CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Approaches to identifying data quality issues: The role of the data broker
Authors	Wibisono, Arif;Sammon, David;Heavin, Ciara
Publication date	2023-11-06
Original Citation	Wibisono, A., Sammon, D. and Heavin, C. (2024) 'Approaches to identifying data quality issues: the role of the data broker', Information Systems Management, 41(3), pp. 226–237. Available at: https://doi.org/10.1080/10580530.2023.2274532.
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1080/10580530.2023.2274532
Rights	© 2023, Taylor & Francis. This is an Accepted Manuscript of an article published by Taylor & Francis in Information Systems Management on 6 November 2023, available online: https://doi.org/10.1080/10580530.2023.2274532. This material is licensed under a Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) licence. - https://creativecommons.org/licenses/by-nc/4.0/
Download date	2026-11-02 13:53:11
Item downloaded from	https://hdl.handle.net/10468/15270



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Approaches to identifying data quality issues: The role of the data broker

Arif Wibisono[a]*, David Sammon[b], Ciara Heavin[b]

[a] *Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia;* [b]*Business Information Systems Department, Cork University Business School, University College Cork, Cork, Ireland*

**Corresponding author. E-mail: wibisono@is.its.ac.id*

Provide full correspondence details here, including e-mail for the *corresponding author

Provide short biographical notes on all contributors here if the journal requires them.

Approaches to identifying data quality issues: The role of the data broker

Data quality issues are problematic and costly for organizations. Employees (termed "Data Brokers") must identify data quality issues before data are used for reporting purposes. In five field studies, we investigate how these employees identify these often-hidden data quality issues. Organizations can execute five "checking" approaches: data templates, supervisor validation, data accuracy, data consistency, and data completeness. We discuss each approach, theorize their inter-relationships, and explain their contributions to research and practice.

Keywords: Data quality issues, field studies, data broker, manual identification, data curation

1 Introduction

Data quality issues can significantly impact organizational decisions and strategies (Cichy & Rass, 2019; Nagle et al., 2020), regardless of the organization's size (Global, 2021). For example, inaccurate financial reporting due to errors in data leads to misinformed business decisions (Hoque, 2013), incorrect budgeting (Hoque, 2013), and financial loss (Lee, 2012). Likewise, research shows that basic quality standards are only met by 3% of companies' data (Nagle et al., 2017). Subsequently, these data quality issues can affect customer satisfaction, operational costs, and organizational trust (Haug et al., 2011; Redman, 1998).

Moreover, research shows that the cost of rework due to data quality and accessibility issues is valued at USD 31.3 billion in the United States, USD 8.4 billion in Australia and New Zealand, and USD 10.2 billion in the United Kingdom. (Krishna et al., 2023). In fact, historically, research has indicated that the cost of data quality issues in the US alone was estimated to be at least $1.5 trillion annually (Fisher,

2009), which could be 8-12% of revenue (Redman, 1998). Nonetheless, the actual costs of data quality issues may be higher since it is difficult to determine their total costs (Nagle et al., 2017; Redman, 1998). Therefore, it is data, its quality, and not technology (although important) *"that represents the management challenge of the 21st century"* (Redman, 2023, p.28).

Data quality research proposes different methodologies to address data quality issues (Batini et al., 2009; Nagle et al., 2020; Pipino et al., 2002; Sebastian-Coleman, 2013). However, it is typically assumed to be a top-down *preventive* approach. For instance, Batini et al. (2009) and Pipino et al. (2002) suggest data-driven and process-driven strategies to avoid data quality issues through quality assessment and improvement. However, while *preventive* assumptions can help with data quality, they can be time-consuming and face obstacles. These can include infrastructure issues (Born & Krönung, 2016), a lack of skilled people (Steinhueser et al., 2017), poor software functionality (Davison et al., 2019; Ibrahim et al., 2019), and usability (Jagannath et al., 2019; van den Hooff & Hafkamp, 2017). These situations can lead to employees producing incorrect or incomplete data. Therefore, addressing data quality issues often involves a curative approach. In this approach, employees manually identify and correct issues after data creation but before data are used (cf. Wibisono et al., 2022a).

Moreover, this study investigates data quality issues related to producing managerial reports (in the context of data use), an essential process in all organizations. Producing such reports is crucial for effectively planning, controlling, and evaluating an organization's operations (Drum et al., 2017); otherwise, the organization may incur high costs and face unmanageable risks (Drum et al., 2017).

We refer to these employees, who manually identify and correct data quality issues to meet managerial reporting needs, as "Data Brokers". So, it is crucial to understand a data broker's approaches to identifying data quality issues. Therefore, our research question is as follows:

RQ. How do employees (performing the role of a data broker) manually identify data quality issues?

This study reports on five field studies in Indonesia to address this research question. Each organization has a different *reporting frequency* (e.g., daily to annually). Research shows that data are often used for shorter and longer reporting cycles (Kaplan & Norton, 1996). With a longer reporting cycle, an organization generates a more comprehensive view; however, it can delay decision-making (Weygandt et al., 2014), whereas a shorter reporting cycle can create a burdensome frequency of producing managerial reports (Pirttimäki & Liew, 2019; Weygandt et al., 2014). Despite these differences, these cycles share similar characteristics for identifying data quality issues. So, we are interested in finding out the shape of these approaches.

We contend that this human dimension (the data broker) to identifying data quality issues before they escalate and negatively impact decision-making is one of the most prominent roles in the modern-day organization, albeit not always captured in an employee 'role' title. Therefore, our results can indirectly aid data brokers by highlighting the approaches used to ensure those at the C-suite have the insights they need, at a tempo they need, and that these insights are based on 'quality' data.

Furthermore, given the 'accidental' data architecture that defines many organizations (often linked to business change through growth) that have data-driven aspirations, the role of the data broker must be both appreciated and understood, especially at the level where organizational strategies are decided.

This study makes three key contributions to data quality research. Firstly, it challenges the traditional belief that resolving data quality issues must be *preventive*. Secondly, it explores how data brokers *curatively* identify data quality issues. Third, it visualizes these data broker approaches as a flow.

The remaining sections are as follows: Section two briefly describes the role of the data broker. Section three considers our research design, while Section four discusses our data gathering and analysis. Section five discusses our findings, and Section six concludes our study.

2 The Role of the Data Broker

A data broker acts as an *intermediary* between data providers and data consumers, like distributors of manufactured goods (see Figure 1). Essentially, a data broker collects data from providers, distributes it to consumers, and provides confidence in their use (as good quality data) (Janssen & Singh, 2022).

A data broker may be an individual or a team in a data processing ecosystem (Janssen & Singh, 2022). They break down data silos that hinder decision-making (Zhu & Marjanovic, 2022), facilitate seamless data sharing (Wernick et al., 2020), add value to existing data for various stakeholders (e.g., producing managerial reports) (Yoon et al., 2018), and enhance communities' data literacy (Yoon et al., 2018). However, very little is known about these data brokers (individual or team).

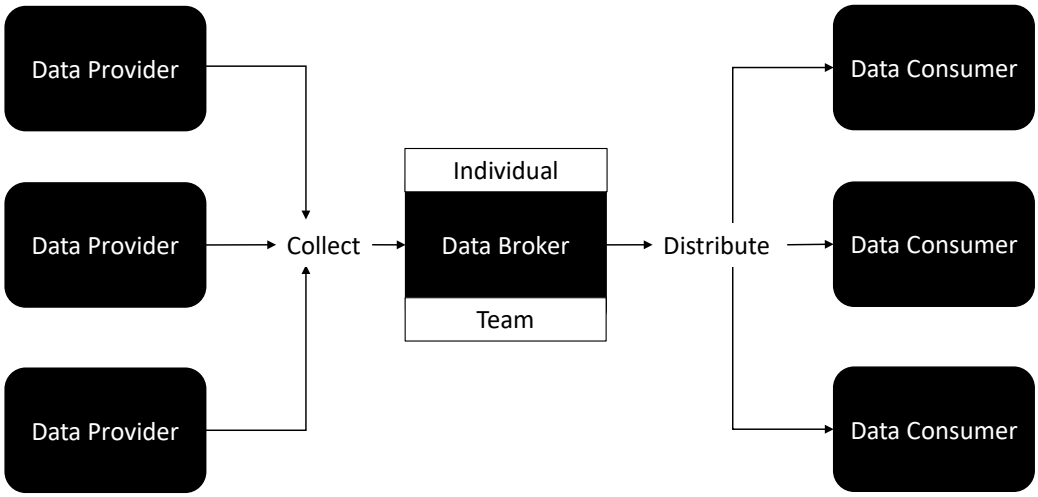


Figure 1. A data broker is a part of the data processing ecosystem in an organization. (Adapted from: (Janssen & Singh, 2022)

Therefore, exploring how data brokers identify data quality issues informs our research design (presented in the next section).

3 Research Design

We conduct five field studies informed by Eisenhardt's (1989) research design to explore how employees (the data brokers) identify data quality issues. This approach helps us generate a valid theory (Eisenhardt, 1989) from multiple field studies and understand the reasons behind similarities and differences across the field studies (Ridder, 2017).

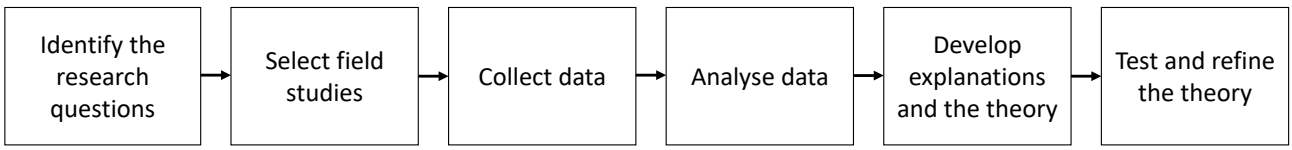


Figure 2. Step-by-step research design

The study involves the following steps: identify the research questions, select field studies, collect data, analyze data, develop explanations and the theory, and test and refine the theory (see Figure 2). The next section outlines the steps followed.

3.1 Identify the Research Questions

This initial stage involves defining the research problem and creating guiding research questions. This study uses research problematization (Alvesson & Sandberg, 2011; Sandberg & Alvesson, 2011) to develop the research question because research problematization is crucial to present different perspectives (Chatterjee & Davison, 2021). Therefore, we develop our research question by challenging assumptions in data quality research: addressing data quality issues must be *preventive*. We posit that addressing data quality issues can also be *curative*. So, we are challenging established research assumptions rather than taking existing findings for granted (Alvesson & Sandberg, 2011). As stated earlier, our research question seeks to learn more about the approaches of data brokers to identifying data quality issues as part of a report production cycle.

3.2 Select Field Studies

The next step is selecting relevant field studies, which can be organizations, individuals, events, or other units of analysis. These field studies must represent the phenomenon and provide comprehensive data for analysis. This study examines five organizational units acting as data brokers (see Table 1).

This strategy aligns with the recommendation made by Eisenhardt (1989), suggesting that a research design should have at least 4 to 10 field studies to be effective. Fewer than four field studies could make it challenging to generate a theory.

In contrast, more than ten studies could challenge data management (Eisenhardt, 1989).

Table 1. Field Studies

No	Field Study	Number of respondents	Employees % (Data Broker)	Organizational Unit	Reporting Frequency
1	Plantation Company	6 respondents	85%	QA (Quality Assurance) Unit (Head Office)	Daily
2	Furniture Company	5 respondents	100%	Administrative Office	Monthly
3	Hospital	5 respondents	100%	Pharmacy Unit	Three Monthly
4	University	3 respondents	75%	Academic Office	Six Monthly
5	Government Agency	3 respondents	100%	Government Bureau	Annually

Each organization requires employees (as data brokers) to manually identify data quality issues when producing managerial reports (see Table 1). The data are not limited to 'centralized systems' data like Enterprise Resource Planning (ERP) or Electronic Health Records (EHR). They may be from other sources, such as paper-based records or spreadsheets. Multiple data sources mean employees must *manually* double-check and determine the most appropriate data when producing reports as per the reporting frequency.

Across the five field studies, the data comes from various sources, including sugar factories, factory operations, pharmacy units, students, and local government agencies. So, by collecting the data from multiple sources, the respective organizational units (data brokers) can provide reports for internal and external stakeholders. Unfortunately, these processes cannot be automated because they involve much human judgment and interpretation. For example, identifying data

quality issues, verifying data accuracy, and making decisions on (how) to present the data in a report (Eisenhardt, 1989).

3.3 Collect Data

Step three involves collecting data from the chosen field studies to answer the research questions. The research focuses on operational employees (accounts of their data work) and examines documents, if available. However, due to COVID-19 restrictions, in-person observations were not possible. Therefore, this study collected data through online semi-structured interviews between June 2021 and June 2022. The interviews lasted between 45 and 60 minutes. They were conducted in Javanese and Indonesian. This interview data was then transcribed into English. See Table 2 for a sample of the excerpts gathered during the interviews.

3.4 Analyse Data

Step four involves analyzing the field study data, which entails identifying patterns and themes and utilizing techniques such as coding and categorization (see Table 2). We employ cross-case analysis to uncover similarities and differences across the five field studies (Ridder, 2017). The advantage of cross-case analysis is that it aids in establishing the "why" of the data patterns and the underlying theoretical arguments that explain them (Eisenhardt, 2021; Sarker et al., 2012). This benefit is critical since it goes beyond merely describing what is happening and explaining why it is occurring. This deeper level of comprehension can inform policy and practice decisions. So, it contributes to developing more effective interventions. An example of how we organized this cross-case comparison is provided in Table 3. While Table

3 shows variability, it also highlights commonality in the approaches used (by the data brokers) to identify data quality issues.

3.5 Develop Explanations and The Theory

Step five involves developing explanations and theories based on data analysis outcomes. This step requires identifying the concepts/categories contributing to the phenomenon being studied and creating a testable and valid yet parsimonious mid-range theory grounded in data (Eisenhardt, 1989). For example, where a particular pattern for identifying data quality issues is consistently observed in our interview data, we should incorporate this pattern into the emerging theory. As such, the theorizing process stops when coherence with the data has been achieved (Eisenhardt, 1989). In this research, we propose a flow diagram (see Figure 3) from a data broker's perspective. This theoretical explanation captures the core insights from our analysis of the five field studies.

3.6 Test and Refine the Theory

The final step is to test and refine the theories developed in the previous step. This step is essential to ensure the theories are valid and applicable to the studied context. It involves additional data collection, analysis, and comparison of the theories with existing literature (Eisenhardt, 1989). Theories are refined by adding new concepts or revising existing ones based on observed patterns. In this paper, throughout the findings section, we reflect on the uniqueness of our work and compare our emerging findings to existing literature. For example, automating data quality issue identification assumes that a single centralized dataset exists. However, this research suggests otherwise, where data existing beyond centralized systems

(necessary for producing managerial reports) demands that manual checking (performed by the data brokers) is a feature of identifying data quality issues.

4 Research Approach: Data Gathering and Analysis

We use open coding to analyze interview texts to identify and label concepts in the data (Böhm et al., 2004; Corbin & Strauss, 1990; Wolfswinkel et al., 2011). First, we analyze the texts line-by-line (Holton, 2012) to identify basic units of meaning. Second, we assign a descriptive code to each concept (Corbin & Strauss, 1990). Third, we categorize codes based on similarities and differences to create related categories, comparing incidents to identify patterns and themes (Corbin & Strauss, 1990). Fourth, we review and improve the categories to ensure that each category covers all cases and that there is no overlap between categories. Thus, we may rename, merge, split, or create new categories (Corbin & Strauss, 1990) to achieve this goal.

Table 2. Coding Sample

No	Excerpt	1st order	2nd order	Organization
1	I asked for Excel to work on. Now, from the Excel, (I) check the stock one by one.	Ask for Excel to work on	Check *data templates*	Hospital
2	I took the reports from my friends. That's what I see. I see this; how strange this is. I checked in more detail. It turns out it is not the same (working) template.	It's not the same (working) template	Check *data templates*	Plantation Company
3	Well, this is not signed by the supervisor. Well, this (letter) is already electronic. (So, we can check)	Check the supervisor's signature	Check *supervisor validation*	Government Agency
4	Then I checked the report (comparison). I checked whether this (document) was signed or not by the head unit.	Check whether this (document) is signed or not by the head unit.	Check *supervisor validation*	Government Agency
5	So, we take the data from the field of each sugar factory. For example, how old is this sugar cane?	Take data from the field	Check *data accuracy*	Plantation Company
6	Why isn't this filled in? How come it's empty?	Why isn't this filled in?	Check *data completeness*	Furniture Company
7	Yes, we immediately (ask) this. To those employees. Why isn't this filled in? How come it's empty?	Data not filled in	Check *data completeness*	Plantation Company
8	We are assisted by conditional formatting likewise when we cross-check with friends, assistant managers, and heads of affairs.	Cross-check with friends	Check *data consistency*	Plantation Company

9	So, in the Google Form, there was a print preview, and the print preview was downloaded, so it was printed and then matched (with the student's data in the system)	Matched data (with student data)	Check *data consistency*	University
10	We'll search. What (we) often find is that the stock is different between system stock and physical stock.	Check the differences between system stock and physical stock.	Check *data accuracy*	Hospital
11	Then, I check if this performance achievement (in the data) follows the population of [name of one regency].	Check whether the data follow the population of [name of one regency].	Check *data consistency*	Government Agency
12	We'll search. It's on the stock card (paper). But the entry (Information Systems) is not there.	Check the stock card and the Information Systems.	Check *data consistency*	Hospital
13	For example, our city has an average of 1000 students. If there are 38 districts, the result is 38000 students. Why does the total number of students in the city data differ from the provincial total?	Check why the city number is different from the local number.	Check *data consistency*	Government Agency

Table 3. Distribution of each excerpt across the field studies

No	Field Study	Checking					
		Data Templates	Supervisor Validation	Data Accuracy	Data Completeness	Data Consistency	Total
1	Plantation Company	5	0	4	6	2	17
2	Furniture Company	0	0	3	3	4	10
3	Hospital	3	3	6	2	3	17
4	University	5	0	0	3	5	13
5	Government Agency	3	3	2	4	5	17
	Total	16	6	15	18	19	74

Emerging from our open coding efforts (see Table 2), we identify five approaches that data brokers can use to *identify data quality issues*: checking (i) data templates, (ii) supervisor validation, (iii) data accuracy, (iv) data consistency, and (v) data completeness. Table 3 shows the occurrence frequency of the different approaches used in identifying data quality issues across the five field studies. These approaches are ordered logically (emerging from our analysis) and are represented visually in the conclusions section (see Figure 3).

Table 3 shows that the most common approach for the Furniture Company and the Government Agency is “Check Data Consistency” (5 and 4 occurrences, respectively). The most common approach for the Plantation Company is “Check Data Completeness” (6 occurrences). While the most common approach for the Hospital is “Check Data Accuracy” (6 occurrences). Finally, the most common

approach for the University is “Check Data Consistency” and “Check Data Templates” (both at 5 occurrences).

Table 3 shows that while there are variations in the approaches used within the field studies, four approaches are the most critical in identifying data quality issues. These approaches are checking *data consistency* (19 occurrences), checking *data completeness* (18 occurrences), checking *data templates* (16 occurrences), and checking *data accuracy* (15 occurrences). We discuss each approach in detail in the following section.

5 Findings and Discussion

This section discusses the advantages and disadvantages of each approach and demonstrates their interconnections.

5.1 Check Data Templates

This approach assumes that data quality issues can be identified by examining the artifact that stores the data (e.g., the *data collection templates*) (Xian et al., 2012). For example, as one employee (the plantation company) states, *“First, I take the report from my friends (on the package). I check (the data) in more detail. It turns out that it comes from a document that doesn’t match the template”*. Therefore, cross-referencing the data with the data collection template is foundational to identifying data quality issues.

Furthermore, utilizing features like conditional formatting in a spreadsheet (where the spreadsheet is the data collection template) is essential. One employee (the plantation company) states, “*I can use conditional formatting to make it faster. Then, we can also make it from the usual graph. Later, I can see if this is wrong or not. For*

example, if I'm wrong, I'll make it red in the conditional formatting. So (my work is) faster". Using data collection templates (provided by a data broker to a data provider) can alleviate the burden of managing extensive data, especially when the data come from various providers.

However, this approach has drawbacks as it assumes a single data collection template (artifact) can support all data types and formats, which may be incorrect. It also relies on algorithmic procedures, which may be prone to errors. These errors are difficult to spot by novice employees when using the artifact.

It is essential to acknowledge that this approach also has benefits. First, data brokers can reject data not submitted in the correct data collection template, avoiding future checks. Second, spreadsheet formulas can detect data anomalies, allowing providers to identify data quality issues earlier. This action can prevent data quality issues from spreading and allows for a more consistent routinized style of data evaluation (each time a data collection template is received). Third, this approach empowers employees to identify data quality issues independently using domain-specific software.

5.2 Check Supervisor Validation

This approach assumes that the data must be checked to ensure they are 'correct'. It is part of a *validation mechanism* (Wiggins et al., 2011) performed by the data provider. The supervisor is an employee who acts as the data provider's 'data quality' gatekeeper. This approach occurs when the data broker acquires data from a data provider (who can be another data broker). The cycle of providing data repeats until the desired 'quality' of data are obtained. This situation creates *a chain of brokers* that introduces uncertainty, as the length of the chain and the time it takes to

collect the data, cannot be pre-determined. Despite this, supervisor validation is essential to legitimize the data being used.

As one employee (the government agency) comments, *"We cross-check whether the head has (already) checked (submitted documents) or not. The head checks whether t*hese (data) are appropriate or not. Then, we (can) continue the *process"*. This statement indicates that the supervisor (as a senior manager) agrees with the data. Hence, their approval is necessary to ensure data quality. There are advantages and disadvantages to this approach.

We observe that this approach may lengthen the data management process. It requires employees to send data for approval and wait for the supervisor to approve it. This situation can cause delays in obtaining critical data if the supervisor is unavailable. Additionally, this approach can introduce bias to the data quality. The supervisor may *ignore* or *shortcut* the approval process if there is too much data to approve.

The approach promotes clear data ownership for the data provider. It fosters a culture of data responsibility (Vilminko-Heikkinen & Pekkola, 2019). Also, it specifies that certain data should come from specific providers, avoiding legal problems caused by versioning. Therefore, organizations can determine which data version has been approved for distribution.

5.3 Check Data Accuracy

This approach assumes that the original data sources are internal to the organization and not owned by external parties. For instance, if the data capture a machine's attributes, the machine must be present in the factory. If the stock is reported in a warehouse, the stock should be present in the warehouse. Data brokers

and providers can visit the site to compare the reported data with the observable reality and directly access the sources for further investigation. Therefore, access to the original data source(s) is necessary to check data accuracy.

Without this assumption, it would be challenging to conduct on-site visits. For example, a plantation company employee states, *"Are the farmers still (planting) sugarcane in the field? Please do something if it's not sugar cane (planted in the field). However, if the farmers say they still want to partner with us, we must know in which location. We must Check! Where?"*.

This approach has both advantages and disadvantages. Data accuracy means that the data correctly represents the real world (NCES, 1997). However, it can lead to additional costs due to checking multiple data sources or investing in technology. For example, drone technology can check the 'realities' without visiting the site in person. However, this solution requires a significant upfront investment, including software licenses, hardware, and IT staff to manage the technology. Overall, this approach is factual. It can result in more reliable data by capturing previously unreported phenomena. Previous research has confirmed that data accuracy is a crucial aspect of data quality (Nagle et al., 2020; Strong et al., 1997; Wang et al., 1995).

5.4 Check Data Consistency

This approach assumes that multiple data sources are reliable and accessible for comparison when checking data consistency. However, unlike data accuracy, data consistency focuses on consistency across various sources, even if the data does not represent the real world precisely. For example, one employee (the University) states, *"We contact (students) concerned. We let them know there is a difference between*

the ID card and their last diploma. So, we suggest to the student to look at the history. We ask the student to check from their elementary school diploma to the most recent one. Well, where are the mistakes?".

Two approaches are used to check for data inconsistencies. The first involves comparing the data with non-aggregated data from other sources. In contrast, the second involves comparing the data with *aggregate* data from other sources. Therefore, the data can be used if the data are consistent across multiple data sources.

This approach can lead to cognitive overload for employees due to data inconsistencies, different formats, and the need to find credible artifacts (data sources). However, the approach can also help find the best data by examining multiple sources. Employees can filter unrelated data versions and assess the importance of the data based on the data outputs and the priority of the data sources. Therefore, data consistency refers to ensuring that "*all clients can see the same data*" regardless of their location or access method (Rodger, 2011, p. 180). This approach assesses if the data are consistent across multiple accessible sources, such as spreadsheets and documents. Therefore, aligning with previous research, data consistency is significant as a data quality attribute (Wang & Strong, 1996).

5.5 Check Data Completeness

This approach assumes that cells of *blank data* are unacceptable. It highlights that each data attribute must hold a data value so that a data record can be considered complete, not partial, in capturing the observable 'reality'. For example, one employee (the plantation company) states, "*I'll return the spreadsheet if it's empty*". This approach has several positives and negatives (aligned with its simplicity as an approach).

Checking for data completeness alone cannot explain why the data are missing. Organizations must gather all the relevant data together to create a more complete explanation, even though all columns are filled with specific values (not empty). These columns may not always provide an accurate account. For example, the reality captured by the data can change over time. Therefore, it is not accessible (at a time in the future) for fact-checking and completing a data record.

Furthermore, checking for data completeness does not require checking other sources or complex interpretations to identify data quality issues. It also reduces the need for advanced statistical knowledge/techniques in detecting these issues, making it useful when there is a high volume of data to check. Therefore, *data completeness* refers to the extent to which we have *"stable data to preserve the information we need"* (Johnston, 2014, p. 278). As suggested in previous research, organizations use data completeness as a critical component of data quality assessment (Nagle et al., 2020; Redman, 1998).

6 Conclusion

This study explores the role of a data broker (as an individual or a team) in identifying data quality issues in organizations through the insights gained from five field studies in Indonesia. Our analysis emphasizes the importance of identifying data quality issues. It discusses the five "checking" approaches used by data brokers, namely (checking): *data templates*, *supervisor validation*, data *accuracy*, *data consistency*, and *data completeness*. This research supports data brokers by visualizing their strategies for identifying data quality issues as a flow. This flow is now discussed (see Figure 3).

6.1 A Flow Diagram to Identify Data Quality Issues

Our flow diagram assumes several things. First, the data examined is a snapshot of all operational data available. Second, the primary role of a data broker is to identify data quality issues. Third, human subjectivity is used to determine the 'quality' of the data (Kooper et al., 2011). Fourth, a supervisor (within a data provider) has sufficient knowledge and skills to assess data quality. Fifth, human intervention is necessary to address specific data quality issues. Sixth, identifying data quality issues is only relevant in the context of the 'known' intended use of the data.

Furthermore, we posit that the flow must start with check *data templates*. This "checking" approach sets the context and triggers all other approaches (as required) in the flow (see Figure 3). All five approaches exist on a continuum, enabling data brokers to identify the locus of the data quality issue(s) (seeing as data brokers typically do not produce the data, this continuum is essential). Therefore, a data broker may be unfamiliar with the data nuances and idiosyncrasies. For example, one employee (the plantation company) states, "*The loss in factory A is unreasonable. It is too high compared to other factories. That's when we go down for supervision (to the field). We check the problem: wrong data entry, analysis, or something else*." By identifying data quality issues through this continuum of "checking" approaches, organizations can take *curative* measures to resolve the data quality issues.

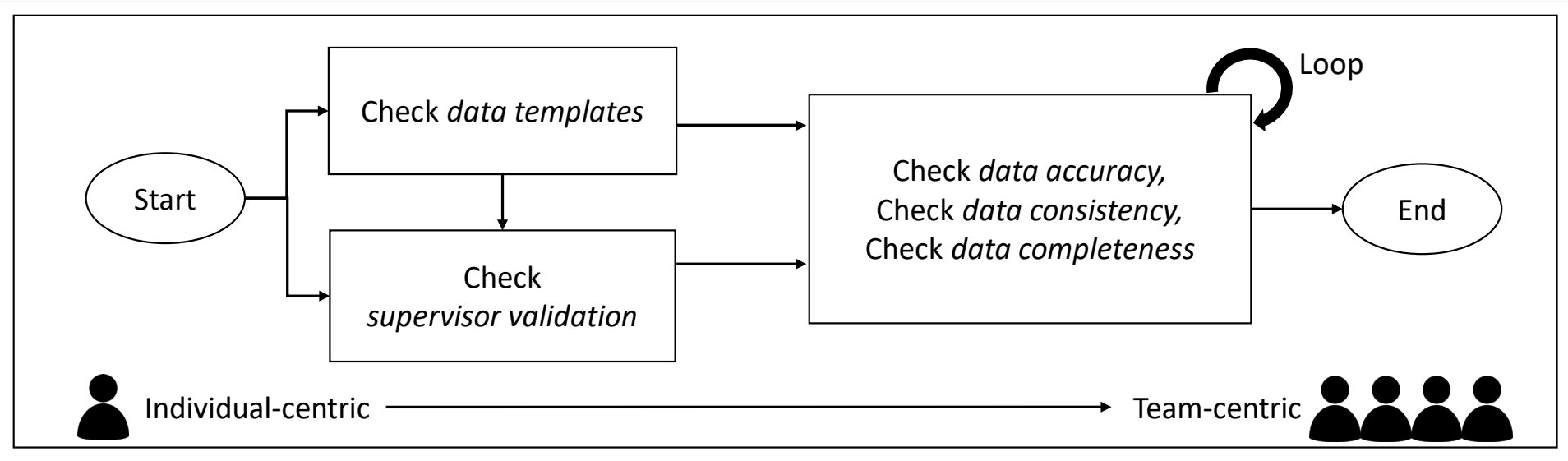


Figure 3. Identifying data quality issues: A data broker's perspective

Checking *data accuracy*, checking *data consistency*, and checking *data completeness* are understood to be both individual- and team-centric activities. As a flow, these approaches can also be recursive (see Figure 3), thereby producing multiple interpretations of the data quality issue(s) (as more individuals become involved in these "checking" approaches). This notion equates well to the idea that data is a *"team sport"* (Redman, 2023, p. 125). Furthermore, identifying data quality issues does not ensure sustained data quality. *First*, the sources and phenomena may become inaccessible over time, making it hard to benchmark current data with reality. *Second*, the flow does not address problems in the original data sources, leading to repeated mistakes and data quality issues in the future.

Identifying data quality issues stops if employees feel they have achieved order, clarity, and rationality (Weick, 1995) against the data. Therefore, employees can demand *more* or *different information* from other sources to achieve these. It resolves any ambiguity and equivocality experienced by earlier identification flow.

Novice data brokers in an organization need training to understand data types and ensure their order, clarity, and logic. This process takes time, and these novices may misunderstand data and non-data quality issues. Experienced data brokers should review and assist novices in developing the right mindset to prevent this situation. One employee (the Hospital) states, *"Yes, (we) must (continue to) train employees, and it must be unified with the service. (After some time), this way of thinking becomes a reflex"*.

Through continuous training, data brokers can develop data maturity where they deeply understand the data and its production process. For instance, an employee from the furniture company states, "*So I have to (repeatedly) look again and look*

again. It's always. So, (I) understand. (I) have memorized them - because every day I observe the data one by one". This finding confirms previous research suggesting that an employee's work, training, and life experiences can contribute to a better understanding of a phenomenon (cf. Weick et al., 2005).

6.2 Theoretical Contributions

We outline several contributions to data quality research. First, we challenge the idea that addressing data quality issues must be *preventive* (e.g., *before* the data are produced). We show that tackling data quality issues can also be *curative* (e.g., after the data are produced). This contribution can result in a new strand of research in data governance structures, policies, and rules focused on *curative* measures.

Second, we outline five "checking" approaches for identifying data quality issues. Each of these are mutually exclusive and exhaustive. Each approach can serve as a theory for analyzing – Theory Type 1 (Gregor, 2006, p. 622). These approaches can be utilized as a theoretical framework to analyze how organizations address data quality issues such as completeness, accuracy, consistency, and availability.

Third, we connect these "checking" approaches as a flow and thereby reveal the "evaluate data" process as a workaround-cfrodgerentric data activity (Wibisono et al., 2020, 2022b, 2022a) (c.f. Wibisono et al., 2020, 2022b, 2022c). Other data activities (WCDA) can be combined to solve data quality issues, such as *compiling*, *transferring*, *retrieving*, and *storing* data (Wibisono et al., 2022a). Therefore, we argue that identifying data quality issues is part of the "evaluate data" process, followed by "formulating data solutions."

Fourth, we address a call in the existing literature to unpack how data brokers operate in practice (Janssen & Singh, 2022). By doing so, we can reveal data brokers' governance mechanisms, which often tend to be non-transparent to other parties (Janssen & Singh, 2022). As they are non-transparent, research shows that an uncontrolled data broker can produce "*hub-and-spoke cartels*" (Wernick et al., 2020). This situation leads to unintended abuse of collective dominance (Wernick et al., 2020). Our research directly contributes to the literature by revealing how data brokers operate in a specific context (e.g., identifying data quality issues).

Moreover, identifying data quality issues can be complex. We describe the conditions that enable each "checking" approach, and the flow diagram serves as a theory to explain, proposing a Theory Type 2 (Gregor, 2006, p. 624). This diagram helps in theorizing on data consequences and their economic effects. This theorization includes identifying and categorizing costs arising from data quality issues.

6.3 Managerial Implications

We conceptualize the 'flow' to identify data quality issues. It is part of a more extensive intertwined socio-technical system that comprises four components: people, process, technology, and data (Alter, 2013). To prevent negative impacts on the overall system, a data broker must intervene in all components of the socio-technical system. This action is necessary because issues in one component can be closely linked with issues in another (see Table 4).

While data quality issues are often the primary concern for organizations, addressing them requires consideration of the processes and technologies used to manage the data. For instance, a company with a robust data management system

may still experience data quality issues if employees lack proper training. Likewise, a skilled data analysis team may encounter data quality issues using outdated or inadequate technology. Therefore, a comprehensive approach to addressing data quality issues requires a holistic approach, which involves considering the socio-technical system rather than focusing on one component (e.g., technology) in isolation.

Table 4. How to improve identifying data quality issues

No	Approach	Possible Issue	Possible Solution
1	Data Templates	A universal data artifact that can accommodate all types and formats may not be possible.	Train employees to use the right artifact for data collection.
			Procure necessary artifacts that can support the identification process.
2	Supervisor Validation	Prolong the data management process.	Recognize and reward supervisors who show a strong work ethic.
			Set clear expectations of data and their process.
3	Supervisor Validation	Delays in obtaining critical data - if the supervisor is unavailable.	Establish clear communication channels.
			Create backup procedures when supervisors are not available.
4	Supervisor Validation	Compromised data quality if the supervisor ignores or shortcuts the approval process.	Implement automated approval systems (if possible).
5	Data Accuracy	Extra costs for checking multiple data sources.	Prioritise data sources.
6	Data Consistency	Cognitive overload for employees - due to data inconsistencies.	Encourage employees to communicate when they find inconsistencies.
7	Data Consistency	Cognitive overload for employees - due to different data formats.	Standardise data formats.
			Provide tools for data transformation.
8	Data Consistency	Cognitive overload for employees - due to the need to find credible artifacts.	Create a list of trusted sources.
			Provide training for employees.
9	Data Completeness	Failure to explain why data are missing.	Investigate the reasons for missing data.
10	Data Completeness	Failure to provide an accurate account because reality can change or become inaccessible.	Establish a process of continuous monitoring of data.

6.4 Limitation and Further Research

The current research has limitations that researchers should consider in future studies. The research focuses on the perspective of a data broker. So, it may affect the generalisability of the findings in non-brokering contexts. For example, a small business owner who uses their data for decision-making and does not involve external parties is an example of a context without a data broker. Another example could be a researcher who collects and analyses data without involving other individuals or organizations.

Secondly, this study fails to account for the impact of emotions and power dynamics (Maitlis & Christianson, 2014) on identifying data quality issues. Future research should consider these factors to improve data governance and mitigate potential data quality issues. Also, developing a taxonomy of data quality issues and their chain of action could help organizations identify the root cause of issues and the actions necessary to prevent them.

Furthermore, a data broker's diverse knowledge and skills, such as domain-specific knowledge, statistical analysis, programming, and data visualization skills, can significantly impact identifying data quality issues. Individuals with varied expertise can bring different perspectives to identify potential issues in the data. Future research should focus on strategies organizations can use to encourage collaborative identification of data quality issues and ensure employees have the necessary knowledge and skills.

A 'solutions to data quality issues' taxonomy can aid organizations in developing effective strategies for resolving data quality issues. Therefore, future research can examine the types of solutions used, their effectiveness, and common

barriers to implementation. These research opportunities lead to a sustainable framework for data issue identification, improving data quality, and governance initiatives.

Ultimately, effectively identifying data quality issues requires a mix of manual and automated approaches. Although technology plays a part, it is not enough to detect all data quality issues. Therefore, human involvement (e.g., data brokers) is crucial to maintaining data quality over time. For example, data brokers are vital links between data producers and consumers. They meticulously evaluate all avenues to identify data quality issues. They communicate their findings and recommendations to stakeholders. Moreover, failure to monitor data quality issues can cause organizations to over-rely on automated processes or overlook critical issues that require manual processes. By implementing a more hands-on approach, data brokers help organizations stay ahead of data quality issues.

References:

Alter, S. (2013). Work system theory: Overview of core concepts, extensions, and challenges for the future. *Journal of the Association for Information Systems*, *14*(2), 72–121. https://doi.org/10.17705/1jais.00323

Alvesson, M., & Sandberg, J. (2011). Generating Research Questions through Problematization. *Academy of Management Review*, *36*(2), 247–271. https://www-jstor-org.libezproxy.open.ac.uk/stable/pdf/41318000.pdf?refreqid=excelsior%3A6fbc0ddf365a66bdceded563c806e0ff

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3). https://doi.org/10.1145/1541880.1541883

Böhm, A., Glaser, B., & Strauss, A. (2004). Theoretical Coding: Text Analysis in Grounded Theory. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A Companion to Qualitative Research* (Issue January 2004, pp. 270–275). Sage Publications Ltd.

Born, F., & Krönung, J. (2016). Deviant cloud usage in public institutions - A matter of personal innovativeness? *Twenty-Fourth European Conference on Information Systems (ECIS)*. http://aisel.aisnet.org/ecis2016_rip%0Ahttp://aisel.aisnet.org/ecis2016_rip/19

Chatterjee, S., & Davison, R. M. (2021). The need for compelling problematisation in research: The prevalence of the gap-spotting approach and its limitations. *Information Systems Journal*, *31*(2), 227–230. https://doi.org/10.1111/isj.12316

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

Corbin, J. M., & Strauss, A. L. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Davison, R. M., Wong, L. H. M. M., Alter, S., & Ou, C. X. J. (2019). Adopted Globally But Unusable Locally : What Workarounds Reveal About Adoption , Resistance , Compliance and Non-Compliance. *Twenty-Seventh European Conference on Information Systems (ECIS2019)*, 0–16.

Drum, D., Pernsteiner, A., & Revak, A. (2017). Workarounds in an SAP environment: Impacts on accounting information quality. *Journal of Accounting and Organizational Change*, *13*(1), 44–64. https://doi.org/10.1108/JAOC-05-2015-0040

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, *14*(4), 532–550. https://www.jstor.org/stable/pdf/258557.pd

Eisenhardt, K. M. (2021). What is the Eisenhardt Method, really? *Strategic Organization*, *19*(1), 147–160. https://doi.org/10.1177/1476127020982866

Fisher, T. (2009). *The Data Asset: How Smart Companies Govern Their Data for Business Success*. Wiley. https://www.wiley.com/en-ca/The+Data+Asset%3A+How+Smart+Companies+Govern+Their+Data+for+

Business+Success+-p-9780470508008
Global, D. (2021). *The DQ Guide to Data Quality in 2021*. https://www.dqglobal.com/wp-content/uploads/2021/11/The-DQ-Guide-to-Data-Quality-in-2021-1.pdf
Gregor, S. (2006). The nature of theory in Information Systems. *MIS Quarterly*, *30*(3), 611–642. https://aisel.aisnet.org/misq/vol30/iss3/5/
Haug, A., Zachariassen, F., & van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, *4*(2), 168–193. https://doi.org/10.3926/jiem.2011.v4n2.p168-193
Holton, J. A. (2012). The Coding Process and Its Challenges. *The SAGE Handbook of Grounded Theory*, *9*(1), 265–289. https://doi.org/10.4135/9781848607941.n13
Hoque, Z. (2013). *20 Essential Questions Corporate Directors Should Ask About Strategy*. Business Expert Press.
Ibrahim, S., Donelle, L., Regan, S., & Sidani, S. (2019). A Qualitative Content Analysis of Nurses' Comfort and Employment of Workarounds With Electronic Documentation Systems in Home Care Practice. *Canadian Journal of Nursing Research*, *52*(1), 084456211985550. https://doi.org/10.1177/0844562119855509
Jagannath, S., Sarcevic, A., Young, V., & Myers, S. (2019). Temporal Rhythms and Patterns of Electronic Documentation in Time-Critical Medical Work. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300564
Janssen, H., & Singh, J. (2022). Data intermediary. *Internet Policy Review*, *11*(1), 1–9. https://doi.org/https://doi.org/10.14763/2022.1.1644 Published:
Johnston, T. (2014). *Bitemporal Data: Theory and Practice*. Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-408067-6.00021-8
Kaplan, R. S., & Norton, D. P. (1996). *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business Press.
Kooper, M. N., Maes, R., & Lindgreen, E. E. O. R. (2011). On the governance of information: Introducing a new concept of governance to support the management of information. *International Journal of Information Management*, *31*(3), 195–200. https://doi.org/10.1016/j.ijinfomgt.2010.05.009
Krishna, C. M., Ruikar, K., & Jha, K. N. (2023). Determinants of Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework. *Buildings*, *13*(944).
Lee, T. A. (2012). *Financial and Managerial Accounting for School Administrators: Superintendents, School Business Administrators and Principals*. R\&L Education.
Maitlis, S., & Christianson, M. (2014). Sensemaking in Organizations : Taking Stock and Moving Forward. *The Academy of Management Annals*, *8*(1), 57–125. https://doi.org/10.1080/19416520.2014.873177
Nagle, T., Redman, T., & Sammon, D. (2017). *Only 3% of Companies' Data Meets Basic Quality Standards*. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards#
Nagle, T., Redman, T., & Sammon, D. (2020). Assessing data quality: A managerial call to action. *Business Horizons*, *63*(3), 325–337. https://doi.org/10.1016/j.bushor.2020.01.006

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Pirttimäki, V., & Liew, A. (2019). Does shortened reporting cycle improve quality of external financial reporting? *Journal of Applied Accounting Research*, *20*(3), 321–341. https://doi.org/10.1108/JAAR-02-2018-0035

Redman, T. (1998). The impact of Poor Data Quality on the Typical Enterprise. *Communications of the ACM*, *41*(2), 79–82.

Redman, T. (2023). *People and Data: Uniting to Transform Your Business*. KoganPage.

Ridder, H. G. (2017). The theory contribution of case study research designs. *Business Research*, *10*(2), 281–305. https://doi.org/10.1007/s40685-017-0045-z

Rodger, R. (2011). *Beginning Mobile Application Development in the Cloud*. John Wiley & Sons. https://www.wiley.com/en-us/Beginning+Mobile+Application+Development+in+the+Cloud-p-9781118203354

Sarker, S., Xiao, X., & Beaulieu, T. (2012). Towards An Anatomy of "Successful" Qualitative Research Manuscripts in IS: A Critical Review and Some Recommendations. *Thirty Third International Conference on Information Systems*, 1–21.

Sebastian-Coleman, L. (2013). *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework* (L. Sebastian-Coleman (ed.)). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-397033-6.00020-1

Steinhueser, M., Waizenegger, L., Vodanovich, S., & Ritcher, A. (2017). Knowledge management without management- Shadow IT in Knowledge-Intensive Manufacturing Practices. *Twenty-Fifth European Conference on Information Systems (ECIS)*, *2017*, 1647–1662. https://pure.itu.dk/portal/files/82377908/2017_Steinh_ser_et_al_ECIS_KM_without_Mgt.pdf

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context DQ problems. *Communications of the ACM*, *40*(5), 103–110. http://delivery.acm.org/10.1145/260000/253804/p103-strong.pdf?ip=141.59.82.181&id=253804&acc=ACTIVE SERVICE&key=2BA2C432AB83DA15.6118EF98202DC727.4D4702B0C3E38B35.4D4702B0C3E38B35&__acm__=1519660767_f4fa7913c4dfb8d4987d460a4b761e1b%0Ahttp://portal.acm.org

van den Hooff, B., & Hafkamp, L. (2017). Dealing with Dissonance: Misfits between an EHR System and Medical Work Practices. *Thirty-Eighth International Conference on Information Systems*, 0–17.

Vilminko-Heikkinen, R., & Pekkola, S. (2019). Changes in roles, responsibilities and ownership in organizing master data management. *International Journal of Information Management*, *47*(April 2018), 76–87. https://doi.org/10.1016/j.ijinfomgt.2018.12.017

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.2307/40398176

Weick, K. E. (1995). *Sensemaking in Organizations*. Sage Publications.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wernick, A., Olk, C., & von Grafenstein, M. (2020). Defining Data Intermediaries: A Clearer View through the Lens of Intellectual Property Governance. *Technology and Regulation*, 65–77. https://doi.org/10.26116/techreg.2020.007

Weygandt, J. J., Kimmel, P. D., & Kieso, D. E. (2014). *Managerial Accounting: Tools for Business Decision Making* (7th ed.). Wiley.

Wibisono, A., Sammon, D., & Heavin, C. (2020). Modelling data activities in workarounds: a narrative network approach. *Journal of Decision Systems*, *29*(Sup1), 1–12. https://doi.org/10.1080/12460125.2020.1778874

Wibisono, A., Sammon, D., & Heavin, C. (2022a). Data availability issues : decisions as patterns of action. *Journal of Decision Systems*, 1–14. https://doi.org/10.1080/12460125.2022.2070945

Wibisono, A., Sammon, D., & Heavin, C. (2022b). Plausible Pictures for Data Governance: A Narrative Network Approach. *30th European Conference on Information Systems*. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1109&context=ecis2022_rp

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). Mechanisms for data quality and validation in citizen science. *Proceedings - 7th IEEE International Conference on e-Science Workshops, EScienceW 2011*, 14–19. https://doi.org/10.1109/eScienceW.2011.27

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2011). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Xian, Y., Fonarow, G. C., Reeves, M. J., Webb, L. E., Blevins, J., Demyanenko, V. S., Zhao, X., Olson, D. M., Hernandez, A. F., Peterson, E. D., Schwamm, L. H., & Smith, E. E. (2012). Data quality in the American Heart Association Get with the Guidelines-Stroke (GWTG-Stroke): Results from a National Data Validation Audit. *American Heart Journal*, *163*(3), 392-398.e1. https://doi.org/10.1016/j.ahj.2011.12.012

Yoon, A., Copeland, A., & McNally, P. J. (2018). Empowering communities with data: Role of data intermediaries for communities' data utilization. *Proceedings of the Association for Information Science and Technology*, *55*(1), 583–592. https://doi.org/10.1002/pra2.2018.14505501063

Zhu, J., & Marjanovic, O. (2022). A Taxonomy of Data Cooperatives. *Pacific Asia Conference on Information Systems (PACIS)*.