Bus Inf Syst Eng 61(5):575–593 (2019)
https://doi.org/10.1007/s12599-019-00608-0

RESEARCH PAPER

Discovering Data Quality Problems

The Case of Repurposed Data

Ruojing Zhang · Marta Indulska · Shazia Sadiq

Received: 3 October 2017 / Accepted: 28 June 2019 / Published online: 22 July 2019

Abstract Existing methodologies for identifying data quality problems are typically user-centric, where data quality requirements are first determined in a top-down manner following well-established design guidelines, organizational structures and data governance frameworks. In the current data landscape, however, users are often confronted with new, unexplored datasets that they may not have any ownership of, but that are perceived to have relevance and potential to create value for them. Such repurposed datasets can be found in government open data portals, data markets and several publicly available data repositories. In such scenarios, applying top-down data quality checking approaches is not feasible, as the consumers of the data have no control over its creation and governance. Hence, data consumers – data scientists and analysts – need to be empowered with data exploration capabilities that allow them to investigate and understand the quality of such datasets to facilitate well-informed decisions on their use. This research aims to develop such an approach for discovering data quality problems using generic exploratory methods that can be effectively applied in settings where data creation and use is separated. The approach, named LANG, is developed through a Design Science approach on the basis of semiotics theory and data quality dimensions. LANG is empirically validated in terms of soundness of the approach, its repeatability and generalizability.

Keywords Data quality · Open data · Design science

1 Introduction

In contemporary societies and organizations, data is both a resource and an asset (Fisher 2009). For individual and organizational processes that depend on data, data quality has become a key determinant of the quality of decisions and actions (Stvilia et al. 2007). Poor data quality affects analytical results from Business Intelligence (BI) tools and Data Warehouses and causes severe losses to organizations (English 2009). As a result, in public and private sectors, several related initiatives have been launched, with data quality playing a leading role. Examples include the Data Quality Act enacted by the United States government (OMB 2002) and the Data Quality Assessment Methods and Tools (DatQAM) promoted by the European Commission (Ehling and Körner 2007).

Data quality has been an area of research for over 2 decades (Sadiq et al. 2011), with contributions from computer science, statistics, information systems, and respective domain areas such as health, transport and administrative data. It has been widely acknowledged that one cannot manage data quality without first being able to measure it meaningfully (Stvilia et al. 2007). Therefore, discovering the quality of a dataset is a fundamental task in

Accepted after two revisions by Matthias Jarke.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12599-019-00608-0) contains supplementary material, which is available to authorized users.

R. Zhang · S. Sadiq
School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia
e-mail: r.zhang3@uq.edu.au

S. Sadiq
e-mail: shazia@itee.uq.edu.au

M. Indulska (✉)
UQ Business School, The University of Queensland, St Lucia, QLD 4072, Australia
e-mail: m.indulska@business.uq.edu.au

most, if not all, data quality management and improvement projects (Batini et al. 2009). Recent research has called for a shift in focus from data providing/collection to issues of coverage, openness and data quality (Corsar and Edwards 2017).

Quality of data is typically assessed against certain stated requirements (English 2009; ISO 2011; Loshin 2001), which are elicited from data users. Hence, most data quality assessment approaches are user-centric and 'top-down', following well established design guidelines, organizational structures and data governance frameworks. In the current data landscape, however, users are often confronted with new, unexplored datasets that may have relevance and potential to create value. This situation is evident from the proliferation of publicly available datasets (Duus and Cooray 2016), commonly referred to as open data, and emergence of data markets (Elbaz 2012). These developments present an unprecedented opportunity to governments, business and entrepreneurs to harness the power of data for economic, social and scientific gains. However, since the creators and subsequent users of such *repurposed* data are often disconnected, there is a lack of knowledge on the quality characteristics of the data. Thus, the time-to-value from these datasets remains prohibitively long primarily due to the effort required in making the data ready for use (Belkin and Patil 2013). At the same time, the meta-data as well as the underlying data quality for these datasets is known to be deficient. For example, many open datasets have duplicate, inconsistent, and missing data (see Fig. 1), and generally lack easily accessible schema descriptions, e.g., the MusicBranz.org public dataset consists of 324 schema-less CSV files with a data volume of 35.1 GB.

The body of knowledge on how to evaluate the quality of datasets that exhibit characteristics typical of repurposed data is critically lacking (Clarke 2016). Indeed, an ad-hoc manual approach is the most common approach used in real applications (Hey and Trefethen 2003). Data curation tasks are completed manually by those who need to use the data (e.g., data scientists), in an ad-hoc manner without following well-defined processes or guidelines (e.g., they fix an error when they encounter it during their analysis for their own benefit). Such an approach, which falls within the data preparation stage of the analytics cycle and cannot be fully automated due to the essential nature of human judgement (Prat 2019), can introduce biases attached to the specific needs of the data scientists. The lack of transparency (i.e., performed actions may not be recorded) further creates problems of reusability of the new version of the dataset. Such a manual approach is also not scalable and generalizable to different datasets and use cases. Thus, data scientists need data exploration capabilities that will

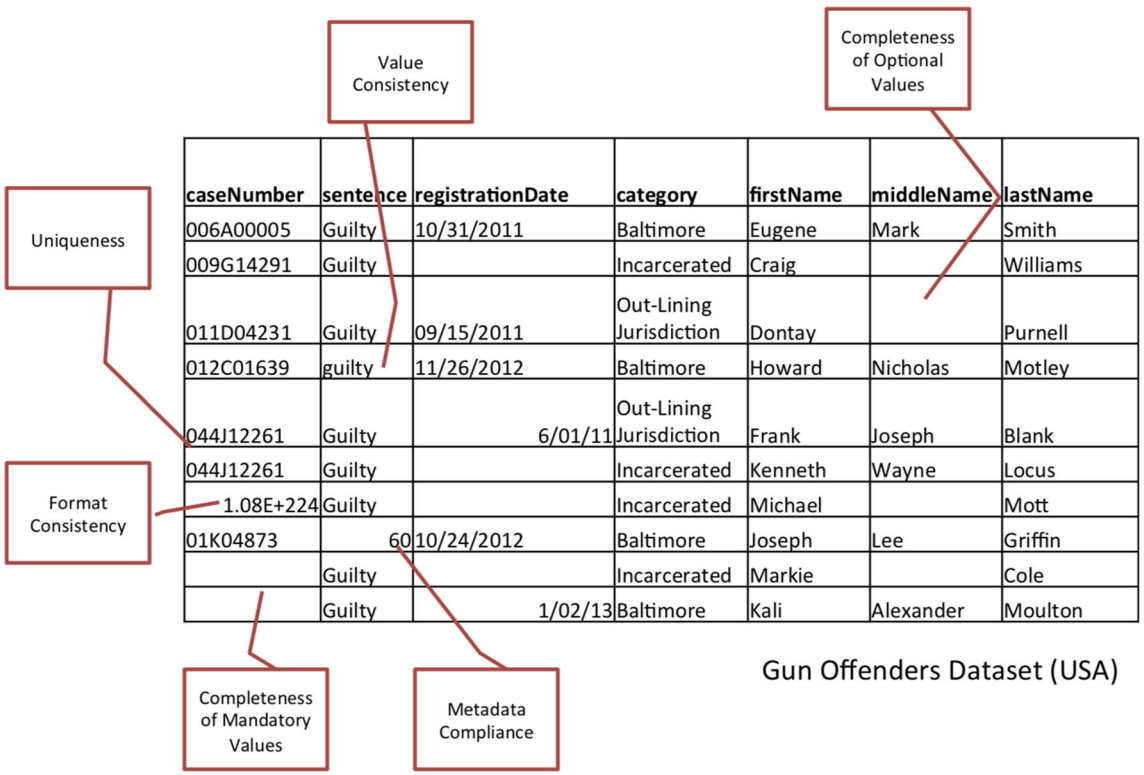

Fig. 1 Data quality issues identified in the Gun Owners open dataset. Source: Sadiq and Indulska (2017), data available from https://catalog.data.gov

allow them to consistently and systematically investigate the quality of the datasets in a bottom-up manner and, subsequently, understand the implications of their use to guide their decisions to use the data. We argue that a bottom-up approach has the potential to discover data quality problems using exploratory and/or analytical methods. The identification of data quality problems in repurposed datasets then allows data scientists to consider whether the identified problems are likely to affect the value that the business could generate from the use of the data. Indeed, the lack of awareness of data quality in repurposed datasets is considered a key challenge, and often results in unanticipated data exploration and preparation cycles that consume up to 80% of data scientist's efforts in analytical projects (Sadiq and Indulska 2017). Such problems have endured, despite several advances in sophisticated computational and statistical methods and tools to explore, transform and repair data. Accordingly, in this paper we aim to address this gap in knowledge and address the following research question: How can data consumers identify data quality problems, in a systematic and repeatable manner, in structured datasets that are unknown to them?

To answer this question, we develop a bottom-up approach for discovery of data quality problems in structured datasets with minimal meta-data. We refer to this approach as LANG.[1] LANG builds on our initial proposal of such an approach (Zhang et al. 2014) and is developed through a Design Science approach using semiotic theory and data quality dimensions as the theoretical foundations underpinning its design. Its development is a direct response to the critical lack of approaches for evaluating data quality of unknown datasets (Clarke 2016), thus aiming to advance practice that is typically based on ad-hoc approaches (Hey and Trefethen 2003). Our approach is refined on the basis of focus groups with data custodians and curators and on the basis of results from our repeatability evaluation. The refined approach is tested for generalizability through an application on 20 open datasets. The results of our evaluation show that LANG has the capacity to accelerate the data quality discovery process for a number of data quality problems that require minimal or no knowledge of the data. Our findings also provide insight into the requirements for extending data quality discovery.

In the remaining sections of this paper we first present a synopsis of foundational concepts from data quality research and practice, specifically data quality dimensions, and a summary of approaches related to data quality assessment. We also briefly present the foundations of the semiotic theory and how it relates to our work. Section 3 describes the methodology used for conducting this research. Section 4 presents the LANG approach. Section 5 describes the empirical evaluation of LANG for soundness, repeatability and generalizability. Section 6 provides a summary of insights from the design and evaluation process. Finally, Sect. 7 summarizes the contributions and limitations of LANG and discusses future research directions.

2 Data Quality Foundations and Related Work

2.1 Data Quality Dimensions

Data quality dimensions are a central notion to assessing data quality and modelling data quality requirements. The general idea of data quality assessment is to make a judgment about data quality dimensions (Caballero et al. 2007, 2008). Various data quality methodologies are based on identifications of data quality dimensions. For example, Storey and Wang (2001), Tu and Wang (1993) and Wang et al. (2001) extended the Entity Relationship model (ER) with quality characteristics to measure data quality dimensions and model data quality requirements. Naumann and Rolker (2000) proposed three assessment-oriented data quality criteria classes: subject-criteria, object-criteria and process-criteria which respectively relate to a source of data quality criteria, viz. the user, the information itself and the query process on the information. Lee et al. (2002) empirically derived data quality dimensions that are important to information consumers and grouped the dimensions into four categories: intrinsic, contextual, representational, and accessibility. Pipino et al. (2002) categorized commonly used data quality dimensions into two categories, viz. objective and subjective categories, and focused on the assessment of the objective dimensions by proposing three functional forms. Price and Shanks (2004) derived data quality criteria from the three levels of a semiotic framework of data quality, and hence their quality criteria can be seen as consisting of three categories, i.e. syntactic, semantic and pragmatic. Stvilia et al. (2007) identified the sources of Information Quality (IQ) variance, and developed a taxonomy of data quality dimensions that allows an evaluation of the data quality variance caused by these sources. Based on an analysis of literature, their taxonomy consists of 22 data quality dimensions organized into three categories, i.e. intrinsic, relational or contextual and reputational.

Typically, data quality requirements in organizations are documented following a hierarchy of data quality dimensions and associated metrics (ISO 2012), thus data quality

[1] The researchers named the approach as 'LANG' – 'Lang' conveys the meaning of 'becoming clear' in the Chinese language, which fits with the aim of the approach, that is, to make clear the data quality requirements of a dataset.

Table 1 Relevant data quality dimensions. Source: Adapted from Jayawardene et al. (2013a)

Dimension	Description
Completeness	
Completeness of mandatory attributes	The attributes which are mandatory for a complete representation of a real-world entity must contain values and cannot be null
Completeness of optional attributes	Optional attributes should not contain invalid null values
Accuracy	
Precision	Attribute values should be accurate as per linguistics and granularity
Validity	
Business rule compliance	Data should comply with business rules
Meta-data compliance	Data should comply with its meta-data
Consistency	
Uniqueness	The data is uniquely identifiable
Non-redundancy	The data is recorded in exactly one place
Semantic consistency	Data is semantically consistent
Value consistency	Data values are consistent and do not provide conflicting or heterogeneous instances
Format consistency	Data formats are consistently used
Referential integrity	Data relationships are represented through referential integrity rules

dimensions are foundational to data quality management and data governance. However, many disparate classifications of data quality dimensions have proliferated over the years (Byrne et al. 2008; English 2009; Gatling et al. 2007; HIQA 2011; Loshin 2006; McGilvray 2008; Stvilia et al. 2007; Wand and Wang 1996), with overlaps and contradictions. The disparity in the definition, and consequently understanding and interpretation of the dimensions, results in significant variance in the quantification of data quality problems depending on which type of benchmark is applied, e.g., 5 Star Open Data Rating System (Kim and Hausenblas, 2012) vs the Open Data Institute Certification Badge (Corsar and Edwards, 2017).

Recent research has offered an empirically validated consolidation of these dimensions – specifically, Jayawardene et al. (2013a) analyzed and consolidated sixteen published sources and consolidated 129 definitions of data quality dimensions to develop a comprehensive list of 33 dimensions, grouped across eight categories namely Completeness, Accuracy, Validity, Consistency, Currency, Availability and Accessibility, Reliability and Credibility, and Usability and Interpretability. The consolidation of Jayawardene et al. (2013a) incorporates literature on data quality dimensions definitions and taxonomies from both researchers and practitioners over the past 20 years, and hence is considered comprehensive in scope. Jayawardene et al. (2013a) distinguish two classes of dimensions: dimensions which are inherent in the data, that is, can be determined without knowledge of the use-context; and those which are dependent on the use of data. For example, *uniqueness* of an attribute can be determined by the presence of duplicates in the data, however, *usefulness and relevance* can only be determined when the use context is known. The work reported in this paper relies on the data quality dimensions identified by Jayawardene et al. (2013a) that are not dependent on the use context. Table 1 identifies the 11 relevant dimensions from Jayawardene et al. (2013a) that have been used to develop LANG.

2.2 Data Quality Assessment

A wide range of techniques have been proposed to assess and improve the quality of data, such as record linkage, and integrity constraints. The diversity and complexity of these techniques led to an effort from research and practitioner communities in data management, to define methodologies that help select, customize, and apply data quality assessment and improvement techniques. A *data quality methodology* is thus a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data (Batini and Scannapieco 2006). We note that data quality assessment, which measures the quality of data collections along relevant quality dimensions, is an inevitable component in all data quality methodologies.

Data quality assessment is generally comprised of several steps, as follows:

- *data analysis*, which examines data schemas and performs interviews to reach a complete understanding

Table 2 Summary of user-centric data quality assessment methodologies. Source: Authors' own summary

Methodology	Summary
A methodology for information quality assessment (Lee et al. 2002)	A data quality assessment and improvement methodology that consists of three components, the PSP/IQ model (Product and Service Performance model for Information Quality), an Information Quality Assessment (IQA) instrument and Information Quality (IQ) Gap Analysis Techniques. The assessment of information quality is conducted through a *user survey*. The IQ dimensions covered by the PSP/IQ model are *empirically* derived based on the perspective of information consumers
Assessment methods for information quality criteria (Naumann and Rolker 2000)	A new classification of IQ criteria based on the source of the IQ score, which are *perception of the users*, the data source and the query process of assessing the information. The assessment methods are subjective to *individual user's experiences* and understanding of certain criteria
Data quality assessment methodology (Pipino et al. 2002)	A methodology that focuses on identifying the general quality measurement principles for a data quality assessment and improvement effort. It distinguishes between subjective measures pertaining to the *perceptions, needs and experiences of data consumers*, and objective measurements based on the dataset itself. For comprehensive understanding of the data quality regarding a metric, both objective scoring and subjective assessment on the data are conducted and then compared and combined.
Total data quality management (Wang 1998)	A methodology for IQ assessment that includes a *survey-based diagnostic instrument* for IQ assessment; a related software tool to collect data and plot IQ dimensional scores for the *individual*, organizational role, and overall averages once data has been collected; and a methodology for IQ management
Comprehensive methodology for data quality management (Batini and Scannapieco 2006)	A business process oriented methodology. It comprises of three phases: (1) state reconstruction phase reconstructs the relationships among organizational units, processes, services, and data. (2) Assessment phase involves *interviewing internal and final users* to identify the most relevant problems and their causes, locate the critical variables affected by poor data quality, analyses data quality requirements and select relevant dimensions and metrics, as well as undertake data quality assessment based on an *understanding of the organization* and its data. (3) Choice of the optimal improvement process phase identifies the optimal improvement process. This approach emphasizes gathering and modelling of *contextual knowledge*, compared to many other methodologies that implicitly assume that contextual knowledge is known
InfoQ methodology (Kenett and Shmueli 2014)	A top-down approach that focuses on the utility of the data, and is entirely goal (user)-centric. The InfoQ methodology is oriented towards the *usage context* of statistical analysis. The methodology takes on a statistical perspective in defining and assessing information quality since its goal is to determine whether or not the data facilitates the statistical analysis task at hand. While it provides guidance to IQ assessment for data use in scenarios such as hypotheses testing, population effects quantifying and data summarization, its potential for general use in data quality assessment is limited

of data and related architectural and management rules (English 1999; Lee et al. 2002; Pipino et al. 2002; Scannapieco et al. 2004; Wang 1998),

- *data quality requirements analysis*, which surveys the opinion of data users and administrators to identify quality issues and set new quality targets (English 1999; Loshin 2001; Pipino et al. 2002),
- *identification of critical areas*, which selects the most relevant databases and data flows to be assessed quantitatively (English 1999; Lee et al. 2002; Loshin 2001; Pipino et al. 2002; Scannapieco et al. 2004; Wang 1998),
- *process modeling*, which provides a model of the processes producing or updating data (English 1999; Loshin 2001; Scannapieco et al. 2004; Wang 1998), and
- *measurement of quality*, which selects the quality dimensions affected by the quality issues identified in the data quality requirements analysis step and defines corresponding metrics; measurement can be *objective* when it is based on quantitative metrics, or *subjective*, when it is based on qualitative evaluations by data administrators and users (English 1999; Lee et al. 2002; Loshin 2001; Pipino et al. 2002; Scannapieco et al. 2004; Wang 1998).

Batini et al. (2009) provide a comprehensive analysis of existing approaches for data quality assessment and requirements identification. Our further analysis of the notable contributions to data quality assessment indicates that most, if not all, of these approaches follow a user–centric, top-down approach, where requirements are solicited from users before the data is explored. In Table 2 we

provide a summary of these contributions highlighting the user-centric nature where relevant.

Despite significant advances in data quality management attained through the introduction of methodologies such as the above, the translation of these methodologies in the new context of repurposed data is not obvious. This limitation stems from the lack of access to data creators and custodians who can provide experiential and domain knowledge on the data characteristics. In these settings, data-driven methods need to be utilized. There are two existing areas where data-driven methods have been considered for data quality assessment – data exploration and data profiling.

Data exploration, where statistical methods are used to reveal facts about data, has been researched over more than a decade (Dasu and Johnson 2003). These facts are used to formulate quality criteria and, thereby, evaluate quality, followed by data cleansing activities to improve quality. Dasu and Johnson (2003) provide a comprehensive list of existing statistical methods for data exploration. While the authors emphasize the possibility of using these methods for the purpose of data quality problem detection, there is a lack of methodology or guidelines for conducting such an exploration of an arbitrary dataset.

Data profiling is a related concept to data exploration (Abedjan et al. 2015) and has a significant commercial tool market. Gartner estimates that this market reached $1.4 billion in software revenue back in 2014 (Saul and Friedman 2015). The market's growth was forecast to accelerate, with estimates of $2.1 billion value by the end of 2017 (Saul and Friedman 2015), and compound annual revenue growth of 10% is predicted in this market through to 2021 (Selvage et al. 2017).

Approximately 60% of the market is dominated by several large and well-established vendors, such as IBM, Informatica, Pitney Bowes, SAP and SAS. The remaining 40% is divided among a large number of providers, including Microsoft, Oracle, Talend, Ataccama, Human Inference and Experian QAS, to name a few. These profiling tools focus on a wide range of capabilities including statistical distribution analysis of data, data redundancy checks, detecting data glitches, outlier detection, functional dependency analysis, column correlation analysis, validity checks etc. However, such tools generally lack reliance on a well-defined set of data quality dimensions and instead focus on niche, even esoteric definitions, thus reducing the explanatory power of the profiling reports and potentially resulting in different outcomes for the same dataset.

While there have been several contributions towards measuring data quality against specific dimensions through data quality profiling (Abedjan et al. 2015), statistical approaches (Dasu and Johnson 2003), as well as work on assessing data quality through the discovery of data dependency constraints (Fan and Geerts 2012), these solutions are specialized towards specific dimensions (such as consistency, or freshness). One or two dimensions alone are inadequate to capture an accurate and complete picture of the overall data quality, which can span a large number of dimensions (Jayawardene et al. 2013b). Additionally, these solutions are generally underpinned by assumptions relating to the availability of significant meta-data [e.g., data distributions (Dasu and Johnson 2003), thresholds (Song and Chen 2011) and probabilities (Köhler et al. 2015)], which may not be readily available for open or repurposed datasets.

Accordingly, current methodologies and tools lack the capacity to comprehensively evaluate the quality of datasets that exhibit characteristics typical of repurposed datasets, i.e. where data created for one purpose, is used by a new community of users. This necessitates generic approaches that facilitate assessment of data quality in a data-driven manner, but without the overhead of specialist algorithmic/statistical knowledge and tools and/or extensive knowledge of the meta-data. Indeed, there is "a critical need for exploratory tools and approaches that allow users to become aware of the data's shortcomings in terms of their intended use" (Sadiq and Indulska 2017). This paper advances the current body of knowledge and practice by developing such an approach for structured data in a tool agnostic manner and with minimal meta-data requirements.

2.3 Foundations of Semiotic Theory

The semiotic theory (semiotics), the philosophical theory of signs, has a natural and clear correspondence with data. It derives the quality categories and the corresponding criteria as presented below. Thus, in relation to the development of a data quality approach, a semiotic approach provides a rigorous and internally coherent theoretical basis, compared to other research approaches, i.e. empirical, e.g., (Wang and Strong 1996), practitioner, e.g., (Maydanchik 2007), or literature-based (Eppler 2001). The semiotic theory has also been previously utilized in the information systems area, including on quality management i.e. application of semiotics to understand Information Systems by Stamper (1992, 1993) in the context of systems analysis, then in the context of evaluating the quality of data models (Krogstie 2002; Krogstie et al. 1995a, b; Lindland et al. 1994), and finally application for evaluating information as well as data model quality (Shanks and Darke 1998; Shanks and Tansley 2002).

The design of LANG is based on modern semiotics (Morris 1938; Pierce 1931–1935), wherein three semiotic levels are studied: syntactic, semantic and pragmatic. Respectively, these levels examine the relationship between (sign) representations, the relationship between representation and referent, and the relationship between

representation and interpretation. A datum stored in a database or data warehouse can be seen as a sign, which has a stored representation, reflecting a certain external referent, with its own interpretation, which is to be carried out by a human or machine based on a certain context. Quality of data has been frequently measured from the perspective of format, meaning and use (Price and Shanks 2004, 2005a; Shanks and Tansley 2002; Shanks and Darke 1998). Therefore, a correspondence between semiotics and data quality can be observed.

Price and Shanks (2004, 2005a) argue that objective measures evaluate data quality by assessing the degree of the data's conformance to predefined requirement specifications, integrity rules, or through its correspondence to external phenomena. Subjective measures, on the other hand, continuously survey information consumers' task-dependent quality perceptions (Price and Shanks 2004, 2005a). Thus, the syntactic and semantic levels correspond with objective (user-independent) quality measures, whereas the pragmatic level of semiotics corresponds with subjective (user-dependent) quality measures. Further, Price and Shanks (2004, 2005a) derive a set of criteria for data quality for each semiotic level. Since this research aims to develop a data-driven approach which explores the dataset in a bottom-up manner, the objective quality measures viz. the syntactic and semantic levels, are closely relevant to the development of the approach, whereas the user-dependent measures viz. the pragmatic level, are outside the scope of this research.

3 Research Approach

Our development of LANG follows the Design Science (DS) approach (Hevner et al. 2004) because it guides the development of artefacts that are both practice-inspired and theory-ingrained. As an approach, DS provides the right balance between research rigor and relevance in research (Rosemann and Vessey 2008), which is important given our aim to develop a bottom-up approach for data quality assessment – a process artefact (Venable et al. 2012) that aims to help data scientists systematically discover and assess the quality of repurposed datasets.

We are guided by DS guidelines (Hevner et al. 2004) and specifically by Peffers' et al. (2007) DS process model (see Fig. 2) in our development. Our artefact is inspired by the lack of bottom-up data quality assessment approaches and by evidence in the existing body of knowledge that this void causes ad-hoc approaches to data quality assessment (Clarke 2016; Hey and Trefethen 2003). To address this need, we use semiotic theory and consolidated data quality dimensions as kernel theories that inform our artefact (Gregor and Jones 2007) and create a mapping between these to provide a foundation for our design.

We therefore set out, on the basis of these kernel theories, to develop an approach (a *method*) that can be used by data scientists to identify data quality problems in datasets unknown to them and with minimal meta-data present. Similarly to Chakraborti and Dey (2019), we combined multiple different evaluations of our artefact. An SQL instantiation of our initial LANG design is evaluated for soundness (Peffers et al. 2007 – demonstration stage) in a real-world context with two datasets. Our application of LANG is then verified with practitioners familiar with those datasets. As it is important the approach result in consistent outcomes, we also conduct a repeatability evaluation in a laboratory setting (Peffers et al. 2007 – evaluation stage). The soundness and repeatability evaluations provide input that allows us to refine LANG, following which we further evaluate the refined version of

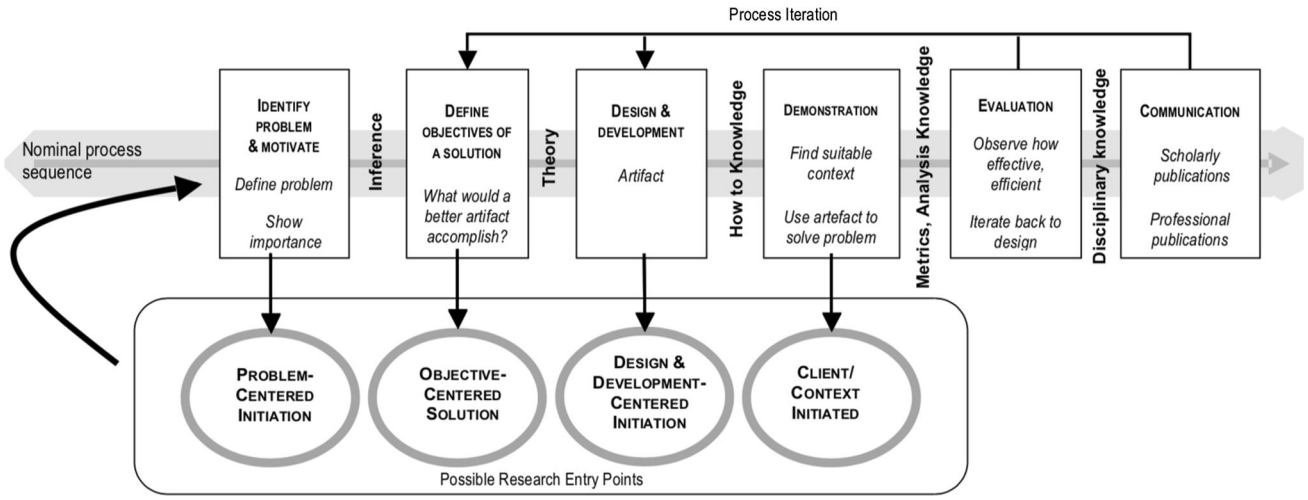

Fig. 2 Design science process. Source: Peffers et al. (2007)

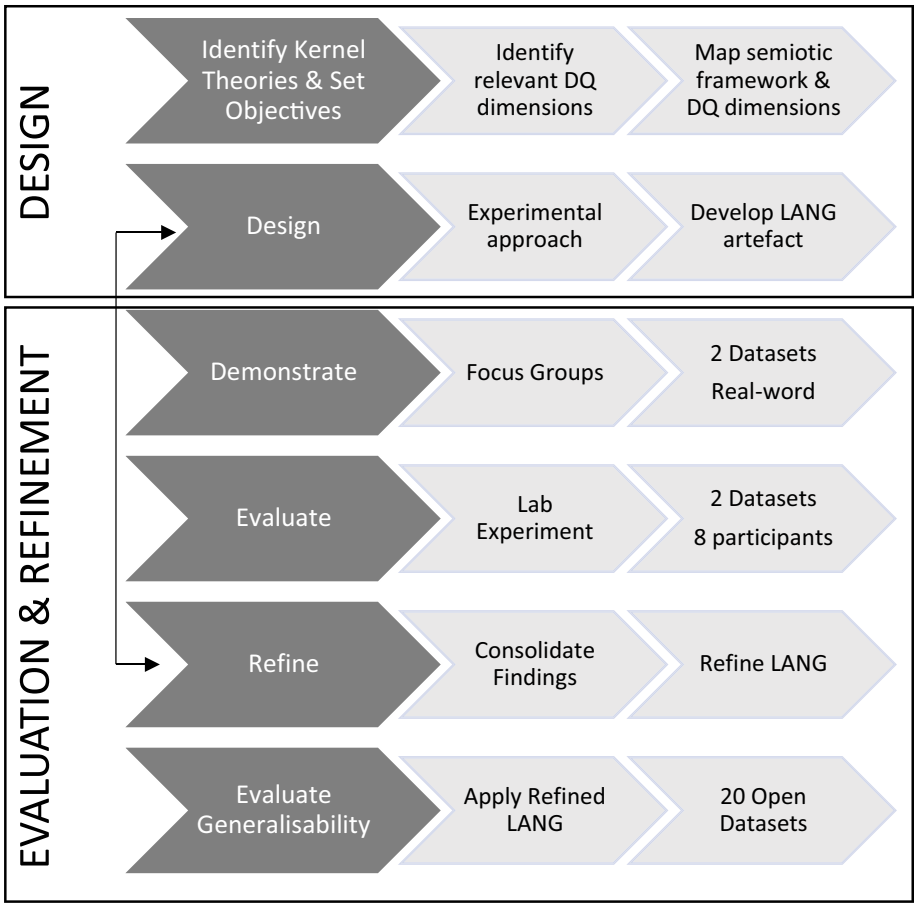

Fig. 3 Development of LANG. Source: Authors, informed by Peffers et al. (2007)

LANG through its application to 20 open datasets. Our evaluations are informed by the methodological guidance of Venable et al. (2012, 2016). The research approach is summarized in Fig. 3.

In this section we first present the theoretical basis of design for the proposed methodology LANG – our chosen kernel theories – followed by a description of the design and evaluation stages of LANG.

3.1 Design of LANG

Our design is founded on Price and Shanks' (2004, 2005a, b) semiotic framework due to its rigor and internal coherence. However, the quality criteria used by Price and Shanks (2004, 2005a, b) are not consistent with the notion of 'data quality dimensions', which is central to all data quality management efforts. Accordingly, our first step in the design required a mapping of Price and Shanks's (2004, 2005a, b) quality criteria to the consolidated data quality dimensions (Jayawardene et al. 2013a, b) to derive an operationalizable and rigorous basis for LANG (Zhang et al. 2014). This process was done by two researchers on the basis of matching definitions. Building on the semiotic theory, the mapping[2] served as a basis for the development of LANG.

Following semiotic theory, and the mapped data quality dimensions relating to each semiotic level, LANG was developed experimentally based on a real public transport smart card data set. For an arbitrary dataset, the approach provides a set of investigation guidelines/steps in two stages. In line with the foundation of the semiotic theory, LANG consists of two stages, viz. a syntactic stage, which focuses on identification of data quality issues relating to Uniqueness, Format Consistency, Referential Integrity, Meta-data Compliance, Business Rule Compliance; and a semantic stage, which focuses on detection of data quality issues associated with data semantics which include Completeness of Mandatory Attributes, Completeness of Optional Attributes, Semantic Consistency, Value Consistency, Precision, Redundancy.

3.2 Evaluation and Refinement

According to Hevner et al. (2004), a design artefact is effective when it satisfies the requirements and constraints

[2] The mapping is omitted due to length considerations but is available from the authors upon request.

of the problem it was meant to solve. Venable et al. (2012) further indicates that Design Science artifacts can be evaluated for (at least) three main reasons, viz: rigor, efficiency and ethics. In our study, we focus on evaluation of rigor in terms of LANG's effectiveness. Specifically, we conduct several evaluations that fall under the category of "Human Risk and Effectiveness" because we need to "rigorously establish that the utility/benefit will continue in real situations and over the long run" (Venable et al. 2016). We therefore evaluate an SQL instantiation of LANG to ensure its utility, repeatability and generalizability. In line with the research method guidance for evaluation (Venable et al. 2012), we conduct a soundness evaluation through a focus group approach (ex-ante naturalistic approach). Further, we conduct a repeatability evaluation through an experiment with 8 research students, as a means of ex-ante artificial evaluation. The results of these two evaluations inform the refined LANG approach described in this paper. Finally, we offer a generalizability evaluation as further evidence of LANG's effectiveness. This last ex-ante naturalistic evaluation allows us to reason about LANG's applicability in repurposed data settings.

3.2.1 Soundness Evaluation

Soundness of LANG was evaluated via focus group studies with the custodians of data (practitioners), in line with the methodological recommendations of Venable et al. (2012) for ex ante naturalistic evaluation. A focus group involves carefully selected individuals who are assembled to discuss a topic that is the subject of the study (Powell and Single 1996). As it allows clarification and justification of opinion, a focus group enables the researcher to gather from the target audience both the general opinion as well as the rich, detailed perspectives that could not be obtained through other methodological strategies (Krueger and Casey 1994; Morgan 1993).

In this study, the focus groups were designed to reveal the extent to which the identified data quality problems were perceived by the users of the data as actual data quality problems. Three criteria were used to examine the extent of alignment: *true positives*, which refer to correctly identified data quality problems; *false positives*, which refer to data quality problems that are identified but are not considered as genuine problems by the data users; and *false negatives*, which are genuine problems that the approach failed to identify. Results from the focus groups allow an authentic assessment of the utility of the approach, and accordingly identify opportunities for refining the approach and facilitate a better understanding of its limitations.

Two datasets were used for soundness evaluation of LANG, viz. the public transport smart card data and the Online Learning data. The Smart card data is a real public transport smart card dataset consists of 16.9 GB, 69,572,902 records, and contains minimum documentation. The data is mainly used for analysis by the public transport service company itself and its BI provider – a small-sized Australian Business Intelligence provider to the transport industry (referred to as company A in the following text). Minimal meta-data documentation describing the attributes in the data was provided with the dataset. The second dataset is one of Online Learning data - the log data of an Australian university's online learning system. The data is used for analysis of students' learning behavior by the university to facilitate better teaching outcomes. Data is generated through access to learning resources. The data was accompanied only by field names and without meta-data documentation.

The researchers first investigated the two datasets using LANG, and derived a set of data quality problems for each dataset. Two focus groups were conducted, one for each data set, respectively. Results from the application of LANG to the respective dataset were presented at the respective focus group. For the smart card focus group, the three participants were IT practitioners from company A. The participants were the director and the core technicians who hold the data warehousing and post-load aggregation responsibilities for the smart card data. For the Online Learning data, the focus group of six participants included administrative, managerial, and technical database users with IT backgrounds from the university's in-house data custodian team. To increase the reliability and ensure consistency during data collection between focus group studies, a semi-structured protocol was devised and employed that guided the research team during the focus groups. Each focus group lasted approximately 1 h. Field notes were taken during the sessions, and the discussion was transcribed prior to analysis.

The protocol dictated that each focus group commence by having a member of the research team introduce the motivation for the study and explain the data quality dimensions to participants. Handouts of data quality dimension definitions were also distributed to participants for convenience of reference at later stages of the focus group. For each identified problem, the relevant dimension definition was presented as an explanation of the diagnosis (e.g., Format Consistency: Data formats are consistently used), and a snapshot of one or more examples showing the potential problem records (e.g., different formats in the data for the same bus stop), were presented on a projector screen. The participants were then asked about their opinions regarding four aspects of the problem being discussed so as to gain insights into the problem and ensure correct interpretation: whether they recognize the problem reported by the approach as a real problem, and (if relevant): what was the cause of the problem; whether the

problem had since been fixed; and how the problem was addressed. According to the protocol, for each session, after being presented with all identified problems, the participants were asked if they could recall any other quality problems that had not been identified in the session. This question was posed in relation to any data quality dimensions, which were shown on the screen (i.e. to identify any false negatives).

3.2.2 Repeatability Evaluation

Apart from the soundness validation, a repeatability lab experiment was conducted to evaluate whether multiple users achieve consistent results using LANG on the same dataset. LANG is designed for the current data landscape where a wide range of users can be faced with an unfamiliar dataset, therefore this is an important consideration. Eight research students from an Australian university were recruited as participants of this experiment. All eight participants work in the database research area, with exposure to database management or programming experience. On the one hand, the participants' familiarity with database operations qualifies them for the testing of LANG, which guides the user through a series of SQL queries and manual observations in the database. On the other hand, the relative unfamiliarity of the participants with data quality concepts demands the LANG guidelines to be clear and definite.

The experiment dataset consisting of 2249 records was extracted from the original public transport smart card dataset. Data records from the original dataset with quality problems pertaining to each data quality dimension (with the exception of referential integrity, due to a single table design) were inserted into the extracted dataset and constituted 250 records. Further records were selected from one date in November 2012 to another date in April 2013 to guarantee complete journeys within the extracted records.

The experiment environment was deployed as a web application. The dataset was loaded into a MySQL database on the server so that participants would be able to connect to the database and query the dataset via a webpage. A pilot study was conducted with a fellow researcher with previous experience in data quality research. Based on an interview after the pilot study, adjustments were made to the presentation format to make information easier to find.

Following the pilot study, the eight participants were sent emails with the following documents: documentation of LANG; an SQL query template in a text file, which contained the template SQL queries in the approach that are ready-to-run once column names and table names are filled in by participants; the available minimal meta-data documentation; and a results document for recording the outcome of applying LANG on the dataset. In the email, participants were provided with instructions to connect to the database, follow the approach to investigate the dataset, and document their results. Particularly, they were required to complete the experiment independently, without discussing the experiment procedures with other participants. The independence of participants' application of LANG was guaranteed in that participants were divided into three groups, and applied the approach during three consecutive sessions and with the researchers' presence.

3.2.3 LANG Refinement

The soundness and repeatability evaluations were followed by an analysis of the collected data to assess the design artefact. Based on the analysis results, reflections were made on the limitations of LANG and refinements were proposed and implemented. For example, for format consistency, which usually involves specification of data format in the meta-data documentation accompanying the dataset, the empirical evaluations highlighted that a case where no meta-data is available should be considered, and guidelines were adjusted for detection of potential inconsistency problems. Similar changes were made for other dimensions including a number of minor changes to the improve clarity. The refined LANG approach is presented in Sect. 4.

3.2.4 LANG Generalizability

Following the refinement, we applied LANG to 20 open datasets to show LANG's generalizability. To reduce bias, we hired two researchers with only basic SQL experience, to apply LANG's SQL instantiation to, independently of each other discover the data quality problems in the datasets. Appendix B (available online via http://link.springer.com) provides a list of the datasets used. The refined LANG approach documentation and the generic SQL template (as provided to the participants in the repeatability evaluation) were provided and the researchers were guided to use MySQL as the environment for applying LANG.

Each researcher was asked to document for each data set, including which problems were found, and where, by highlighting the records and/or attributes. The documentation results of the two researchers were then compared. A sample of the results was checked by the authors to ensure correctness of applying LANG and the documentation guidelines.

All datasets were downloaded[3] from the respective open data portals to ensure that both researchers worked on the

[3] The download period is between June and August 2016. We note that the datasets are frequently updated in the respective open data portals including change of meta-data, such as adding or removing columns as well as providing or removing other documentation related to the dataset. Hence, the current versions of the datasets may not have the same data quality problems as those identified in our study.

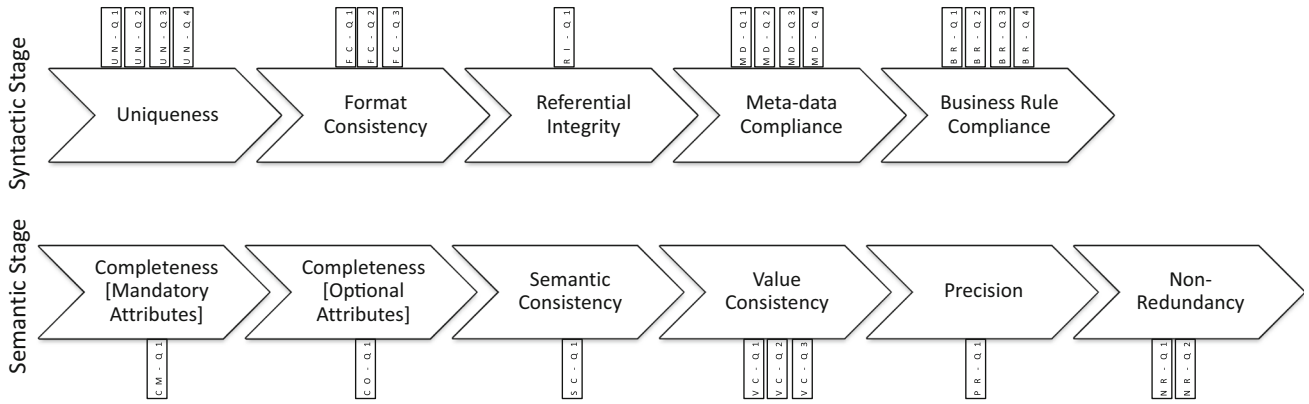

Fig. 4 LANG summary of stages and checks. Source: Authors

same version of the dataset. Since all datasets available from the open data portals consisted of a single table (provided for download as a.csv file), hence the dimension of referential integrity which required cross-table evaluation was not considered in this evaluation.

4 The LANG Approach

LANG takes a given dataset with minimum meta-data documentation as input and, through the below documented approach, produces a set of data quality problems. It consists of two core stages, viz. the *syntactic stage* which focuses on identification of data quality issues involving syntactic constraints on the data; and the *semantic stage* which focuses on detection of data quality issues associated with data semantics. LANG takes a variety of source data formats, such as a flat text file, tab delimited, comma separated values (csv) or an SQL export, as input. It can also be applied using a preferred data query language, such as SQL[4] or Python. LANG does not support high-dimensional data such as multimedia (images, videos) data, but is capable to support any form of data that can be exposed in a structured format beyond typical transactional data (customer, product, sales etc.), for example spatio-temporal or timeseries data, sensor/measurement data, and event/log/network data. Use of LANG is thus bounded by the structural format of the data and availability of minimum meta-data. This is further evidenced by the diversity of datasets used in the evaluation of LANG (see Sect. 5 and Appendix B).

LANG assumes that the most basic meta-data, i.e., the specification of a primary key (identifiers or mandatory attributes) is present or at least inferable. Similarly, value ranges for enumerated discrete values, continuous values, a type of value, etc. are also assumed to be present or inferable. Many open, or publicly available datasets contain such basic meta-data. For example, the Bristol Bus Stops dataset from data.gov.uk provides documentation to indicate that there are only three values possible for the attribute Shelter Pole or Lamp Post, namely 'Shelter', 'Pole' or Lamp Post'. In general, the approach can still be applied if the above documentation is not available, however the extent to which the documentation provides this or additional information (such as business rules) determines the extent to which the approach thoroughly discovers the quality of the data.

In the following sections we provide a high-level description of the two stages, and provide an instantiation of the approach to SQL in Appendix A (available online via http://link.springer.com). Figure 4 visually summarizes the two-stage approach, including further detail from Appendix A in relation to the number of checks (shown as rectangles, cross-referenced to the specific checks outlined in Appendix A) involved at each step, highlighting the structured and systematic nature of the approach.[5]

4.1 Syntactic Stage

The syntactic stage is concerned with the data's conformance to data quality dimensions of uniqueness, format consistency, referential integrity, meta-data compliance and business rule compliance. An outline of the investigation required to discover each data quality dimension is presented below. The procedure outlined corresponds to the requisite definition of the data quality dimension as

[4] In this paper we have demonstrated the application of LANG with the help of relational database (MySQL). We present the overall approach in the body of the paper, and present the SQL instantiation of the method in Appendix A.

[5] Some detail is abstracted in this figure for visual simplicity; in particular sequences between some of the individual checks, which may result in skipping certain checks/stages (as relevant on the basis of analysis results).

provided in Table 1. For detailed steps instantiated in SQL, please refer to Appendix A.

1. Uniqueness specifies that the data be uniquely identifiable. In a relational database uniqueness is ensured by enforcing primary key constraints. Inspection of uniqueness issues involves comparison of the count of all data records against the count of distinct values of a (existing or inferred) primary key attribute.
2. Format Consistency specifies that data formats be consistently used. Format (in)Consistency is detected by examining whether values consistently follow a standard format within a column. Investigation of Format Consistency issues involves comparison of the observed format of a data field against the format pattern specified or summarized for that field.
3. Referential Integrity specifies that data relationships be represented through referential integrity rules. Checking if Referential Integrity constraints are satisfied requires cross-table investigation. Investigation of Referential Integrity involves checking the correspondence between the (existing or inferred) foreign key column of the referencing table and the (existing or inferred) primary key column of the referenced table.
4. Meta-data Compliance specifies that data comply with its meta-data (where available). This means that data values should be consistent with the attribute definition according to meta-data. Investigation of Meta-data Compliance primarily involves inspection of domain constraints, i.e. compliance with the expected data types.
5. Business Rule Compliance specifies that data values conform to specified business rules. Investigation relating to this data quality dimension involves examination of the compatibility of attribute values with business rules (where available from the meta-data or provided documentation).

By the end of the syntactic stage, the user has completed the documentation of data quality problems found pertaining to the dimensions of *Uniqueness, Format Consistency, Referential Integrity, Meta-data Compliance and Business Rule Compliance*, and for each problem found, has identified relevant attributes and records.

4.2 Semantic Stage

The semantic stage investigation examines problems concerning the data quality dimensions of completeness, semantic consistency, value consistency, precision and redundancy, as outlined below. For detailed steps instantiated for SQL please refer to Appendix A.

1. Completeness of Mandatory Attributes specifies that attributes necessary for a complete representation of a real-world entity must contain values and cannot be null. Investigation relating to Completeness of Mandatory Attributes involves checking null values for all mandatory attributes in the dataset, identified explicitly in available meta-data or inferable.
2. Completeness of Optional Attributes pertains to non-mandatory attributes for which null values should be replaced with meaningful default values so that invalid cases of null values can be identified clearly. Investigation relating to Completeness of Optional Attributes involves identification of null values for all attributes that are not checked in the previous step.
3. Semantic Consistency specifies that the meaning of the data should be consistent across tables and datasets. Investigations relating to Semantic Consistency involve looking for fields in different tables with conflicting semantics. For example, a listed service for transport card retailers may be represented as ‘Change Expiry Date’ in one place and ‘Renewal’ in another, both of which may refer to the same service.
4. Value Consistency requires values be consistent – i.e. values cannot provide conflicting or heterogeneous instances. This consistency may be checked between records and between datasets. Investigation relating to Value Consistency involves comparison of data values between records, or against specified standard values, for example from a reference source (where available from provided meta-data documentation).
5. Precision specifies that data values should be correct to the right level of detail or granularity. Investigation relating to Precision involves checking for data values that are lexically, syntactically or semantically wrong, according to a stable reference source such as a dictionary or a set of domain norms.
6. Non-Redundancy specifies that only one record exist in a given data store that represents a single real-world entity. Investigations relating to Redundancy involve checking for entire duplicate records or duplication of all attributes except the identifying (key) attribute.

Similar to the syntactic stage, by the end of the semantic stage, the user has documented data quality problems found that relate to the data quality dimensions of *Completeness of Mandatory Value, Completeness of Optional Value, Semantic Consistency, Value Consistency, Precision, Non-Redundancy*, and identified the relevant attributes and records where the problem was found.

5 Evaluation

This section presents the evaluation results, refinement of LANG, and its generalizability evaluation. Section 5.1

presents the result of soundness evaluation of LANG through two focus group studies. Section 5.2 presents the outcome of repeatability evaluation of LANG. On the basis of these evaluations, LANG refinements are explained in Sect. 5.3, which is then followed by evaluation of generalizability through 20 open datasets.

5.1 Soundness Evaluation

This section outlines our findings relating to the validity of the data quality problems identified through applying LANG. The findings are based on focus groups conducted with IT practitioners and data custodians, as described in Sect. 3. Table 3 illustrates the validation results of the data quality problems in the Public Transport Smart Card dataset and the Online Learning dataset. Based on the focus group sessions, pertaining to the validated problems, various categories of causes of data quality were observed based on the data users' articulations. Those include: standardization, mechanisms of automatic data processing, data transformation, and manual error. The users' articulation of the problem causes by the users confirm that the participants understand the meaning of the data quality dimensions, as utilized by and central to LANG.

Table 3 provides a summary of the findings under three categories, viz. true positives, false positives and false negatives. *True Positives* refer to data quality problems found in the dataset with LANG and confirmed by the participants as genuine problems. *False Positives* refer to data quality problems found by LANG that participants indicated were not errors attributable to the data quality dimension. *False negatives* refer to data quality problems that were not identified by LANG but that were identified by participants as problems they face in their use of the data.

For the smart card data, ten out of eleven data quality problems identified through the application of LANG were validated as genuine through the focus group. One was found to be a (partial) false positive – the requirement that the 'passenger' column should only take a value of '1' because values of '0', '− 1', '2' are invalid. However, according to the focus group participants, while 'passenger' value of '2' is invalid, value '− 1' and '0' are compliant with their value derivation rule. Since the rule was not available to the researchers, the compliance of the values of '− 1' and '0' could not be confirmed.

For the Online Learning dataset, two out of the five data quality problems were identified as false positives. Discussion in the focus group indicated that no standard format was specified at the design stage of the system, therefore the value format varied significantly as the system evolved and expanded. Because of this, the inconsistent values discovered by LANG were considered valid due to the way the data was generated. Similarly, the identified redundancy problem was not considered to be a data quality problem per se, but redundancy was a result of the log generation system design.

The evaluation also identified 3 false negative data quality problems relating to business rule compliance in the smart card data. For example, one of the problems LANG did not discover was related to the bus route number having to be within a valid range. The major factor that hinders the approach from identifying those problems is the

Table 3 Data quality assessment results. Source: Authors' analysis

	Smart card data	Online learning data
# True positives		
Completeness of mandatory attributes	0	0
Completeness of optional attributes	1	0
Precision	0	0
Business rule compliance	2	0
Meta-data compliance	3	1
Uniqueness	0	1
Semantic consistency	0	0
Value consistency	1	0
Format consistency	3	1
Referential integrity[a]	0	0
# False positives		
Business rule compliance	1	0
Format consistency	0	1
Non-redundancy	0	1
# False negatives		
Business rule compliance	3	0

[a]Not expected as both datasets contained a single table

Table 4 Repeatability evaluation results. Source: Authors' analysis

Data quality dimension	Percentage of successful problem identifications
Completeness of mandatory attributes	100%
Completeness of optional attributes	100%
Precision	100%
Business rule compliance	100%
Meta-data compliance	100%
Uniqueness	100%
Non-redundancy	100%
Semantic consistency	100%
Value consistency	50%
Format consistency	62.5%
Referential integrity	N/A[a]

[a]Not applicable as the dataset contained a single table

inaccessibility of established business rules and a complete view of the captured data.

5.2 Repeatability Evaluation

Two criteria were used to study the repeatability of the approach – the successful identification rate of each quality problem found, and the overall identification rate of all problems within a data quality dimension (shown in Table 4). These two criteria serve to verify that the approach facilitates the identification of all types of data quality problems and that it produces consistent outcomes from multiple users. Eight participants were involved in this evaluation. Based on the experiment, the percentage of participants who successfully identify data quality problems pertaining to each of the quality dimensions is presented below.

All eight participants successfully and consistently identified data quality problems pertaining to Completeness for Optional Attributes, Business Rule Compliance, and Meta-data Compliance. No false problems were detected for the above or for any of the dimensions including Completeness of Mandatory Attributes, Precision, Uniqueness, Semantic Consistency and Non-Redundancy, resulting in 100% identification. All participants were also able to observe the existence of the Format Consistency problem. However, there were three format consistency problems in the experiment dataset (see Table 3). One of these, related to inconsistency in stop formats, was identified by all participants, whereas the two others were not identified by all of the eight participants, resulting in 15/24 successful identifications and an overall success rate of 62.5%. For Value Consistency, only half of the participants successfully documented the problem of the stop value inconsistency, which manifests itself in many forms. Since the experiment dataset consisted of a single table, referential integrity was not evaluated.

The post experiment interviews revealed that the participants generally felt their familiarity with regular expressions was inadequate for using complex regular expressions in SQL queries (e.g., for representing syntactically complex stops). This limitation partly explains the relatively low identification rate for Format Consistency and Value Consistency issues. This observation had the dual consequence of identifying minimal background knowledge for users of LANG while also helping refine the Format Consistency steps with further details on how to build regular expressions. The low success rate in identification of data quality problems regarding Value Consistency also exposed a limitation of LANG pertaining to the case where even the basic meta-data is lacking. The identification resulted in a refinement of the steps relating to Value Consistency.

5.3 LANG Refinement

The two sets of evaluations helped us refine LANG in three ways. First, specific instructions in the guidelines for Format and Value Consistency were adjusted to add further guidance and improve clarity on what needs to be done. Second, the evaluations allowed us to get clarity on the minimum expectations of knowledge for LANG users. Third, the implications of lack of requisite meta-data have become evident. We observed three levels of meta-data needs. Notably, problems pertaining to 6 data quality dimensions require minimal meta-data, which is mostly, if not always available, such as attribute names. With additional meta-data, LANG extends coverage to two additional dimensions, namely Precision and Value Consistency which may require further information, such as desired number of decimal points for a numeric attribute. The remaining three dimensions, namely Referential Integrity, Business Rule Compliance and Semantic Consistency generally require additional documentation to afford full opportunity for data quality problem detection.

Because LANG is framed within semiotics, it relates to the syntactic and semiotic levels – or user-independent levels of the semiotic theory (for which stipulation of use specific domain knowledge is minimal). However, the pragmatic level of the semiotic framework (Price and Shanks 2004, 2005a, b), which relates to the interpretation of data, is closely related to the realization of data's value. Therefore, integration of user needs, e.g., prioritization of data quality dimensions, into the data quality assessment is critical to the success of data quality management. Validation of the identified data quality problems is only a first step towards the pragmatic level of the approach, and needs the convergence of data quality investigation with user needs. Further, although LANG is shown to produce a valid set of data quality problems, how the results can be used to clean and improve data quality of the source datasets requires further research.

5.4 LANG Generalizability

LANG provides a guided approach to discover data quality problems in datasets for which users have minimal control or knowledge of underlying features and rules. Open data is a clear example of such repurposing of data. In this section we demonstrate the application of LANG to 20 Open Datasets sourced from data.gov portals. In Table 5 we present the results of our findings. The complete list of datasets, together with their descriptions, is provided in Appendix B. Each X indicates that problems relating to the particular dimension were found in the dataset.

Collectively, the evaluated datasets constitute an examination of over 185 thousand records, with an average number of 7 columns, over 1.2 million error opportunities. Additionally, the datasets examined were sourced from a diverse number of domains and included data relating to geo-codes, dates, categorical and coded data, location and address data, and measurement and numeric data. The extensive evaluation allowed us to review the potential and limitations of LANG generalizability at a fine level of detail.

Consistent with the results from the soundness and repeatability evaluations, we also observed in the generalizability evaluation, difficulties in discovery of problems relating to Precision and Value Consistency, which often required further information that was lacking in the dataset documentation. This problem was more significant for Business Rule Compliance and Semantic Consistency, for which there was an even higher need for additional documentation, e.g., data constraints, or business rules.

To further analyze the generalizability of LANG, we investigated if the guidelines for each of the 11 data quality dimensions have the potential to be automated. We developed a proof of concept of such a tool (Almars 2016).

Table 5 Data quality problem discovery for open data using LANG. Source: Authors' analysis

Datasets	CM	CO	PR	BR	MD	UN	NR	SC	VC	FC
1	X	X			X				X	X
2		X			X					X
3		X								X
4		X							X	X
5		X								
6										X
7		X								X
8		X								X
9										
10										
11		X			X					
12			X		X	X	X			X
13		X								X
14		X			X	X			X	X
15		X	X		X	X			X	X
16	X	X			X	X			X	X
17		X								X
18		X					X		X	X
19	X	X			X					
20	X	X			X	X	X			X

CM completeness of mandatory attributes, *CO* completeness of optional attributes, *PR* precision, *BR* business rule compliance, *MD* meta-data compliance, *UN* uniqueness, *NR* non-redundancy, *SC* semantic consistency, *VC* value consistency, *FC* format consistency

A brief usability study conducted on the tool indicated that, given the right background of the users (e.g., familiarity with writing SQL using regular expressions' conditions), the manual and automated approaches had close to 100% similar results. We further observed that the automated tool does increase the efficiency (time taken) of finding the data quality problems, especially for large datasets, but unsurprisingly does not significantly change the accuracy of the results. This indicates that LANG is sufficiently clear to be programmable as well as interpretable by humans.

6 Discussion

Through our evaluations, we were able to consider the utility and the limitations of the LANG approach. We conclude that the validity and completeness of the data quality problems discovered through LANG is evident. The soundness, repeatability and generalizability evaluations have demonstrated that in settings where structured datasets are repurposed and disconnected from the original use and creation, LANG provides an effective and consistent means of discovering a variety of data quality problems with minimal meta-data, thus addressing the critical gap in

the current body of knowledge (Clarke 2016) and advancing mostly manual and ad-hoc practice (Hey and Trefethen 2003).

Specifically, LANG provides the capability to discover data quality problems relating to eleven dimensions. In our study we observed that out of these, problems relating to seven dimensions, namely Completeness of Mandatory Attributes, Completeness of Optional Attributes, Meta-data Compliance, Uniqueness, Non-Redundancy, Format Consistency and Referential Integrity (in datasets with more than one table), were discovered with minimal meta-data. With additional information on value types and ranges, Precision and Value Consistency could also be discovered. Business Rule Compliance and Semantic Consistency required information relating to domain specific rules and semantic constraints, which was less evident in open (repurposed) data settings.

While the evaluations demonstrate the utility of LANG, they do highlight, however, that missing meta-data leads to false negative or false positive data quality problem identification, thus delineating the scope of LANG's capabilities and highlighting the requirement for some meta-data presence. In cases where meta-data is entirely missing, or cases with complex schemas, generally more advanced methods are needed, many of which have been individually studied in prior research. For example, if primary keys (or any form of unique record identifiers) are not specified and it is difficult to infer possible primary key candidates from the meta-data, more advanced approaches may be needed. Köhler et al.'s (2013) method for discovery of possible and certain keys of a dataset based on a semantic perception of entity integrity constraint could be used in such cases. Alternatively, Sismanis et al.'s (2006) algorithm for efficient and scalable discovery of keys in large datasets could be used. Similarly, advanced techniques for duplicate detection exist, for example Hernández and Stolfo (1998) uses a rule-based knowledge base to implement an equational theory for resolving duplicates. Such advanced techniques also exist for complex format patterns that are not easily detected through Regex mechanisms, for example Bohannon et al. (2007) propose methods that can be used to detect violations to expected formats of data specified in the form of constraints. Advanced methods also exist for detection of non-compliance to meta data, where Raman and Hellerstein (2001) propose an approach for detecting domain discrepancies and structural conflicts.

We note that the value of LANG was realized in the guidance it provides with respect to what data quality considerations need to be factored in the analysis of a repurposed and unfamiliar new dataset, rather than tool support or automated techniques for complex analysis of individual problems. Indeed, we observe that the struggles of data scientists who spend upwards of 80% effort in data science projects in understanding the fitness of the data (Belkin and Patil 2013) have persisted despite a plethora of data profiling tools. In contrast, the mapping of the semiotic framework against a consolidated and well-defined set of data quality dimensions equip LANG with a strong foundation upon which data scientists can reliably build their strategy for objectively and transparently assessing the quality and fitness of a repurposed dataset for their intended use. To the best of our knowledge such a guided approach does not exist in research or practice, which prohibits us to do a comparative study, but at the same time highlights the novelty of LANG as an approach for bottom-up identification of data quality problems in repurposed datasets.

Finally, we share some lessons learnt in the application of design science research towards the development of LANG. The lack of methodological guidance for evaluation of Design Science artefacts has been lamented by researchers in the past, and several studies provide guidelines and principles (e.g., Venable et al. 2012; Sonnenberg and vom Brocke 2012). Similarly to Sturm and Sunyaev (2019), we found it beneficial to combine advice from a variety of Design Science research articles to guide our design process and also evaluation. However, while we have been guided by these studies, we noted a lack of guidance of what is *sufficient* in terms of evaluation as opposed to how to methodologically conduct the evaluation. In our case, we found that the soundness and generalizability evaluations alone were not sufficient, as we were also motivated to show that data scientists can use the method to achieve consistent results. Consequently, we see an opportunity for a better articulation of methodological guidance dependent on not just the type of artefact developed but also on its purpose.

7 Conclusions

This paper is motivated by the lack of systematic approaches that data scientists can use to evaluate the quality of repurposed datasets – unfamiliar datasets with minimal meta-data. Following a Design Science approach, and using semiotic theory and data quality dimensions as our theoretical foundations, we present LANG – a novel approach for discovering data quality problems to address this gap in the body of literature and in practice.

LANG consists of two stages relating to syntactic and semantic aspects of data quality and can be instantiated into a query language of choice. The two stages provide guided steps that allow eleven data quality problems to be discovered. The identified problems relate to data quality dimensions that are independent of the use context – a setting that is exceedingly common due to repurposing of

datasets. By following the steps in the approach and identifying a set of data quality problems, data scientists will be able to assess the quality of the data and hence the implications of its use. While most existing data quality methodologies are user-centric and top-down in nature, requiring deep domain knowledge and investment in the understanding of the use context, LANG empowers users to explore an arbitrary dataset with minimum documentation. The three evaluations of LANG to assess its soundness (ability to discover genuine problems), repeatability (ability to consistently discover the same problems) and generalizability (ability to discover problems for any arbitrary dataset), demonstrate that LANG is effective in its ability to discover data quality problems in repurposed datasets.

We acknowledge, however, that our work has some limitations. First, LANG was developed based on one dataset with a single table (representing a large join of multiple tables from the underlying relational database). Although the application of LANG on schemas with multiple tables has been undertaken without issue, further investigation may be needed in the application of LANG where there are complex relationships between tables. Second, LANG is able to facilitate identification of problems belonging to a subset of all known data quality dimensions. While it is comprehensive in its coverage in terms of dimensions that are independent of the use context, further investigations into the remaining dimensions will extend the scope of LANG into the pragmatic level of the semiotic framework.

In our immediate future work, we are investigating an instantiation of LANG using Jupyter notebooks[6] and Python programming. We are further aiming to study the impact of LANG on productivity of data scientists when tackling repurposed data sets with minimal meta-data. Lastly, we hope to extend and improve the number and effectiveness of dimensions LANG can currently handle through a deeper analysis of the relationship between meta-data and capacity to detect data quality problems.

References

Abedjan Z, Golab L, Naumann F (2015) Profiling relational data: a survey. VLDB J Int J Very Large Data Bases 24(4):557–581

Almars A (2016) Automated data quality discovery tool. Master Thesis, The University of Queensland

Batini C, Scannapieco M (2006) Data quality—concepts, methodologies and techniques. Springer, Heidelberg

Batini C, Francalanci C, Cappiello C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Comput Surv 41(3):1–52

Belkin R, Patil D (2013) Everything we wish we'd known about building data products. http://firstround.com/review/everything-we-wish-wed-known-about-building-data-products/. Accessed 14 Nov 2018

Bohannon P, Fan W, Geerts F, Jia X, Kementsietsidis A (2007) Conditional functional dependencies for data cleaning. In: IEEE 23rd international conference on data engineering, pp 746–755

Byrne B, Kling J, Mccarty D, Sauter G, Smith H, Worcester P (2008) The information perspective of SOA design, part 6: the value of applying the data quality analysis pattern in SOA. IBM Corporation

Caballero I, Verbo E, Calero C, Piattini M (2007) A data quality measurement information model based on ISO/IEC 15939. In: Proceedings of the 12th international conference on information quality, pp 393–408

Caballero I, Verbo E, Calero C, Piattini M (2008) MMPRO: a methodology based on ISO/IEC 15939 to draw up data quality measurement processes. In: Proceedings of the 13th international conference on information quality, pp 326–340

Chakraborti S, Dey S (2019) Analysis of competitor intelligence in the era of big data. Bus Inf Syst Eng 61(3):345–355

Clarke R (2016) Big data, big risks. Inf Syst J 26(1):77–90

Corsar D, Edwards P (2017) Challenges of open data qality: more than just license, format, and customer support. ACM J Data Inf Qual 9(1):3:1–3:4

Dasu T, Johnson T (2003) Exploratory data mining and data cleaning. Wiley, New York

Duus R, Cooray M (2016) The future will be built on open data—here's why. http://theconversation.com/the-future-will-be-built-on-open-data-heres-why-52785. Accessed 14 Nov 2018

Ehling M, Körner T (2007) Handbook on data quality assessment methods and tools. European Commission, Eurostat

Elbaz G (2012) Data markets: the emerging data economy. http://techcrunch.com/2012/09/30/data-markets-the-emerging-data-economy/. Accessed 14 Nov 2018

English LP (1999) Improving data warehouse and business information quality. Wiley

English LP (2009) Information quality applied. Best practices for improving business information, processes and systems. Wiley, New York

Eppler MJ (2001) The concept of information quality. Stud Commun Sci 1(2):167–182

Fan W, Geerts F (2012) Foundations of data quality management. Synth Lect Data Manag 4(5):1–217

Fisher T (2009) The data asset: how smart companies govern their data for business success. Wiley, New York

Gatling GCBR, Champlin R, Stefani H, Weigel G (2007) Enterprise information management with SAP. Galileo, Boston

Gregor S, Jones D (2007) The anatomy of a design theory. J Assoc Inf Syst 8(5):312–335

Hernández MA, Stolfo SJ (1998) Real-world data is dirty. Data cleansing and the merge/purge problem. Data Min Knowl Discov 2(1):9–37

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105

Hey AJG, Trefethen AE (2003) The data deluge. An e-science perspective. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf. Accessed 3 July 2019, pp 809–824

HIQA (2011) International review of data quality. Health Information and Quality Authority (HIQA), Ireland. http://www.hiqa.ie/

[6] "The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more." (jupyter.org).

press-release/2011-04-28-international-review-data-quality. Accessed 2 Oct 2017
ISO (2011) ISO/TS 8000-1 data quality part 1: overview. ISO
ISO (2012) ISO 8000-2 data quality-part 2-vocabulary. ISO
Jayawardene V, Sadiq S, Indulska M (2013a) An analysis of data quality dimensions. School of Information Technology and Electrical Engineering, The University of Queensland, ITEE Technical Report
Jayawardene V, Sadiq S, Indulska M (2013b) The curse of dimensionality in data quality. In: 24th Australasian conference on information systems. RMIT University, pp 1–11
Judah S, Friedman T (2015) Magic quadrant for data quality tools. Gartner
Kenett RS, Shmueli G (2014) On information quality. J R Stat Soc Ser A 177(1):3–38
Kim J, Hausenblas M (2012) 5 * Open Data. https://5stardata.info/en/. Accessed 14 Nov 2018
Köhler H, Leck U, Link S (2013) Possible and certain SQL keys. Department of Computer Science, The University of Auckland
Köhler H, Link S, Zhou X (2015) Possible and certain SQL keys. Proc VLDB Endow 8(11):1118–1129
Krogstie J (2002) A semiotic approach to quality in requirements specifications. In: Proceedings of the IFIP TC8/WG8 (1), pp 231–249
Krogstie J, Lindland OI, Sindre G (1995a) Defining quality aspects for conceptual models. In: Falkenberg ED, Hesse W, Olivé A (eds) Information system concepts. Springer, Boston, pp 216–231
Krogstie J, Lindland OI, Sindre G (1995b) Towards a deeper understanding of quality in requirements engineering. In: International conference on advanced information systems engineering. Springer, Heidelberg, pp 82–95
Krueger R, Casey M (1994) Focus groups. A practical guide for applied research. Sage Publications, Thousand Oaks
Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inf Manag 40(2):133–146
Lindland OI, Sindre G, Solvberg A (1994) Understanding quality in conceptual modeling. IEEE Softw 11(2):42–49
Loshin D (2001) Enterprise knowledge management. The data quality approach. Morgan Kaufmann, Burlington
Loshin D (2006) Monitoring data quality performance using data quality metrics. Informatica Corporation, Redwood City
Maydanchik A (2007) Data quality assessment. Technics Publications, New Jersey
McGilvray D (2008) Executing data quality projects: ten steps to quality data and trusted information. Morgan Kaufmann, Burlington
Morgan DL (ed) (1993) Sage focus editions. Successful focus groups: advancing the state of the art, vol 156. Sage Publications, Thousand Oaks
Morris CW (1938) Foundations of the theory of signs. In: Langford CH (ed) International encyclopedia of unified science. University of Chicago Press, London
Naumann F, Rolker C (2000) Assessment methods for information quality criteria. Humboldt-Universität zu Berlin, Informatik-Berichte, Berlin
OMB U (2002) Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies, part IX. Office of Management and Budget
Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. J Manag Inf Syst 24(3):45–77
Pierce CS (1931–1935) Collected papers. Harvard University Press, Cambridge
Pipino L, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45(4):211–218
Powell RA, Single HM (1996) Focus groups. Int J Qual Health Care 8:499–504. https://doi.org/10.1093/intqhc/8.5.499
Prat N (2019) Augmented analytics. Bus Inf Syst Eng 61(3):375–380
Price R, Shanks G (2004) A semiotic information quality framework. In: Proceedings of the international conference on decision support systems, pp 658–672
Price R, Shanks G (2005a) A semiotic information quality framework: development and comparative analysis. J Inf Technol 20(2):88–102
Price R. J, Shanks G (2005b) Empirical refinement of a semiotic information quality framework. In: Proceedings of the 38th annual Hawaii international conference on system sciences, Big Island, pp 216a
Raman V, Hellerstein JM (2001) Potter's wheel: an interactive data cleaning system. In: Proceedings of the 27th VLDB conference, Rome, pp 381–390
Rosemann M, Vessey I (2008) Toward improving the relevance of information systems research to practice: the role of applicability checks. MIS Q 32(1):1–22
Sadiq S, Indulska M (2017) Open data: quality over quantity. Int J Inf Manag 37(3):150–154
Sadiq S, Yeganeh NK, Indulska M (2011) 20 years of data quality research: themes, trends and synergies. In: 22nd Australasian database conference, Perth, pp 153–162
Scannapieco M, Virgillito A, Marchetti C, Mecella M, Baldoni R (2004) The Daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. Inf Syst 29(7):551–582
Selvage M, Saul J, Jain A (2017) Magic quadrant for data quality tools. Gartner
Shanks GG, Darke P (1998) Understanding data quality and data warehousing: a semiotic approach. IQ, pp 292–309
Shanks G, Tansley E (2002) Data quality tagging and decision outcomes. An experimental study. IFIP Working Group, pp 399–410
Sismanis Y, Brown P, Haas PJ, Reinwald B (2006) Gordian: efficient and scalable discovery of composite keys. In: Proceedings of the 32nd international conference on very large data bases, VLDB Endowment, pp 691–702
Song S, Chen L (2011) Differential dependencies Reasoning and discovery. ACM Trans Database Syst 36(3):16
Sonnenberg C, vom Brocke J (2012) Evaluations in the science of the artificial. Reconsidering the build-evaluate pattern in design science research. In: Peffers K, Rothenberger M, Kuechler B (eds) Design science research in information systems, vol 7286. Advances in theory and practice. DESRIST. Lecture notes in computer science. Springer, Heidelberg
Stamper RK (1992) Review of Andersen "Theory of Computer Semiotics". Comput J 1
Stamper R (1993) A semiotic theory of information and information systems/applied semiotics. In: Invited Papers for the ICL/University of Newcastle Seminar on "Information", September 6–10
Storey V, Wang R (2001) Extending the ER model to represent data quality requirements. Kluwer, Dordrecht
Sturm B, Sunyaev A (2019) Design principles for systematic search systems. Bus Inf Syst Eng 61(1):91–111
Stvilia B, Gasser L, Twidale MB, Smith LC (2007) A framework for information quality assessment. J Am Soc Inf Sci Technol 58(12):1720–1733
Tu SY, Wang Y-YR (1993) Modeling data quality and context through extension of the ER model. Total Data Quality Management Research Program, Sloan School of Management, Massachusetts Institute of Technology, Cambridge
Venable J, Pries-Heje J, Baskerville R (2012) A comprehensive framework for evaluation in design science research. In: Peffers

K, Rothenberger M, Kuechler B (eds) Design science research in information systems, vol 786. Advances in theory and practice. Springer, Heidelberg, pp 423–438

Venable J, Pries-Heje J, Baskerville R (2016) FEDS: a framework for evaluation in design science research. Eur J Inf Syst 25(1):77–89

Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. Commun ACM 39(11):86–95

Wang R (1998) A product perspective on total data quality management. Commun ACM 41(2):58–65

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12(4):5–33

Wang R, Ziad M, Lee Y (2001) Data quality. Kluwer, Dordrecht

Zhang R, Jayawardene V, Indulska M, Sadiq S, Zhou X (2014) A data driven approach for discovering data quality requirements. In: 35th international conference on information systems, Auckland