Reflections on quality requirements for digital trace data in IS research

Gregory Vial

Submission title: Reflections on Quality Requirements for Digital Trace Data in IS Research

Corresponding author:
Gregory Vial (gregory.vial@hec.ca)
HEC Montréal
3000 Chemin-de-la-Côte-Ste-Catherine
Montréal, CANADA
H3T 2A7
Phone: 1.514.340.1467
Fax: 1.514.340.6411

Gregory Vial is Assistant Professor of Information Technology (IT) at HEC Montreal. His research interests are in the areas of outsourcing, systems development practices and methodologies, and databases. His work has been published in the *European Journal of Information Systems*, *IEEE Software* as well as in several IS conferences.

Reflections on Quality Requirements for Digital Trace Data in IS Research

Abstract

In recent years an increasing number of academic disciplines, including IS, have sourced digital trace data for their research. Notwithstanding the potential of such data in (re)investigations of various phenomena of interest that would otherwise be difficult or impossible to study using other sources of data, we view the quality of digital trace data as an underappreciated issue in IS research. To initiate a discussion of how to evaluate and report on the quality of digital trace data in IS research, we couch our arguments within the broader tradition of research on data quality. We explain how the uncontrolled nature of digital trace data creates unique challenges for IS researchers, who need to collect, store, retrieve, and transform those data for the purpose of numerical analysis. We then draw parallels with concepts and patterns commonly used in data analysis projects and argue that, although IS researchers probably apply such concepts and patterns, this is not reported in publications, undermining the reader's ability to assess the reliability, statistical power and replicability of the findings. Using the case of GitHub to illustrate such challenges, we develop a preliminary set of guidelines to help researchers consider and report on the quality of the digital trace data they use in their research. Our work contributes to the debate on data quality and provides relevant recommendations for scholars and IS journals at a time when a growing number of publications are relying on digital trace data.

Keywords

Digital trace data; Data quality; GitHub.

1. Introduction

Data quality has been acknowledged as an enduring issue in research and in practice (Khatri & Brown, 2010; Strong, Lee, & Wang, 1997; Wand & Wang, 1996). Considering the rise of digital technologies, big data, analytics and artificial intelligence, all of which are large producers and consumers of data in various formats, discussions about the veracity of data – a notion related to data quality – have become common in teaching, practice as well as in research in computer science and software engineering (Demchenko, Grosso, De Laat, & Membrey, 2013). Notwithstanding, we concur with Marsden and Pingry's (2018) arguments and view data quality *in IS research* as an issue that deserves much-needed attention for the advancement of scientific knowledge.

In this short paper, we build on Marsden and Pingry's taxonomy of numerical data types and apply it to digital trace data (DTD) (Berente, Seidel, & Safadi, 2018; Howison, Wiggins, & Crowston, 2011) — which we define as digital records of activities and events that are produced, stored and retrieved using information technologies — as a variation of the "third party" data type discussed by the authors. We explain why data quality issues are particularly relevant in the context of DTD, and we provide a current illustration of those issues before offering guidelines for authors working with DTD. Although our work is by no means an authoritative list of data quality requirements for DTD research, it encourages IS researchers and journals to value discussions of data quality as the subject of DTD continues to grow in popularity.

2. Boundary Assumption

As an initial step, we wish to establish an important contextual boundary. We assume that high-quality data are "fit for use by data consumers" (Wang & Strong, 1996, p. 6). In the

context of this paper, this means that high-quality data are fit for use in the numerical analyses[1] supporting the study of a research question. As we develop our arguments, we refer to the four main categories of data quality and their underlying dimensions, which we have summarized in Table 1. Our core argument is that DTD quality is an important topic because DTD often exhibit low quality within each category, rendering their use in science challenging.

Table 1. Data Quality Categories and Dimensions (adapted from Strong et al., 1997; Wang & Strong, 1996)

Data Quality Category	Description	Data Quality Dimensions
Intrinsic data quality	Fitness for use based on the inherent properties of the data	Accuracy, objectivity, believability, reputation
Accessibility data quality	Fitness for use based on the ability to retrieve the data	Accessibility, access security
Contextual data quality	Fitness for use based on the task at hand	Relevancy, value-added, timeliness, completeness, amount of data
Representational data quality	Fitness for use based on the format and the presentation of data for the task at hand	Interpretability, ease of understanding, concise representation, consistent representation

3. The Rise of Digital Trace Data in (IS) Research

DTD is broadly defined as "digital records of activity and events that involve information technologies." (Berente et al., 2018, p. 1) Although some authors argue that DTD are both "produced through and stored by an information system" (Howison et al., 2011, p. 770), for the purposes of this discussion we would add that DTD are also *retrieved* by researchers using information systems (the relevance of this point is discussed below). In recent years, researchers have begun to use DTD for a variety of phenomena and research questions that had been otherwise difficult, or sometimes even impossible, to study. This includes questions related to distributed coordination in complex knowledge work environments (Dabbish, Stuart, Tsay, & Herbsleb, 2012), participation in online communities (Faraj & Johnson, 2011), organizational routines (Pentland, Haerem, & Hillison, 2009), and the diffusion of

[1] We use the term "numerical analysis" as a general term referring to any form of data analysis technique, whether qualitative or quantitative, exploratory, explanatory, or predictive.

information across social media (Vosoughi, Roy, & Aral, 2018), among others. The growing availability of DTD has mirrored the rise of digital technologies in our lives, driving new research opportunities and even, as some have argued, theory development (Berente et al., 2018).

Our reading of articles that have used DTD has led us to concur with Marsden and Pingry's (2018) observation that IS articles often suffer from the fact that "the amount of space devoted to data collection, validation, and/or quality details pales in comparison to the space devoted to detailing and explaining why a relatively sophisticated model form and estimation technique(s) are employed." (p. A1) In our opinion, this does not imply that IS researchers do not consider DTD quality an important aspect of their work. Rather, the final result of their work (i.e. scientific publications) fails to incorporate information that could help us evaluate the quality of the data used in their analysis. This may be due to the fact that the IS community has not valued having such information in publications. Yet, as we discuss below, such information is in fact highly relevant.

4. Digital Trace Data as a Raw Material

4.1. Public and Private Digital Trace Data

At this point, it is important to distinguish between the two main types of DTD. *Private* DTD are essentially data that are provided to a researcher but that remain inaccessible to the general public (e.g., available for a fee). Private DTD can be sourced from application logs (Pentland et al., 2009), private source code repositories (Zimmermann & Nagappan, 2008), databases or documentation, among others. The key issues here are that another researcher would not be able to retrieve the same data to replicate the findings of a given study and that researchers may have few means to assess the quality of the data they have obtained (Marsden & Pingry, 2018).

Public DTD are accessible to the general public. The rise of digital platforms and ecosystems and the open data movement have helped make public DTD increasingly available (Dabbish et al., 2012; Tsay, Dabbish, & Herbsleb, 2014). Unlike private DTD, public DTD offer an unprecedented opportunity for researchers to build research that is replicable and transparent. Although our examples are based on public DTD, our arguments are equally relevant to private DTD.

4.2. Digital Trace Data as an Uncontrolled Source of Data

By emphasizing the use of increasingly complex methods and modeling techniques to make significant methodological contributions, we often forgo explanations of the quality of data in general, and that of DTD in particular. Yet, as the old adage says, "garbage in, garbage out." Without high-quality data, a model will have poor explanatory and/or predictive power, regardless of whether sophisticated imputation techniques are used. By definition, the quality of most DTD is *a priori* questionable because researchers have little to no control over how the data are generated. Compared to other forms of numerical data, this creates a significant challenge. For instance, there is a long-standing tradition reflected in articles offering guidelines (Churchill, 1979; Gerbing & Anderson, 1988) to help researchers design survey instruments to generate numerical data fit for their purpose. Instruments are designed with a purpose in mind, and the constructs are operationalized based on the requirements of numerical analysis. In contrast, DTD are taken as a given, and researchers must work with the data without any control over how they are generated. This raises issues about the intrinsic quality of such data as well as the alignment between the research question, methods, and data.

4.3. Extract, Transform, Load Pattern in Digital Trace Data Research

As researchers, we must often perform a number of steps to retrieve, transform and manipulate raw DTD, in order to generate numerical data (a curated data set) in a format that are fit for numerical analysis. For example, we may *scrape* reviews from an online source, extract sentiment features from the textual data contained in those reviews, and transform those features into numerical data points to investigate our research question. In many ways, the process followed in these steps mirrors the Extract, Transform, Load (ETL) pattern commonly used in data warehousing (Kimball & Ross, 2011) (see Table 2). The ETL pattern is for extracting data from sources that are not designed for analysis (e.g., data from operational systems, logs), transforming them into a structure and a format that are fit for analysis, and consolidating them into a single source of high-quality data that will be readily available to users. When we teach the ETL process to our analytics students, whose other courses focus primarily on applied statistics, they realize that the creation of curated data sets requires effort. ETL processes must not only be followed; they must also be documented, audited and tested to ensure that the quality of data loaded into a data warehouse/mart remains consistent over time. In theory, a properly designed ETL process is therefore both transparent and reproducible. Such efforts are paramount in order to draw meaningful deductions from numerical analysis. Referring to the data quality dimensions presented in Table 1, it is through these steps that data become *accessible* to the researcher, who can then *interpret* them and use them as *relevant* and *accurate* in the investigation of a research question.

Table 2. Applying the ETL Pattern to DTD Processing

ETL Pattern	DTD Processing
Extract the source data	Retrieve DTD from information system(s)
Transform the source data	Transform DTD into a format that is fit for analysis (e.g., convert variable types, assess quality/cleanse, and join across multiple sources)
Load the transformed data	Save the transformed data as a curated data set (e.g., in a CSV file) to be used as input in formal data analysis and modeling

5. Beyond a Vision of Data as "Given"

To understand the underlying issues in DTD quality, it is useful to consider the original meaning of the word "data". The etymology of the word data (the plural of *datum*[2]) means "given" in Latin. In IS, like in many other fields such as computer science and software engineering, the concept of data refers to something that is raw and largely devoid of meaning on its own before it has been used and processed. Simple examples of data are the height of Mount Everest, or the current temperature, as recorded by instruments. There is often an implicit assumption that data, as things that are given, are neutral and unbiased. This assumption is problematic because it creates a false sense of confidence in data as objective means for describing phenomena. Considering the above examples, faulty instruments may very well generate data, but such data will be inaccurate (Alkhalil & Ramadan, 2017). Whether they are generated by machines and calibrated instruments or by human actors (e.g., from a customer relationship management system), DTD should be treated as highly susceptible to data quality issues.

5.1. From Data to Capta

Seeking to address this assumption, some researchers have advocated use of the term *capta*—"[things] taken" in Latin—as a more accurate depiction of the objects we use in numerical analysis (Checkland, 1999; Drucker, 2011). More specifically, social constructivist approaches and scholars in the humanities acknowledge that the decisions we make to enable the collection, analysis and graphical representation of data are not neutral. Rather, "humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact" (Drucker, 2011). While this position has some

[2] While the term "data" is often conjugated as a singular noun, we treat it as a plural noun, in accordance with its etymology.

significant ontological and epistemological implications that may not be universally accepted, it calls upon us to question: (1) how things are *given* to us as researchers and (2) how we *take* (or retrieve) those things before using them in our research.

This is precisely the type of question that statisticians and data scientists are expected to ask before using raw data in their analyses (Cox, 2007). In data analysis projects, as in data warehousing, *data profiling* is an important activity carried out iteratively to assess the fitness for use of data. However important data profiling may be, IS researchers spend little time describing the process used to assess the quality of their DTD and the transformations they performed to render those DTD amenable to their research question (Clarke, 2016). This issue is becoming more relevant as an increasing amount of external data is gathered from sources that provide no guarantee as to their quality (e.g., there is no public data available on the number of robots or fake accounts on Twitter). Even in contexts where the widespread adoption of the "Internet of Things", sensors, and other machine-based data generation mechanisms has the potential to remove human intervention from the data generation process, this issue remains. For example, specific decisions made on the use and placement of sensors or RFID tags can influence the DTD generated by physical objects and the inferences and predictions we can make from these DTD.

6. A Concrete Illustration: The Case of GitHub Data

6.1. Collecting Digital Trace Data from GitHub

To illustrate DTD quality issues, consider the case of GitHub. GitHub is the largest host for public and private software code repositories and has become an important source of DTD in IS (e.g., Zhang, Yoo, Wattal, Zhang, & Kulathinal, 2014) and computer science (e.g., Dabbish et al., 2012). Each repository hosted on GitHub contains a source code repository (Git) as well as tools that are used to help manage and coordinate work among project

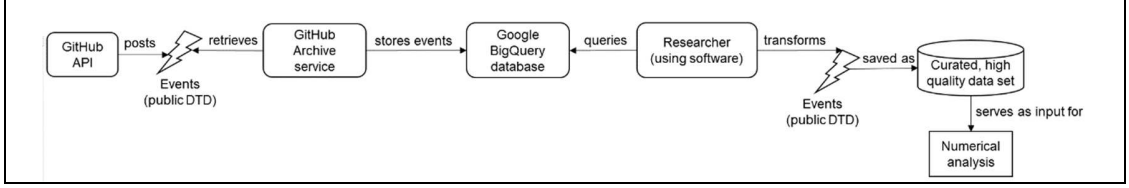
contributors and with the general public (e.g., announcements, bug reports, feature requests etc.). To access GitHub data, researchers typically have four options:

- They can manually retrieve and code the data from project web pages.
- They can scrape HTML pages and extract pertinent data using software libraries (e.g., BeautifulSoup in Python).
- They can use the GitHub Application Programming Interface (API) to retrieve the DTD required for the research project. This is achieved by using a pre-written piece of software or by writing custom software to call the GitHub API over the Internet.
- The final and most widely used option (e.g., Zhang et al., 2014) is to retrieve the data from a third party provider. For instance, GitHub data is freely available through the GitHub Archive project, hosted on Google BigQuery, a cloud database service.

6.2. GitHub Digital Trace Data Quality Issues

Each of these options present researchers with different data quality issues affecting all four data quality categories. The first option is time-consuming and likely to be error-prone due to the manual data entry process (intrinsic data quality). The second option can also yield intrinsic data quality issues if the HTML page structure is not uniform across repositories. Due to API quotas and historical data restrictions, the third option may not allow for completeness and accessibility, thereby hindering contextual data quality as well as other dimensions of intrinsic data quality. Although the fourth option appears to be the most viable, it also has drawbacks. To explain these issues, we have summarized the DTD extraction and transformation process in Figure 1, where every arrow has a potential impact on data quality.

Figure 1. Using DTD from GitHub



In theory, every event generated by a user or a machine interacting with GitHub generates an API event that is intercepted by the GitHub Archive service. The data contained within the event are saved at a temporary destination. At frequent intervals, the data contained in the temporary destination are archived in a Google BigQuery database. In practice, data quality issues can arise in the processing and saving of data at every step of this process, whether in the production and storage of DTD by third parties or in the retrieval of DTD by the researcher, even *before* the DTD are transformed and manipulated for numerical analysis. Turning to the GitHub Archive project's repository (which is hosted on GitHub), there are indeed a number of reports of DTD that are either not saved properly or not saved at all, often with no explanation provided as to why problems occurred. In other instances, changes to GitHub's API went unnoticed by the GitHub Archive's team, who later modified their code to accommodate these changes. Beyond these issues, the researcher is still responsible for ensuring that the data are retrieved, transformed and manipulated in an appropriate manner and that the processes used for this purpose (or the software packages used) are free of bugs. In every one of these steps, we should have confidence that appropriate decisions were made (e.g., by discarding certain classes of events, or certain types of projects) based on the research objectives rather than on convenience, so that each dimension of the data quality categories can be addressed.

7. Addressing Quality Issues in Digital Trace Data

Considering that sets of DTD may be rather large, it could be argued that some level of data quality issues (e.g., accuracy) may not impact the final results of a numerical analysis. Although this argument may hold true from a statistical standpoint, our main concern is not that some of the data may be inaccurate or incomplete. Rather, it is the fact that we have no

clear idea as to how much of the data is of low quality, and which dimension(s) of data quality are affected. It is only when these aspects are assessed and explained *within the context of the research question at hand* that we can make an informed decision about the data's suitability for the analysis and the potential impacts of low data quality on the power of the numerical analyses.

7.1. Applying the Seven W's

This raises the issue of whether we are able to propose practical criteria for assessing the quality of the numerical data extracted or derived from DTD. Although desirable, the high degree of variability in DTD across different sources means that we may not be able to establish clear-cut, universal criteria. Rather, we argue that researchers should make a number of considerations explicit in their work, to show that they were mindful of data quality issues and took appropriate steps to address them. Marsden and Pingry (2018, p. A2) argue that a scientific article should strive to answer the Seven W's (what, when, where, how, who, which, and why) about data quality. In our view, these seven questions provide a solid foundation for making statements about the quality of DTD more explicit in IS research. Table 3 summarizes these seven questions and their application to DTD.

Table 3. Asking the Seven W's for DTD

Consideration (from Marsden & Pingry, 2018, p. A2)	**Relevance to DTD (private or public)**
"*What* provides an explanation of exactly what is captured in the data."	• Provide an explanation of the nature of DTD (private, public), the elements of the DTD, their intrinsic quality, as well as their relationship to real-life activities or events as they relate to the research question. • Justify mismatches or approximations that are loosely associated with core aspects of the research question based on empirical or conceptual fitness rather than convenience.
"*When* refers to the time at which the data is collected."	• Provide a detailed account of data collection period(s) as well as the period when access to the data was granted, if applicable.
"*Where* refers to the location (virtual or real) of the data collection."	• Although we expect that DTD are always virtual, researchers can sometimes triangulate DTD with real-life observations to increase confidence that the DTD reflect real-life processes. Virtual sources should be clearly identified (e.g., data providers).
"*How* describes the	• Provide comprehensive coverage of the techniques used to retrieve the data

precise process(es) of data collection."	and manipulate it from their raw format for use as a curated data set for numerical analysis. This includes extraction, transformation and loading techniques, as well as data profiling and quality assurance processes that support the generation of high-quality data. • Explain in detail which criteria were used to exclude any data from the original data set and justify the impact(s) of this decision on the curated data set and the analysis (e.g., reduced sample size, higher intrinsic quality). • Whenever possible, explain how the DTD were originally generated.
"*Who* details the individual(s) involved in the data collection."	• Explain how many individuals were involved in the collection, storage, retrieval, manipulation and analysis of the raw data set in order to generate the curated data set. • Provide detailed information on the various roles of those individuals, whether they implemented computational data collection processes (e.g., writing a service, an API) or performed quality assurance for those processes. • In some instances, these processes can be automated and performed by machines (e.g., quality assurance through continuous integration).
"*Which* details instruments or artifacts used in collecting the data."	• Provide a detailed account of the artifacts used to collect and manipulate the original DTD. This includes third party services, custom software, APIs and any other form of IT artifact used to create the curated data set. When relevant, provide version numbers as well.
"*Why* provides the set of reasons or goals for collecting the data."	• Ensure that there is a clear and well justified link between the data collected and the research question. Strive for fitness rather than convenience.

7.2. Bridging Data Quality Categories and the Seven W's

Researchers working with DTD spend a significant amount of time studying their data to identify and address outstanding quality issues. Returning to the specific case of GitHub, Kalliamvakou et al. (2016, p. 2041) provide an excellent description of what they aptly refer to as the *perils* of GitHub data. Without proper knowledge of *how* the GitHub platform and Git, its underlying source control system, work (along with their known bugs at the time of data collection) and how they are configured for a given repository, it is easy to make assumptions that can impact representational data quality and adversely affect findings.

In our view, readers should be able to find answers to these seven questions and understand how they relate to the four data quality categories. A comprehensive description of the data set (e.g., *what* type of data are contained within event payloads for different types of events) helps increase intrinsic, contextual data quality (e.g., completeness, amount of data), as well as representational data quality (e.g., interpretability, consistent representation).

Dimensions of accessibility (accessibility, access security) as well as other dimensions of contextual data quality (e.g., timeliness, completeness) are addressed by identifying *when* and *where* the data collection took place. *How, who,* and *which* questions contribute to all four data quality categories because they help us understand what data were collected as well as how they were transformed and manipulated to generate a curated data set in response to a specific research question. Last but not least, the *why* question is of paramount importance in order to contextualize the data collection process.

7.3. Encouraging Transparency and Replication

In the natural sciences, replication and transparency are key elements of the scientific process and the accumulation of knowledge. In IS, the push for replication remains largely absent, save for a few notable exceptions such as the journal *AIS Transactions on Replication Research*. However, we argue that DTD, and especially public DTD, are viable candidates for transparency and replication. Indeed, since a large portion of the data collection and transformation process is aided by IT, detailed explanations of the steps performed with these IT enable transparency over the process used to generate the curated data set. It should even be possible, in many instances, to share any piece of written software used to generate this curated data set. For example, if a program is written to retrieve and manipulate data from the GitHub Archive in order to generate a curated data set that is then used to perform some form of numerical analysis, the researcher could upload the source code to a repository on GitHub. This repository could be made public using an open source software license, or kept private but shared with reviewers. One could then re-execute the source code, and/or actually validate the code, to ensure that the quality of the curated data set satisfies the requirements of the study.

In this manner we could build better, more accessible tools to help us engage with DTD more efficiently, e.g., by using design science research principles to create artifacts that help the IS research community. Replication is an important aspect of scientific progress. Public DTD lend themselves to replication, but their inherent lack of quality means that, even if they are publicly available, they need to be collected and transformed before they can be used. Given that such collection and transformation processes are often extensive, raw DTD alone will not be sufficient to foster transparency and replication.

8. Concluding Remarks

DTD carry tremendous potential for the advancement of scientific knowledge in IS and other disciplines. In order for this potential to be realized and in the interests of producing replicable research, we need to engage in discussions on the quality of DTD. How DTD are collected, stored, retrieved and transformed into data that are fit for numerical analysis are important matters in the context of building methodological and theoretical contributions. With this short paper, we hope to encourage researchers to reflect on DTD quality and its relevance to their work. At a higher level, we hope to motivate the IS community to value DTD quality in research. Current efforts by publishers[3] in this area often go unnoticed or are perceived by researchers as not adding value. In our view this is far from the truth, since data quality, especially in the context of DTD, is an issue that we address on a constant basis in our research. It is high time that these efforts begin to percolate into our publications.

Acknowledgments

This work was supported by the Fonds de recherche du Québec – Société et culture (FRQSC). The author is grateful to the editor and the reviewers for their comments and suggestions.

[3] https://www.elsevier.com/authors/author-services/research-data

References

Alkhalil, A., & Ramadan, R. A. (2017). IoT data provenance implementation challenges. *Procedia Computer Science, 109*, 1134-1139.

Berente, N., Seidel, S., & Safadi, H. (2018). Research Commentary—Data-driven computationally intensive theory development. *Information Systems Research, forthcoming*.

Checkland, P. (1999). Systems thinking. In W. L. Currie & R. Galliers (Eds.), *Rethinking management information systems* (pp. 45-56). Oxford, United Kingdom: Oxford University Press.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research, 16*(1), 64-73.

Clarke, R. (2016). Big data, big risks. *Information Systems Journal, 26*(1), 77-90.

Cox, D. R. (2007). Applied statistics: a review. *The Annals of Applied Statistics, 1*(1), 1-16.

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). *Social coding in GitHub: Transparency and collaboration in an open software repository.* Paper presented at the 2012 Conference on Computer Supported Cooperative Work.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). *Addressing big data issues in scientific data infrastructure.* Paper presented at the Collaboration Technologies and Systems Conference.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly, 5*(1), 1-21.

Faraj, S., & Johnson, S. L. (2011). Network exchange patterns in online communities. *Organization Science, 22*(6), 1464-1480.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of marketing research, 25*(2), 186-192.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767-797.

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., & Damian, D. (2016). An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering, 21*(5), 2035-2071.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(1), 148-152.

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: The complete guide to dimensional modeling*. Indianapolis, IN: John Wiley & Sons.

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems, 115*, A1-A7.

Pentland, B., Haerem, T., & Hillison, D. W. (2009). Using workflow data to explore the structure of an organizational routine. In M. C. Becker & N. Lazaric (Eds.), *Organizational routines: Advancing empirical research* (pp. 47-67). Northampton, MA: Edward Elgar Publishing.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Tsay, J., Dabbish, L., & Herbsleb, J. (2014). *Influence of social and technical factors for evaluating contribution in GitHub.* Paper presented at the 36th International Conference on Software Engineering.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

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.

Zhang, Z., Yoo, Y., Wattal, S., Zhang, B., & Kulathinal, R. (2014). *Generative diffusion of innovations and knowledge networks in open source projects.* Paper presented at the International Conference on Information Systems, Auckland, New Zealand.

Zimmermann, T., & Nagappan, N. (2008). *Predicting defects using network analysis on dependency graphs.* Paper presented at the International Conference on Software Engineering, Leipzig, Germany.