What is Data Quality? Defining Data Quality in the Age of AI

Roman Lukyanenko
McIntire School of Commerce
University of Virginia
romanl@virginia.edu

Abstract

The value of data hinges on its quality, which is not solely defined by accuracy or completeness but also by ethical, legal, and contextual considerations. This article reviews the concept of data, examines the evolution of definitions of data quality (information quality), and introduces the FACT+ Framework - Fairness, Accuracy, Completeness, Timeliness, and other contextually relevant dimensions (PLUS) - as a comprehensive approach to understand and improve data quality. FACT+ provides a long-overdue update to understanding data quality to support data-driven developments, such as analytics, artificial intelligence and smart products and services.

Keywords

Data, Data Quality, Information Quality, FACT+ Framework, Fairness, Accuracy, Completeness, Timeliness, Information Systems, Ethical Data Use, Data Representation, Decision-Making, Artificial Intelligence, Data Integrity, User Interface Design, Data Ethics, Data Governance.

1. The Costly and Elusive Data Quality

In today's digital world, data has become the foundation upon which our most consequential decisions rest. From the algorithms powering healthcare diagnostics to financial systems managing trillions in assets, from smart city infrastructure to personalized customer experiences - all depend fundamentally on the quality of their underlying data. Yet despite this pivotal importance, a surprising ambiguity surrounds “data quality.”

Organizations worldwide invest billions in data collection, storage, and analysis, yet many struggle with the most fundamental question: how do we know if our data is any good? The consequences of poor data quality are far-reaching and often devastating. Medical misdiagnoses, financial losses, algorithmic bias, and failed strategic initiatives can all trace their roots to compromised data. As artificial intelligence systems increasingly make autonomous decisions affecting human lives, the stakes of data quality have never been higher.

In the US alone, the estimated cost of poor data quality is $3 trillion [26, 80]. Other assessments suggest organizations spend up to 30% of their revenue handling data quality issues [26]. These shocking figures are a vivid testimony of how important data has become. With so much at stake, it is outrageous that we lack the understanding of data quality that considers all aspects of the diverse digital landscape.

The challenge lies not merely in technical implementation but in conceptual clarity around data quality. Traditional academic definitions of data quality focus on the needs of data consumers, rather than intrinsic properties of data. Often definitions merely list attributes, such as accuracy and completeness, but even so, they fail to capture the ethical, legal, and contextual dimensions that determine whether data is actually usable. Things are not better in popular sphere. When we ask an AI system like ChatGPT to define data quality, we receive technically sound but incomplete answers that reflect the very confusion plaguing the field.

What makes this conceptual murkiness particularly troubling is that we cannot effectively manage what we cannot define. In a world increasingly defined by data-driven decision-making, understanding data quality is essential to responsible technological advancement and organizational success. The question remains: in a landscape where data underpins virtually every aspect of modern life, how do we know what makes data truly "good"?

We introduce the FACT+ Framework, an integrated approach to data quality that addresses not only the technical aspects of data (accuracy, completeness, timeliness) but also the ethical and legal foundations through fairness.

The framework acknowledges that high-quality data must faithfully represent underlying reality (Accuracy), include all relevant aspects of what it purports to represent (Completeness), remain current and reflect any change (Timeliness), while being collected and used in ways that respect ethical principles and legal requirements (Fairness). This holistic approach recognizes that data cannot be "high quality" if it was obtained through deceptive means or violates privacy norms, regardless of its technical perfection.

By elevating ethical considerations to the same level as technical attributes, the FACT+ Framework provides a comprehensive foundation for evaluating data quality in an increasingly complex digital landscape. It shifts the conversation from merely technical specifications to the broader question of how data serves human needs while respecting human rights and values. This integrated perspective is crucial as organizations grapple with responsible AI development and deployment, where data quality issues can amplify into significant debacles.

2. Understanding Data Quality

2.1. Briefly about data and information

Before discussing data quality, let us first establish a general understanding of data itself. Data can take various forms - numbers, text, images, or even sensory inputs - and can be structured (like entries in a database) or unstructured (like a collection of social media posts). Understanding what data is and how it functions is essential before we deal with its quality.

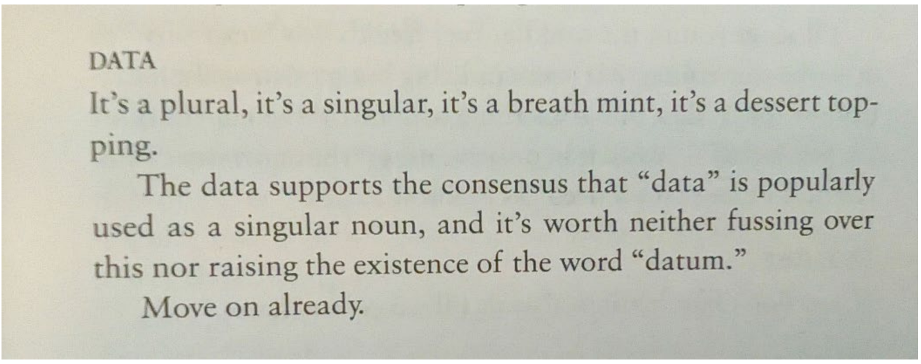
Data is representation of any object or event in some physical medium. The value of data is in its ability to convey something about the object that we care to know. By using data, rather than interacting directly with the object itself, we can be more efficient and effective [103]. For example, a surveillance camera installed on the corner of the parking garage can be used to monitor the garage remotely. This can be safer, cheaper, healthier (perhaps, even, more enjoyable), than physically being inside the garage, rain or shine, and inhaling vehicle exhausts.

Sometimes data makes it possible to know something about the object represented, which would be impossible without data. It is only through data (such as words, drawn images), that we can know the contents of somebody else’s mind. Only data opens a window into human past. It is only through data (e.g., words, videos, text), that we can know about anything we have not directly experienced.

It is only through data that we can know about anything we have not directly experienced

Often when you hear the word data, the word **information** comes to mind. Information is sometimes distinguished from data as data that is placed into a meaningful context, that is: information=data + meaning. We choose not to make this distinction. We fail to find cases when data is devoid of meaning. Try this as a mental exercise: is there a data that is absolutely without meaning? And conversely, when you think of some information, are you always fully confident in its meaning? If so, what does misinformation mean? By treating data and information as synonyms, we follow many scholars and practitioners who do the same [10, 53, 88, 100, 107].

We use data as a singular noun. Data is the plural form of Latin *datum* - given, but the singular *datum* is rarely used (have you ever heard it in speech?). In most cases singular usage sounds more natural. Refuse to take my word for it? Heed the advice of Benjamin Dreyer [31], a fixture in the publishing industry and authority on modern English style and grammar (Figure 1).


DATA

It's a plural, it's a singular, it's a breath mint, it's a dessert topping.

The data supports the consensus that "data" is popularly used as a singular noun, and it's worth neither fussing over this nor raising the existence of the word "datum."

Move on already.

Figure 1. Data to back up my singular usage of data: Page 153 in Dreyer [31]

By manipulating data as a proxy for the objects it represents, we are able to understand and act upon the world in a way that would be impossible or difficult to do without data. This premise of data underlies the basic value of information systems, and information technology.

The very value of information technology is in the ability to create information systems to handle data efficiently, at scale, in a way that would be difficult for humans to do manually. A calculator is a simple example of this efficiency, whereas a driverless car is a more advanced example. Recently, this impressive efficiency has been provided by generative artificial intelligence, through tools such as ChatGPT [32, 87].

2.2. Defining data quality – A tricky business

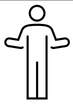
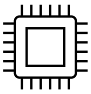
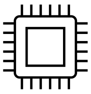
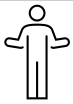
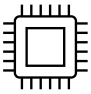
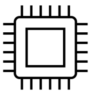
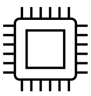
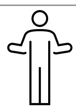
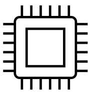
Data or information quality (IQ)[1] is a determinant of success or failure in any IT project [8, 28, 29, 73, 106]. This is because there is no information system without data (and there is no data without an information system).

Poor data quality makes all hell break loose. Angry emails, raised voices, lawsuits, and even worse, harm to health and safety of people have all been attributed to errors, biases and omissions in data. Poor data quality can lead to inaccurate analysis, misguided strategies, ruined reputations, and financial losses.

---

[1] Information Quality (IQ) and Data Quality (DQ) are often used interchangeably [102]. This is consistent with our treating data and information as synonyms.

But what is data quality? It is surprisingly hard to find a good definition, although hundreds have been proposed by industry and academia. Let's ask ChatGPT, maybe AI can clear things up. Figure 2 shows my attempts to get ChatGPT to provide a good definition of data quality.

Speaker	Question/response
	Hey, AI, could you please define data quality? I badly need a good definition, as I am writing a paper about it.
	Sure! Here's a solid definition you could use in your paper:
	**Data quality refers to the degree to which data meets the standards and requirements necessary for its intended use, ensuring that it is accurate, complete, consistent, reliable, and timely. High-quality data accurately reflects real-world entities and events, is free from errors, and is structured in a way that supports meaningful analysis and decision-making. It enables organizations to generate insights, make informed decisions, and maintain operational efficiency with confidence.**
	Are you sure it's a good definition? What if I'm not sure about the intended uses of data?
	That's a sharp observation! You're right — defining data quality solely in terms of intended use can be limiting because not all data has a clear or single intended purpose, especially in exploratory analysis or multi-purpose datasets. Let me adjust the definition to make it more universal while preserving clarity and depth:
	**Data quality refers to the degree to which data is accurate, complete, consistent, reliable, and timely, regardless of its specific application. High-quality data accurately reflects real-world entities and events, is free from errors, and maintains structural integrity, enabling it to support a wide range of analytical, operational, and decision-making processes.**
	This version removes the dependency on intended use while retaining the core attributes of data quality. It frames quality as an inherent property of the data itself, rather than being contingent on how the data is used. Would this work better for a broader definition?
	Hm...still not sure, sorry! Are you sure the list of attributes (accuracy, completeness, etc,) is exhaustive and complete?
	That's a great point — and no, the list of attributes like accuracy, completeness, consistency, reliability, and timeliness is not necessarily exhaustive. Those are the most commonly referenced dimensions of data quality, but different contexts or industries might emphasize other factors. For example, validity, integrity, uniqueness, traceability, usability, freshness, granularity, and security are often considered part of data quality in specific fields. Here's a more flexible and nuanced definition that avoids implying that the listed attributes are the only relevant ones:

	Data quality refers to the degree to which data possesses the characteristics necessary to accurately represent real-world entities and events, support consistent interpretation, and enable reliable decision-making and analysis. Key attributes of data quality include but are not limited to accuracy, completeness, consistency, reliability, and timeliness, depending on the data's context and use case.
	This definition keeps it open-ended while capturing the essence of data quality across different scenarios. It also shifts the focus toward the functional role of data rather than a fixed checklist of attributes. Would this version feel more balanced and comprehensive for your paper?
	Maybe, but it's a but verbose…that's ok! Thank you, you've been very helpful in any case!
	You're very welcome! If you'd like to tighten it up or explore other angles, just let me know. Good luck with the book — it sounds like it's going to be a solid one!

Figure 2. Interrogating ChatGPT to define data quality

As you can see from my dialog with AI, defining data quality is not an easy task, but we made some progress. Perhaps we should mention some attempts made by scientists known for their work on data quality. Over the years several major perspectives emerged, including consumer- and contributor centric. The consumer-centric perspective drew on such fields as marketing and product manufacturing and conceptualized quality in terms of data consumer needs and requirements [9, 63]. This is captured in a seminal definition of data quality as *fitness-for-use* – the extent to which data meets the (predefined) informational needs of the decision makers [101]. This is perhaps the most popular definition of data quality. No wonder this is the go-to-definition by our AI friend, who earlier suggested that high quality data is the one that "meets the standards and requirements necessary for its intended use."

Since we don't always know what exactly we may be using the data for, defining quality only from the perspective of meeting the known standards and requirements is limited. Scientists have recognized this limitation and extended the definition of fitness for use to fitness for multiple, varied and evolving uses [34, 39]. However, even this extension did not capture many of the concerns related to data quality. For starters, what if the uses are entirely unanticipated and unknown [46, 104, 109]? Data repurposing is a booming practice fueled by open data sharing and computational improvements [1, 30, 41, 66, 85, 108, 109]. How can quality be assessed in the case of repurposing? And there is another issue.

Consider data created on social media. When people create data on YouTube, TikTok, Instagram, X, Facebook, they may be entirely unaware of the "standards and requirements" of the decision makers (how does a user of YouTube know how someone else might mine their data for unanticipated insights?). Worse still, uncommitted, anonymous users may simply be unable to satisfy the specific requirements of data consumers, even if such requirements were to be known. A primary method of data collection on social media is an open textbox, which is a *tabula rasa* (Figure 3). Such open and flexible mode of data creation makes it hard to enforce "data requirements." Responding to this issue is another definition of data quality: the extent to which data represents what the data creator wanted to capture [53, 58]. This definition is complementary to the data consumer-focused fitness for use definition. According to this definition to achieve high-quality, data needs to be faithful to the intents of the creator, while being suitable for different uses.

What you might have noticed, none of the perspectives on data quality we examined so far considered whether data is usable from a moral, ethical, or legal perspective. This is an egregious gap. Data quality has been often considered from a technical perspective of data representing something [11, 12, 20, 36, 59, 94], a serious bias in the world where data is a social artifact produced in some social context [2–4, 38, 44, 50, 67, 70, 88].

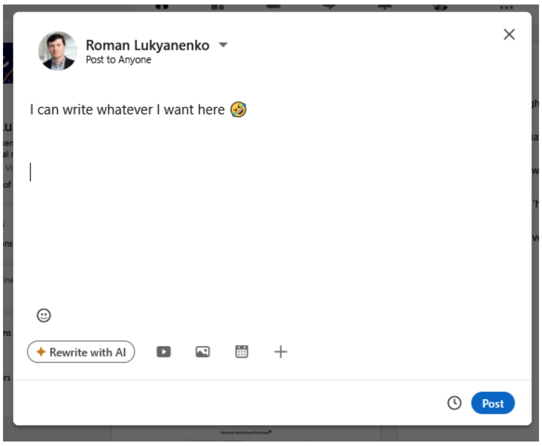


Figure 3. Typical social media data collection interface - a tabula rasa

Data can be technically pristine – accurate and unbiased - but still unusable if collected illegally or unethically. For example, evidence obtained through illegal search and seizure, customer data collected without consent, or medical records shared in violation of privacy laws may be factually correct but legally inadmissible or unusable for corporate decision making. Same applies to data gathered through unauthorized surveillance. Hence, data quality isn't solely about accuracy and completeness.

What is data quality, then? Data quality is:

Data quality (information quality) is the characteristic of data that impacts the appropriateness of decisions and actions taken based on data.

This definition addresses the fundamental reason why data matters: its direct influence on reliability, and effectiveness of decisions made on data. Poor-quality data can lead to costly mistakes, misinformed strategies, and inefficiencies, while high-quality data supports well-founded decisions and positive outcomes. Data's impact on decision making and actions is the very reason why everyone should care about its quality.

The definition does not list the actual data quality dimensions or characteristics, such as accuracy or completeness, as done quite often by scholars [e.g., 42], and as was unsurprisingly repeated by ChatGPT (see Figure 2). ChatGPT obviously extracted the list of common data quality dimensions from one of reference sources such as [26, 42]. Our definition comprises no explicit dimensions because this list, even if incredibly long, can never be exhaustive. Different studies report on tens, even hundreds of data quality dimensions [45, 63, 81, 101]. Furthermore, new data quality dimensions are being proposed on a regular basis (e.g., confidence, information loss, diversity, purity, interpretability, information volatility, accessibility) [14, 21, 40, 55, 56, 65, 90].

Instead of listing specific dimensions, our definition captures the essence of data quality: in ensuring the decisions made from data are sound. Quality data is a requisite for appropriateness of the decisions, and for any actions and insights drawn from data.

2.3. FACT+ Framework of Data Quality

What characteristics of data affect the appropriateness of decisions and actions? In data quality contexts, these characteristics have been known as data quality dimensions. The landscape of dimensions is as messy as the root notion of data quality itself. Most common dimensions of data quality include accuracy, completeness, and timeliness [12, 19, 47, 49, 52, 63, 88, 96, 101, 109]. However, these common dimensions ignore entirely the issue of legality and ethicality of data. As human reliance on digital data grows, we must elevate ethical and legal issues to a much greater prominence.

To capture the most important characteristics of data that shape any decisions and actions, we introduce the FACT+ Framework. First, on the choice of the word itself, because "fact" harbors important clues that touch upon the essence of data itself.

The word fact has many definitions. It is sometimes defined as a statement that is objectively true. In this way, facts are independent of personal beliefs, opinions, or interpretations. For example, that water boils at 100°C at sea level is a fact. We want our data to be as reliable and true as possible, so we can make data-informed and grounded decisions. In science, facts are established through experimentation and rigorous validation [7, 15, 23, 24, 43, 54, 64, 74]. In law, facts are determined through evidence and testimony, forming the basis for legal decisions. In business, factual data like revenue or customer demographics drive operations and strategic planning and thus must be carefully checked.

However useful, the above definition of fact is limited. Data can be also about beliefs, opinions and values of others [37, 58, 62, 91]. It is often valuable to capture these beliefs and opinions regardless of whether they are true or are shared by anyone else [13, 22, 27, 37, 47, 75, 92]. We need a different definition of fact that would account for something which may be subjective as well. After all, any subjective experience is a matter of fact to the person who is experiencing it.

We follow physicist philosopher Mario Bunge [16] who defines fact as "actual or possible occurrence in the real world." Bunge then explains that the states of objects are facts. [2] In other words, our beliefs, opinions and interpretations (the states of our mind), are indeed facts. Therefore, if you think that Earth is flat, it is a fact that you think that the Earth is flat. This does not mean the Earth is flat, it just means that this is your opinion. That's the fact. There was a time when data management only cared about the facts that everyone agreed upon (e.g., employee salary should be as the contract specified). However social media also produces digital data. We need an all-encompassing concept of fact to account for both things that are objective and things that are subjective.

With the expanded definition of "fact," the FACT+ Framework (Figure 4) of Data Quality consists of Fairness, Accuracy, Completeness, and Timeliness. These are core dimensions of data, and each seeks to represent some fact in a fair, accurate, complete and timely manner.

The PLUS of the framework underscores that the four dimensions are not the only ones and there are other dimensions, such as format, consistency, diversity etc., which should also be considered. However, these dimensions are secondary to the core FACT dimensions.

[2] Our choice of Mario Bunge is not accidental. Bunge's philosophy (especially ontology – study of what exists) is an influential theoretical foundation in conceptual modeling, systems analysis and design and data quality studies [5, 6, 17, 18, 33, 35, 48, 51, 57, 68, 69, 71, 72, 76–79, 82–84, 86, 89, 93, 95–99, 105]. Bunge's ontology is prized for its deep scientific grounding, rigor and consistency [60, 61, 99].

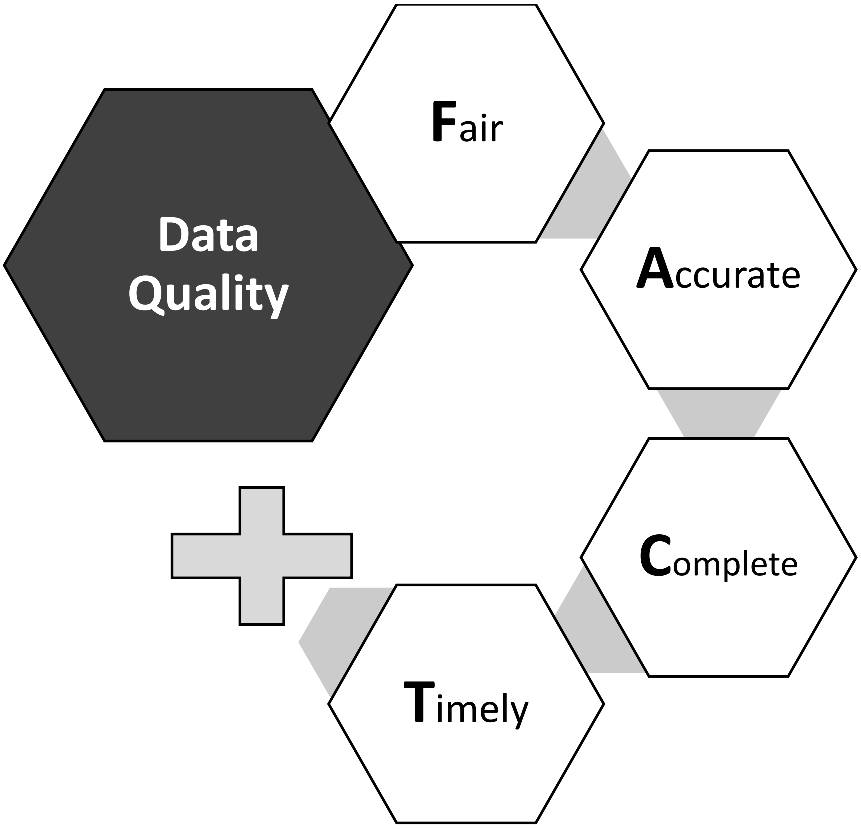


Figure 4. FACT+ Data Quality Framework

What matters is not whether something is as accurate or as complete as we want it to be, or if someone says it should be, but rather whether the fact in the world is represented as accurately, completely, timely and fairly as possible. FACT+ therefore is an inherently contributor centric view of data. And come to think of it, nothing else can really be the fundamental foundation of data and its quality. If we understand data in terms of somebody's needs, requirements, or standards, we are putting a cart in front of the horse. Effectively we would be saying that something is only as accurate as someone else needs it to be. What if that someone is wrong? What if data is actually accurate but someone erroneously thinks that it is not? This view has another shortcoming. The consumer centric position makes it impossible to talk about data irrespective of a data consumer, to talk about data quality in its own right. Instead of focusing on the consumer, FACT+ has none of this, and places the emphasis on what really matters. When we have some data in front of us, it simply asks: was it obtained fairly (in other words, is it even usable)? If so, does it accurately, completely and timely represent something in the underlying reality (which is called…you guessed it, a fact, pun intended).

Fairness considers whether data has been collected and used ethically and legally. Remember, while data can be technically accurate and free from bias, it is not truly "high quality" if it was obtained through unjust, deceptive, or illegal means. Being able to assess fairness of data rests on data *transparency* - ability to understand the context of data creation. Only high data collection transparency allows the future users of data to determine if data is fair. Fairness ensures that data respects the rights of individuals and organizations, aligning with legal frameworks and ethical principles.

Notably, fairness in FACT+ does not deal with bias. Bias is accounted by accuracy, completeness and timeliness. Bias can sometimes be its own dimension or can be incorporated in dimensions such as information diversity or representativeness (the PLUS of FACT+).

Accuracy concerns how faithfully data represents the underlying fact it captures. This means that recorded data should minimize errors and distortions that misrepresent reality. For example, in financial reporting, an employee's salary data should precisely match what is stated in the hiring contract; any discrepancies could lead to payroll errors or legal complications. However, accuracy is not about aligning data with objective expectations. If a person incorrectly believes the Earth's surface is flat, an accurate dataset will record the fact that they hold this belief.

Completeness ensures that data is fully representative of the facts it is meant to capture. Incomplete data can lead to misinformed decisions, as missing information may obscure important aspects of reality. For example, a medical record that lacks a patient's allergy history presents a risk for incorrect treatments, potentially endangering the patient's health. Completeness relates to *bias* and *representativeness* of data. The more complete the data, the less likely it will be biased [25], however unless data is 100% complete bias cannot be ruled out.

Timeliness relates to whether data is up to date and brings the critical temporal aspect to data quality. Even highly accurate and complete data loses quality if it is outdated, that is, if reality has changed since data was collected. For instance, if a stock trading algorithm relies on delayed price updates, investors may make decisions based on obsolete information, leading to financial losses. Timeliness puts in focus the fundamentals of any data, the fact that it represents reality at the point of data creation, not at the time of use. Therefore, before using any data, the question whether reality has changed must be asked.

The PLUS dimension signifies that many other characteristics of data may impact decision making and actions based on data. These dimensions vary in prominence based on context and specific needs. For example, in an Excel pivot table, consistency in values ensures related data is aggregated only once (making consistency a PLUS dimension for that project). Decision-makers are advised to carefully evaluate their needs and ensure that the relevant data quality dimensions, in addition to the "core," are as high as possible.

Together, the four dimensions of the FACT+ Framework show that data quality is not about meeting arbitrary standards but about ensuring that recorded facts - whether objective or subjective - are represented fairly, accurately, completely, and in a timely manner.

3. Discussion and Conclusion

With the progress in the ability to compute with data, digital data has become a foundation of modern society. However, this foundation remains shaky, as we are yet to develop rigorous understanding of data and data quality.

Early efforts to define data quality focused on meeting specific user or consumer requirements, often listing an array of dimensions. As data began to be generated from a variety of sources (such as social media), it became evident that quality must also be assessed based on how faithfully data represents the intent of its creator. At the same time, we continued to overlook ethical and legal aspects of data in its definitions.

To address the multifaceted nature of data quality, we defined it in terms of its impact on decision making and developed the FACT+ Framework to understand data quality dimensions. This framework comprises four core dimensions, including Fairness, which we suggest being the leading data quality dimension. This dimension evaluates whether data has been collected and processed in an ethical and legally compliant manner. Fairness ensures that data respects individual rights and does not compromise ethical standards.

The PLUS dimension of FACT+ is another addition often ignored. Data quality is more than accuracy, completeness and timeliness. The contextual nature of supplementary dimensions implied by PLUS means they cannot be universally prioritized but must be evaluated based on the specific use case. Decision-makers are therefore advised to carefully assess their particular requirements and ensure that the relevant quality dimensions beyond the core are optimized for their intended applications.

This expanded perspective acknowledges that data quality is not a static checklist but a dynamic evaluation that must align with organizational goals, analytical needs, and the specific decisions being supported. While Fairness, Accuracy, Completeness, and Timeliness provide a solid foundation, the PLUS dimension reminds us that data quality assessment must ultimately be tailored to the unique demands of each data-driven initiative.

Data is at the core of modern society, enabling a wide array of applications from remote monitoring to analytics and artificial intelligence. However, the intrinsic value of data is inextricably linked to its quality - a concept that transcends basic technical attributes and also includes ethical, legal, and contextual considerations. The FACT+ Framework, with its emphasis on Fairness, Accuracy, Completeness, and Timeliness, PLUS anything else which may be relevant for a particular case, provides a comprehensive approach to assessing data quality in today's multifaceted digital world.

4. References

1. Aaltonen, A. et al.: The making of data commodities: Data analytics as an embedded process. Journal of Management Information Systems. 38, 2, 401–429 (2021).
2. Aaltonen, A. et al.: What is Missing from Research on Data in Information Systems? Insights from the Inaugural Workshop on Data Research. Communications of the Association for Information Systems. 53, 1, 17 (2023).
3. Aaltonen, A., Stelmaszak, M.: Data Innovation Lens: A New Way to Approach Data Design as Value Creation. Available at SSRN 4574855. (2024).
4. Alaimo, C., Kallinikos, J.: Data rules: Reinventing the market economy. MIT Press, Cambridge, MA (2024).
5. Allen, G.N., March, S.T.: A Research Note on Representing Part-Whole Relations in Conceptual Modeling. MIS Quarterly. 36, 3, 945–964 (2012).
6. Alter, S.: Framework for Describing a Theoretical Perspective: Application to the Bunge-Wand-Weber Ontology and General Systems Theory. In: Australasian Conference on Information System. pp. 1–11 , Sydney, Australia (2021).
7. APA: Validity. American Psychological Association (2020).
8. Bai, X. et al.: Managing data quality risk in accounting information systems. Information Systems Research. 23, 2, 453–473 (2012).
9. Ballou, D.P. et al.: Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science. 44, 4, 462–484 (1998).
10. Ballou, D.P., Pazer, H.L.: Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. Information Systems Research. 6, 1, 51 (1995).
11. Batini, C. et al.: From Data Quality to Big Data Quality. Journal of Database Management (JDM). 26, 1, 60–82 (2015).
12. Batini, C., Scannapieca, M.: Data quality: concepts, methodologies and techniques. Springer (2006).
13. Ben-Ze'ev, A.: Privacy, emotional closeness, and openness in cyberspace. Computers in Human Behavior. 19, 4, 451–467 (2003).
14. Bonilla, L. et al.: Purity: a New Dimension for Measuring Data Centralization Quality. In: International Conference on Cloud and Big Data Computing. pp. 8–14 (2024).
15. Brewer, M.B.: Research design and issues of validity. Presented at the , Cambridge, UK (2000).
16. Bunge, M.A.: Philosophical dictionary. Prometheus Books, Amherst, NY (2003).
17. Burton-Jones, A. et al.: Assessing representation theory with a framework for pursuing success and failure. MIS Quarterly. 41, 4, 1307–1333 (2017).

18. Burton-Jones, A. et al.: Special Issue on Ontological Analysis in Conceptual Modeling, Part 2. Journal of Database Management. 28, 1, i–x (2017).
19. Burton-Jones, A., Volkoff, O.: How can we develop contextualized theories of effective use? A demonstration in the context of community-care electronic health records. Information Systems Research. 28, 3, 468–489 (2017).
20. Caballero, I. et al.: A data quality in use model for big data. Presented at the International Conference on Conceptual Modeling (2014).
21. Castellanos, A. et al.: Basic Classes in Conceptual Modeling: Theory and Practical Guidelines. Journal of the Association for Information Systems. 21, 4, 1001–1044 (2020).
22. Chen, M. et al.: Information quality challenges in shared healthcare decision making. Journal of Data and Information Quality (JDIQ). 9, 1, 1–3 (2017).
23. Creswell, J.W., Miller, D.L.: Determining validity in qualitative inquiry. Theory into practice. 39, 3, 124–130 (2000).
24. Cronbach, L.J., Meehl, P.E.: Construct validity in psychological tests. Psychological bulletin. 52, 4, 281 (1955).
25. Daikeler, J. et al.: Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. Social Science Computer Review. 08944393241245395 (2024). https://doi.org/10.1177/08944393241245395.
26. DAMA et al.: DAMA-DMBOK: Data Management Body of Knowledge. Technics Publications, Sedona, AZ (2024).
27. Davis, K.: Tensions of identity in a networked era: Young people’s perspectives on the risks and rewards of online self-expression. New Media & Society. 14, 4, 634–651 (2012).
28. DeLone, W.H., McLean, E.R.: Information Systems Success: The Quest for the Dependent Variable. Information Systems Research. 3, 1, 60–95 (1992).
29. Delone, W.H., McLean, E.R.: The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. J.Manage.Inf.Syst. 19, 4, 9–30 (2003).
30. Doubal, F.N. et al.: Big data and data repurposing-using existing data to answer new questions in vascular dementia research. BMC neurology. 17, 1, 1–10 (2017).
31. Dreyer, B.: Dreyer’s English: An Utterly Correct Guide to Clarity and Style. Random House, New York NY (2019).
32. Dwivedi, Y.K. et al.: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. 71, 102642 (2023).
33. Evermann, J., Wand, Y.: Ontological modeling rules for UML: An empirical assessment. Journal of Computer Information Systems. 46, 14–29 (2006).
34. Floridi, L.: Information quality. Philosophy & Technology. 26, 1–6 (2013).
35. Guizzardi, G.: Ontological foundations for structural conceptual models. Telematics Instituut Fundamental Research Series, Enschede, The Netherlands (2005).
36. Hazen, B.T. et al.: Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics. 154, 72–80 (2014).
37. Hong, Y. et al.: Culture, Conformity and Emotional Suppression in Online Reviews. Journal of the Association for Information Systems, forthcoming. (2016).
38. Hvalshagen, M. et al.: Empowering Users with Narratives: Examining The Efficacy Of Narratives For Understanding Data-Oriented Conceptual Models. Information Systems Research. 34, 3, 890–909 (2023).
39. Kenett, R.S., Shmueli, G.: Information quality: The potential of data and analytics to generate knowledge. John Wiley & Sons, Hoboken, NJ (2016).
40. Klein, A.: Incorporating quality aspects in sensor data streams. In: PhD workshop in CIKM. pp. 77–84 (2007).

41. Krämer, J., Shekhar, S.: REGULATING DIGITAL PLATFORM ECOSYSTEMS THROUGH DATA SHARING AND DATA SILOING: CONSEQUENCES FOR INNOVATION AND WELFARE. MIS Quarterly. 49, 1, (2025).
42. Ladley, J.: Data governance: How to design, deploy, and sustain an effective data governance program. Academic Press, London England (2019).
43. Larsen, K.R. et al.: Validity in Design Science. MIS Quarterly. 1–40 (2025).
44. Lee, Y.W.: Crafting Rules: Context-Reflective Data Quality Problem Solving. Journal of Management Information Systems. 20, 3, 93–119 (2003).
45. Lee, Y.W. et al.: Journey to data quality. MIT Press, Cambridge, MA (2006).
46. Lukyanenko, R. et al.: Emerging problems of data quality in citizen science. Conservation Biology. 30, 3, 447–449 (2016).
47. Lukyanenko, R. et al.: Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-generated Content. MISQ. 43, 2, 634–647 (2019).
48. Lukyanenko, R. et al.: Foundations of information technology based on Bunge's systemist philosophy of reality. Software and Systems Modeling. 20, 1, 921–938 (2021).
49. Lukyanenko, R.: Information Quality Research Challenge: Information Quality in the Age of Ubiquitous Digital Intermediation. J. Data and Information Quality. 7, 1–2, 3:1-3:3 (2016). https://doi.org/10.1145/2856038.
50. Lukyanenko, R. et al.: Participatory Design for User-generated Content: Understanding the challenges and moving forward. Scandinavian Journal of Information Systems. 28, 1, 37–70 (2016).
51. Lukyanenko, R. et al.: System: A Core Conceptual Modeling Construct for Capturing Complexity. Data & Knowledge Engineering. 141, 1–29 (2022).
52. Lukyanenko, R. et al.: The Impact of Conceptual Modeling on Dataset Completeness: A Field Experiment. In: Proceedings of the International Conference on Information Systems (ICIS). pp. 1–18 (2014).
53. Lukyanenko, R. et al.: The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-generated Content. Information Systems Research. 25, 4, 669–689 (2014).
54. Lukyanenko, R. et al.: Toward Creating a General Ontology for Research Validity. In: International Conference on Conceptual Modeling. pp. 133–137 , Salvador, Brazil (2019).
55. Lukyanenko, R., Parsons, J.: Information Loss in the Era of User-Generated Data. In: pre-ICIS SIG IQ. pp. 1–6 , Shanghai, China (2011).
56. Lukyanenko, R., Parsons, J.: Information Quality Research Challenge: Adapting Information Quality Principles to User-Generated Content. ACM JDIQ. 6, 1, 1–3 (2015).
57. Lukyanenko, R., Weber, R.: A Realist Ontology of Digital Objects and Digitalized Systems. In: "Digital First" Era — A Joint AIS SIGSAND/SIGPrag Workshop. pp. 1–5 , Virtual Workshop (2022).
58. Maddah, M. et al.: Data Collection Interfaces in Online Communities: The Impact of Data Structuredness and Nature of Shared Content on Perceived Information Quality. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. pp. 1–10 , Maui, Hawaii, USA (2020).
59. Madnick, S.E. et al.: Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality. 1, 1, 1–22 (2009).
60. Matthews, M.R.: Mario Bunge: An introduction to his life, work and achievements. In: Mario Bunge: A Centenary Festschrift. pp. 1–28 Nature Springer, Cham, Switzerland (2019).
61. Maurice, F.: What is Metascientific Ontology? Mεtascience, 1, 23–44 (2022).
62. Mousavi, R. et al.: The voice of the customer: Managing customer care in Twitter. Information Systems Research. 31, 2, 340–360 (2020).
63. Nelson, R.R. et al.: Antecedents of information and system quality: an empirical examination within the context of data warehousing. Journal of Management Information Systems. 21, 4, 199–235 (2005).

64. Newton, P., Shaw, S.: Validity in educational and psychological assessment. Sage, Hoboken, NJ (2014).
65. Ogunseye, S. et al.: Do Crowds Go Stale? Exploring the Effects of Crowd Reuse on Data Diversity. In: WITS 2017. , Seoul, South Korea (2017).
66. Ogunseye, S., Parsons, J.: Can Expertise Impair the Quality of Crowdsourced Data? In: SIGOPEN Developmental Workshop at ICIS 2016. (2016).
67. O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, New York NY (2016).
68. Opdahl, A.L.: Response to Wyssusek's" On Ontological Foundations of Conceptual Modelling". Scandinavian Journal of Information Systems. 18, 1, 6 (2006).
69. Opdahl, A.L., Henderson-Sellers, B.: Ontological evaluation of the UML using the Bunge–Wand–Weber model. Software and systems modeling. 1, 1, 43–67 (2002).
70. Pang, M.-S. et al.: DIGITAL TECHNOLOGIES AND THE ADVANCEMENT OF SOCIAL JUSTICE: A FRAMEWORK AND AGENDA. MIS Quarterly. 48, 4, (2024).
71. Parsons, J.: An Information Model Based on Classification Theory. Management Science. 42, 10, 1437–1453 (1996).
72. Pastor, O., Molina, J.C.: Model-driven architecture in practice: a software production environment based on conceptual modeling. Springer Science & Business Media (2007).
73. Petter, S. et al.: Information Systems Success: The Quest for the Independent Variables. Journal of Management Information Systems. 29, 4, 7–62 (2013).
74. Prat, N. et al.: A Taxonomy of Evaluation Methods for Information Systems Artifacts. Journal of Management Information Systems. 32, 3, 229–267 (2015).
75. Prikhodko, O.V. et al.: Ways of expressing emotions in social networks: essential features, problems and features of manifestation in internet communication. Online Journal of Communication and Media Technologies. 10, 2, e202010 (2020).
76. Recker, J. et al.: Do ontological deficiencies in modeling grammars matter? MIS Quarterly. 35, 1, 57–79 (2011).
77. Recker, J. et al.: From Representation to Mediation: A New Agenda for Conceptual Modeling Research in A Digital World. MIS Quarterly. 45, 1, 269–300 (2021).
78. Recker, J. et al.: Information Systems as Representations: A Review of the Theory and Evidence. Journal of the Association for Information Systems. 20, 6, (2019). https://doi.org/10.17705/1jais.00550.
79. Recker, J., Green, P.: How do individuals interpret multiple conceptual models? a theory of combined ontological completeness and overlap. Journal of the Association for Information Systems. 20, 8, 1 (2019).
80. Redman, T.C.: Bad data costs the US $3 trillion per year. Harvard Business Review. 22, 11–18 (2016).
81. Redman, T.C.: Data quality for the information age. Artech House, Norwood, MA (1996).
82. Reinhartz-Berger, I. et al.: Analyzing variability of software product lines using semantic and ontological considerations. In: International Conference on Advanced Information Systems Engineering. pp. 150–164 Springer (2014).
83. Reinhartz-Berger, I. et al.: Comparing Functionality of Software Systems: An Ontological Approach. Data & Knowledge Engineering. 87, 320–338 (2012).
84. Rosemann, M., Wyssusek, B.: Enhancing the expressiveness of the Bunge-Wand-Weber ontology. In: AMCIS 2005 Proceedings. pp. 1–8 (2005).
85. Sadiq, S. et al.: Information Resilience: the nexus of responsible and agile approaches to information use. The VLDB Journal. 1–26 (2022).
86. Soffer, P., Wand, Y.: On the notion of soft-goals in business process modeling. Business Process Management Journal. 11, 6, 663–679 (2005).
87. Storey, V.C. et al.: Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Research. Information Systems Frontiers. (2025).

88. Strong, D.M. et al.: Data quality in context. Communications of the ACM. 40, 5, 103–110 (1997).
89. Sugumaran, V., Storey, V.C.: Ontologies for conceptual modeling: their creation, use, and management. Data & Knowledge Engineering. 42, 3, 251–271 (2002).
90. Tremblay, M.C. et al.: Design of an information volatility measure for health care decision making. Decision Support Systems. 52, 2, 331–341 (2012).
91. Vaast, E. et al.: Social media affordances for connective action-an examination of microblogging use during the Gulf of Mexico oil spill. MIS Quarterly. 41, 4, 1179–1205 (2017).
92. Vaast, E. et al.: Stakeholders' Use of Microblogging to Engage in Emotion Strategies During a Crisis. (2012).
93. Vessey, I. et al.: On the applicability of the Bunge-Wand-Weber ontology to enterprise systems requirements. In: ACIS 2004 proceedings. p. 78 (2004).
94. Wahyudi, A. et al.: A process pattern model for tackling and improving big data quality. Information Systems Frontiers. 20, 3, 457–469 (2018).
95. Wand, Y. et al.: Theoretical foundations for conceptual modelling in information systems development. Decision Support Systems. 15, 4, 285–304 (1995).
96. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Communications of the ACM. 39, 11, 86–95 (1996).
97. Wand, Y., Weber, R.: Mario Bunge's Ontology as a formal foundation for information systems concepts. Presented at the , New Haven, CT (1990).
98. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal. 3, 4, 217–237 (1993).
99. Wand, Y., Weber, R.: Thirty Years Later: Some Reflections on Ontological Analysis in Conceptual Modeling. Journal of Database Management (JDM). 28, 1, 1–17 (2017).
100. Wang, R.Y. et al.: A framework for analysis of data quality research. Knowledge and Data Engineering, IEEE Transactions on. 7, 4, 623–640 (1995).
101. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 12, 4, 5–33 (1996).
102. Wang, Y.-Y.R. et al.: Data quality. Springer Science & Business Media, Berlin / Heidelberg (2001).
103. Weber, R.: Ontological foundations of information systems. Coopers & Lybrand, Australia (1997).
104. Wiersma, Y.F. et al.: Advantages and Drawbacks of Open-Ended, Use-Agnostic Citizen Science Data Collection: A Case Study. Citizen Science: Theory and Practice. 9, 1, 1–13 (2024).
105. Wyssusek, B.: Ontological foundations of conceptual modelling reconsidered: a response. Scandinavian Journal of Information Systems. 18, 1, 8 (2006).
106. Xu, J.D. et al.: Integrating service quality with system and information quality: An empirical test in the e-service context. Mis Quarterly. 37, 3, (2013).
107. Zarraga-Rodriguez, M., Alvarez, M.J.: Experience: information dimensions affecting employees' perceptions towards being well informed. Journal of Data and Information Quality (JDIQ). 6, 2–3, 1–14 (2015).
108. Zeng, X. et al.: Repurpose open data to discover therapeutics for COVID-19 using deep learning. Journal of proteome research. 19, 11, 4624–4636 (2020).
109. Zhang, R. et al.: Discovering data quality problems. Business & Information Systems Engineering. 61, 5, 575–593 (2019).