Dirty Data in the Newsroom: Comparing Data Preparation in Journalism and Data Science

Stephen Kasica
kasica@alumni.ubc.ca
The University of British Columbia
Vancouver, BC, Canada

Charles Berret
charles.berret@liu.se
Linköping University
Norrköping, Sweden

Tamara Munzner
tmm@cs.ubc.ca
The University of British Columbia
Vancouver, BC, Canada

ABSTRACT

The work involved in gathering, wrangling, cleaning, and otherwise preparing data for analysis is often the most time consuming and tedious aspect of data work. Although many studies describe data preparation within the context of data science workflows, there has been little research on data preparation in data journalism. We address this gap with a hybrid form of thematic analysis that combines deductive codes derived from existing accounts of data science workflows and inductive codes arising from an interview study with 36 professional data journalists. We extend a previous model of data science work to incorporate detailed activities of data preparation. We synthesize 60 dirty data issues from 16 taxonomies on dirty data and our interview data, and we provide a novel taxonomy to characterize these dirty data issues as discrepancies between mental models. We also identify four challenges faced by journalists: diachronic, regional, fragmented, and disparate data sources.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**.

KEYWORDS

data journalism, data science, data wrangling, data cleaning, thematic analysis

ACM Reference Format:
Stephen Kasica, Charles Berret, and Tamara Munzner. 2023. Dirty Data in the Newsroom: Comparing Data Preparation in Journalism and Data Science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581271

1 INTRODUCTION

A large body of research addresses data preparation in data science, where studies show the work of wrangling, cleaning, and otherwise preparing data for analysis can be responsible for 80% of the time and cost of data warehousing projects [17]. We seek to understand how closely the abundant research on data scientists applies to data journalists, who use computational tools and techniques to leverage data for the production of news. Data journalists often gather and analyze datasets using structured, quantitative information as an additional source to fact check claims, supplement material gathered through traditional methods like interviewing, and support in-depth investigations. Previous studies have suggested a close relationship exists between data journalists and data scientists with regards to tool usage, data sources, and work practices [39, 71].

Although we see clear parallels between the activities of these two groups, the relationship between data journalism and the adjacent field of data science requires further study. While anecdotal evidence points to the prevalence and difficulty of data preparation in data journalism [31], we lack empirical data on the specific challenges faced by data journalists in comparison to data scientists. A significant body of interview-based research has attempted to understand the daily workflows of data scientists by studying the lived experience of practitioners across diverse domains [2, 32, 35, 37, 40, 41, 46, 52, 53, 69, 78, 79, 82, 84]. However, none of these studies include a journalist among their participants. In this paper, we will use *data worker* as an umbrella term to mean both data scientists and data journalists, particularly when emphasizing their commonalities. Considering the needs of data workers, broadly construed, may help to narrow the research to reporting gap [74], the frequent ineffectiveness of tools and techniques proposed by computer scientists when applied to the problems that journalists actually encounter.

We examine the extent to which accounts of data preparation among data scientists match the preparatory process of data journalists, featuring a semi-structured interview study with 36 data journalists analyzed qualitatively. Our hybrid thematic analysis incorporates both deductive *a priori* codes and inductive *a posteriori* codes [77]. We construct an initial *a priori* codeset by analyzing 16 research papers on data science workflows that address data preparation, allowing us to note where our *a posteriori* interview findings diverge or overlap with previous work. We also analyze 16 taxonomies of dirty data from the database and data warehousing literature to compare and contrast the conventional wisdom in those fields with our interview findings. Our work provides four contributions. First, the results of a semi-structured interview study with 36 data journalists: This interview study addresses a longstanding research gap, offering a novel perspective of data journalism that contributes to a more complete and pluralistic understanding of data work as a whole [19]. The results of this study are two-fold: a set of activities undertaken by data journalists during data preparation, and the set of data quality issues they face. We situate these results within the research literature through additional analysis, leading to three additional formalism contributions:

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581271

Second, an augmented model of preparatory activities: In Section 4, we present a synthesis set of 23 data preparation activities. We combine and consolidate the activities revealed in our interview study with those articulated in previous work, which we identified through a thematic analysis of 16 papers documenting data science workflows. We situate these activities by extending the process model of data science work proposed by Crisan et al. [16] to an additional level of fine-grained activities, grouping them according to preparation subprocesses (initiate, gather, create, profile, wrangle) and communication subprocesses (disseminate and document).

Third, a new model-discrepancy taxonomy of dirty data issues: In Section 5, we propose a new taxonomy to classify 60 dirty data issues as discrepancies between mental models among different data workers. Our dirty-data taxonomy features a two-dimensional design space, with an axis of four data objects (table, item, attribute, and value) and an axis of six data qualities (accuracy, completeness, form, granularity, relation, and semantics). It reconciles the top-down, domain-focused perspective of our practitioner participants and the bottom-up, theory-focused perspective commonly used by computer science researchers to define and describe data issues. We incorporate the latter by analyzing 315 instances of dirty data documented in 16 previous data warehousing papers.

Finally, four challenges in multi-table data integration: In Section 6, we identify four data integration challenges: *regional* inconsistencies from independent spatially dispersed data sources, *diachronic* inconsistencies from tables recording the same phenomena that evolve over time, *fragmented* tables containing different yet related items that must be re-assembled, and *disparate* tables that are topically dissimilar yet must be related. We identify these challenges from our analysis of the multi-table data integration nightmare stories described by participants, which illuminate both activities and issues.

We also provide extensive supplemental materials documenting our qualitative process, with both backing spreadsheets and detailed prose discussions about each table at https://osf.io/nbtvm.

2 RELATED WORK

Our interdisciplinary work is broadly related to studies in journalism and mass communication as well as human-computer interaction. We divide the most relevant areas of research into data preparation in data science, measuring and classifying errors in data, and data preparation in journalism.

2.1 Data preparation in data science

The importance and ubiquity of data preparation is well known in data science [1]. Many researchers proposing end-to-end process models for conducting data science include specific stages for data preparation [16, 81] or synonymous labels such as such as wrangling [35, 50, 82], scrubbing [47], or preprocessing [23]. While this body of research characterizes the entire data science process, our work focuses exclusively on data preparation. We choose to build on the model of data science work from Crisan et al. [16] for three reasons. First, this model provides a clear distinction between preparation and analysis . Second, its comprehensiveness exceeds other models due to being synthesized from a systematic literature review. Finally, this model provides a sufficiently high-level characterization of workflows that it generalizes to our broad category of data workers, which includes data journalists. Our study provides additional levels of detail for two higher-order processes identified in their model, preparation and communication. We do not address its other stages, namely analysis and deployment.

Many researchers report tasks, challenges, pain points, and tool usage during data preparation via broader inquiries into the general workflows of data scientists. Artificial intelligence practitioners working in high-stakes domains discard potentially valuable data due to missing metadata [70]. Muller et al. [53] characterizes data wrangling along dimensions of intervention. While studying the workflows of data scientists in software engineering teams, Kim et al. [40, 41] identifies specific participant activities typically associated with data preparation, such as merging, cleaning, and shaping data. We identify a larger set of preparation activities and discuss those activities within the context of data issues.

Integrating data is a common challenge during data preparation [15, 35, 37, 41, 53]. Kandel et al. [35] finds that missing and inconsistent identifiers between tables impede data integration. Kandogan et al. [37] addresses the necessity and absence of semantic metadata when integrating tables. Our study corroborates these findings in the context of data journalism and identifies further data issues that make this activity challenging.

Several studies examine the use of visualization tools for data preparation and the role of exploration in data science workflows . Wongsuphasawat et al. [82] finds that assessing the quality of data is an exploratory goal. Batch & Elmqvist [5] identifies a "visualization gap", meaning that visualization is under-utilized beyond the final checking and dissemination stages despite research showing its benefits. Milani et al. [52] observes this visualization gap among data analysts cleaning and standardizing data. Alspaugh et al. [2] finds that exploratory activity in the overall data analysis process involves understanding semantics, identifying structure, characterizing data, and assessing quality during data preparation. Our study also finds a visualization gap when assessing the quality of raw data, and identifies other areas where visualization could be leveraged in data work.

2.2 Dirty data

The term "dirty data" is used at two levels. At a low level, it means problematic individual items within a dataset, and at a higher level it means the properties of a dataset that degrade its quality; we use the latter definition. While dirty data is an under-researched subject in journalism and mass communication [48], database researchers have studied this subject in depth. Companies can lose 20% of revenue from errors that propagate through a system due to dirty data [22]. There are many descriptive models for dirty data to frame the data issues that a proposed technical contribution addresses, evaluate data cleaning tools, or measure data warehouse quality [4, 12, 18, 27, 42–44, 55, 58, 59, 63, 66].

Different taxonomies frame the same atomic types of dirty data according to different schemes. One common scheme involves structural vs. semantic distinctions between dirty data. Chatterjee & Segev [12] applies this scheme to catalog problems arising from

data heterogeneity, the differences between independently maintained data stores. Likewise, Kim & Seo [43] uses this scheme in a taxonomy of multi-database system conflicts. Finally, Barateiro & Galhardas [4] uses this convention when organizing data quality issues by which one to evaluate using various data cleaning tools.

Another scheme involves high-level distinctions between issues involving a single data source or multiple data sources. Gschwandtner et al. [27] applies this scheme to classify types of dirty time-oriented data. Oliveria et al. [59] derives a taxonomy from an analysis of production databases in the retail sector, distinguishing between single vs. multiple source issues. Rahm & Do [63] incorporates both schemes: classifying data quality problems according to single source vs. multiple source origins and at the schema vs. instance (semantic) level. Our framework also addresses issues involving multiple tables but does not make high-level distinctions between single-source and multiple-source problems. We find that many issues involving one data source are compounded when working with multiple data sources.

Kim et al. [42] contributes an extensive taxonomy of dirty data organized into eight permutations of three binary categories: missing, wrong, or unusable data. From 33 types of dirty data, at least 25 require some form of human intervention. Our work further describes the ways in which data workers intervene to remedy these and other issues.

Outside of data warehousing research, taxonomies of dirty data generally provide a high-level classification of issue categories. Dasu & Johnson [17] names four challenges in exploratory data mining: heterogeneity, quality, scale, and paradigm. Hellerstein [28] reports four sources of data errors in a survey of quantitative data cleaning strategies: entry, measurement, distillation, and integration. When detecting anomalies in univariate data, Kandel et al. [36] identifies five specific categories of data anomalies guided by these taxonomies: missing, erroneous, inconsistent, extreme, and key violations. Finally, Wickham [80] describes the five common problems with messy data involving mismatches between data variables and observations with their representation in rows, columns, and tables that are addressable through data tidying procedures.

Our work is most similar to a group of past taxonomies that enumerate properties of data quality and frame dirty data as a threat to these properties. The taxonomy from de Almeida et al. [18] organizes data quality problems into five categories of compromised data and maps each problem to where it manifests in a model of multidimensional data warehousing. Likewise, Oliveira, et al. [58] provides formal definitions of data quality issues according to the multidimensional model. Müller & Freytag [55] classifies data anomalies into one of three categories that affect nine data quality criteria. Finally, Li et al. [44] proposes a rule-based taxonomy that classifies dirty data into violations of 13 data quality rules. In contrast, our taxonomy classifies dirty data according to six data qualities and a simplified four object model that more accurately describes the way data journalists discuss issues with data.

2.3 Data preparation in journalism

Preparing data has been an important part of data-oriented newswork long before the term "data journalism" was coined in the early 2000s [60]. Understanding the context around data, or lack thereof, is a longstanding and important part of data preparation in economic journalism [3]. Professional organizations for data-oriented newswork have been formalizing and disseminating practical knowledge on data cleaning since at least the early 1990s [60]. Today, data wrangling is one topic where applied artificial intelligence research can have an immediate impact for journalists [75].

While there is generally limited empirical research on data journalists' workflows [13, 75], among some extant process models related to producing data journalism, data preparation is an integral component under labels such as "Clean" [10, 13] or simply "Spreadsheets" [67]. Skills for this stage are valued in the profile of a data journalist, yet secondary to traditional reporting skills [11] and other aspects of data work, such as analysis and visualization skills [57]. Rogers et al. [68] finds that the prevailing view of data processing skills as a specialization is an organizational barrier that limit the use of data in newsrooms.

Many data journalism skills involve tools and techniques familiar in data science [8, 71, 76]. Data journalists rely on general purpose tools, such as Excel and OpenRefine[1] [73] as well as specialized, open-source tools built by other journalists [61].

Our interviews provide substantially more detail about activities and issues when journalists prepare data, and our analysis carefully situates our findings with respect to the research literature.

Showkat & Baumer [71] compares and contrasts practices in data-driven investigative journalism and data science. While our work shares a similar line of inquiry, it is distinct in two important ways. First, our work includes non-investigative reporting practices. Our participants describe their processes for both accountability reporting involving investigations as well as day-to-day reporting. Second, we recruit a broader participant pool of journalists, spanning multiple newsrooms. From this diverse perspective, we provide a broader characterization of the challenges data journalists face when preparing data. Our findings refute one claim of this work: we do not find that the use of unstructured documents vs structured data is a salient difference between data journalists and data scientists, respectively. Our participants often worked with structured data, including on investigative pieces.

Although no previous interview study addresses data journalists' workflows during data preparation, Table Scraps [39] is grounded upon a study of technical artifacts created by data journalists: their code notebooks and wrangling scripts. It answers questions related to what journalists do when working with data through a taxonomy of actions and how they do it through a taxonomy of processes. That study does not address the question of why data journalists do what they do when preparing data, which our work seeks to answer by identifying issues inherent in raw data that affect its quality. Our taxonomy of actions does partially overlap with this previous model of activities, providing a complementary triangulation between knowledge gained from two sources: the artifacts journalists produce vs. their direct statements within interviews.

3 METHODS

We conducted our study in four sequential phases:

(1) Analysis of data science workflow literature
(2) Analysis of novel interview data

[1] Formerly known as Google Refine

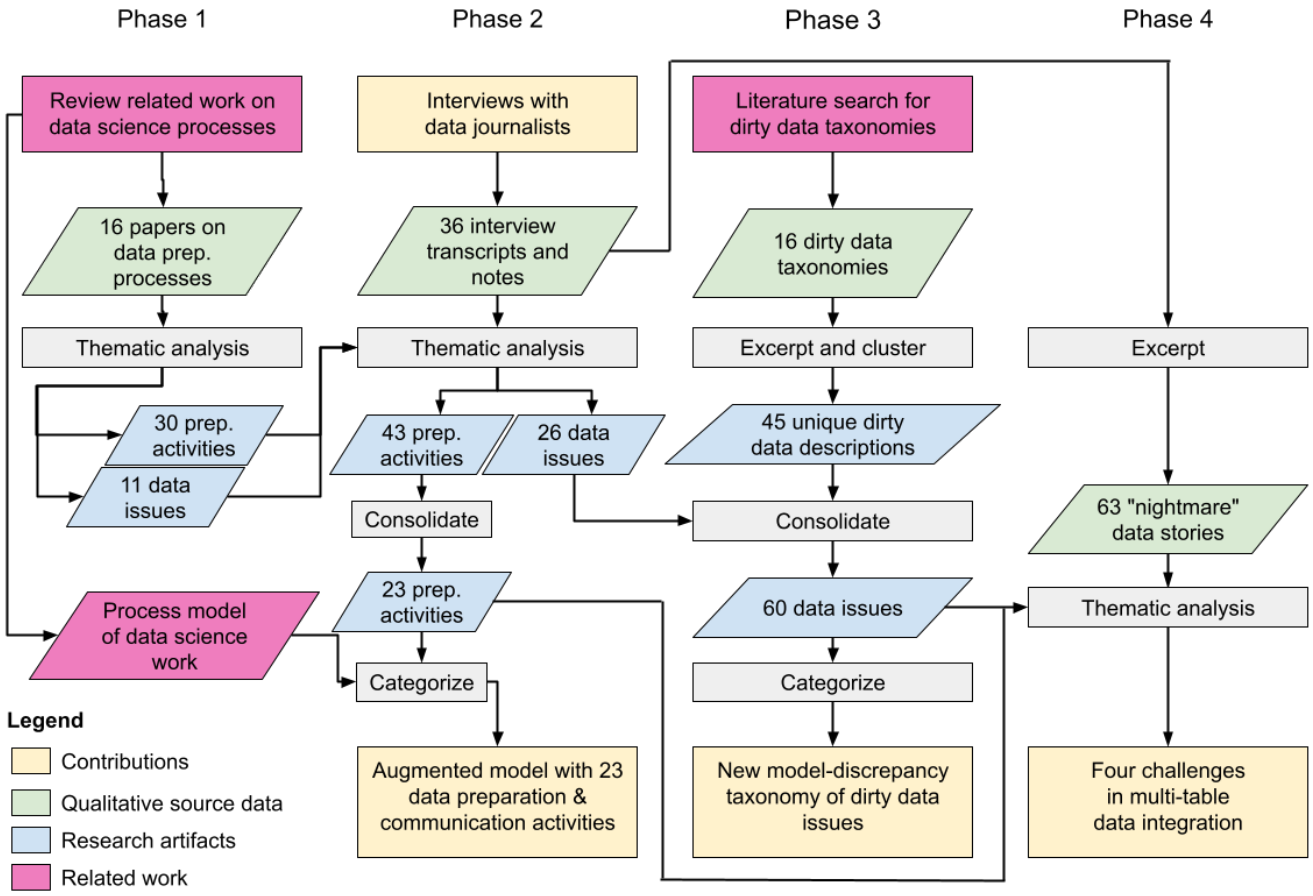

Figure 1: Process, products, and contributions: Our hybrid deductive-inductive thematic analysis [77] began by analyzing 16 studies of data science workflows to generate *a priori* codes pertaining to data preparation (Phase 1). We then conducted an interview study with 36 data journalists on their preparation processes, generating *a posteriori* codes from those transcripts (Phase 2). The resulting artifacts yielded combined code sets of preparation activities and data quality issues. Our categorization of these activities extended a previous model of data preparation activities. We then analyzed 16 taxonomies of dirty data issues (Phase 3), noting disparate coverage compared to our interview data. We produced a new model-discrepancy taxonomy for classifying dirty data issues to encompass them all. Finally, we reflected upon emergent patterns of data issues and preparation activities within the nightmare stories section of our interviews to identify four challenges for data integration (Phase 4).

(3) Analysis of dirty data issue taxonomy literature
(4) Further analysis of integration nightmare stories from interviews

The first two phases followed general guidelines for conducting hybrid thematic analysis, incorporating deductive approaches to create *a priori* codes from previous work and inductive approaches to create *a posteriori* codes [77]. In the third phase, we analyzed an additional data corpus to contextualize our intermediate results. In the final phase, we further analyzed the nightmare stories provided by participants to identify four types of data integration challenges.

3.1 Phase 1: Data science workflow literature

We constructed an initial codeset by analyzing accounts of data science workflows from previous interview, observation, and survey studies of data scientists. We began with the set of 31 papers previously identified in a recent systematic literature review of data science workflows as being relevant to data preparation [16]. We excluded papers that do not directly derive their results from the lived experience of practicing data scientists. The remaining 16 papers that we analyzed are listed in Table 1; Supp. Section/Sheet 1 provides additional information on each of these papers, including study size, methods, and application domain.

These papers cover data scientists occupied in a diverse set of domains, described at different levels of abstraction. The most prominent domains were technology [2, 35, 37, 40, 41, 52, 53, 78, 79, 82], including software engineering and social media; business [35, 37, 52, 53, 78, 78, 82], including finance; and healthcare [2, 35, 46, 53, 79, 82].

From each paper, the first author excerpted relevant sections on data preparation (resulting in 150 excerpts), then consolidated

Data science process papers	
Study	**Year**
Kandel, Paepcke, Hellerstein, Heer [35]	2012
Kandogan, Balakrishnan, Haber, Pierce [37]	2014
Kim, Zimmermann, DeLine, Begel [40]	2016
Batch & Elmqvist [5]	2018
Kim, Zimmermann, DeLine, Begel [41]	2018
Alspaugh et al. [2]	2019
Battle & Heer [6]	2019
Kaggle [32]	2019
Mao et al. [46]	2019
Muller et al. [53]	2019
Rule, Tabard, Hollan [69]	2018
A. Wang et al. [79]	2019
D. Wang, Mittal, Brooks, Oney [78]	2019
Wongsuphasawat, Liu, Heer [82]	2019
Milani, Paulovich, Mannssour [52]	2020
Zhang, Muller, Wang [84]	2020

Dirty data taxonomies	
Study	**Year**
Chatterjee & Segev [12]	1991
Kim & Seo [43]	1991
Rahm & Do [63]	2000
Dasu & Johnson [17]	2003
Kim et al. [42]	2003
Müller & Freytag [55]	2003
Barateiro & Galhardas [4]	2005
Oliveria et al [58]	2005
Oliveria et al. [59]	2005
Hellerstein [28]	2008
Li et al. [44]	2011
Gschwandtner et al. [27]	2012
Kandel et al. [36]	2012
de Almeida et al. [18]	2013
Wickham [80]	2014
Roeder et al. [66]	2020

Table 1: Related work: (Left) In Phase 1, we analyze 16 data science process papers relevant to data preparation, a subset of those identified in a systematic literature review [16]. (Right) We also analyze 16 taxonomies of dirty data in order to better contextualize the data issues described by our participants.

related excerpts into coherent groups through affinity diagramming [30]. Each group was given an *a priori* code. The resulting 41 codes were categorized into two higher-level *families* [77]: preparation activities (30) and data issues (11).

3.2 Phase 2: Novel interview data

We conducted 36 one-on-one, semi-structured interviews with data journalists from 31 different news organizations on their experience preparing data in the newsroom. We thematically analyzed these interview materials using the codeset generated deductively from related research in the previous phase, while generating new codes inductively from the interview data.

3.2.1 Recruitment. To recruit participants, we used purposive [65] and snowball [25] sampling. We solicited interviews from a curated list of more than 100 contacts in our professional networks, considering the criteria of organization size (large, small), publication medium (print, broadcast, online), and business model (for-profit, non-profit, academic) in this purposive sampling to fill our participant pool with a representative cross-section of data journalists. We also used snowball sampling to request interviews from a few journalists (2/36) recommended by participants. Because many data journalists do not have a formal job title connoting their expertise in data work, we used the inclusion criterion that participants should fit at least one of three personas:

- *Practitioner* (86%): actively demonstrates data-oriented newswork through publishing articles, graphics, or applications at a media organization.
- *Educator* (19%): holds faculty or staff position at an institution of higher education and teaches classes on skills relevant to data journalism.
- *Tool builder* (8%): develops computational tools to assist in data-oriented newswork.

We did not use country as a criterion, but the final set of participants discussed experiences working at newsrooms based in Canada, India, the United Kingdom, and the United States. Full details on the 36 participants are available in Supp. Sheet 2.

3.2.2 Procedure. Prior to each interview, participants were asked to provide their informed consent and share artifacts related to specific data projects that were challenging with regard to preparing data. All participants complied, and this pre-interview background research primed the interviewer on subject material. The first author conducted each interview via video conference. See Supp. Section 2 for our interview script. All participants gave permission to record conversation audio, and the first author also took extensive notes during each interview. The average interview length was 49 minutes, and the 36 interviews yielded over 29 hours of recorded audio. The first author reviewed the recorded audio to revise the interview notes, transcribe salient portions as passages, and build familiarity with the data [77].

3.2.3 Analysis. The first author applied the 41 *a priori* codes generated in the previous phase to appropriate passages and developed a total of 28 new *a posteriori* codes inductively from the interview data. After the final interview, the first author returned to earlier interviews to apply codes developed in subsequent interviews. For both kinds of codes, passages were selected that demonstrated qualitative richness [9]. A total of 566 passages were extracted and coded.

3.2.4 Termination. We concluded gathering data upon reaching theoretical saturation after 36 interviews, using the growth of our

codebook's cardinality as a proxy for saturation. Notably, this number of participants conforms with sample size guidelines for qualitative studies with in-depth interviews [21].

3.2.5 Reflective synthesis. We combined *a priori* and *a posteriori* codes and searched for higher-level structure.

The *activity* codeset contained 30 *a priori* codes from data science workflow papers and 13 *a posteriori* codes from the interviews, totalling 43 activity codes (see Supp. Section/Sheet 3). Through reflective synthesis, we consolidated these into a final set of 23 activity codes, renaming some for clarity. We realized that the activity codes could be used to extend the process model of data science work proposed by Crisan et al. [16] by adding an additional level of detail for the processes of data preparation and communication. We categorized all activity codes according to preparation subprocesses (initiate, gather, create, profile, wrangle) and communication subprocesses (disseminate, document). We present these results in Section 4.

The *issues* codeset contained 30 *a priori* codes from previous workflow papers and 15 *a posteriori* codes from the interviews, totalling 26 issue codes (see Supp. Section/Sheet 4). Our first attempt at categorization through reflective synthesis did not lead to fruitful results. The high proportion of *a posteriori* codes in this family (over 50%) was one methodological indicator that the data sources in the first two phases contained highly divergent information. We thus chose to add another data source for the next analysis phase in hopes of bridging this gap.

3.3 Phase 3: Dirty data taxonomy literature

Many researchers studying data warehousing, data cleaning, and statistics have proposed taxonomies of dirty data. We reviewed this research literature with a snowball sampling approach. We started with a set of four such papers already familiar to us [12, 36, 42, 80], then followed references and forward citations. We repeated this process until we discovered no further taxonomies of dirty data. Our final set of 16 papers that contain dirty data taxonomies are listed in Table 1; Supp. Section/Sheet 1 enumerates the number of leaf nodes in the taxonomy trees, each corresponding to a dirty data issue we considered distinct.

Our analysis collated 330 concrete *instances of dirty data*: the union of all leaf nodes in these taxonomy trees. We excluded 15 items that did not describe dirty data issues, were related to non-tabular forms of data, or whose descriptions we judged to be overly broad. We consolidated the remaining 315 issues by grouping together identical or essentially similar instances of dirty data into 45 *clusters*, listed in Supp. Section/Sheet 4. We then compare and synthesize these clusters of previously identified issues with our 26 data issues from the previous phase, reconciling our self-generated labels to use existing terminology when applicable. This synthesis resulted in a set of 60 issue codes, with 13 unique to our interview analysis, 16 unique to the previous work, and 31 overlapping. Our reflective synthesis of this material led to the new taxonomy for dirty data that we present in Section 5.

3.4 Phase 4: Interview integration nightmares

Finally, we conducted further analysis of the nightmare stories told by the 36 participants, describing their difficulties combining data from multiple sources during data preparation. In this case, multi-table integration was the desired end, not a means to another end. All participants describe at least one such project, with 69 in total across all interviews. There were 104 coded passages from these stories, out of 566 total passages extracted; see Supp. Sheet 6 details. These passages had already been assigned activity and issue codes in Phase 2. Revisiting these passages, we found four emergent patterns occurring in 63 out of the 69 stories. We present and discuss these four data integration challenges in Section 6.

4 PREPARATION ACTIVITIES

Our thematic analysis in Phases 1 and 2 results in a set of 23 data preparation activities, shown in Figure 2. Our hybrid analysis approach allows us to distinguish between three cases, which we color code in figures and text: those performed by data scientists but not reported by our journalist participants (blue); activities that emerged directly from our journalist interviews that were not recorded in previous papers about data scientists (green); and activities performed by both groups (no highlight).

To increase the utility of our results, we categorize these activities within the process model of data science work proposed by Crisan et al. [16]. Their model posits four higher order processes (preparation, analysis, deployment, communication). Our activities all map to two of these: the preparation process and its five constituent subprocesses (initiate, gather, create, profile, wrangle) and the communication process with its two subprocesses (disseminate and document). By mapping activities to each of these, we extend their model to an additional level of detail.

4.1 Initiate

Data scientists often begin preparing a dataset by defining the needs of the project; outlining project objectives; and identifying requirements with colleagues, collaborators, and external stakeholders [16]. We call this process *Initiate*, with the following three activities:

- *Establish goals*: define the overall objectives for a data project, including questions to answer, statistics to calculate, and final deliverables.
- *Make a plan*: draft a proposal for a data project that specifies implementation details, monetary costs, and a rough timeline to achieve the established objectives.
- *Test proof of concept*: implement a small-scale test or pilot study before conducting a full-scale data project.

Our findings: We find that establishing goals and making a plan to achieve those goals can be challenging for data journalists when preparing an unfamiliar dataset. For example, the goal of generating new story ideas often requires a significant amount of data preparation, which may be prohibitively expensive. One participant explains:

> Cleaning data takes so long, and here's the gamble: I don't know what the stories are in the data. But my track record indicates that there are stories in there...For a lot of media outlets that can't afford to free up people to do this kind of thing [data journalism], they're not necessarily going to take that

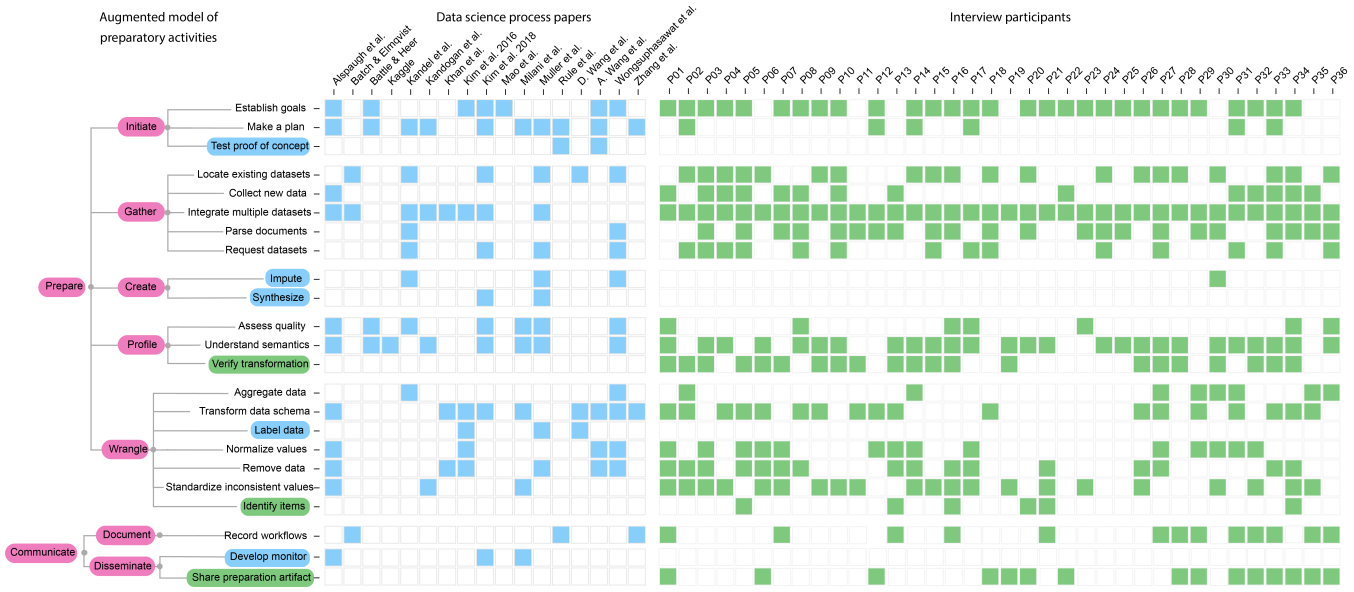

Figure 2: Data preparation activities: From our thematic analysis, we identify 23 activities that data scientists and data journalists perform when preparing data; blue and green backgrounds highlight divergences. Legend: processes from related work [16], prominent activities in data science, and prominent activities in data journalism.

> gamble. If part of your pitch to your editor is 'I can spend weeks and weeks and weeks wrangling and cleaning this data, but I have no idea what the stories are.' Guess your odds of getting an editor to sign off on that? Virtually nil. That's the problem we face.
> — P09

An unclear return on investment (ROI) is one barrier to the adoption of data journalism [68], and a few participants (3/36) lament time spent preparing data that did not yield publishable stories. In data science work, ROI can be clarified before investing significant time and resources by test proof of concept. Notably, no participant describes conducting this activity. However, a few participants (2/36) describe identifying a "minimum viable story" in raw data with the expectation that further story ideas will appear during the preparation processes.

Throughout data preparation, journalists often discover limitations within their data that affect the goals initially established for the project. Many participants (15/36) report abandoning a data project due to issues such as cleanliness, complexity, and reliability; see Supp. Section 6 for details. One participant reports that a factor affecting their ability to achieve initial objectives is whether the data is an "closed or open universe". With a closed universe, they are sufficiently confident in the data's completeness to make absolute claims, but in an open universe they would always couch specific claims with disclaimers, such as "at least".

4.2 Gather

Data gathering includes the process of identifying existing data [16]. We expand this definition to include activities related to obtaining data. Both data journalists and data scientists perform these five gathering activities:

- *Locate existing data*: find and identify data of interest either within their organization, publicly via the Internet, or from an external organization.
- *Collect new data*: record data from observed phenomena or processes in the world when existing data are not available.
- *Integrate multiple data*: combine multiple tables into one (including *schema matching* [62]).
- *Parse documents*: create structured data by parsing data found in unstructured or semi-structured documents.
- *Request data*: request data from an organization, formally or informally.

Our findings: While both data scientists and data journalists request data and parse documents, every participant in our interview study reports issues that uniquely characterize these activities in data journalism.

Data scientists often work with data collected or maintained by clients or other divisions within a company, and may also make requests for data. Many of our data journalist participants (14/36) also describe requesting data in a unique context not previously identified: formal data requests to government agencies through freedom of information (FOI) requests. It can take months or years for journalists to obtain data from FOI requests. These delays can lead journalists to abandon stories when they are no longer timely, thus less newsworthy. In response, many data journalists tend to gather data on phenomena that are newsworthy regardless of timeliness.

Often, data journalists receive data through FOI in PDF or physical documents, requiring them to further parse documents in order to obtain usable data. Both data scientists and data journalists obtain

these data by parsing unstructured or semi-structured documents, especially when scraping data from the web. However, parsing activities involving PDF documents, typically from FOI requests, is a unique context reported by our interview study participants. In this situation, journalists transform data populated in paper forms into tabular format or extract tables of data embedded in documents into a programatically accessible format.

Some participants (5/36) express the belief that FOI requested data has been deliberately returned in inaccessible forms that require extracting data, a sentiment that also been reported in other journalism studies [24]. One participant explains:

> Sometimes it's just what they're used to doing [supplying data in image-based PDFs], like they want to stamp it or they want to redact it. Sometimes I feel like they're just being ornery and don't want to be responsive to public information requests. I feel that way sometimes. I can't ever say that it's true, but I've definitely gotten data that way.
> — P03

All participants (36/36) integrate multiple tables, especially by supplementing one table with additional demographic data, such as local COVID-19 cases with demographic information from census data. Another common scenario is to detrend population-affected data by integrating data to calculate per capita rates, an established practice in precision journalism [49]. We discuss more challenging integration scenarios in Section 6.

4.3 Create

When data cannot be collected or directly observed, data scientists may fabricate placeholder data [16]. From our analysis of the data preparation literature, we identify two activities within this subprocess, neither of which was prevalent with our journalist participants:

- Impute : replace missing data with values derived from other attributes.
- Synthesize : fabricate data of hypothetical or approximate values that simulate data from observed phenomena.

Our findings: Almost no data creation instances appear in our interview data. Every participant describes preparing data that represented observation of real-world processes or phenomena, but only one participant reports imputing missing values, in this case for six days out of an entire year. We attribute the extreme reluctance of journalists to impute or synthesize data to the professional norm to work only with material that might yield a publishable story and caution surrounding legal concerns if placeholder data accidentally appeared in print [7].

Data journalists are also cautious about using data that contained estimates rather than observed values. One journalist discusses an instance investigating how the digital divide intersects with the COVID-19 pandemic as public education moved online, focusing on families in rural areas without access to high-speed internet. The journalist eventually abandoned the story because an official government dataset detailed hypothetical coverage rather than actual coverage, making it a poor measure of internet connectivity.

4.4 Profile

Profiling describes the subprocess of assessing, understanding, and examining data [16]. While checking for understanding is also a part of data exploration [2], we treat it as part of profiling due to the integral role it plays in other preparation processes, especially when removing data [26]. We identify three profiling activities:

- Assess quality: ascertain the quality, identify issues, and any apparent limitations within a dataset.
- Understand semantics: uncover or reveal the underlying meaning or context surrounding data.
- Verify transformations : ensure that recently applied data transformations did not have any unintended consequences.

Our findings: With regard to profiling data, we note that data journalists exhibit a similar behavior to data scientists when assessing data, spending a significant amount of time understanding basic information about datasets, and using the same tools and techniques for other profiling activities to verify the effects of their transformations when wrangling and integrating data.

While visualization can be a powerful tool for assessing data, many data scientists assess their data numerically with summary statistics [41, 52]. Many participants acknowledge that visualization could be useful in this activity but rely on numerical summary assessments of their data, such as counting the number of null values in an attribute.

Data scientists often devote significant time to understanding the nuances, underlying semantics, and subtle limitations of a dataset during data preparation. This activity is sometimes called "becoming one with the data" [20] or "building intuition" [52]. Many participants (26/36) also stress the importance of developing a deep understanding of the dataset; they often spend significant time developing basic understanding because the data had inadequate documentation, if any. According to one participant:

> Understanding, that's pretty big, especially when there's not enough documentation. You may request data but column names are 'ODCNLYTT' and you're like what is that? So there's a lot of incomplete documentation at all levels, but states, governments, have a way of either not providing or not properly documenting the data they collect in the first place.
> — P20

One new code to emerge from our journalist interviews involves verifying the effects of applied data transformations to confirm that no unintended side effects found their way into the transformed data. Our participants describe using profiling techniques to assess the quality of the transformed data. While they sometimes use visualization methods, they gravitate towards non-visualization methods, such as spot checks, summarizing attributes, and counting null or missing values. Journalists would also compare individual data items against previous versions of the same dataset to verify transformation effects.

4.5 Wrangle

Wrangling is defined elsewhere as the process of making data usable for analysis [33]. However, as many other preparation subprocesses are also aimed at this objective, we adopt a narrower definition

of wrangling: modifying, refining, or otherwise altering a single table into an alternative form that is amenable to analysis. Many participants used synonyms for wrangling, such as "munging", "massaging", and "cleaning". We identify seven wrangling activities:

- *Aggregate data*: decrease the size and granularity of a dataset by summarizing or grouping items in a table.
- *Transform data schema*: modify the underlying schema of the data.
- *Label data items*: annotate data items with semantically meaningful labels.
- *Normalize values*: adjust values measured on different scales to a common one.
- *Remove data*: decrease the size of the dataset by taking away, discarding, or filtering items or attributes from a table.
- *Standardize inconsistent values*: resolve inconsistency involving how the same entity is represented.
- *Identify items*: distinguish unique items within a table or identify the same entities between multiple tables.

Our findings: While we find that data journalists mostly engage in the same wrangling activities as data scientists, we note that removing and normalizing data were especially prominent codes in our interview data in a context not addressed by related work on data science workflows. Additionally, identify items is a frequent and new code that emerged directly from our interview data. We speculate that this activity is not unique to data journalism, but is under-reported in data science workflows. Some data wrangling applications, such as Wrangler [34], address this need through support for skolemization.

While the choice to remove data often addresses noise, errors, and large datasets in data science, many journalists describe removing entire sections of extraneous data, items, and attributes that are not relevant to their inquiry during the initial steps of data preparation. "I'll get a large dataset", says P05, "and 80% of it is just stuff that I don't want".

Some participants (6/36) describe creating unique keys in a new attribute to uniquely identify items or groups of items. Creating an attribute that identifies groups of items within a table is often a prerequisite activity for aggregating data within a single table. Journalists often craft this attribute as a soft key, with no guarantee of uniqueness [17], by concatenating ostensibly unique attributes of names, addresses, birth dates, phone numbers.

One participant describes encountering a table with what appeared to be duplicate data. The names and addresses matched. However, in reality they were father and son living in the same home, and the one differentiating attribute, birth date, was excluded from the dataset.

Data scientists often normalize data to satisfy model assumptions downstream in their workflow [40, 82]. Some participants (7/36) report normalizing quantitative data by calculating per capita rate, facilitating fair comparisons [49]. Some participants (7/36) describe normalizing qualitative data labels, mapping categorical attribute values into an ontology representing a different mental model. Participants rarely distinguish quantitative from qualitative data explicitly, but many journalists will create categorizations. For example, when preparing criminal justice data, one participant describes normalizing more than a dozen of a court judge's sentencing descriptions into three categories understandable by the general public, such as "convicted" or "dismissed". Arbitrary decisions around the definition of these labels can lead to differing conclusions in downstream analysis, as occurs in other phases of end-to-end data analysis [45].

Notably, no participant reports performing the activity label data, or marking items as ground truth to train machine learning models, even though this activity is common in accounts of data science work. We believe preparing data for descriptive modeling, instead of predictive modeling itself, explains this difference. However, future work is needed to test this hypothesis.

4.6 Document

Documentation, or creating a record that describes performed work, is a communication process that intersects with data preparation and other high-level processes in data science [16]. While there may be other documentation activities across the entire data science process, including archiving digital artifacts through documentation [29], we identify one documentation activity relevant specifically to data preparation:

- Record workflow: log the steps taken to prepare a dataset.

Our findings: Both data journalists and data scientists document aspects of data preparation by recording workflows; however, data journalists contend with two distinct aspects of documenting the preparation process. First, in order to consolidate separate preparation processes performed across many different tools, some participants (5/36) created a *data diary*, a separate document containing data provenance information typically composed with a word processor. While the data diary may include a list of steps made while preparing data, it may also include relevant preparation details beyond a simple data transformation log, such as data collection details, a contact phone number for questions about the data, and its limitations. Second, while data scientists often communicate their work to a variety of stakeholders [82, 84], data journalists focus on a unique stakeholder, the public. Thus, many data journalism articles post code, data, and methodological processes publicly [39].

4.7 Disseminate

Dissemination, or sharing insights into the data science process, is another cross-cutting data science communication process that intersects with data preparation [16]. We identify two activities where the two populations diverge:

- Develop monitor: create a means of checking the quality of a dataset as new items are ingested by the system.
- Share preparation artifacts: distribute byproducts of the data preparation process.

Our findings: Data scientists may develop dashboards and other visualization artifacts to monitor the data preparation process, often for dynamic datasets. However, only one of our journalist participants describes a single instance where they continuously maintained a dataset. Our participants share artifacts, executable snippets of code, intermediate data products for less technically

adept colleagues, and reports on datasets after rounds of cleaning and vetting.

5 MODEL-DISCREPANCY TAXONOMY OF DIRTY DATA ISSUES

As we discuss in Section 3, our initial attempt to categorize dirty data issues was unsatisfying. The Phase 1 material of workflow papers from the data science literature and the Phase 2 material of journalist interviews did not sufficiently overlap. We thus extended our analysis to include data quality issues discussed in the database literature, leading to a set of 60 data quality issues that encompasses all three corpora of material, shown in Figure 3. This large list requires some kind of hierarchical categorization to be useful, but previous data quality issue taxonomies lacked enough breadth to cover them all. Our novel taxonomy does, with a different lens than the others.

Previous taxonomies of dirty data issues all characterize dirty data as falling short of some perfect kind of ideal data. In contrast, we view datasets as design artifacts made by data workers: people who collect, store, maintain, and prepare these datasets. Therefore, the data model represents the synthesis of mental models from the data workers involved. In this framing, dirty data constitutes an instance of a gulf where the existing data model does not match a data worker's mental model of the dataset.

We propose a new taxonomy that classifies data quality issues as discrepancies between the user's model and the existing data model along two dimensions of a dataset: objects and qualities. Both dimensions are orthogonal with regard to the specific data issues they categorize.

5.1 Data objects

The first axis of our taxonomy is built upon the four main concepts used by our participants to describe their data issues, using terminology following Munzner [54]. It is considerably simpler than the more complex dirty data models from the database literature that handle multidimensional data. The four data objects are:

Table: a collection of items and attributes. Dirty data at the table levels affects multiple items and/or attributes. Tables are represented in rows and columns, but we use the term table to include other representations of tabular data, including a single relation in relational, JSON-structured, and XML data.

Item: a collection of different attribute values that describe a specific observation or entities within a table. Dirty data at the item level affects one or more attributes with regard to a single item. Items are uniquely identified through a combination of attribute values, a candidate key, or one unique attribute, a primary key. Equivalent terms: spreadsheet rows, tuple in a relation, or database records.

Attribute: a specific, measurable property shared by items within a table. Dirty data at the attribute level affects multiple items along a single attribute. Unless explicitly specified in the design of the data schema through integrity constraint mechanisms, describable properties of attributes are often emergent qualities of the values associated with all items, including domain or range, semantic meaning, and associated data type. Equivalent terms: columns in spreadsheets and databases.

Value: the amount or variety of a specific item with regard to a specific attribute in a table. Dirty data at the value level affects a single item-attribute pair. Values often carry many implicit assumptions that may not be expressed as attributes elsewhere in the table, such as units for quantitative measurement.

5.2 Data qualities

The second axis of our taxonomy contains six data qualities, abstract characteristics of data objects. They are:

Completeness: whether a data object has all the necessary and appropriate components. Instances of dirty data involving discrepancies in completeness can be characterized along a dual spectrum with opposing sides. Underlying missing data are discrepancies involving under-completeness. However, data with too much extraneous, irrelevant information also constitute discrepancies characterized by over-completeness. We concur with Müller & Freytag [55] that removing incomplete data, instead of correcting the issue, artificially inflates the completeness of a dataset.

Accuracy: the degree to which data objects are correct and precise with regard to the phenomena they represent in the world. While accuracy and precision are two separate measurements of observational error, we consider them together.

Form: the arrangement, format, or configuration of data objects. Dirty data with discrepancies in form affect how data objects appear rather than what they mean or represent. Examples: pivot tables vs. tidy data [80]; formatting attributes containing phone number, dates, or currencies; the order of attributes within a table.

Granularity: a data object's scale or level of detail. As with completeness, the granularity of items and attributes may be above or below the expectation of a user's mental model.

Relation: the connection or relationship between multiple data objects of the same class. Examples: Multiple tables containing the same type of item [80]; multiple attributes containing values with logical dependencies, such as ages and dates of birth.

Semantics: the underlying meaning behind individual data objects. Undocumented or under-documented data can cause semantic discrepancies involving every data object. Attributes often carry high-level semantic types, such as people's names and social security numbers. These high-level types are often extensions of low-level types such as integers and character strings. Therefore, we consider conflicts involving primitive data types to be semantic discrepancies. Examples: multiple interpretations for the same value, such as homophones; conflicts between primitive data types, such as integer vs. character string; and duplicate items.

Many of the data qualities we propose are not considered in previous models; the two that overlap with previous work are completeness and accuracy [18, 44, 55]. We describe many related, high-level classification schemes for dirty data in Section 2.2.

5.3 Results

We note a substantial difference between the data issues that had received previous attention from database researchers and those uncovered by our interviews with data journalists: 16 issues were unique to the database literature, 13 issues were unique to data journalists' accounts, and only 31 overlapped. Figure 3 shows how this taxonomy covers these three groups of the 60 dirty data issues,

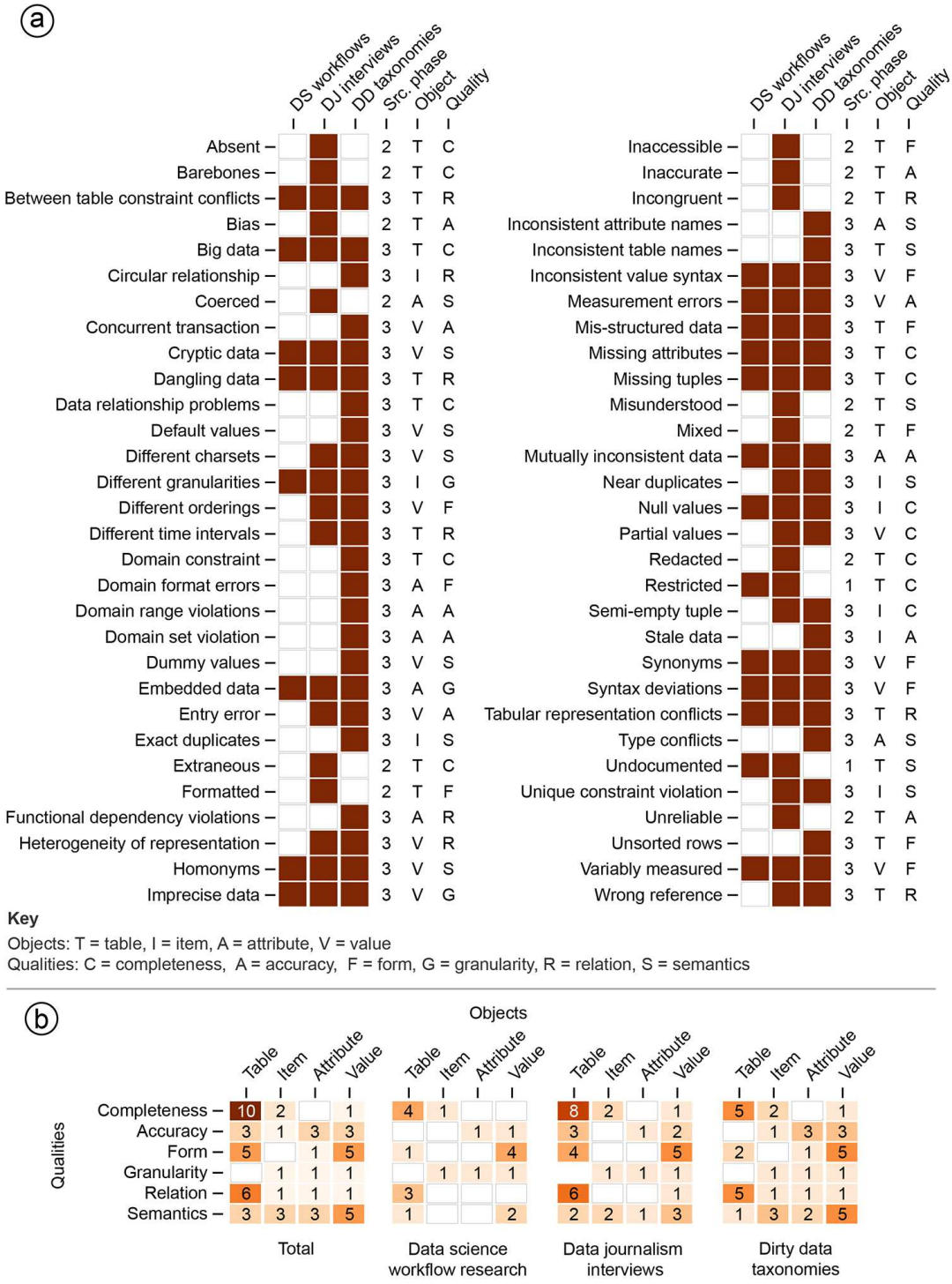

Figure 3: (a) Sixty data issues and which source of data they occur in (data science workflows, data journalism interviews, or dirty data taxonomies), the source phase they were identified in (1-3), and the object and quality the issue corresponds to within our model-discrepancy framework. See Supp. Section 4 for a detailed explanation of each data issue. (b) The distribution of issues above in total and in each group of qualitative source data according to our new taxonomy for classifying dirty data.

illustrating the need for a new taxonomy with adequate coverage of the full breadth of dirty data issues.

We note that database researchers have a theory-focused perspective based on the concerns of people storing the data, whereas both data scientists and data journalists have a domain-oriented perspective focused on their data needs. We conjecture that the view of data issues found in our data journalist interviews may also pertain to other domain-oriented populations of people who use rather than store data. Our broad category of data workers encompasses all such consumers of data.

6 MULTI-TABLE DATA INTEGRATION CHALLENGES

In the final phase of our analysis, we re-analyze the 69 nightmare stories told by all participants about their difficulties integrating multiple data sources. From these stories, we identify four recurring data integration challenges:

- Regional: tables with inconsistencies due to independent, spatially dispersed data sources.
- Diachronic: tables on the same phenomena that evolve over time.
- Fragmented: tables on a similar topic that contain different yet related items.
- Disparate: tables that are topically dissimilar and seemingly unrelated.

These challenges are not mutually exclusive; more than one challenge occurred in 11 of the 69 nightmare projects. In an exceptional incident, one participant (P12) reports using cloud computing services to process more than 10,000 individual tables for a story that involved a regional and diachronic dataset, combining a decade of monthly policing data from across the United Kingdom.

6.1 Regional datasets

Our participants report often working with open government data that federal regulatory or legislative bodies require be disclosed to the public, yet delegate the implementation of this mandate to constituent state, provincial, and municipal governments with little standardization guidance concerning how this data is collected, organized, or disseminated. These constituent data collectors are often dispersed across disjoint geographic regions and institutional bodies. These conditions often produce regional datasets: multiple tables on the same phenomena from data collectors who are dispersed spatially and institutionally. Because many participants work on a level spanning the territory of multiple data collectors, our interviews reflect many issues with preparing regional data, including COVID-19 infection rates, political campaign expenditures, and crime statistics.

Due to the independence these data collectors exercise, a dataset may contain tables representing the same topic but structured differently in ways that impede the integration of these tables and make data preparation time consuming. One participant elaborates:

> If I want to write a pan-Canadian story about a topic, it means I have to go to ten different provinces and ask them for data. No two of them will have the data in the same way ... There's different ways of recording the data and storing it. So to standardize this dataset into one single thing I can use takes a whole lot of time.
> — P06

However, local data journalists are also affected by regional data, sometimes to a greater degree than their national counterparts. One participant (P14) based in Missouri reports consolidating data from 90 municipalities to report on stories concerning a single county within the state.

The distributed nature of the dataset is often the most tractable issue with regional data, where many issues are perceived to stem from the independence of regional data stewards [16] in collecting, storing, and publishing data. One related issue is reconciling different classification ontologies for the same data items between multiple tables [35]. Our participants discuss reconciling incompatible ontologies on food hygiene ratings or types of business license. In our model-discrepancy taxonomy of issues, this data-related challenge represents a VALUE-RELATIONALITY discrepancy, and many participants describe the activity of standardizing data in response. When standardizing dissimilar ontologies, a common strategy is to derive abstract categories that logically describe different categories. One participant describes reconciling different ontologies on use-of-force incidents by police departments across the United States:

> Let's say a police department has 10 categories for use of force and another one has six...Deadly force is a less ambiguous category, but physical force that is non-lethal might be a broad category. Where one department has it broken down as like pushing and shoving and tasers and hitting with a baton, another department has it broken down as like push and shoving and everything else, you can then turn the three categories into one and be able to match them up.
> — P07

Participants report receiving regional data in many formats, including PDFs, spreadsheets, and flat text files, but also within email and sometimes values spoken over the phone. Other issues related to regional data are similar to those caused by data heterogeneity, conflicts in structure and representation arising from independently operated databases [43]. The data schema of one table may not conform to the data worker's mental model or the models of other tables. For example, data may be represented as one attribute or many. Some tables may be pivoted or cross-tabulated while others may be in tidy format [80]. In the thorniest cases that participants describe, attributes may be intermittently present across tables, and the structure of multiple tables may not conform with the user's mental model.

The most insidious issues involve differences in data collection between regions that lead to TABLE-SEMANTIC discrepancies regarding table items. For example, one participant (P20) covering the opioid crisis found inconsistencies in the counts of fatal opioid overdoses due to different definitions of resident and cause of death.

6.2 Diachronic datasets

While difficulties surrounding the preparation of a regional dataset can be due to multiple, non-coordinating data sources on the same

phenomena, a dataset published by the same source can still be difficult to prepare. Perennial news stories may involve analysis of civic data from a single source published at regular intervals, such as reports of government spending. However, data may evolve in subtle or dramatic ways in subsequent publications, leading to a diachronic dataset: a set of tables on the same phenomena that structurally or semantically change over time. Some examples of diachronic datasets discussed by participants include: salaries of high-earning public sector employees, annual listings of donations to registered charities, and campaign contributions between election cycles. Some participants discuss preparing dynamic data, especially for dashboards and individual charts related to COVID-19 pandemic statistics. As the majority of participants discuss static data, we consider diachronic datasets a property of chiefly static data. However, many inherent issues may also extend to instances of preparing dynamic data.

Data issues associated with preparing diachronic datasets extend beyond discrepancies in TABLE-RELATIONALITY, stemming from data on the same phenomena separated into multiple tables [80]. Changes due to the evolution of the dataset over time involve many other preparation subprocesses, especially profiling [36] and wrangling [34].

Schema drift informally refers to changes in the data schema over time [51]. This data issue represents a TABLE-FORM discrepancy in our dirty data framework. Our participants report a common form of schema drift is the inclusion of additional attributes over time. As in data science, these attributes may be redundant [35], but they may also represent new information. Moreover, participants describe addressing changing attribute names or meanings through transforming data.

Another related issue involves the evolution of codes for a specific attribute. One entity may be referenced by two or more codes as the classification ontology evolves. Inflation is a common example involving quantitative data, and journalists derive index values to address this issue [49]. A more difficult issue involves the evolution of categorical value meanings, a form of VALUE-SEMANTIC discrepancy. One participant (P20) preparing economic data from the Bureau of Labor Statistics explains that while the occupation "computer analysts" is present in data from 1990, the meaning is not the same as in 2020.

Missing data is another common issue among participants, which we consider a TABLE-COMPLETENESS discrepancy, with regard to both attributes and items that represent continuous time periods. Some diachronic datasets may not be published at regular intervals, such as those released by hospitals. Other times, regularly published sources of data inexplicably dry up, according to a criminal justice reporter (P13) who analyzed prison population data. "They're required by law to provide this data", he says. "But there's no punishment when they don't provide it".

Participants report that anomalies within the data stem from undocumented methodological changes. One participant (P08) says documented changes in the data collection methods are the exception rather than the rule. These methodological changes may result in anomalies may be detected when assessing the data and require further understanding.

Finally, participants describe a particularly difficult issue with shifting geographic boundaries for diachronic data representing a specific region. Cities grow. Smaller population areas amalgamate. Legislative districts are redrawn. These changes make fair comparison of the same area over time prohibitively complicated. A similar issue occurs when preparing fragmented datasets, and methods used to address this issue may also apply to diachronic datasets.

6.3 Fragmented datasets

Both regional and diachronic datasets describe sets of tables with items that share the same meaning. However, another challenge many participants (17/36) describe involves preparing tables with items that are semantically distinct yet logically related: a *fragmented dataset*. When requesting data, many participants receive exported data that was previously organized into multiple related tables for efficient storage and retrieval. Examples of fragmented datasets include data on: rejected vote-by-mail applications and voter demographics; state hospitals and hospital procedures; and delinquent mine safety violations implicating multiple mines, operators, and owners.

Preparing a fragmented dataset is like assembling a puzzle. The challenge involves understanding how the pieces fit together. Many data preparation activities can involve combining a primary table with an auxiliary table containing an area's demographic or population data. What distinguishes preparing a fragmented dataset from standard data integration is that combined tables need not include all constituent components. Successfully prepared fragmented data may shed light on a particular aspect of the data, or it may reveal enough of the final picture to generate leads for traditional news reporting methods.

Fragmented datasets may have opaque codes from being originally stored in relational databases. Entities that represent categorical data may be represented as integers or other shortened codes in relational databases [12], and a related wrangling activity involves translating entity codes [35]. Journalists may approach resolving this issue as a form of standardization or as an integration activity involving a lookup table, also known as a *crosswalk* [83], a map that converts data to a new or different standard. This lookup table may have to be manually constructed by journalists from a data dictionary, textual descriptions for attributes accompanying published datasets [64]. In some cases, the "pieces" may not align. Different tables may use different identifiers for the same entity, or the items in separate tables may represent overlapping, but not identical, geographic regions.

Another area of difficulty is matching election results with demographic data, especially from national censuses. In many cases, demographic data must be aggregated into larger areas equivalent to election precincts. However, some areas use idiosyncratic regions that census data cannot be aggregated into. One participant describes encountering this problem with Philadelphia's system of wards.

> This stuff isn't limited to just election data. The inability for different geographies to match up with each other is a well known problem that I think everyone who works with spatial data will encounter at some point in their lives, and we all have different ways of dealing with it.
> — P30

In this case, P30 was able to address this issue by apportioning values by area, a technique used by an election blogger she consulted. Weighting overlapping regions by area or population may also be useful when integrating incongruent geo-political boundaries for the same areas, as seen when preparing diachronic datasets.

Reassembling related data accurately can be particularly challenging, even for veteran data journalists. Two participants coincidentally describe preparing data that combines the Debt by Age dataset and data on delinquent mine safety fines originally obtained by a FOI lawsuit filed against the US Mine Safety and Health Administration (MSHA). For the participant (P36) who prepared the original raw data from MSHA, under-documented and duplicate data impeded data preparation; an error in the received data added significant time to the preparation process. When one table was missing the ownership end date for some mines, it resulted in contradictory numbers that added months to the data preparation time. Later, the other participant (P20), who works at a different news organization, used the cleaned Debt by Age dataset to report on safety fines from a specific mine owner. Despite being previously cleaned, this participant still encountered difficult aspects of preparing this data due to the complicated relationship between entities:

> Delinquencies are applied to mines, but mines over time change ownership...if you're trying to find out who accrued the most violations in terms of ownership, you have to understand that you can't just pull from the violations and look at the owner's dataset or vice versa. They own mines that have violations that are unpaid that they're not responsible for, regulatorily.
> — P20

6.4 Disparate datasets

The three previous challenges describe datasets where individual items represent the same or similar topics. These challenges may involve spatial inconsistencies (regional), temporal variations (diachronic), or else inconsistencies that arise from idiosyncratic features of different source databases (fragmented). However, data journalism has a long history of gleaning insights by combining seemingly unrelated datasets [14, 60], and many participants (14/36) describe preparing data in order to integrate tables on dissimilar topics. These *disparate datasets* are topically dissimilar, but contain reference to a common entity, such as attributes representing names, addresses, or phone numbers. These attributes can often serve as linkages between tables, and the intersection of these tables can reveal latent insights during an investigation, often implicating the subject in some form of wrongdoing.

The most common disparate dataset participants describe involves tables with items that semantically represented the same type of entity, such as people or companies, and specific items potentially referencing the same entity between tables. Hence, this process of combining multiple tables on common entities equates to entity resolution, reconciling multiple distinct references to the same real-world entity [38]. Interesting examples include:

- Investigating healthcare workers dying of opioid overdoses by integrating tables of state health care provider licenses and death records.
- Identifying companies that laid off workers even though they were loaned funds from the federal Paycheck Protection Program designed to encourage small businesses to retain employees during the COVID-19 pandemic.

Many participants report two compounding issues when preparing disparate data that make it challenging to integrate datasets. First, there typically exist discrepancies in the identity of individual items. Identifying items within a single table is a common preparation activity, which we identify in Section 4. When preparing disparate datasets, the difficulty of the activity is compounded by the additional requirement to craft keys that correctly reference the same entity between tables, serving as a makeshift foreign key. Second, inconsistent values can further impede data integration by complicating the process of creating keys. The same entity may go by multiple names, and different entities may use the same name. Hence, some journalists describe standardizing data to reconcile inconsistencies, and a few describe circumventing standardization to some extent by relying on fuzzy match algorithms.

These issues can create uncertainty in the accuracy of match results. Journalists often deal with this uncertainty by re-evaluating the goals established in the Initiate process. All journalists who report matching disparate datasets describe tuning their matching parameter in order to minimize or eliminate the rate of false positives in their combined data, which they acknowledge increases the percentage of false negatives in the results. Even a handful of correct matches can support multiple stories; however, publishing an incorrect match could end a career.

Disparate datasets may share another commonality that can be exploited to integrate two seemingly unrelated datasets: geography. While some instances of related data can be combined on equivalent geography, disparate datasets that are geographically related represent overlapping, but not equivalent, geographic regions. Often this type of disparate dataset involves census data, other data that do not use the same area definitions, or one area where the geographic boundaries change over time. A few participants (2/36) describe cases where this issue stopped them from pursuing a story, but one participant (P30) describe resolving this issue though apportioning values by area.

7 DISCUSSION

We discuss the prominence of accountability journalism in our interviews, the role of tool usage in the capabilities of our participants, and the implications of our work for the design of future tools.

7.1 Accountability journalism

Investigative journalists serve an essential role in democratic society, acting as a counterbalance to those who wield economic and political power by revealing corruption, dysfunction, abuse, and other forms of wrongdoing.

We note that although data journalism may include other genres such as sports and entertainment reporting, participants in our study focused primarily on investigative journalism, also known as accountability or watchdog reporting. We conjecture that data preparation is most difficult for this type of journalism because it involves bringing transparency to unknown or deliberately concealed matters of concern. While sports organizations have an incentive

to provide clean, readily usable statistics about games, teams, and players, corporations and governments often have the opposite incentive, leading to more laborious data preparation for journalistic investigations of business practices, health, labor, the environment, and other highly political subjects.

Instances of wrongdoing by powerful figures and institutions are seldom readily apparent in the contents of a database or spreadsheet, especially when deliberate measures are in place to conceal this information. As a result, many works of investigative journalism require weeks, months, or even years of both traditional reporting and data-driven investigation, often straining the resources of news organizations whose budgets are already strained.

7.2 Tool-based archetypes and MacGyvering

In Phase 1 and Phase 2 of our analysis, in addition to preparation activities and data quality issues, we also created a third family of codes for the usage of tools. We present these codes in Supp. Section/Sheet 5. Our initial analysis did not yield the rich results of the other two code families, so we did not continue in search of formalisms. However, we did find one intriguing aspect of tool usage that both aligns and diverges from previous work, which we discuss here.

Related work [35] proposes three archetypes of data worker based on the tools they use when performing data work: *application users*, who use spreadsheets or other click-based applications (Excel); *scripters*, who use software packages for data analysis (R or Matlab); and *hackers* who are fluent in the same analysis packages as Scripters but also proficient in scripting languages (Python, Perl) and data processing languages (SQL).

Our participants align with this model of data worker expertise, especially with respect to how each archetype correlates with a set of tools used to prepare data. Many participants (13/36) were application users, employing only tools such as Excel, Google Sheets, or Microsoft Access.

Participants discuss working with colleagues who are proficient spreadsheet users but not data specialists, who would also fit into this archetype. Other participants fit the scripter archetype because they know basic Python (16/36). Finally, the most advanced users (7/36) were fluent in multiple programming languages and familiar with querying and creating databases, fitting with the hacker archetype.

However, in contrast to previous work [35], we do not find that a data worker's preference for click-based vs. code-based tools necessarily restricted the preparatory activities they perform. We find that some application users implement a creative and improvisational approach to accomplishing preparation activities with the tools at their disposal. Following the term in widespread use among data journalists, we call this behavior *MacGyvering*, after a 1985 American television series where the protagonist routinely escapes life-threatening scenarios through creative, even implausible, engineering feats using whatever objects happen to be nearby. One participant who fits the application user archetype describes the practice of MacGyvering in data journalism:

> At some point, I might feel like the way I do this isn't sophisticated enough. I just don't know how to do this in a way that someone who knows how to program would. Are you just going to throw up your hands and give up? The point is not how beautiful your steps look. The point is, can you get there? Can you get there in a way that's accurate? If you have to MacGyver your way there with tape and spit, but it's accurate, then it's a success. — P02

Application users MacGyver when re-appropriating existing data tools for unintended users. One participant uses tools for data removal that provide summary statistics to initially assess and verify data transformation as so called "filter checks". Some participants (5/36) without the experience to implement common data join operations supported in database applications or scripting languages still integrated data using copy-and-paste or chained calls to spreadsheet macro functions, such as `VLOOKUP`.

Similarly, some data journalists with enough technical expertise to satisfy the hacker archetype, and who predominantly manipulate data with scripting and database languages, will MacGyver when they incorporate click-based applications into their preparation process. A few participants standardize data in OpenRefine due to the iterative control this application provides when performing this activity. One participant succinctly summarizes his reason for using multiple tools. "I care about getting a story that somebody else doesn't have", P18 says. "That's the job of journalists. I don't care what the tool is that lets me do it".

7.3 Implications for design

Based on our results, we outline three recommendations for the development of data preparation tools that address the needs of data journalists.

7.3.1 Support for verification activities. We find that verifying the effects of recently applied data transformations is a profiling activity previously unidentified in the research literature. Participants describe using the same methods in other profiling activities to confirm that their mental model of a table matches the data model, via spot checks or visual exploration and assessment. All user archetypes engage in verifying with a variety of tools.

From a design perspective, verifying describes one way in which users attempt to understand the state of data throughout the preparation process. Hence, its presence reveals a gulf of evaluation [56] with regard to data states represented in preparation tools. With many tools used participants, this gulf is big. But designers could shrink this gulf by incorporating better feedback about the system state. Features that leverage data visualization can provide feedback at a scale that is easier to interpret.

7.3.2 Support application users when integrating. As mentioned in Section 7.2, we do not find the same limiting relationship between data worker archetypes and data preparation activities reported in related work [35], especially concerning data integration. Despite not using the programming languages that implement join operations in relational algebra, application users who expressed *MacGyvering* tendencies still integrate data through creative uses of available tools.

To better support data journalists, data preparation applications need to offer better support for combining data from multiple

sources, especially those in our taxonomy of data integration scenarios (Section 6). While many applications used by participants, such as Tableau Prep [72], do attempt to implement join operations, scalability is still an issue. The technical capability or interface design of these applications limits their usability when integrating data at the scale described by many participants.

We also find that proficient programmers will use applications for specific activities because they provide the same utility but greater usability than code-based tools. Therefore, we believe users who currently prefer to integrate data using code-based tools may also consider applications for this activity if they offer the same utility but better usability.

7.3.3 Support for preparation documentation. Our study finds that both data scientists and data journalists create a record of data provenance when preparing data. While some preparation tools support provenance recording, version management, and workflow annotation, participants still perform this documentation activity in an external document, commonly called a data diary. This documentation artifact serves as a workaround in the absence of integrated data provenance tracking and reporting tools, and also reflects the extent to which participants used a very heterogeneous tool environment and could not rely on the internal capabilities of any single tool. We identify two limitations with the status quo that future tools should address.

First, preparation processes lack a cohesive medium to document workflow among the diverse set of preparation tools. Both data scientists and data journalists use a variety of tools when preparing data and often deploy their own idiosyncratic conventions for documenting data provenance. To the best of our knowledge, no system exists to ingest and unify provenance information from various applications, nor are there standards around the structure of data provenance information to promote interoperability between tools. Therefore, data workers must perform this consolidation manually in a word processor or text editor.

Second, data journalists report to a unique category of external stakeholder, the public, often publishing methodological posts documenting their data preparation in a so-called "nerd box" [39]. The step-by-step detail recorded in integrated documentation features used in data science are too granular for such public explanations, even for motivated citizens such as subject matter domain experts. Moreover, other stakeholders who are not directly working with the data such as supervising newsroom editors and collaborating journalists, also require a higher-level view of the data preparation processes. Journalists have a unique external audience in the form of motivated citizens, including subject matter domain experts, who may also be interested in high-level provenance information in the methodological posts that accompany many published instances of data journalism.

8 CONCLUSION

To understand how data preparation practices of data journalists compare to data scientists, we conduct an interview study with 36 data journalists and situate these results within research on data science workflows and dirty data. From these results, we propose a general taxonomy that considers data as a design artifact and dirty data as discrepancies between users' mental models, and we synthesize a process model of data preparation activities that data workers perform in pursuit of conforming data to one's mental model. We argue for the benefits of a more inclusive, pluralistic definition of data workers that includes both data scientists and data journalists. Although they perform many of the same preparation activities, we find important differences, including four challenges faced by journalists when combining tables during the preparation process: regional, diachronic, fragmented, and disparate datasets. Our findings can inform future work on the development of data preparation software. We encourage researchers to study and address the needs of all data workers, including data journalists.

ACKNOWLEDGMENTS

We thank the UBC InfoVis group, especially Sam Fraser, for their useful feedback on paper drafts; our anonymous reviewers for their helpful comments; and the journalists who participated in our study for volunteering their time. This work is funded in part by NSERC RGPIN-2014-06309 and the Wallenberg AI, Autonomous Systems and Software Program (WASP).

REFERENCES

[1] Timo Aho, Outi Sievi-Korte, Terhi Kilamo, Sezin Yaman, and Tommi Mikkonen. 2020. Demystifying Data Science Projects: A Look on the People and Process of Data Science Today. In *International Conference on Product-Focused Software Process Improvement* (Turin, Italy) *(PROFES '20)*. Springer, New York, NY, USA, 153–167. https://doi.org/10.1007/978-3-030-64148-1_10

[2] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. 2019. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *Transactions Visualization and Computer Graphics* 25, 1 (Jan. 2019), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

[3] Ángel Arrese. 2022. "In the Beginning Were the Data": Economic Journalism as/and Data Journalism. *Journalism Studies* 23, 4 (Feb. 2022), 487–505. https://doi.org/10.1080/1461670X.2022.2032803

[4] José Barateiro and Helena Galhardas. 2005. A Survey of Data Quality Tools. *Datenbank-Spektrum* 4, 14 (Aug. 2005), 15–21. http://dc-pubs.dbs.uni-leipzig.de/files/Barateiro2005ASurveyofDataQuality.pdf

[5] Andrea Batch and Niklas Elmqvist. 2018. The Interactive Visualization Gap in Initial Exploratory Data Analysis. *Transactions Visualization and Computer Graphics* 24, 1 (Jan. 2018), 278–287. https://doi.org/10.1109/TVCG.2017.2743990

[6] Leilani Battle and Jeffrey Heer. 2019. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum* 38, 3 (July 2019), 145–159. https://doi.org/10.1111/cgf.13678

[7] Charles Berret and Cheryl Phillips. 2016. *Teaching Data and Computational Journalism.* Columbia Journalism School, New York, NY, USA.

[8] Eddy Borges-Rey. 2021. *Journalism with Machines? From Computational Thinking to Distributed Cognition.* Amsterdam University Press, Amsterdam, Netherlands. 92–95 pages. https://doi.org/10.1515/9789048542079-023

[9] Richard E Boyatzis. 1998. *Transforming Qualitative Information: Thematic analysis and code development.* SAGE, Thousand Oaks, California.

[10] Paul Bradshaw. 2011. The Inverted Pyramid of Data Journalism. Retrieved Aug. 13, 2021 from https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism

[11] Suzana Guedes Cardoso. 2022. *The Practice of Data Journalism and Changes in the Professional Profile of Journalists in Newsrooms in the United States, United Kingdom, and Brazil.* Springer International Publishing, New York, NY, USA. 17–29 pages. https://doi.org/10.1007/978-3-030-74428-1_2

[12] Abhirup Chatterjee and Arie Segev. 1991. Data Manipulation in Heterogeneous Databases. *ACM SIGMOD Record* 20, 4 (Dec. 1991), 64–68. https://doi.org/10.1145/141356.141385

[13] Fanny Chevalier, Melanie Tory, Bongshin Lee, Jarke van Wijk, Giuseppe Santucci, Marian Dörk, and Jessica Hullman. 2018. From Analysis to Communication: Supporting the Lifecycle of a Story. In *Data-Driven Storytelling.* CRC Press, Boca Raton, FL, USA, 151–183. https://doi.org/10.1201/9781315281575-7

[14] Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational Journalism. *Commun. ACM* 54, 10 (Oct. 2011), 66–71. https://doi.org/10.1145/2001269.2001288

[15] Anamaria Crisan and Brittany Fiore-Gartland. 2021. Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop. In *Proceedings of the CHI*

Conference on Human Factors in Computing Systems (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445775

[16] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2020. Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. *Transactions Visualization and Computer Graphics* 27, 2 (Oct. 2020), 1860–1870. https://doi.org/10.1109/TVCG.2020.3030340

[17] Theordore Dasu, Tamraparni & Johnson. 2003. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Hoboken, NJ, USA.

[18] Wesley Gongora de Almeida, Rafael Timóteo de Sousa, Flávio Elias de Deus, Georges Daniel Amvame Nze, and Fábio Lúcio Lopes de Mendonça. 2013. Taxonomy of Data Quality Problems in Multidimensional Data Warehouse Models. In *Proceedings 8th Iberian Conference on Information Systems and Technologies* (Lisbon, Portugal) *(CISTI '13)*. IEEE, New York, NY, USA, 1–7. https://ieeexplore.ieee.org/document/6615784

[19] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press, Cambridge, Massachusetts.

[20] David Donoho. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26, 4 (Oct. 2017), 745–766. https://doi.org/10.1080/10618600.2017.1384734

[21] Shari L. Dworkin. 2012. Sample Size Policy for Qualitative Studies Using In-Depth Interviews. *Archives Sexual Behavior* 41, 6 (Sept. 2012), 1319–1320. https://doi.org/10.1007/s10508-012-0016-6

[22] Marc A Feldman. 2020. Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. *Quality Progress* 53, 2 (2020), 54–54.

[23] U.M. Feyyad. 1996. Data Mining and Knowledge Discovery: Making Sense Out of Data. *Expert* 11, 5 (Oct. 1996), 20–25. https://doi.org/10.1109/64.539013

[24] Katherine Fink and C. W. Anderson. 2015. Data Journalism in the United States. *Journalism Studies* 16, 4 (July 2015), 467–481. https://doi.org/10.1080/1461670X.2014.939852

[25] Bruce Frey (Ed.). 2018. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE, Thousand Oaks, California. https://doi.org/10.4135/9781506326139

[26] Garrett Grolemund and Hadley Wickham. 2014. A Cognitive Interpretation of Data Analysis: A Cognitive Interpretation of Data Analysis. *International Statistical Review* 82, 2 (Aug. 2014), 184–204. https://doi.org/10.1111/insr.12028

[27] Theresia Gschwandtner, Johannes Gärtner, Wolfgang Aigner, and Silvia Miksch. 2012. A Taxonomy of Dirty Time-Oriented Data. In *Multidisciplinary Research and Practice for Information Systems* (Prague, Czech Republic) *(CD-ARES '12)*. Springer, New York, NY, USA, 58–72. https://doi.org/10.1007/978-3-642-32498-7_5

[28] Joseph M Hellerstein. 2008. *Quantitative Data Cleaning for Large Databases*. Technical Report. University of California, Berkeley, Geneva, Switzerland.

[29] Bahareh Heravi, Kathryn Cassidy, Edie Davis, and Natalie Harrower. 2022. Preserving Data Journalism: A Systematic Literature Review. *Journalism Practice* 16, 10 (March 2022), 2083–2105. https://doi.org/10.1080/17512786.2021.1903972

[30] Karen Holtzblatt and Hugh Beyer. 2015. *Contextual Design: Evolved*. Spinger Nature, Berlin, Germany.

[31] Sarah Hutchins. 2020. Data Dive: School's Out. *The Investigative Reporters & Editors Journal* 43, 1 (Feb. 2020), 6–7.

[32] Kaggle. 2019. *State of Data Science and Machine Learning*. Kaggle. Retrieved May 15, 2022 from https://www.kaggle.com/kaggle-survey-2019

[33] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Puono. 2011. Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data. *Information Visualization* 10, 4 (Oct. 2011), 271–288. https://doi.org/10.1177/1473871611415994

[34] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, Vancouver, Canada, 3363–3372. http://dl.acm.org/citation.cfm?doid=1978942.1979444

[35] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *Transactions Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

[36] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12)*. ACM, New York, NY, USA, 547–554. https://doi.org/10.1145/2254556.2254659

[37] Eser Kandogan, Aruna Balakrishnan, Eben M. Haber, and Jeffrey S. Pierce. 2014. From Data to Insight: Work Practices of Analysts in the Enterprise. *Computer Graphics and Applications* 34, 5 (Sept. 2014), 42–50. https://doi.org/10.1109/MCG.2014.62

[38] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, and L. Licamele. 2008. Interactive Entity Resolution in Relational Data: A Visual Analytic Tool and Its Evaluation. *Transactions Visualization and Computer Graphics* 14, 5 (Sept. 2008), 999–1014. https://doi.org/10.1109/TVCG.2008.55

[39] Stephen Kasica, Charles Berret, and Tamara Munzner. 2020. Table Scraps: An Actionable Framework for Multi-Table Data Wrangling From An Artifact Study of Computational Journalism. *Transactions Visualization and Computer Graphics* 27, 2 (2020), 957–966. https://doi.org/10.1109/TVCG.2020.3030462

[40] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The Emerging Role of Data Scientists on Software Development Teams. In *Proceedings of the 38th International Conference on Software Engineering* (Austin, Texas) *(ICSE '16)*. ACM, New York, NY, USA, 96–107. https://doi.org/10.1145/2884781.2884783

[41] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *Transactions Software Engineering* 44, 11 (Nov. 2018), 1024–1038. https://doi.org/10.1109/TSE.2017.2754374

[42] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. 2003. A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery* 7, 1 (Jan. 2003), 81–99. https://doi.org/10.1023/A:1021564703268

[43] Won Kim and Jungyun Seo. 1991. Classifying Schematic and Data Heterogeneity in Multidatabase Systems. *Computer* 24 (Dec. 1991), 12–18. https://doi.org/10.1109/2.116884

[44] Lin Li, Taoxin Peng, and Jessie Kennedy. 2011. A Rule Based Taxonomy of Dirty Data. *International Journal of Computing* 1, 2 (Feb. 2011), 140–148.

[45] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376533

[46] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec. 2019), 1–23. https://doi.org/10.1145/3361118

[47] Hilary Mason and Chris Wiggins. 2010. A Taxonomy of Data Science. Retrieved February 9, 2021 from https://sites.google.com/a/isim.net.in/datascience_isim/taxonomy

[48] Marcus Messner and Bruce Garrison. 2017. Journalism's "Dirty Data" Below Researchers' Radar. *Newspaper Research Journal* 28, 4 (Aug. 2017), 88–100. https://doi.org/10.1177/073953290702800408

[49] Philip Meyer. 2002. *Precision Journalism: A Reporter's Introduction to Social Science Methods* (4th ed.). Rowman & Littlefield, Lanham, MD, USA.

[50] Microsoft. 2017. What is the Team Data Science Process? https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

[51] Microsoft. 2022. Schema Drift in Mapping Data Flow. https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-schema-drift

[52] Alessandra Maciel Pax Milani, Fernando V Paulovich, and Isabel Herb Manssour. 2020. Visualization in the Preprocessing Phase: Getting Insights From Enterprise Professionals. *Information Visualization* 19, 4 (Jan. 2020), 273–287. https://doi.org/10.1177/1473871619896101

[53] Michael Muller, Lange Ingrid, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Glasgow, UK) *(CHI '19)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[54] Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press, Boca Raton, FL, USA.

[55] Heiko Müller and Johann-Christoph Freytag. 2003. *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Technical Report. Humboldt Universität, Berlin, Germany.

[56] Donald A. Norman. 2013. *The Design of Everyday Things*. Basic Books, New York, NY, USA.

[57] Adegboyega Ojo and Bahareh Heravi. 2018. Patterns in Award Winning Data Storytelling. *Digital Journalism* 6, 6 (Nov. 2018), 693–718. https://doi.org/10.1080/21670811.2017.1403291

[58] Paulo Oliveira, Fátima Rodrigues, and Pedro Henriques. 2005. A Formal Definition of Data Quality Problems. In *Proceedings of the International Conference Information Quality* (Cambridge, MA, USA) *(ICIQ '05)*. MIT, Cambridge, MA, USA, 14. http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf

[59] Paulo Oliveira, Fátima Rodrigues, Pedro Henriques, and Helena Galhardas. 2005. *A Taxonomy of Data Quality Problems*. Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento. Retrieved April 6, 2022 from https://www.inesc-id.pt/ficheiros/publicacoes/2455.pdf

[60] Sylvain Parasie. 2022. *Computing the News: Data Journalism and the Search for Objectivity*. Columbia University Press, New York, New York, USA.

[61] Ryan Pitts and Lindsay Muscato. 2021. *Open-Source Coding Practices in Data Journalism*. Amsterdam University Press, Amsterdam, Netherlands. 191–193 pages. https://doi.org/10.1017/9789048542079.041

[62] Erhard Rahm and Philip A. Bernstein. 2001. A Survey of Approaches to Automatic Schema Matching. *The International Journal on Very Large Data Bases* 10 (Dec.

2001), 334–350. https://doi.org/10.1007/s007780100057

[63] Erhard Rahm and Hong Hai Do. 2000. Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.* 23, 4 (2000), 3–13.

[64] Sabbir M Rashid, James P McCusker, Paulo Pinheiro, Marcello P Bax, Henrique O Santos, Jeanette A Stingone, Amar K Das, and Deborah L McGuinness. 2020. The Semantic Data Dictionary: An Approach for Describing and Annotating Data. *Data Intelligence* 2, 4 (Oct. 2020), 443–486. https://doi.org/10.1162/dint_a_00058

[65] Rebecca S. Robinson. 2014. Purposive Sampling. In *Encyclopedia of Quality of Life and Well-Being Research*, Alex C. Michalos (Ed.). Springer, New York, NY, USA, 5243–5245. https://doi.org/10.1007/978-94-007-0753-5_2337

[66] Jan Roeder, Jan Muntermann, and Thomas Kneib. 2021. Towards a Taxonomy of Data Heterogeneity. *Band-1* 1, 1 (July 2021), 16. https://doi.org/10.30844/wi_2020_c6-roeder

[67] Simon Rogers. 2013. *Data Journalism Broken Down: What We Do to the Data Before You See It.* The Guardian. Retrieved August 16, 2022 from https://www.theguardian.com/news/datablog/2011/apr/07/data-journalism-workflow

[68] Simon Rogers, Jonathan Schwabish, and Dainelle Bowers. 2017. *Data Journalism in 2017: The Current State and Challenges Facing the Field Today.* Technical Report. Google News Lab, Mountain View, CA, USA.

[69] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Montréal, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173606

[70] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to do the Model Work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518

[71] Dilruba Showkat and Eric PS Baumer. 2021. Where do Stories Come From? Examining the Exploration Process in Investigative Data Journalism. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–31. https://doi.org/10.1145/3479534

[72] Tableau Software. 2018. Tableau Prep Builder. https://www.tableau.com/products/prep

[73] Daniele R de Souza, Lorenzo P Leuck, Caroline Q Santos, Milene S Silveira, Isabel H Manssour, and Roberto Tietzmann. 2018. Interacting with Data to Create Journalistic Stories: A Systematic review. In *International Conference on Human Interface and the Management of Information* (Las Vegas, NV, USA) *(HCII '18)*. Springer, New York, NY, USA, 685–704. https://doi.org/10.1007/978-3-319-92043-6_54

[74] Jonathan Stray. 2017. Making NLP Work for Investigative Journalism. Retrieved November 13, 2021 from https://www.youtube.com/watch?v=yRP9DL8E36A

[75] Jonathan Stray. 2019. Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism* 7, 8 (July 2019), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289

[76] Jonathan Stray. 2021. *Making Algorithms Work for Reporting.* Amsterdam University Press, Amsterdam, Netherlands. 90–91 pages.

[77] Jon Swain. 2018. *A Hybrid Approach to Thematic Analysis in Qualitative Research: Using a Practical Example.* SAGE, Thousand Oaks, FA, United States. https://doi.org/10.4135/9781526435477

[78] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–30. https://doi.org/10.1145/3359141

[79] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359313

[80] Hadley Wickham. 2014. Tidy Data. *Journal of Statistical Software* 59, 10 (Sept. 2014), 1–23. https://doi.org/10.18637/jss.v059.i10

[81] Rüdiger Wirth and Jochen Hipp. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (Manchester, UK) *(PAKDDM '00)*. Practical Application Company, Lancashire, UK, 29–39. http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

[82] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. 2019. Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study. *arXiv* 1, 1 (2019), 10 pages. https://doi.org/10.48550/arXiv.1911.00568 arXiv:arXiv:1911.00568

[83] Mary S Woodley. 2008. *Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information* (3rd ed.). Getty Research Institute, Los Angeles, CA, USA.

[84] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–23. https://doi.org/10.1145/3392826