the content and can be linked, shared and searched [53]. schema.org[10] is a vocabulary to annotate web pages - these annotations facilitate web search and ranking. Similar technology is available for other media, for instance videos [32], or in the form of sticky note systems. Methods to annotate data are emerging (for instance in specific domains such as in statistics with the SDMX model,[11] or for specific data types such as CSV)[12] but they are not used widely by data producers or supported by the tool developers.

User Need: Conversation

A common issue reported by most interviewees was the limited ability to communicate with the data producer, as well as with other users. For the reuse of data, participants mentioned the need for targeted discussions that would enable them to assess the data effectively.

> P5: I think that a pain point is someone, you work with someone on a project and then they say "Here's the data, take the data" and I think it's important to have a discussion with the other person, try to get them to communicate to you their understanding of the data because coverage.

While most of the points participants made in relation to conversation could theoretically be solved by more structured documentation, capturing everything in a way that serves all types of user tasks and their background knowledge is often unfeasible [23, 52]. When asked what they would want to communicate about, the main themes included the coverage of the data; hidden limitations or caveats attached to it; clarification of names and labels; details on the original purpose of the data; and cleaning choices.

[10]https://schema.org/

[11]https://sdmx.org/

[12]https://www.w3.org/TR/tabular-data-primer/#cell-annotations

Data conversations, whether within the user community or with the data producer, has value for all actors in the collaborative activity - knowledge about data use is communicated back to the producers and this feedback can be used to correct the data and better understand its possible use cases. One participant mentioned how conversation could be used to clarify caveats:

> P17: It is about all the stuff that looks like junk, that it looks like junk to you but it's not junk to someone else, because it might mark a very serious problem.

The interviews with data practitioners provide insights into *what* data users might want to talk about when working together with data. Our exemplary discussion analysis, carried out on comments from public discussion threads on Kaggle (K) and Data.world (D) helps us understand how such discussions look like. A common usage of the discussion threads is simply to congratulate other users on the data or a related data task. More interestingly the rest of the comments were centred around the following themes: *analysis*, *documentation*, *connections*, *issues*, *requests* and *context*. We review each of them in the following.

Analysis includes asking for help with data handling, pointing to already completed analyses, as well as giving or asking for feedback on them. Users also exchanged observations about patterns or trends in the data and related visualisations.

> K: It would be interesting to weigh each pollutant with the maximum recommended value of each pollutant per unit volume. You currently add all the pollutants, but it is very likely that this sum is a bad approximation since not all pollutants may be equally harmful.

Another popular topic was *documentation*, which includes asking for additional details on data variables, methodology for data collection, data provenance etc.

> K: I want to know what POS, AF_TGP and AF_EXAC actually stand for and what are they used for in genetic variant classification.
>
> D: Will you be adding descriptions to the data dictionary so it's easier to know what dimensions capture/represent? Kinda hard to decipher the abbreviations/labels.

Another theme emerged around *connections*, which includes references to other sources that could augment the dataset (documents, code, source files), pointers to other data or analyses, as well as relations between columns.

> K: Can you guide me how can we incorporate Local League Data to this as well to refine the results?
>
> K: I augmented the data set with geocode information. This allows for plotting things. Check it out.

A common discussion point was around *issues* with the data. People discuss outliers, raise questions about particular data

records, and point out errors and inconsistencies in the data. This includes data quality aspects such as missing values and machine readability.

> K: The Area field should be in square km, not square miles.
>
> D: Did anyone else realize there are a ton of dupes in the data set too? There are some that have 10+ records all for the same thing?

Requests are sometimes raised for different or more data. Users suggest other variables that could be added to extend the dataset (for instance, geospatial information) or ask for a different level of granularity for specific columns.

> D: Is there any way to get a "City/State" column in the dataset? Would be helpful for a geo map we're trying to build. Thanks!

Finally, some conversations are around the *context* of the data, pointing out potential biases. This includes questions about provenance and data collection, for instance about the area of where the particular data is from.

> D: I'm also looking to gain more insight on how this score is developed.

Although these themes cannot be a seen as an exhaustive representation of discussion topics around datasets online, they show the range and variety of topics collaborators find worth discussing, reinforcing the importance of feedback and conversation channels in data teamwork. Thorough documentation can provide explanations and to some extent context to data, however our findings suggest a role for communication and engagement capabilities, which could provide data producers with valuable guidance into how to improve on their documentation to address a wide range of data usage scenarios which is difficult to prepare for in advance.

Such capabilities are needed beyond the interaction between producers and consumers. This is reflected in the design of the tools we reviewed, which offer some level of conversation support for a wide range of data-centric activities

Comments Most tools support comments at the level of datasets. A subset of online tools allow users to attach comments to cells (e.g. Google Sheets, Excel Online) and fewer to rows (e.g. Airtables). GitHub facilitates discussions about code related to a specific line. The level of conversation currently supported does not appear to be sufficient for what people expressed as their needs. Participants wanted to highlight the exact subset of data they want to refer to (as it is possible for text in e.g. comments in Google Docs).

Filtering, sorting and categorising comments [66] is essential in data projects that involve multiple participants and interactions. Many existing solutions supporting conversations can be reused by either being incorporated in the tool, or by providing a link that directs users to a shared conversation space. For example, Wikipedia provides separate discussion pages, connected to an article, which help editors coordinate changes, discuss procedures, and get assistance from other editors [33]. Similarly, conversations in Kaggle take place in a discussion thread attached to a particular dataset. Real-time chats are, for instance, supported in Google Sheets, Airtables and Floobits. These can be beneficial especially for the co-creation of data and tight collaboration, but also for reuse (e.g. in the same organisation). [20] point out that conversation about data often happens alongside sharing of data or data fragments which was confirmed by our participants.

Feedback can be seen as a special type of communication. It can be provided in the form of issuing tickets (e.g. GitHub) or instructions of what to do when an error is found. Feedback enables correction of data and so can increase data quality and value. This can be supported in a structured (e.g. feedback forms or pull requests) or unstructured way (forums, third party communication). Public feedback can also save time by making others aware that a dataset is unsuitable for a specific task or by documenting known errors and allowing users to ask in the community for workarounds.

User Need: Change Control

The ability to access earlier versions of the dataset, for example before particular cleaning methods have been applied, was reported as a need by some participants. Versioning systems, which highlight differences between versions of a dataset as supported by GitHub or Dat project; and the opportunity to fork datasets were also discussed:

> P9: I'd like to apply some sort of GIT version control to that sort of approach, which work very well for text, images it would be horrible for and I don't know how that would work, just because of the file sizes and the spatial data.

The usefulness of notification services when a collaborator makes a change in the dataset, or when an updated version of a dataset is published was noted by participants.

In an example of open collaboration, one participant referred to notifications and described how in that instance all changes had to be approved by those in charge of the project.

> P13: You have to make a picture and then I get an email alert, I get information that somebody added a new plant and I look at the picture and then I say, okay, it's the plant or it's wrong.

Tasks in asynchronous collaboration are often split into sub-tasks that can be worked on in parallel [66]. Collaborators need to be aware of changes that are made by others [13], merge individual contributions and resolve potential conflicts. This aligns with the need to design for data transparency[21, 68] and with existing efforts to capture

provenance trails for data e.g.[40, 45]. Enabling change control is very common for textual documents. A range of methods are in place to track a history of changes, automatic merging, e.g.[15, 55], customisable notifications [9], or pull requests. Systems offer specialised merging interfaces that enable users to filter changes based on their type and help solve conflicts (e.g. [2]). Similar capabilities for data work are still in their infancy.

Version control is supported by GitHub and Git, which are not tailored for data, by Dat project, which focuses on providing access to data, but not exploring it, and by OSF, which is a research project management platform that manages access to previous versions of project files. Some tools provide a history of changes, but restoring older versions of the data or comparing versions is still tedious. Notifications were supported by more tools, e.g. Figshare.

For data, not many tools support automatic notifications when a dataset is updated or a new version gets released. Users who have reused data could benefit from such notifications as they can update their individual projects. Users who have co-created a dataset might want to be notified if one collaborator made a change, such as in Google Sheets for tagged users. This highlights the fluid boundary between data consumers and producers. Wikidata is an example where every data operation is recorded and can be analysed.

From the tools we reviewed, GitHub shows differences between versions of datasets at a row level, which works for some types of data formats, but not others. However, comparisons between versions of a datasets should also be provided at a column and cell levels. Other approaches propose systems which record a history of the process a user followed when exploring the data (e.g. Vis trails[13]). They visually illustrate a sequence of steps while inspecting the data and support a visual representation of differences of graphics (or their underlying data) [46]. The area of visual exploration was outside the scope of this work, however it points to the need to find meaningful representations and experiences with data evolution.

User Need: Custom Data Access

Our findings suggest a number of functionalities that can improve the ease of use of data. This benefits collaboration because teams working with data have shown to be highly heterogeneous [14]. We include the ability to access subsets of the data through the tool, data being provided in alternative formats, filtering and plot-generating functionalities, as well as custom visualisations. Additionally, being able to access the *'raw' data*, not just a aggregation or a statistical analysis of the data, was reported as a common need by participants:

> P2: It does depend on the task but usually I will almost immediately get rid of the dataset if it is already turned into statistics. In other words, I want raw data.

Being able to easily plot data, access subsets of data or create relations between tables facilitates a shared understanding of data by collaborators. Some participants mentioned the need for specific data formats, either due to their skill set or depending on what they want to use the data for:

> P16: First thing is make sure I can download the format and that I can read the format, probably in Excel because I'm fairly unsophisticated [..].

Several tools support data downloads in multiple formats (Google Sheets, Excel Online), via APIs (Wikidata), or work directly through programming languages which let one change the file format (e.g. Jupyter, Data.world). Some offer functionalities for downloading subsets of the data (e.g. CKAN, Data.world) and filters.

Analytical features such as graph plots, custom visualisations [38] and meaningful pivot tables emerged as important both in the interviews and in the exemplary discussion analysis. At the same time, allowing input mechanisms on different levels can be seen as supporting custom data access. Such capabilities are supported by OpenStreetMap, as users have the possibility to enter data through a graphical user interface and edit the map directly. Fusion tables [26] allows map-based plotting as an alternative input form [57]. Tools could provide support for the creation of different data formats and representations. These could either be shared between collaborating data users, or with the data owner.

A common time-consuming task in data-related activities is cleaning data and understanding problematic rows. Being able to share code that filters out noise and access a cleaned subset of the data can benefit collaboration if done in a transparent way.

Both the analysis of discussion threads on Kaggle and Data.world, as well as the interviews have shown that users can be interested in merging datasets, for instance to add reference points, such as zip codes to make existing data easier to analyse or more relevant.

5 IMPLICATIONS

We present guidelines according to the four high-level user needs resulting from the thematic analysis of the interviews (Figure 1). Their focus is on how specific functionalities in data tools can support people to work together with data across the spectrum of collaborative activities. Some of the suggested functionalities (e.g. dataset overviews, history of changes) could be considered at design time, while others (e.g. missing values, original purpose of data) are more dependent on the data producer. In these, tools need to focus on support and semi-automation.

[13] www.vistrails.org

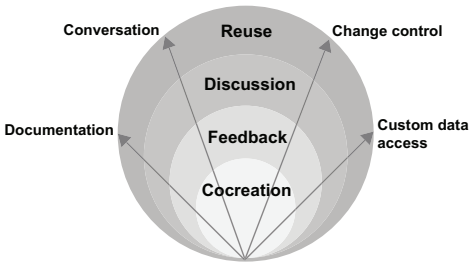

Figure 1: User needs are relevant across a spectrum of collaborative activities from tight to loose types of collaboration. Arrows represent different user needs based on findings from this study; circles represent collaborative activities with data

Guidelines for create more useful documentation. Functionalities supporting users in understanding data and its context are needed along the whole spectrum of collaboration, from co-creation to reuse of data.

The less explicit the collaboration is, the more users have to rely on documentation, due to the lack of background information, and the inability to gain additional insight through discussions and exchange. However, even in co-creation, documentation is crucial to maintain consistency, record decisions, and develop a common understanding between collaborators. If we look at data as a source that can be used in different social, cultural and organisational contexts for different purposes even within a team, similar to the notion of boundary objects [36], the need to augment data with additional contextual information in the form of documentation becomes obvious. In line with existing literature (e.g. [23]) we describe guidelines for collaborative data tools that aim to support the workflow of documentation tasks.

We recommend supplying more **contextual information** in an automated or semi-automated way. This can be in the form of metadata, which can for instance include seasonal or other external events that influenced the data, which is likely to be especially useful on a column level. Where applicable, we suggest adding information about changes in sample size, temporal or geographical coverage; units of measurements; and definition of categories to reduce ambiguities. Data producers need to be guided through the process of supplying contextual information where automatic extraction is not possible yet.

Creating an **overview of the dataset** can be achieved through summarising statistics or text summaries, which can be automatically generated using natural language generation approaches. Data visualisations and previews of core subsets of a dataset can further support users in understanding aspects such as content, coverage, and granularity of data. Temporal and geospatial boundaries of a dataset can be enabled as timelines and map views.

We found evidence that data collaborations require **annotations at different levels**, beyond the level of an entire dataset to cover columns, rows, and individual values. Data portals should provide a description of headers provided by the creators of the data. More research needs to be done in understanding and characterising custom dataset regions and in managing annotations that apply to one or several levels of the data.

We propose providing information about the creation of the data in a formalised way, as a **creation strategy** of the data. That includes everything that helps others to understand methodological choices during the creation of the data, such as steps to reproduce, experiment design, sample size, etc. The provision of code used to create, clean or analyse the data, additionally supports people's understanding of its creation.

A description of the **original purpose of the data** should be provided as well as any known **limitations or caveats** of the data.

Guidelines to facilitate conversation. Our findings suggest that conversation among data users, as well as between users and producers is not well supported by the tools under review. Tools should include capabilities for embedded conversation and integrate with other, general-purpose channels and facilitate both synchronous and asynchronous discussions. This is paramount for a shared understanding to emerge and for collaborative sensemaking [50]. Communication enables a necessary transfer of knowledge to use data that is not easily documented [5, 22].

We recommend allowing **comments** on different subsets of the data, such as a whole dataset, columns and rows, individual cells and a range of cells. Tools could facilitate **targeted discussions** about the data, including third-party conversation, by reusing existing conversation channels. Providing direct **feedback channels** between collaborators can increase the quality of the data,[14] and make collaboration more efficient [24]. We recommend providing contact information, as well as instructions on what to do when an issue with the data is found. The possibility to flag errors or ambiguities by opening a public issue, pull requests or conversation channels provide further opportunities for feedback. For direct types of collaboration, we suggest real-time chats that support **sending data snippets** or screen sharing.

Guidelines for better change control. Change control enables collaborators to track and understand the evolving nature of collaborative data projects. The benefits change and version control are well-established for code and text and have been shown to increase trust and minimise barriers to working

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

together with others [**?**]. Our findings suggest this applies to the whole spectrum of collaborative activities with data.

Depending on the type of collaboration, we recommend providing **version control** for co-creation and maintenance of data. For reuse of data we recommend providing an easily accessible **history of changes and other versions of the dataset**. In both instances, users can benefit from receiving **notifications of updates** when a change is made. In co-creation these notifications should be possible on a more granular level. For data reuse, users would benefit by being notified of new releases of the data or major corrections or changes in the metadata. Hence users should be able to choose for which type of changes or when a notification is sent. **Pull requests** should be possible in order to increase data quality and provide feedback. Access to these mechanisms should be customisable to third parties, allowing data users to decide the level of detail collaborators will be notified of. More research is needed to understand the scope and representation of these reports for different task contexts and audiences, similar to approaches from software or wiki-based projects [61].

Guidelines for custom data access. Custom data access supports collaboration across the spectrum mainly by allowing easier access and more user friendly and seamless workflows with data. As people are collaborating with data in different contexts, they need different data formats and representations. This can include custom visualisations, as well as the ability to select a subset of the data. Access to data in alternative formats allows users with different skill sets and preferences to work with data.

The more data consumers can explore and understand data in a way that matches their personal skill sets and mental models, the easier it will be to collaborate amongst heterogeneous teams. There are many ways to work with data even amongst data professionals. Data as an information source is uniquely mutable and filterable, arguably more than sources like documents or code. This explains why functionalities supporting personalised access to the data representation are currently still limited, or come with a high technical barrier. An open area of research in this context is related to the best ways to communicate additional information, analyses and observations about a dataset in the same customised, accessible way, including quality statements (e.g. missing data, noisy data, uncertain data) and notes, comments and feedback from users. Furthermore, tools need to become better at providing a wide range of mechanisms and experiences to **discover new data** and **merge multiple datasets**, which do not make strong implicit assumptions about the technical skills of the users.

Exploring synergies with related initiatives. The guidelines identified in this paper complement existing efforts such as Share PSI 2.0 and the W3C's Data on the Web Best Practices working group,[15] which offer technical advice for data producers to release their data in ways that facilitate reuse.

A number of initiatives have developed collaborative infrastructure to support interdisciplinary data sharing e.g. DataNet [38] or data archives for access and preservation of digital data [7]. Our work adds to them by exploring collaboration practices that leverage common tools such as wikis and web-based table management software.

High-level efforts for standardising data sharing practices, such as the FAIR data principles [65] or the five stars of linked open data[16] focus on interoperability as a means to encourage reuse rather than on user needs and collaborative experiences. For instance, the FAIR principles promote standardisation, but there is no further guidance on what metadata to capture to support a user's understanding. To limit the burden on the data publisher we would like to see solutions for implementation of functionalities that e.g. support documentation of provenance and the use of controlled vocabularies for metadata from a user point of view.

Recently ideas for supplementary information (datasheets) that should be shared together with datasets have been suggested by [25]. While the development towards meaningful standardisation for data publishing is very relevant, in the context of this work we focus on how different elements of supplementary information can be communicated and facilitated through tool functionalities rather than through extensive mandatory documentation. We further believe that the varied contexts of data collaborators can potentially make information relevant that the data publisher might not be able to anticipate.

Limitations

The majority of interview participants were male ($n = 17$) and working in the UK ($n = 16$). We interviewed a particular type of professional, though working in a wide range of sectors and roles. As this was meant as an initial study into professional practices with data, having a large number of participants per sector was less of a priority. For the tools reviewed, we are aware that these only represent a subset of the options available. While we have selected our sample carefully, it is clear that this list cannot be exhaustive and that more in-depth studies are needed for particular categories of tools supporting different user needs in different collaboration contexts. We are also aware of the limitations of our choice of review procedure, as it does not reflect how the individual functionalities are used in reality. However, as we combined them with the interviews, we believe that for the purpose of this work a more structured approach would

[15] https://www.w3.org/2013/share-psi/bp/

[16] https://www.w3.org/DesignIssues/LinkedData.html

not have resulted in many more insights. We discussed the results of a small scale additional analysis of discussions about datasets on Kaggle and Data.world. This is done to add depth [11] to the discussion of the user need 'conversation' but is done on an exemplary basis for two data platforms. This cannot therefore be seen as representative for discussions on all types of collaborative data tools. Due to the nature of Kaggle as a platform for a machine learning community the discussion focuses on data analysis and the unit of data, related code and its output. This might not be representative for the general population but it is likely to represent the more technical spectrum of data users. We included discussions on Data.world as the focus of the platform is more about sharing and communicating about data and less on pure data analysis.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we identify user needs when collaborating with data and how those are supported by existing tools. We discuss functionalities that support collaboration across a spectrum of activities with data for four core user needs. Our findings contribute to formalising collaborative practices with data and to a better understanding of peoples' motivations and barriers. The results of this work can be used for the development of functionalities that support data collaboration, from co-creation to sharing and reuse, to offer more effective user experiences, independently of domain and data type. This can improve developer experience and help inform the design of data-centric tools to reflect current work practices, which are often a collaborative effort.

We believe there is a gap in research that investigates how existing methods, which facilitate collaboration for documents or code, could be applied more efficiently in collaboration with data. At the same time, we believe these are unique challenges and needs when working with data due to its structured and mutable nature.

Key areas for future work include investigating the actual use of specific functionalities in collaborative activities by users and analysis of online discussions around datasets on a larger scale. This can include exploring whether features of datasets are predictable, based on the conversations that happen around them.

We hypothesise that similar high-level user needs apply when working with various types of dataset from different domains, which come with specific task types. For instance machine learning datasets, a rapidly growing area in which the reuse of data is very common; or geospatial datasets, which use specific formats, which would need to be validated in further research.

The results of our work and future research exploring conversations and collaborative work practices with data can further be used for the development of user-centred and transparent reporting practices for data sharing and reuse. Better tools that enable collaboration with data can help engage communities, reduce efforts in data management and create useful and high-quality datasets that can be used across sectors and domains. At the same time such insights can inform the development of collaborative data tools, for instance by feeding into the design of APIs for data access or arguing for comprehensive solutions to publish and manage data annotations at multiple levels.

REFERENCES

[1] Michael J. Albers. 2015. Human-Information Interaction with Complex Information for Decision-Making. *Informatics* 2, 2 (2015). https://doi.org/10.3390/informatics2020004

[2] Arwa Alsubhi and Ethan Munson. 2016. Design and Usability Testing of a User Interface for Three-way Document Merging. In *Proceedings of the 4th International Workshop on Document Changes: Modeling, Detection, Storage and Visualization (DChanges '16)*. ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/2993585.2993590

[3] Georgia Bafoutsou and Gregoris Mentzas. 2002. Review and Functional Classification of Collaborative Systems. *Int. J. Inf. Manag.* 22, 4 (Aug. 2002), 281–305. https://doi.org/10.1016/S0268-4012(02)00013-0

[4] Liam Bannon. 2000. Understanding common information spaces in CSCW. In *Workshop on Cooperative Organisation of Common Information Spaces, Technical University of Denmark.*

[5] Jeremy P. Birnholtz and Matthew J. Bietz. 2003. Data at Work: Supporting Sharing in Science and Engineering. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work (GROUP '03)*. ACM, New York, NY, USA, 339–348. https://doi.org/10.1145/958160.958215

[6] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*. IGI Global, 205–227.

[7] Christine L. Borgman, Andrea Scharnhorst, and Milena S. Golshan. 2018. Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse. *CoRR* abs/1802.02689 (2018). arXiv:1802.02689 http://arxiv.org/abs/1802.02689

[8] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3645–3656. https://doi.org/10.1145/3025453.3025738

[9] A. J. Bernheim Brush, David Bargeron, Jonathan Grudin, and Anoop Gupta. 2002. Notification for Shared Annotation of Digital Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 89–96. https://doi.org/10.1145/503376.503393

[10] A. J. Bernheim Brush, David Bargeron, Anoop Gupta, and J. J. Cadiz. 2001. Robust Annotation Positioning in Digital Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 285–292. https://doi.org/10.1145/365024.365117

[11] Alan Bryman. 2006. Integrating quantitative and qualitative research: how is it done? *Qualitative research* 6, 1 (2006), 97–113. https://doi.org/10.1177/1468794106058877

[12] J. J. Cadiz, Anop Gupta, and Jonathan Grudin. 2000. Using Web Annotations for Asynchronous Collaboration Around Documents. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 309–318. https://doi.org/10.1145/358916.359002

[13] John M. Carroll, Mary Beth Rosson, Gregorio Convertino, and Craig H. Ganoe. 2006. Awareness and Teamwork in Computer-supported Collaborations. *Interact. Comput.* 18, 1 (Jan. 2006), 21–46. https://doi.org/10.1016/j.intcom.2005.05.005

[14] Joohee Choi and Yla Tausczik. 2017. Characteristics of Collaboration in the Emerging Practice of Open Data Analysis. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 835–846. https://doi.org/10.1145/2998181.2998265

[15] Stephen M. Coakley, Jacob Mischka, and Cheng Thao. 2015. Version-Aware Word Documents. In *Proceedings of the 2Nd International Workshop on (Document) Changes: Modeling, Detection, Storage and Visualization (DChanges '14)*. ACM, New York, NY, USA, Article 2, 2:1–2:4 pages. https://doi.org/10.1145/2723147.2723152

[16] Tim Davies and Mark Frank. 2013. 'There's No Such Thing As Raw Data': Exploring the Socio-technical Life of a Government Dataset. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*. ACM, New York, NY, USA, 75–78. https://doi.org/10.1145/2464464.2464472

[17] Tim Davies and Mark Frank. 2013. 'There's No Such Thing As Raw Data': Exploring the Socio-technical Life of a Government Dataset. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*. ACM, New York, NY, USA, 75–78. https://doi.org/10.1145/2464464.2464472

[18] Brenda Dervin. 1997. Given a context by any other name: Methodological tools for taming the unruly beast. *Information seeking in context* 13 (1997), 38.

[19] AnHai Doan, Raghu Ramakrishnan, and Alon Y Halevy. 2010. Mass collaboration systems on the world-wide web. *Commun. ACM* 54, 4 (2010), 86–96.

[20] Paul N Edwards, Matthew S Mayernik, Archer L Batcheller, Geoffrey C Bowker, and Christine L Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41, 5 (2011), 667–690. https://doi.org/10.1177/0306312711413314

[21] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces, IUI 2018, Tokyo, Japan, 2018*. 211–223. https://doi.org/10.1145/3172944.3172961

[22] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Marie Fassett, and Sarah Duda. 2016. Storytelling with Data: Examining the Use of Data by Non-Profit Organizations. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1273–1283. https://doi.org/10.1145/2818048.2820068

[23] Ixchel M. Faniel and Trond E. Jacobsen. 2010. Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work* 19, 3-4 (2010), 355–375. https://doi.org/10.1007/s10606-010-9117-8

[24] Andrea Forte and Cliff Lampe. 2013. Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist* 57, 5 (2013), 535–547. https://doi.org/10.1177/0002764212469362

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[26] Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. 2010. Google Fusion Tables: Web-centered Data Management and Collaboration. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD '10)*. ACM, New York, NY, USA, 1061–1066. https://doi.org/10.1145/1807167.1807286

[27] Mordechai (Muki) Haklay and Patrick Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7, 4 (Oct. 2008), 12–18. https://doi.org/10.1109/MPRV.2008.80

[28] Jeffrey Heer and Maneesh Agrawala. 2008. Design Considerations for Collaborative Visual Analytics. *Information Visualization* 7, 1 (March 2008), 49–62. https://doi.org/10.1145/1391107.1391112

[29] Fawzi Fayez Ishtaiwa and Ibtehal Mahmoud Aburezeq. 2015. The impact of Google Docs on student collaboration: A UAE case study. *Learning, Culture and Social Interaction* 7 (2015), 85–96. https://doi.org/10.1016/j.lcsi.2015.07.004

[30] Ruogu Kang, Aimée A. Kane, and Sara B. Kiesler. 2014. Teammate inaccuracy blindness: when information sharing tools hinder collaborative analysis. In *Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, 2014*. 797–806. https://doi.org/10.1145/2531602.2531681

[31] Arvind Karunakaran, Madhu C. Reddy, and Patricia Ruma Spence. 2013. Toward a model of collaborative information behavior in organizations. *Journal of the American Society for Information Science and Technology* 64, 12 (2013), 2437–2451. https://doi.org/10.1002/asi.22943

[32] Isaak Kavasidis, Simone Palazzo, Roberto Di Salvo, Daniela Giordano, and Concetto Spampinato. 2014. An Innovative Web-based Collaborative Platform for Video Annotation. *Multimedia Tools Appl.* 70, 1 (May 2014), 413–432. https://doi.org/10.1007/s11042-013-1419-7

[33] Aniket Kittur and Robert E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 37–46. https://doi.org/10.1145/1460563.1460572

[34] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1277–1289. https://doi.org/10.1145/3025453.3025838

[35] Steve Lavalle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. 2011. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review* 52, 2 (2011).

[36] Charlotte P. Lee. 2007. Boundary Negotiating Artifacts: Unbinding the Routine of Boundary Objects and Embracing Chaos in Collaborative Work. *Computer Supported Cooperative Work (CSCW)* 16, 3 (2007), 307–339. https://doi.org/10.1007/s10606-007-9044-5

[37] Gwanhoo Lee and Young Hoon Kwak. 2011. Open government implementation model: a stage model for achieving increased public engagement. In *Proceedings of the 12th Annual International Conference on Digital Government Research, DG.O 2011, College Park, MD, USA, 2011*. 254–261. https://doi.org/10.1145/2037556.2037598

[38] Jae W Lee, Jianting Zhang, Ann S Zimmerman, and Angelo Lucia. 2009. DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. *AIChE Journal* 55, 11 (2009), 2757–2764.

[39] Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*. https://doi.org/10.1145/2872518.2889386

[40] John Lucocq. 2012. Can data provenance go the full monty? *Trends in cell biology* 22, 5 (2012), 229–230. https://doi.org/10.1016/j.tcb.2012.03.001

[41] Narges Mahyar and Melanie Tory. 2014. Supporting Communication and Coordination in Collaborative Sensemaking. *IEEE Trans. Vis. Comput. Graph.* 20, 12 (2014), 1633–1642. https://doi.org/10.1109/

TVCG.2014.2346573

[42] Gary Marchionini. 1997. *Information seeking in electronic environments*. Number 9. Cambridge university press. https://doi.org/10.1017/CBO9780511626388

[43] G. Marchionini, S. W. Haas, J. Zhang, and J. Elsas. 2005. Accessing government statistical information. *Computer* 38, 12 (Dec 2005), 52–61. https://doi.org/10.1109/MC.2005.393

[44] Matthew Mayernik. 2011. Metadata realities for cyberinfrastructure: Data authors as metadata creators. (2011).

[45] Paolo Missier, Bertram Ludäscher, Saumen C. Dey, Michael Wang, Timothy M. McPhillips, Shawn Bowers, Michael Agun, and Ilkay Altintas. 2012. Golden Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository. *IJDC* 7, 1 (2012), 139–150. https://doi.org/10.2218/ijdc.v7i1.221

[46] Benoît Otjacques, Mickaël Stefas, Maël Cornil, and Fernand Feltz. 2012. Open Data Visualization Keeping Traces of the Exploration Process. In *Proceedings of the First International Workshop on Open Data (WOD '12)*. ACM, New York, NY, USA, 53–60. https://doi.org/10.1145/2422604.2422612

[47] Sharoda A. Paul and Madhu C. Reddy. 2010. Understanding together: sensemaking in collaborative information seeking. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010, Savannah, Georgia, USA, February 6-10, 2010*. 321–330. https://doi.org/10.1145/1718918.1718976

[48] Hanspeter Pfister and Joe Blitzstein. 2015. cs109/2015, Lectures 01-Introduction. https://github.com/cs109/2015/tree/master/Lectures.

[49] Alessandro Piscopo and Elena Simperl. 2018. Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 141.

[50] Yan Qu and Derek L Hansen. 2008. Building shared understanding in collaborative sensemaking. In *Proceedings of CHI 2008 Sensemaking Workshop*.

[51] Colin Robson and Kieran McCartan. 2016. *Real world research*. John Wiley & Sons.

[52] Betsy Rolland and Charlotte P. Lee. 2013. Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In *Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, 2013*. 435–444. https://doi.org/10.1145/2441776.2441826

[53] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2016. Web Annotation Data Model. https://www.w3.org/TR/2016/CR-annotation-model-20160906/

[54] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)* 1, 1 (01 Mar 1992), 7–40. https://doi.org/10.1007/BF00752449

[55] Svante Schubert, Sebastian Rönnau, and Patrick Durusau. 2015. Interoperable Document Collaboration. In *Proceedings of the 2Nd International Workshop on (Document) Changes: Modeling, Detection, Storage and Visualization (DChanges '14)*. ACM, New York, NY, USA, Article 6, 6:1–6:4 pages.

[56] Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, et al. 2012. Linked open government data: Lessons from data. gov. uk. *IEEE Intelligent Systems* 27, 3 (2012), 16–24.

[57] Robert Shepard. 2014. Map-based Input with Google Fusion Tables. *Cartographic Perspectives* 75 (2014), 49–54.

[58] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

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

[60] Stefaan Verhulst and Andrew Young. 2016. *Open data impact when demand and supply meet*. Technical Report. GOVLAB.

[61] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 575–582. https://doi.org/10.1145/985692.985765

[62] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1121–1128. https://doi.org/10.1109/TVCG.2007.70577

[63] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. https://doi.org/10.1145/2629489

[64] Jillian C Wallis, Elizabeth Rolando, and Christine L Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS one* 8, 7 (2013), e67332. https://doi.org/10.1371/journal.pone.0067332

[65] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016). https://doi.org/10.1038/sdata.2016.18

[66] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: Structured Support for Collaborative Visual Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3131–3140. https://doi.org/10.1145/1978942.1979407

[67] Max L. Wilson, Bill Kules, m. c. schraefel, and Ben Shneiderman. 2010. From Keyword Search to Exploration: Designing Future Search Interfaces for the Web. *Foundations and Trends in Web Science* 2, 1 (2010), 1–97. https://doi.org/10.1561/1800000003

[68] Alyson Leigh Young and Wayne G. Lutters. 2015. (Re)defining Land Change Science through Synthetic Research Practices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, 2015*. 431–442. https://doi.org/10.1145/2675133.2675183

[69] Ann Zimmerman. 2007. Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *Int. J. on Digital Libraries* 7, 1-2 (2007), 5–16. https://doi.org/10.1007/s00799-007-0015-8