AOSIS

Using historical data to explore transactional data quality of an African power generation company

CrossMark
click for updates

Authors:
Patient Rambe[1]
Johan Bester[1]

Affiliations:
[1]Department of Business Support Studies, Faculty of Management Sciences, Central University of Technology, Bloemfontein, South Africa

Corresponding author:
Patient Rambe,
prambe@cut.ac.za

Read online:
Scan this QR code with your smart phone or mobile device to read online.

Background: In developing countries, despite large public companies' reliance on master data for decision-making, there is scant evidence to demonstrate their effective use of transactional data in decision-making because of its volatility and complexity. For the state-owned enterprise (SOE) studied, the complexity of generating high-quality transactional data manifests in relationships between customer call transactional data related to an electricity supply problem (captured by call centre agents, i.e. data creators) and technician-generated feedback (i.e. data consumers).

Objectives: To establish the quality of customer calls transactional data captured using source system measurements. To compare this data set with field technicians' downstream system transactions that indicated incorrect transactional data.

Method: The study compared historical customer calls transactional data (i.e. source system data) with field technician-generated feedback captured on work orders (i.e. receiving system) in a power generation SOE, to ascertain transactional data quality generated and whether field technicians responded to authentic customer calls exclusively to mitigate operational expenses.

Results: Mean values of customer call transactional data quality from the source system and technician-generated feedback on work orders varied by 1.26%, indicating that data quality measurements at the source system closely resembled data quality experiences of data consumers. The SOE's transactional data quality from the source system was 80.05% and that of historical data set from evaluating feedback was 81.31% – percentages that exceeded average data quality measurements in literature.

Conclusion: Using a feedback control system (FCS) to integrate feedback generated by data consumers to data creators presents an opportunity to increase data quality to higher levels than its current norm.

Keywords: feedback control system; power generation; master data; transactional data quality; electricity supply problem; data management capabilities.

Introduction

Master data, which capture core information about an organisation's stakeholders, products and relationships amongst them (Haneem, Kama & Kazmi 2016), have been the staple discourse for explaining organisational operations (Infovest 2018). The concentration on master data has been accentuated by the recognition of big data as a strategic asset for organisations (Hagerty 2016) and increased business and data management capabilities in the digital economy (Bärenfänger, Otto & Gizanis 2015). These capabilities manifest in advanced analytics tools' capacity to generate insightful business information and enhance the provision of digital services to customers (Organisation for Economic Co-operation and Development [OECD] 2016). Such capabilities are also evident in digital technologies that avail and pool masses of data from multiple modes to one central place or distributed locations for in-depth analysis. Whilst these capabilities can increase data types that can be analysed by organisations, master data remain the preferred data sources because of the panoramic picture it renders on business operations (Entity Group 2016). The oversight on less strategic data forms, such as metadata and transactional data quality, is concerning (Palavitsinis 2013) because data quality is a strategic resource in prudent decision-making, especially in public institutions where scarce resources must be strategically harnessed to ensure the long-term sustainability of programmes (Alketbi 2014).

The limited application of transactional data in explaining business operations is attributed to its volume, volatility (Bester 2019) and complex connections with daily operations of organisations. The American Council for Technology-Industry Advisory Council (ACT-IAC) (2015) observes that

the high volume of transactional data that government agencies deal with daily, complex business processes, complex policy requirements and ancient technology systems complicate the management of transactional data quality. The American Council for Technology-Industry Advisory Council (2015) elaborates that even within the same agency, different divisions apply the same data differently and how these data elements are defined within the same agency varies widely.

Literature on transactional data and its applications in large organisations has gained momentum even though such data should be used cautiously because of their complexities (Hand 2018). The increasing prominence of transactional data quality in explaining organisational operations arises from increasing programme funding accountability and effective monitoring of public programmes (Rothbad 2015), the need to render responsible customer experiences and the growing competitive intensity of knowledge-intensive industries (Biegel et al. 2018). Other considerations include the demand for data analytics to steer contemporary business operations and the need for comprehensive business insights (Hagerty 2016), the surging costs of surveys and transactional data's potential to generate insightful discoveries (Hand 2018).

The need for high transactional data quality has been accentuated by large secondary data sets used in secondary analysis (Hand 2018), and organisations' desire for richer understanding of customers to provide relevant, meaningful and efficient interactions (Biegel et al. 2018). Despite these developments, management executives are concerned about the paucity of sufficient, correct, reliable and timely data upon which to ensure sound organisational decisions and bemoan the depletion of quality information (Bester 2019).

One grey area regarding transactional data quality is whether and the extent to which data sets generated at the organisational sources (e.g. by call agents) are consistent with other data sources (e.g. feedback generated by field technicians who deal with customer queries on data quality). Addressing this issue contributes to improving data quality at multiple organisational levels by integrating various information from diverse sources (Rothbad 2015). Quality assurance (QA) processes should commence at the initial data entry stages and progress through the entire process of integrating data from multiple sources to build an integrated data system (Rothbad 2015). The failure to corroborate transactional data quality from multiple agent sources can lead to incongruence of data quality, imprudent decision-making within large publicly owned companies (Rambe & Bester 2016) and catastrophic social consequences such as massive power failures (Fürber 2015).

Providing in-built complementarities and coherence within and across transactional data from multiple agencies ensures appropriate evaluation of public programmes (Rothbad 2015). To establish the quality of transactional data, we examined the coherence of data generated by an African power generation and distribution company (APGDCO[1]) from two fronts: data generated from customer calls received at the call centre (customer calls transactional data) and technician feedback on customer transactions they executed. The study addressed the following questions:

- What is the quality of transactional data captured from customer calls at APGDCO based on source system measurement?
- How many downstream system transactions have feedback of field technicians that indicates incorrect transactional data?

Problem background

The African power generation and distribution company has many customers such as households, corporations, mines and city councils. Larger customers often have complex monitoring systems for measuring quality of electricity received and can alert APGDCO immediately when an interruption in electricity supply occurs. Small customers cannot afford monitoring systems and report electricity supply problem (ESP) cases to APGDCO's call centre whenever they experience electricity outages. Subsequently, call centre agents probe customer queries using call scripts or case-based reasoning to categorise an issue correctly. If agents conceive APGDCO's network as the source of the problem, an ESP case is logged and a field technician is dispatched to restore electricity supply.

Therefore, correct interpretation and classification of customer problems are paramount to utilising APGDCO's resources appropriately. Sometimes, call centre agents may misinterpret a customer's explanation of an issue or are misled by customers. Consequently, a fault caused by a customer's defective equipment can be logged as an ESP case, something that APGDCO should not respond to as it is the customer's responsibility to resolve. For each transaction ESP case logged, a work order is automatically generated in the receiving system and technicians mark on the work order the cause of the ESP and the action performed to restore supply. If the issue was caused by customer's faulty equipment, the technicians will mark the customer side fault, which signifies that call agents captured incorrect transaction data at the source. As a data consumer, the technician is the primary authority on data quality as transaction data quality measured by the contact centre agents may not present a true reflection of the data consumer's experience.

Theoretical framework

Systems theory considers a system as a set of inter-related and interdependent parts (Mele, Pels & Polese 2010). Cybernetics, generally used as a synonym for systems theory, is concerned with the communication and regulation of systems using a feedback control system (FCS) (Skyttner 2001).

1.Because of intense scrutiny from civil society, this state-owned enterprise (SOE) is under pressure owing to its financial losses and allegations of malpractices, and therefore its name has been anonymised as APGDCO.

An FCS is configured to control itself, another system and feedback loops which are the mechanisms for exercising such control by facilitating desired outcomes. The simplest feedback loop is an open-loop system where input influences output via a control system (see Figure 1). Although such open-loop systems are cheap, simple to design and easy to maintain, their configuration is often inaccurate as no feedback is redirected into the system. Consequently, open-loop system can neither act on the output or external environment nor use it to influence subsequent outputs (ed. Liptak 2018).

A closed-loop system is a better configuration for a system to act on its external environment to regulate itself (see Figure 2). Whilst the system replicates an open-loop system, it has an appendage relaying feedback back to the input. Feedback can be used to adjust subsequent input or facilitate a desired output. A closed-loop system supports higher accuracy as it responds quickly to external and internal changes. However, this system is complex and expensive to create, maintain and often overcorrect itself if not designed carefully (ed. Liptak 2018).

Orr (1998) applied cybernetics to data quality by suggesting that an information system's data quality will be 100% if it represented data exactly as it existed in the real world and 0% if no conformity existed. Data quality changes over time and will not stay at a 100% level *ad infinitum*. Orr (1998) warns that the absence of an FCS will prevent continuous alignment between data in a system and the real world because feedback is the only valid mechanism that ensures the tracking of changes in the real world to related information systems. Even though Orr's (1998) application of cybernetics focussed on correcting existing master data records, an FCS can also improve the quality of transaction data.

Literature review
Data types

Data are defined as input of unprocessed items such as facts, figures, text, images, numbers, video and audio typically stored in a digital format (Epstein 2012; Nazim & Mukherjee 2016; Vermaat 2014). Consumer and brand-related data tend to be categorised into various types, namely, digital,
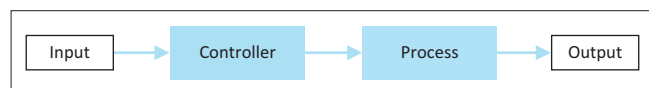
Source: Adapted from Bakshi, U.A. & Goyal, S.C., 2009, *Feedback control systems*, 2nd edn., Technical Publications, Pune.

FIGURE 1: Open-loop system.
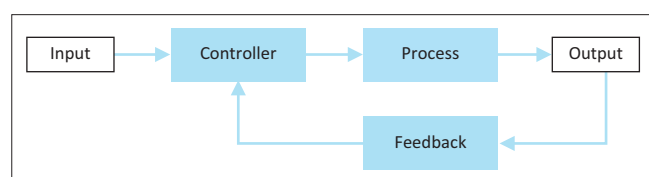
Source: Adapted from Bakshi, U.A. & Goyal, S.C., 2009, *Feedback control systems*, 2nd edn., Technical Publications, Pune.

FIGURE 2: Closed-loop system.

terrestrial, transactional, emergent/speciality and identity data (Biegel et al. 2018). Other classifications include unstructured data such as e-mails, files and videos, and structured data that comprise master, reference, transactional, metadata, history and queue data (Arun & Jabasheela 2014). Master data refer to vital business information about products, suppliers and customers that has a low change frequency. Reference data describe a business entity such as a customer, product or supplier and do not change frequently. Historical data are data that relate to previous transactions, which include master, reference and transactional data and are retained for compliance purposes (Borek et al. 2014). Transactional data are captured during an interaction and are combined with master and/or reference data to form a transaction at a specific time and have high volume and changes frequently (Borek et al. 2014). Overall, master data are often accorded the highest priority in organisational settings as these capture diverse organisational information such as customers, suppliers, products and their inter-relationships and are shareable across multiple business units of public entities (Nelke et al. 2015).

However, the extraction of quality data is often laborious and dissociated with public service quality, such as patient care in emerging contexts (Marais 2017). The high volume and velocity of various data forms imply that new tools and methodologies are required to capture, manage and process them efficiently (Statistics Divisions of United Nations, Department of Economic and Social Affairs [UN/DESA] and UN Economic Commission for Europe 2015). Therefore, huge investments in human expertise, technology and infrastructure are necessary for organisations to leverage fully the benefits of such data. More so, the silo-based approach to the capturing, curation and management of master data in government organisations also increases the duplication of master data across departments, thereby accentuating government costs and wasting resources (Haneem et al. 2016).

Data quality

Wang and Strong (1996) define data quality as data that are fit for use from the consumer's perspective. Although data quality has traditionally been considered from an accuracy perspective (Lin, Gao, Koronios & Chanana 2007), completeness, timeliness and consistency have become over-riding dimensions of data quality in recent times (ACT-IAC 2015). Completeness describes the extent to which data captured all dimensions that they were created to capture. According to Cai and Zhu (2015), completeness denotes that the values of all components of a single datum are valid. They elaborate that completeness is measured by establishing whether the deficiency of a component will impact the use of the data, especially for data with multi-components and whether the deficiency of a component will impact data accuracy and integrity (Cai & Zhu 2015).

Validity implies that data must be generated from reliable sources. Grillo (2018) concurs that the validity of data deals

with the trustworthiness of the data, which increases user engagement, sustains the functioning of a data warehouse and is a critical dimension in the development of a data quality scorecard (DQS) that measures the quality of data in the data warehouse. Consistency means that the same process or source must generate the same data and the same data values describing an object or event should be reflected across the entire organisation (ACT-IAC 2015). For databases, this means that the same data that are located in different storage areas should be considered to be equivalent or possessing equal value (Silberschatz, Korth & Sudarshan 2006). Consistency, therefore, points to whether the logical relationship between correlated data is correct and complete (Cai & Zhu 2015).

Timeliness denotes that data must be adequately current. It is the time delay from data generation and acquisition to its utilisation (McGilvray 2010) to allow for meaningful analysis. It points to the temporal difference between the time at which the data were collected and when they become available for analysis (Brackstone 1999; Keller et al. 2017). Accuracy implies that the appropriate values should be captured at the first time of data entry and should be retained throughout the enterprise (ACT-IAC 2015). Therefore, accuracy measures the degree to which data represent the phenomena they were designed to measure (Brackstone 1999) and whether their numerical value is within a specified threshold (Marev, Compatangelo & Vasconcelos 2018).

However, an improvement of one dimension of data quality may compromise another dimension. For instance, data can be provided timely but at the expense of its completeness. Moreover, whilst the use of data repositories may increase timeliness of access of data to scientists originally not involved in the initial data collection and experiments, access may be restricted if it violates information security and privacy, or accessibility of data may be restricted across organisations because of competitiveness across organisations (Milham 2012). Data completeness may be realised at the expense of their concise representation (Neely & Pardo 2002) and different data consumers may have varying evaluations of the same data quality (Lin et al. 2007). Moreover, the high data expenditures and their exponential growth in relation to media channels they are associated with (Biegel et al. 2018) imply that different organisational departments may undervalue certain data quality dimensions, such as comparability, thereby compromising data quality.

Transactional data quality

Oracle Corporation (2011) observes that transactional data quality captures automated business processes such as sales, service, order management, manufacturing and purchasing. Point-of-sale transactions involve data aspects such as product, place, time, price and name of sale agent and contribute directly to the complexity and volatility of transactional data (Baran & Galka 2016). Research highlights that South African customer data quality is problematic and hampers the quality of decision-making in firms (Burrows 2014). Neil Thorns, Informatica's territory manager

for sub-Saharan Africa, contends that most South African companies have low transactional data quality with an average accuracy of 50% or less. This is lower than the 73% global mark advocated by Experian's (2017) latest survey on customer data. The poor data quality is attributed to few business rules and limited automation during data capturing because of limited high-technology adoption and resource constraints (Burrows 2014). The difficulty of upgrading multiple systems and departmental support systems through which data flow also contributes to low transactional data quality. One conundrum is that whilst the support systems through which data are captured are constantly changing, they may be run on antiquated programming languages, thereby making their interfacing with new programmes problematic, culminating in inaccurate transactional data (ACT-IAC 2015).

World Economics' Data Quality Index on the gross domestic product (GDP) of 154 countries employed five indicators – base year, system of national accounts, informal economy, quality of statistics and corruption – and ranked South Africa 49th, with an overall score of 77.1% (World Economics 2017). This ranking is higher than the 73% global mark set by Experian's (2017) survey on data quality. These contradictory statistics on South Africa's data quality raise perplexing questions about the quality of transactional data, thereby amplifying the need to conduct research on data quality.

Methodology
Research design

Because this research corroborated the perspectives of APGDCO's field technicians who resolved customer ESP queries with source data generated by APGDCO's call centre agents, a combination of a quantitative cross-sectional survey and historical transactional data was considered as the appropriate research design. Cross-sectional surveys serve to establish the prevalence of a phenomenon, attitudes and opinions from individuals to portray an overall picture at the time of conducting the study (Kumar 2014). To ascertain the quality of data, the study needed to compare the technicians' view with customer centre's data quality measurements. As such, the study employed historical data (i.e. work orders from the receiving system that were marked with customer-side fault) as the technicians' view and compared it with the contact centre's data quality measurements captured in the source system. The original data captured by call centre agents (i.e. historical transactional data) relating to client ESP data included call numbers, client locations and nature of ESP type. The historical data sets on transactional data employed in the study covered April 2012 to March 2017, a period that is conceived enough to detect trends in ESPs in the region the study covered.

Instrumentation

The structured questionnaire had two sections, namely, the demographic data section covering technicians' years of experience, the experience of using enterprise digital assistant

(EDA) and transactional data quality section (customer-side fault). Historical transactional data encapsulated the followings: volume of work order distribution across various service centres, source system transactional data quality versus feedback on transactional data per financial year.

Data collection

The second author (J.B.) conducted a census on the 303 technicians who resolved customer queries in APGDCO over 2 months in 2017. The self-administered questionnaire comprised ordinal and ratio scales. Respondents completed (1) a hard copy questionnaire, scanned and emailed it to the second author, (2) an electronic version that they word processed and emailed or (3) a computer- or smartphone-based web browser version of the questionnaire. Despite numerous follow-ups made with APGDCO technicians, the census generated a 35% response rate (106 out of 303), mostly from the scanned copy sent via email. These data were coded and captured on an Excel spreadsheet and exported to Statistical Packages for the Social Sciences (SPSS) for a detailed data analysis.

Data preparation

The historical transactional data were extracted using Structured Query Language (SQL) from the company's ESP application database containing work order records. The normalisation and enrichment of the data involved checking of record descriptions for the wording 'customer fault' and any variations that indicated a customer fault. An interpretation of transaction dates allowed for the addition of fields and categorisation of data according to APGDCO's financial years. A precompiled report that indicated the mean of the transactional call and quality per month was referenced to indicate transactional data quality at the source system per secondary data record. All records were exported to an Excel spreadsheet for secondary data preparation and initial analysis before its exportation to SPSS for detailed analysis.

Data analysis

The study employed descriptive statistics to organise and summarise data based on sample demographics (Holcomb 2016). Frequency distributions, means and percentage analysis were used for the presentation and analysis of descriptive statistics. This study's frequency distribution covered work order amounts and data quality measured at source system and via feedback. Means were generated for the quality of source system data and feedback from technicians. Inferential statistics were instrumental in drawing conclusions about the population by determining the relationship amongst variables and making predictions about the population (McKenzie 2014). Correlation and regression analysis were used to determine the associative and predictive relations of variables examined.

Validity and reliability

Nieuwenhuis (2014) argues that validity requires the researcher to document procedures for evaluating the trustworthiness of the data collection and analysis. The statistician and researchers examined the conciseness and precision of questions to ensure face validity. The technicians' manager also appraised the questionnaire to ensure that the breadth of questions covered data quality issues (i.e. content validity). The researchers also verified raw data through availing questionnaires and field notes to respondents to correct factual errors (Nieuwenhuis 2014). For transactional data at the source, all compiled data sets were exchanged with data custodians (data analyst and manager of call centre agents) to authenticate data sources and ensure the credibility of results.

The researchers and the statistician shared all data sets to ensure consistency in the coding process, and where opinions varied, they met to clarify consistencies and generate consensus. This ensured intra- and inter-coder reliability. The instrument had an internal consistency of 0.638 even though two items had low consistency (see Table 1). The small number of items on transactional data quality perhaps explain the low internal consistency of construct on this instrument.

Presentation of findings

The results on years of experience at an APGDCO Customer Service Centre (CSC) in Figure 3 show that 40.566% of the technicians had 6–10 years, 27.358% had 0–5 years, 21.698% had more than 20 years, 8.491% had 11–15 years and only 1.887% had 15–20 years of experience. It is important to understand that all values were rounded to three decimals so that they add up to 100.000%. Although rounding them to one decimal seemed appropriate mathematically, it was not considered as the values would be too close to but not add up to 100%. Most of the technicians had more than 5 years of work experience and, therefore, had enough work exposure to fulfil their responsibilities.

The African power generation and distribution company adopted EDA devices in the 2011–2012 financial year. Thus, the maximum possible experience a technician had on EDA usage was around 5 years. The distribution analysis of EDA usage, depicted in Figure 4, reveals that only 8.491% of the technicians had EDA usage experience of 1 year or less, whilst 53.774% had more than 4 years of experience. It is important to note that even when expanded to three decimals, exact summation to 100.0000 occurs at fourth decimal only. Therefore, the EDA usage experience was enough for technicians' effective utilisation of these devices judging from the distribution analysis.

TABLE 1: Cronbach's alpha coefficients on questionnaire.

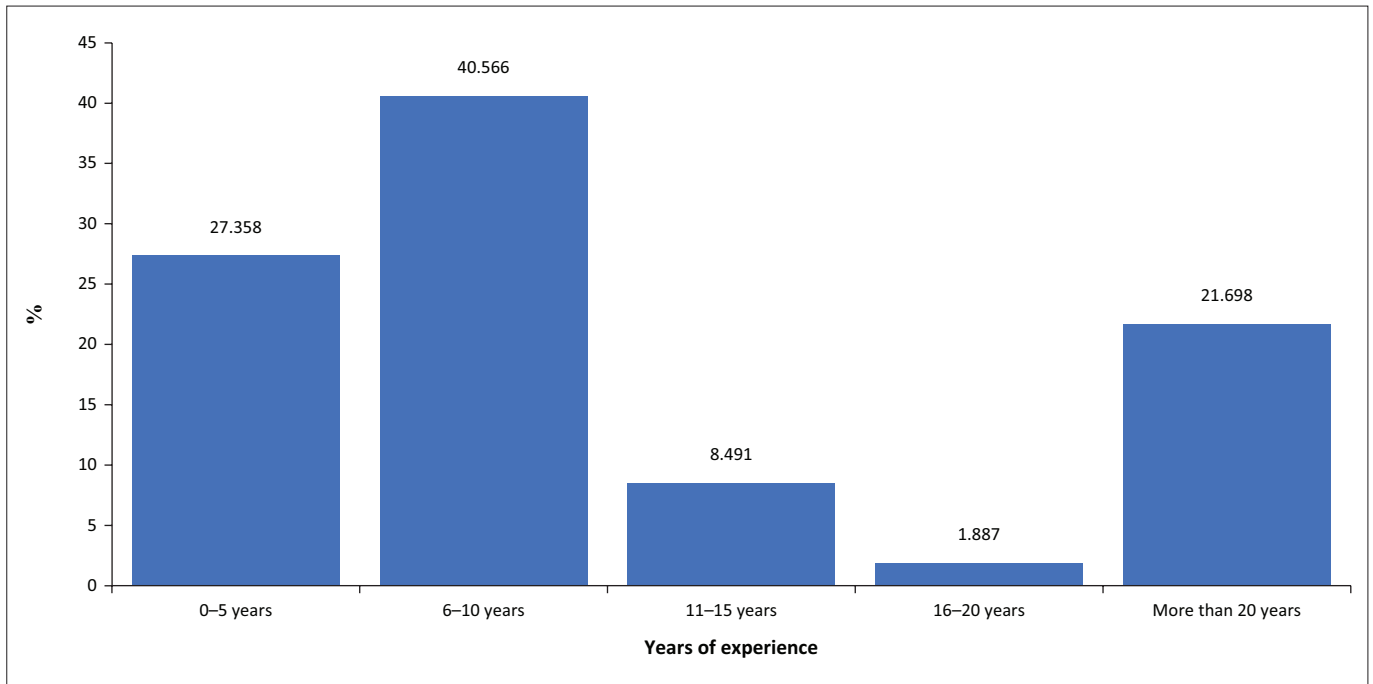| Construct	Question	Number of items	Cronbach's alpha	Comment
Transactional data quality	5 and 8	2	0.114	Low internal consistency
Overall questionnaire	4–18	15	0.638	Moderate internal consistency

FIGURE 3: Distribution of years of experience at African power generation and distribution company Customer Service Centre (CSC).
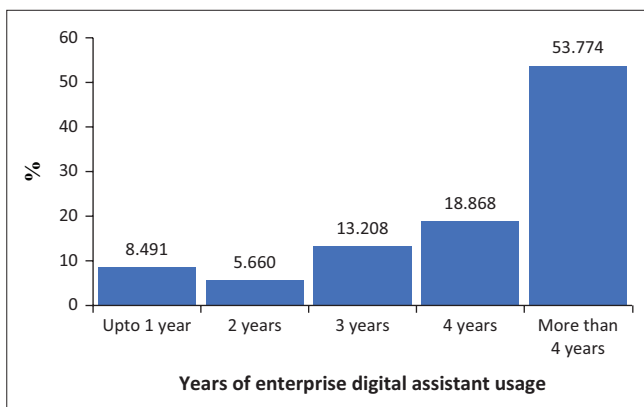

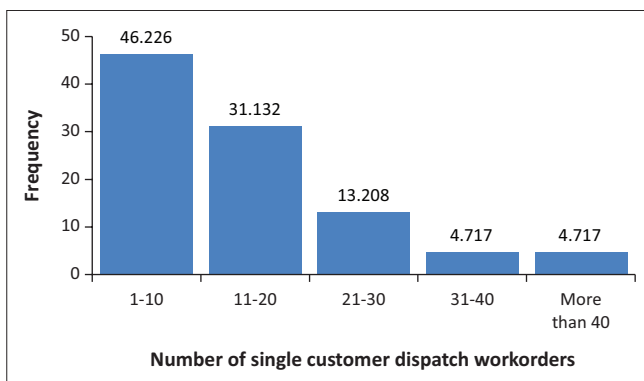**FIGURE 4:** Distribution of years of enterprise digital assistant usage.

FIGURE 5: Distribution of single customer dispatch work orders received per month.

Figure 5 indicates that most technicians received less than 20 single customer dispatch work orders per month, 46.226% had 1–10 and 31.132% had 11–20 work orders. Therefore, 77.358% of technicians received an average of one customer dispatch work order per working day per month. Therefore, technicians were not overburdened with dispatch work orders and could properly evaluate work and provide feedback on the executed work.

Source system and feedback data quality

The descriptive summary in Table 2 shows a hierarchical breakdown of APGDCO's historical data. The three-level hierarchy system has at its lowest level the service centres where technicians are located, who are responsible to restore ESP cases. Service centres are grouped one level up under sectors, which are again grouped under zones as the highest level. It contains columns for the amount of ESP work orders received, the overall percentage of work orders, work order percentage received during work time and overtime, source system data quality as measured by the call centre agents (data creators) at the source system and data quality based on feedback from technicians (data consumers) on work orders.

Source system data quality is measured via a QA process that evaluates 1% of historical call centre system transactional data and expresses it as a quality percentage. Data quality based on technician feedback is determined by dividing the total number of historical work orders at the receiving system by the amount of work orders marked as valid ESP transactions and displaying it as a percentage.

Table 2 reveals that data quality for all hierarchy levels, which data creators measured at the source system, had a very narrow distribution. It consisted of a mean of 80%, a maximum deviation from the mean of 0.2% and a minimum deviation of 0%. Measurement of data quality via technician feedback at the receiving system indicated a mean of 81.3%,

TABLE 2: Historical data on total work order volume distribution.

Hierarchy	Work order	% Work order	% Work time	% Over time	% Source system data quality (measured by data creator)	% Feedback data quality (measured from technician feedback)
Service group 1 zone	92280	39.11	48.0	52.0	80.0	82.6
Service group 1.A sector	60243	25.53	50.1	49.9	80.0	82.0
Service centre 1.A.1	7996	3.39	47.4	52.6	79.9	83.8
Service centre 1.A.2	14775	6.26	49.5	50.5	79.8	80.6
Service centre 1.A.3	9956	4.22	44.9	55.1	80.1	82.1
Service centre 1.A.4	8185	3.47	51.1	48.9	80.1	84.9
Service centre 1.A.5	9204	3.90	58.4	41.6	80.1	79.1
Service centre 1.A.6	10127	4.29	49.7	50.3	80.1	83.2
Service group 1.B sector	32037	13.58	44.1	55.9	80.1	83.6
Service centre 1.B.1	8206	3.48	45.7	54.3	80.1	83.5
Service centre 1.B.2	7129	3.02	42.4	57.6	80.2	84.8
Service centre 1.B.3	4260	1.81	49.4	50.6	80.2	82.4
Service centre 1.B.4	3952	1.67	44.6	55.4	80.0	79.3
Service centre 1.B.5	8490	3.60	41.2	58.8	80.0	85.3
Service group 2 zone	143665	60.89	54.8	45.2	80.0	80.5
Service group 2.A sector	74352	31.51	53.3	46.7	80.0	80.1
Service centre 2.A.1	9001	3.81	45.9	54.1	80.1	78.3
Service centre 2.A.2	4798	2.03	49.2	50.8	79.9	76.7
Service centre 2.A.3	8819	3.74	46.7	53.3	80.1	80.4
Service centre 2.A.4	13574	5.75	50.6	49.4	79.9	77.0
Service centre 2.A.5	5471	2.32	48.3	51.7	80.0	84.1
Service centre 2.A.6	23311	9.88	64.1	35.9	80.0	81.8
Service centre 2.A.7	9378	3.97	48.4	51.6	80.0	81.5
Service group 2.B sector	69313	29.38	56.5	43.5	80.0	80.8
Service centre 2.B.1	1808	0.77	37.5	62.5	80.1	95.0
Service centre 2.B.2	9515	4.03	54.4	45.6	80.0	86.5
Service centre 2.B.3	16886	7.16	48.7	51.3	80.1	79.0
Service centre 2.B.4	22542	9.55	58.9	41.1	79.9	77.2
Service centre 2.B.5	1939	0.82	38.2	61.8	79.8	95.3
Service centre 2.B.6	16623	7.05	66.5	33.5	80.0	81.2
Grand total	235945	100.00	52.2	47.8	80.0	81.3

Source: African power generation and distribution company (APGDCO) historical data

with the biggest deviation of 14% at Service Centre 2.B.5 and the closest deviation of 0.1% observed at Service Centre 2.B.6. The measurements were very similar, implying that measurement by data creators confirmed the data quality experience measured from data consumers' feedback. The higher deviation from the mean found within feedback measurements could be attributed to high granularity in the data analysed (i.e. per transaction) compared to the much lower granularity (i.e. monthly) of data creators' measurements.

Comparison of transactional data quality: Source system with feedback

Data quality measured at the source system by the data creators was based on results from a QA process performed on a 1% sample of the total historical ESP transactions for a month. This QA process consumes much time and only 1% of the total transactions could be assessed. The QA process evaluates the extent of call professionalism when data capturers are answering calls, for instance, whether correct steps were followed to identify a customer, interpret his or her fault symptoms, send the fault to the correct department and if the call was concluded professionally. Most scoring

relates to transactional data quality and hence this score is used as a proxy for transactional data quality at the source system. Data quality measured at the receiving system via technician feedback was calculated by dividing the total number of historical work orders automatically generated from ESP cases logged by the amount of work orders marked as a valid ESP case.

A comparison of historical transactional data quality measured by data creators at the source system and historical data quality measured by evaluating feedback from technicians on work orders at the receiving system reveals whether feedback is impacted by source system data quality or not. Transactional feedback from field technicians indicates the correctness of work order/ transactional data quality if it describes a valid resolution of an ESP.

Descriptive analysis is presented in Table 3 where each financial year's source system transactional data quality and feedback on correct transactions are compared to the average for all years combined. A value higher than the average is denoted with a double dagger, whilst a value lower than the average is marked with a single dagger. The indicated daggers for source system data quality and

TABLE 3: Source system transactional data quality versus feedback on correct transactions per financial year.

Financial year	Source system transactional data quality (%)	Feedback: Correct transactions (%)
2012–2013	77.99†	81.37‡
2013–2014	80.51‡	80.78†
2014–2015	78.26†	80.42†
2015–2016	81.76‡	82.50‡
2016–2017	81.72‡	81.50‡
Average	**80.05**	**81.31**

†, value lower than the average; ‡, value higher than the average.

feedback indicating correct transactions correspond in the financial years 2014–2015, 2015–2016 and 2016–2017.

Figure 6 displays data quality from both source system and transactions, with feedback indicating correct data. Transactional data quality from the source system indicates an upward trend, despite the slump during the 2014–2015 financial year. Data quality based on transactional feedback from the data consumer perspective also displays a general upward trend even though it has more declines than source system data quality. Therefore, on a year-by-year basis, the perspectives on data quality by the data creator and data consumer did not always cohere. Therefore, data creators can focus on improving data quality measurements to ensure its closer reflection of the data consumer's perspective, but this must be considered carefully as the small sample size of the transactions measured by data creators might explain this difference in results.

Correlation analysis between source system data quality (independent variable) and data from transactional feedback (response variable) in Table 4 reveals a weak positive statistically significant (0.01; $p = 0.00002$) relationship between the two variables. The weak correlation could be influenced by the extensive granularity of data quality measured via technician feedback (e.g. measured per individual work order) versus low granularity of data quality measured at the source system that only analyses 1% sample of transactions via the QA process.

The study also sought to establish whether data quality measured at source has an impact on data quality measured via feedback (i.e. whether the nature of data quality at the source predicted the data quality generated by field technicians).

Logistic regression was used to assess inferentially the relationship between transactional data quality from the data creator (i.e. contact centre) and transactional feedback from the data consumer (i.e. field technician). This tool is preferred as the independent variable, source system data quality is a ratio-scale variable measured in percentages and the response variable, transactional feedback, is a binary value where 1 = correct and 0 = incorrect. The results in Table 5 illustrate that transactional feedback is significantly dependent on source system data quality (Wald statistic = 18.209, $df = 1$, $p = 0.000$).
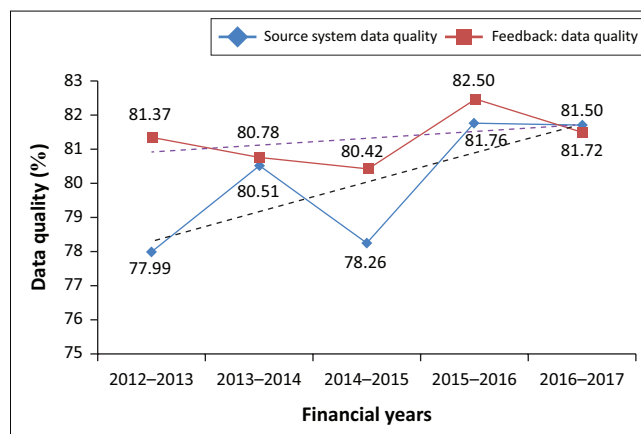
FIGURE 6: Source system and feedback-based data quality per financial year.

TABLE 4: Correlation between source system data quality on transactional feedback.

Correlations	Variable	Source system data quality	Comment
Transactional feedback	Correlation	0.01	Significant
	p	0.00002	Significant
	N	235945	Significant

TABLE 5: Logistic regression of transactional data quality on transactional feedback.

Variables in the equation	Response variable: Transactional feedback					
	B	SE	Wald	df	p	Exp(B)
Source system data quality	0.008	0.002	18.209	1	0.000	1.008
Constant	0.850	0.145	34.338	1	0.000	2.341

SE, standard error; B, coefficient; df, degree of freedon; Exp(B), exponentiation of the B coefficient; Wald, Wald chi-square test.

Discussion

A comparison of percentages of transactional data quality at the source system and feedback on correct transactions gave interesting results. Most of the percentages on feedback on correct transactions were higher than those of source system transactional data quality, creating an impression that data from execution of work orders provided a more accurate picture of transactional data quality than data captured at the source. Nevertheless, the differences between these percentages were marginal and were closer to the averages for the years considered. These differences demonstrate the absence of a clear feedback loop from technicians' feedback on transactions to source system data quality, which can compromise the quality of data at the source. Therefore, despite the importance of data management in addressing data quality challenges, sub-optimal data entry affects the operational efficiency and accurate reporting of business operations (Oracle Corporation 2011). Notwithstanding the minor differences between data quality averages measured from feedback data and source system transactional data in Table 3, at least 18.69% (= 100% – 81.31%) to 19.95% (= 100% – 80.05%) of the dispatched work orders should not have been executed. These minor variances cost APGDCO substantially from a resource and reputation perspective as the resource expended, namely, labour hours and vehicle kilometres driven could have been spent on, for example,

plant maintenance work. Furthermore, customers would wait longer for their supply to be restored because of invalid ESP calls receiving undue attention. The variances between these perspectives on data demonstrate that transactional data constitute large, complex, unstructured data sets that are hard to deal with using conventional tools and techniques and reflect public institutions' incapacity to optimally unlock value from such data (OECD 2016). Despite these discrepancies between source data quality and feedback data quality, the general upward trajectory of data quality between 2012 and 2017 indicates that APGDCO was striving to use its analytical capabilities to optimise data quality, notwithstanding its institutional constraints surrounding productive usage of data. The prudent use of data analytics is considered instrumental in facts-based strategic decision-making, improving operational efficiency and revenue levels (Singh 2018). However, APGDCO has not sufficiently exploited these data-driven capabilities to increase its profit margins and revenue.

The correlation result demonstrates a positive statistically significant but weak relationship between source system data quality and transactional feedback. This somewhat coheres with the upward trajectory reported in the interaction of source system data quality and transactional feedback – a clear indication that as more correct data quality is captured at the source, technicians are more inclined to execute more correct ESP (which are faults of APGDCO), which results in more feedback of correct transactions executed. This finding could give credence to the claim that data integration and transaction analytics are rendering implementable data-driven insight in a visually interactive (to both data agents and technicians) and iterative, agile manner (OECD 2016). However, accepting this claim contradicts the observation that the persistent albeit marginal discrepancies between source system data quality and transactional feedback are clearly indicative of the failure of APGDCO to exploit technicians' feedback to optimally improve successive data capturing processes. These discrepancies between results of data quality measured from data creators and data consumer perspectives demonstrate that processing big data remains a daunting task for public institutions and cannot guarantee optimal decision-making even though it is instrumental in decision support systems (Alketbi 2014).

The results of regression analysis show that transactional feedback is significantly dependent on source system data quality – that is source data significantly predicts feedback quality. This means the frequency of accurate data captured at the source increases the successful resolution of customer ESPs by field technicians. Apart from accurate data capturing, data consolidation and real-time data synchronisation have potential to improve data quality in organisational operations (Oracle Cloud 2015). This finding supports the view that because decision-making in companies can be foreseen and proactive, the initiation of information collection process earlier means that the timeliness of information should

improve the quality of data-driven decision-making (Dillon, Buchanan & Corner 2010).

From a theoretical perspective, the current transactional data system under scrutiny demonstrates an open-loop system as depicted in Figure 7, where:

- Input is the customer call.
- Controller function is performed by the contact centre following the case-based reasoning logic.
- Process such as ESP will be determined via the controlling function.
- Output is a resultant work order executed by the field technician and feedback captured during the completion of the work order.

As evident in the results of the correlation and regression analyses, an increase in correct transactions captured will result in improved transaction feedback, indicating that correct transactions were captured and executed. To improve the quality of transactional data captured, an FCS with a closed-loop configuration shown in Figure 8 would be invaluable. This configuration is the same as Figure 7, but the feedback of the technician is fed back to the controller.

The application of an FCS in this context is different from the one envisioned by Orr (1998) to update an existing master data record. Updating an incorrect transaction after it was executed will not provide any immediate benefit to the organisation as costs (e.g. on mileage, time and overtime payments) would have already been incurred by the organisation to resolve a customer-side fault. This understanding seems incongruent with the claim that large sets of raw data combined with powerful and sophisticated analytics tools provide insights that improve operational performance of large organisations (Henke et al. 2016). This is because recognition of a mistake may not always translate into pre-emptive and corrective future interventions even though
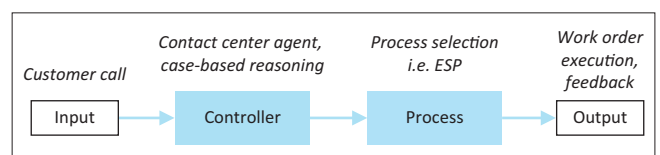
ESP, electricity supply problem.

FIGURE 7: Open-loop transaction data process.
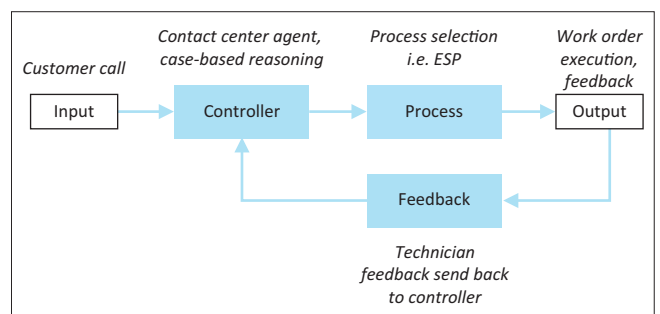
ESP, electricity supply problem.

FIGURE 8: Closed-loop transaction data process.

organisations would be expected to learn from past mistakes to enhance the accuracy of their future transactions. However, when organisations learn, such a learning experience which unfolds via feedback is also known as a self-normative FCS (ed. François 2011). In this example, feedback indicating a customer side fault will be fed back to the data creator/contact centre agent. This feedback can then be used to identify the root cause of the incorrectly logged customer-side fault, such as data creator interpretation, issues in the logic of the case-based reasoning questions or customer providing false information.

Managerial implications

The fact that accurate source data predicts technician feedback implies that effective data management through reducing fragmentation and inconsistencies amongst data agents dealing with data collection can contribute to improved data quality in dynamic environments (Lin et al. 2007). Therefore, organisations can stand to improve their operational efficiency, reduce financial misappropriations and waste if their data capturing at the source improves. This implies that managers should commit more financial and intellectual resources towards continuous on-the-job training in data capturing to reduce inaccurate data capturing which leads to multiple technician assignments to resolve ESP caused by customer mistakes, such as overloading of their residential electrical circuit and faulty appliances causing a circuit breaker to trip.

Transactional data quality can also be improved by an FCS, if (1) a mechanism is available to capture feedback and (2) the feedback is integrated back to the controller of the transactional data. This will allow identification of the root cause of poor-quality data and facilitate the reduction in or eradication of future occurrences thereof. As the complexity of a closed-loop system comes at an increased cost, performing a cost–benefit analysis is critical before deciding whether to implement an FCS strategy to improve transactional data quality (Batini & Scannapieco 2016).

The fact that at least 18.69% – 19.95% of the dispatched work orders should not have been executed is indicative of the financial and time cost that can be avoided should appropriate data capturing techniques be applied. As such, customer misinformation must be prevented by requiring data capturers to sufficiently probe customer queries before dispatching technicians to ensure that correct, comprehensive and reliable data are captured at the source. Because the mistakes made by data capturers also contributed to ESP faults, continuous training on data capturing can contribute to the generation of quality data at the source. Moreover, managers and executives should continue to focus on the critical importance of correct data entry as it forms the basis of a successful data management system.

When this happens, the root cause of faulty data entry can be identified and corrective actions and/or learning and application can be facilitated to reduce or ultimately prevent future occurrences of invalid transactions.

Implications for future research

In view of the consistencies (upward improvements in data quality) and discrepancies (technician feedback quality being slightly higher than source-generated data) between data quality measurements from technician feedback and source-generated data, future studies may need to track the sequence of the generation of each perspective/source on data quality (technician-generated and source-generated) and their synchronisation and integration. As Hand (2018) observes, transaction data are special types of administrative data concerned with the sequences of events relating to transactions, their retention in databases and analysis to improve the understanding of organisations' operations. As such, what happens at each stage becomes critical to improve the quality of transactional data.

Data quality improvement costs should always be an important consideration before deciding on an appropriate quality improvement strategy (Batini & Scannapieco 2016). Therefore, even though an FCS has the potential to facilitate transaction data quality improvements, the cost to implement should be carefully analysed before implementation to ensure a positive return on investment (ROI). Future studies can examine the return on investment potential of various FCS and compare it to the improvement in transaction data quality achieved. This is because although an FCS could be initially expensive to establish, the ROI would rise over time of its application. Future studies may also explore methodologies to accurately determine the cost of incorrect transactional data across industries and different businesses and determine at which level of data quality does the cost to improve negate the benefits derived.

Because some inconsistencies in data quality may be attributed to organisational difficulties in defining master and transactional data, future studies can examine the different ways in which such data types are defined and characterised in organisations to ensure congruence and consistencies in ways the same data pieces are defined, categorised and assessed by different agents within and outside organisations. Recent literature suggests that the quality of a data set and each of its individual components can be influenced by a number of different factors (e.g. inaccuracy, imprecision in definitions, gaps or inconsistencies in measurements), thus necessitating the development of a context-dependent data quality evaluation framework (Marev et al. 2018). For consistency and believability, consistent representation of data partly depends on the precise characterisation and definition of data dimensions in specific contexts (Batini et al. 2009; Pipino, Lee & Wang 2002).

Conclusion and recommendations

In view of the inconsistencies between data quality measurements from technician feedback and source-generated data, the entire value chain of data generation needs scrutiny. Moreover, Broadridge (2017) recommends increased granularity in consolidated data presentation because granular data often demonstrate that surface-level

transactions may be inconsistent with the picture developed from lowest-level detail. Moreover, Alketbi (2014) recommends the development of a complete data quality framework for improved data quality and data-driven decision-making in public organisations. At the level of predictive relations, the positive association between source-generated data and technician feedback data could be indicative of the capacity of tracking different types of data to improve overall data quality. Broadridge (2017) observes that data quality issues are proportionally related to the sequence in which firms tackle data improvement projects (i.e. from valuation or reference, transactional and derived data). Therefore, firms should consider the proper implementation of an FCS to enhance their data quality strategy, as it provides critical information required to drive four of the 12 directives of a continuous improvement programme as suggested by Sebastian-Coleman (2012), namely:

- Understand data quality as defined by its consumers.
- Attend to root causes of data issues.
- Measure data quality.
- Ensure data producers are kept accountable for the quality of their data.

Acknowledgements

Competing interests

The authors have declared that no competing interest exists.

Authors' contributions

All authors contributed equally to this work.

Ethical consideration

This article followed all ethical standards for research without direct contact with human or animal subjects.

Funding information

Data availability statement

Data sharing may not be possible for this article as some of the pieces of data will expose the name of the institution in which the research was conducted, thus compromising participants' anonymity.

Disclaimer

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any affiliated agency of the authors.

References

Arun, K. & Jabasheela, L., 2014, 'Big data: Review, classification and analysis survey', *International Journal of Innovative Research in Information Security* 1(3), 17–23.

Alketbi, O., 2014, *Data quality assurance for strategic decision making in Abu Dhabi's public organisations*, University of Bedfordshire, Luton.

American Council for Technology-Industry Advisory Council (ACT-IAC), 2015, *Data quality: Letting data speak for itself within the enterprise data strategy*, Collaboration & Transformation (C&T) Shared Interest Group (SIG) and Financial Management Committee, Fairfax, VA.

Bakshi, U.A. & Goyal, S.C., 2009, *Feedback control systems*, 2nd edn., Technical Publications, Pune.

Baran, R.J. & Galka R.J., 2016, *Customer relationship management: The foundation of contemporary marketing strategy*, 2nd edn., Taylor & Francis, New York, NY.

Bärenfänger, R., Otto, B. & Gizanis, D., 2015, *Business and data management capabilities for the digital economy White Paper. Capabilities in the digital economy*, Competence Centre Corporate Data Quality (CC CDQ), May 2015.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A., 2009, 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys* 41(3), 1–52. https://doi.org/10.1145/1541880.1541883

Batini, C. & Scannapieco, M., 2016, *Data and information quality: Dimensions, principles and techniques*, Springer, Cham.

Bester, J., 2019, 'The influence of transactional data quality improvements on monetary savings of Eskom distribution, Free State', Master's thesis, Central University of Technology Free State, Bloemfontein.

Biegel, B., Margulies, J., Maggi, G. & Davis, C., 2018, *The state of data 2018: A Winterberry Group Report*, December 2018, Winterberry Group, New York, NY.

Borek, A., Parlikad, A.K., Webb, J. & Wooda, P., 2014, *Total information risk management: Maximizing the value of data and information assets*, Elsevier, New York, NY.

Brackstone, G., 1999, 'Managing data quality in a statistical agency', *Survey Methodology* 25(2), 139–150.

Broadridge, 2017, *A practical guide on data initiatives for public and private investments: Establishing data governance and better data management*, Broadridge Financial Solutions, New Hyde Park, NY.

Burrows, T., 2014, *ITWeb: SA looks to address data quality*, viewed 03 January 2018, from https://www.itweb.co.za/content/JOlx4z7kGXY756km.

Cai, L. & Zhu, Y., 2015, 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal* 14(2), 1–10. https://doi.org/10.5334/dsj-2015-002

Dillon, S., Buchanan, J. & Corner, J., 2010, 'Comparing public and private sector decision making: Problem structuring and information quality issues', *Proceedings of the 45th annual conference of the ORSNZ*, November 2010, pp. 229–237.

Entity Group, 2016, *Why not to put transactional data in an MDM hub*, viewed 07 November 2018, from https://www.entitygroup.com/wp-content/uploads/2016/05/master_data_management_and_transaction_data_whitepaper.pdf. https://www.entitygroup.com/wp-content/uploads/2016/05/master_data_management_and_transaction_data_whitepaper.pdf.

Epstein, E., 2012, *Implementing successful building information modeling*, Artech House, Norwood, OH.

Experian Data Quality, 2017, *The 2017 global data management benchmark report*, viewed 01 April 2020, from https://www.experian.com.my/wp-content/uploads/2017/12/2017-global-data-management-benchmark-report.pdf

François, C. (ed.), 2011, *International encyclopedia of systems and cybernetics*, 2nd edn., Walter de Gruyter, München.

Fürber, C., 2015, *Data quality management with semantic technologies*, Springer, Wiesbaden.

Grillo, A., 2018, 'Developing a data quality scorecard that measures data quality in a data warehouse', PhD thesis, Brunel University, London.

Hand, D., 2018, 'Statistical challenges of administrative and transaction data', *Journal of the Royal Statistical Society: Series A* 181(3), 555–605. https://doi.org/10.1111/rssa.12315

Hagerty, J., 2016, *2017 planning guide for data and analytics*, Gartner Technical Professional Advice, Stamford, CT, 13 October 2016.

Haneem, F., Kama, N. & Kazmi, A., 2016, 'Master data identification in public sector organisations', *Advanced Science Letters* 22(10), 2999–3003.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B. et al., 2016, *The age of analytics: Competing in a data-driven world*, McKinsey Global Institute, McKinsey & Company, San Francisco, CA.

Holcomb, Z.C., 2016, *Fundamentals of descriptive statistics*, Routledge, New York, NY.

Infovest, 2018, *Introduction to data governance*, Rondebosch, Cape Town.

Keller, S., Korkmaz, G., Orr, M., Schroeder, A. & Shipp, S., 2017, 'The devolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches', *Annual Review of Statistics and Its Application* 4, 85–108. https://doi.org/10.1146/annurev-statistics-060116-054114

Kumar, R., 2014, *Research methodology: A step-by-step guide for beginners*, Sage, Los Angeles, CA.

Lin, S., Gao, J., Koronios, A. & Chanana, V., 2007, 'Developing a data quality framework for asset management in engineering organisations', *International Journal of Information Quality* 1(1), 100–126.

Liptak, B.G. (ed.), 2018, *Instrument engineers' handbook, volume two: Process control and optimization*, 4th edn., CRC Press, Boca Raton, FL.

Marais, H., 2017, 'Assessing the data quality of performance information generated by the health sector in the Breede Valley subdistrict for evidence-based decision-making', Master's dissertation, University of Stellenbosch, Cape Town.

Marev, M., Compatangelo, E. & Vasconcelos, W., 2018, *Towards a context-dependent numerical data quality evaluation framework*, Technical Report, University of Aberdeen, Aberdeen.

McGilvray, D., 2010, *Executing data quality projects: Ten steps to quality data and trusted information*, Publishing House of Electronics Industry, Beijing.

McKenzie, S., 2014, *Vital statistics: E-book: An introduction to health science statistics*, Elsevier Australia, Chatswood, NSW.

Mele, C., Pels, J. & Polese, F., 2010, 'A brief review of systems theories and their managerial applications', *Service Science* 2(1–2), 126–135. https://doi.org/10.1287/serv.2.1_2.126

Milham, M.P., 2012, 'Open neuroscience solutions for the connectome-wide association era', *Neuron* 73(2), 214–218. https://doi.org/10.1016/j.neuron.2011.11.004

Nazim, M. & Mukherjee, B., 2016, *Knowledge management in libraries: Concepts, tools and approaches*, Chandos Publishing, Cambridge.

Neely, P. & Pardo, T., 2002, *Teaching data quality concepts through case studies*, Centre for Technology in Government, Albany, NY.

Nelke, S., Oberhofer, M., Saillet, Y. & Seifert, J., 2015, *Method and system for accessing a set of data tables in a source database*, U.S. Patent 2015/0066987 A1, 05 March 2015.

Nieuwenhuis, J., 2014, 'Analysing qualitative data', in K. Maree (ed.), *First steps in research*, pp. 99–103, Van Schaik Publishers, Pretoria.

Oracle Cloud, 2015, *The role of data integration in public, private, and hybrid clouds*, Redwood Shores, CA.

Oracle Corporation, 2011, *Oracle master data management*, An Oracle White Paper, September 2011, Redwood Shores, CA.

Organisation for Economic Co-operation and Development (OECD), 2016, 'Using big data in tax administrations', in *Technologies for better tax administration: A practical guide for revenue bodies*, pp. 47–73, OECD Publishing, Paris.

Orr, K., 1998, 'Data quality and systems theory', *Communications of the ACM* 41(2), 66–71. https://doi.org/10.1145/269012.269023

Palavitsinis, N., 2013, 'Metadata quality issues', PhD thesis, Alcalá de Henares, Madrid.

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

Rambe, P. & Bester, J., 2016, 'Financial cost implications of inaccurate extraction of transactional data in large African power distribution utility', *Problems and Perspectives in Management* 14(4), 112–123. https://doi.org/10.21511/ppm.14(4).2016.14

Rothbad, A., 2015, *Quality issues in the use of administrative data records*, pp. 3–38, University of Pennsylvania, Philadelphia, PA.

Sebastian-Coleman, L., 2012, *Measuring data quality for ongoing improvement: A data quality assessment framework*, Elsevier, Waltham, MA.

Silberschatz, A., Korth, H. & Sudarshan, S., 2006, *Database system concepts*, Higher Education Press, Beijing.

Singh, H., 2018, *Using analytics for better decision-making*, viewed 01 April 2020, from https://towardsdatascience.com/using-analytics-for-better-decision-making-ce4f92c4a025?gi=ea4ae2c019a2

Skyttner, L., 2001, *General systems theory: Ideas and applications*, World Scientific River Edge, NJ.

Statistics Divisions of UN/DESA and UN Economic Commission for Europe, 2015, *Results of the UNSD/UNECE survey on organizational context and individual projects of big data*, Statistics Divisions of UN/DESA and UN Economic Commission for Europe, New York, NY, 03–06 March 2015.

Vermaat, M.E., 2014, *Discovering computers and Microsoft Office 2013: A fundamental combined approach*, Cengage Learning, Boston, MA.

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/10.1080/07421222.1996.11518099

World Economics, 2017, *The data quality index (DQI)*, viewed 03 January 2018, from https://www.worldeconomics.com/Pages/Data-Quality-Index.aspx.