A Data Driven Approach for Discovering Data Quality Requirements

Submission Type: Research-in-Progress

Ruojing Zhang  
School of Information Technology and Electrical Engineering  
The University of Queensland  
Brisbane St Lucia, QLD 4072, Australia  
r.zhang3@uq.edu.au

Vimukthi Jayawardene  
School of Information Technology and Electrical Engineering  
The University of Queensland  
Brisbane St Lucia, QLD 4072, Australia  
w.jayawardene@uq.edu.au

Marta Indulska  
UQ Business School  
The University of Queensland  
Brisbane St Lucia, QLD 4072, Australia  
m.indulska@business.uq.edu.au

Shazia Sadiq  
School of Information Technology and Electrical Engineering  
The University of Queensland  
Brisbane St Lucia, QLD 4072, Australia  
shazia@itee.uq.edu.au

Xiaofang Zhou  
School of Information Technology and Electrical Engineering  
The University of Queensland  
Brisbane St Lucia, QLD 4072, Australia  
zxf@itee.uq.edu.au

Abstract

Existing methodologies for identifying data quality issues are inevitably user-centric, wherein data quality requirements are determined in a top-down manner following organizational structures and data governance frameworks. In the current data landscape, however, users are often confronted with new, unexplored data sets that may have relevance and potential to create value. In such scenarios applying top-down approaches is not feasible. Users need to be empowered with data exploration capabilities that allow them to investigate and understand the quality of data sets and, subsequently, the implications for use. The question is to what extent can the quality of a data set be explored in a bottom up manner without access to well defined data quality measures. Accordingly, in this paper we present an approach for discovering data quality issues using generic exploratory methods, which we derived through experimentation with a real data set based on public transport.

Keywords: Data quality, Database management systems (DBMS), Evaluation methods and criteria, Requirements analysis
Introduction

Discovering the quality of a data set is a fundamental task in most, if not all, data quality improvement projects (Batini et al. 2009). Because quality of data is assessed against a certain stated requirement (English 2009; ISO 2011; Loshin 2001), most data quality assessment approaches are user centered, and hence also ‘top-down’ following organizational structures and data governance frameworks. Here, identifying the requirements comes first, followed by defining metrics to measure the data quality and, finally, exploring the actual data sets based on the defined metrics.

In the current data landscape, however, it can be observed that users are often confronted with new, unexplored data sets that, presumably, have relevance and potential to create value. In such scenarios applying top-down approaches is not feasible. Users need data exploration capabilities that will allow them to investigate the quality of the data sets and, subsequently, the implications of their use. The question is to what extent one can explore the quality of a data set in a bottom up manner without having the knowledge of well-defined data quality requirements. There are two existing areas where data-driven methods have been considered for data quality assessment – data exploration and data profiling.

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

Data profiling is a related concept to data exploration and has a significant commercial tool market. Gartner (Friedman 2013) estimates that this market reached $960 million in revenue by the end of 2012. Approximately 50% of the market is dominated by a few large and well-established vendors, such as IBM, Informatica, Pitney Bowes, SAP and SAS. The remaining 50% is divided among a large number of providers, including Microsoft, Oracle, Talend, Ataccama and 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, functional dependency analysis, column correlation analysis, validity checks etc. Such tools are generally not accompanied by guidelines on how the profiling reports can be used for identification of actionable data quality requirements.

To the best of our knowledge, there is no generic bottom-up data driven approach that can be used on an arbitrary data set to assess data quality and, consequently, identify data quality requirements. Accordingly, to address this gap in the current body of data quality management knowledge, this paper is focused on developing such an approach. Figure 1 below contrasts the two approaches, viz. top-down and bottom-up. In the top-down approach, derivation of user requirements comes first, and prescribes the metrics for the later stages of assessment and improvement. We argue that a bottom-up approach has the potential to discover (at least part of) the data quality requirements using exploratory and/or analytical methods. The bottom-up approach allows requirements to be dynamically discovered and adapted as the use and understanding of the data set expands, which can be instrumental in creating value from large and unfamiliar data sets.

In our work, we have endeavored to develop such an approach that can provide a generic set of guidelines for data driven discovery of data quality requirements. Through a systematic series of experiments with a real data set based on public transport, we are able to inform and derive an approach using user independent, generic exploratory methods. In the sections that follow we first present a brief background on data quality assessment and requirements modeling. We then present the methodological and theoretical underpinnings of the developed approach. The details of the approach accompanied by illustrations from an example are then provided, followed by a discussion on the current limitations and future research.
A Data Quality Requirements Discovery Approach

Background

Batini et al. (2009), provide a comprehensive analysis of existing approaches for data quality assessment and requirements identification, indicating that such approaches typically include three core aspects, viz. data and process analysis, data quality requirements analysis, and, data quality analysis. Data and process analysis includes examination of data schemas, performing interviews, and meetings with data users to reach a complete understanding of data, related constraints and rules, and processes creating or consuming the data. Data quality requirements analysis often includes surveys of data users and administrators to identify quality issues, with an aim to identify critical data sets, define data quality metrics, and set quality targets. Data quality analysis then pertains to activities related to data sets exploration, assessment and profiling against the defined data quality metrics.

Notable contributions to data quality assessment and requirements identification include (Lee et al. 2002), who propose 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. Similarly, Naumann and Rolker (2000), present 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. For example, the criteria ‘interpretability’ and ‘concise-representation’ are both assigned the assessment method of ‘user sampling’. While concise representation is constrained by business rules in certain application contexts, the degree of interpretability of data is subject to the individual user’s perception. It is evident that most, if not all, of these approaches follow a user–centered, top-down approach, where requirements are solicited from users before the data is explored.

Another closely related work is InfoQ, a recently proposed top-down approach for information quality assessment (Kenett and Shmueli 2014). InfoQ evaluates its components, i.e. data quality and analysis quality, using statistical methods in terms of their potential to achieve the elicited goal. It focuses on the utility of the data, is entirely goal(user)-centric and defines eight assessment dimensions from a statistical and high level perspective. InfoQ thus relates closely to the subjective (user-dependent) aspects of data quality rather than declarative (user-independent) aspects. While most InfoQ dimensions can only partially be mapped to existing, widely recognized data quality dimensions, i.e. InfoQ dimension 1 data resolution (which describes ‘measurement scale’ and ‘aggregation level’ of the data), maps to accuracy (precision and granularity aspects of the data); InfoQ dimension 2 data structure, the definitive descriptions ‘data characteristics’ and ‘wrong data’ relate to several dimensions including completeness, consistency, accuracy and validity; InfoQ dimension 3 data integration, maps to consistency and availability; InfoQ dimension 4 temporal relevance, maps to volatility and currency; and InfoQ dimension 5 chronology of data and goal partly relates to availability of data. The remaining dimensions pertain to aspects beyond data quality dimensions e.g. InfoQ dimension 6 generalizability, 7 construct operationalization, 8 communication and part of 5 chronology of data and goal relate to generalizability, experiment design and communication aspects. Meanwhile, data quality dimensions (Wang and Strong 1996) are a central notion in defining data quality requirements. Typically data quality requirements are
documented following a hierarchy of data quality dimensions and associated metrics. While many disparate classifications of data quality dimensions have proliferated over the years, we rely on the work of Jayawardene et al. (2013a, 2013b) who have considered and consolidated a large pool of academic and practitioner data quality dimensions, resulting in eight dimensions, each with several themes. The eight dimensions are, completeness (existence and adequacy of data in an application context), availability and accessibility (ease of use, maintainability and control of the data from end users’ perspective), currency (time sensitivity of data applicable in data usage), accuracy (correspondence to real world object or reference source, to a right level of granularity), validity (compatibility of data with metadata, business rules, definition and conventions), reliability and credibility (trust worthiness and relevance of data from end users’ point of view), consistency (no data anomalies within single or multiple datasets), usability and interpretability (ease of interpretation and applicability of data for use).

Methodology

Our aim in this paper is to develop an artifact – a data quality requirements identification approach. Given our focus on developing an IT artifact, the overarching methodology approach that governs our research is that of Design Science (Hevner et al. 2004). The Design Science approach has been widely used in Information Systems research that focuses on creating artifacts (solutions, which can be in the form of methods, guidelines, frameworks, etc.) and is a ‘received view of the IS field’s conception’ of design research (Venable 2010). Important aspects of Design Science include the investigation of the problem space to inform rigorous and relevant artifact design, as well as artifact evaluation, through a variety of research methods, to prove its utility in the real world (an aspect of our study that is currently in progress).

The development of our artifact is based on observations and investigations performed on a public transport trajectory data provided by Translink (translink.com.au) between November 2012 and April 2013, namely goCard data. The goCard dataset, consisting of 16.9GB, 69572902 records, was made available with minimal documentation and hence was ideal for use in the development of the approach. The documentation received for the goCard data defines the level of minimum information required and includes meaningful table and column names and the domain of columns. Primary keys and foreign keys though not provided for the goCard data were easily determined through reverse engineering and are considered part of the minimum requirements. The proposed approach inspects data through a combination of systematic manual observations and SQL queries as the means of practical data quality exploration and discovery, making the approach easy to understand and conduct.

We frame our approach within modern semiotics (Morris 1938; Pierce 1931-1935), wherein three semiotic levels are studied: syntactic, semantic and pragmatic levels, that, respectively 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, 2005; Shanks and Tansley 2002; Shanks and Darke 1998). Therefore a correspondence between semiotics and data quality can be observed.

Price and Shanks (2004, 2005) 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, 2005). 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, 2005) derive a set of criteria for data quality for each semiotic level as follows (pragmatic level is omitted as it is outside the scope of a user-independent approach for discovering data quality):

**Syntactic level criteria:**

1. Conformance to Metadata: Data follows specified integrity rules.
Semantic level criteria:
1. Mapped completely: Every external phenomenon is represented.
2. Mapped unambiguously: Each identifiable data unit represents at most one specific external phenomenon.
3. Phenomena mapped correctly: Each identifiable data unit maps to the correct external phenomenon.
4. Properties mapped correctly: Non-identifying attribute values in an identifiable data unit match the property values for the represented external phenomenon.
5. Mapped consistently: Each external phenomenon is either represented by at most one identifiable data unit or by multiple identifiable units whose inconsistencies are resolved within an acceptable time frame.
6. Mapped meaningfully: Each identifiable data unit represents at least one external state.

Although the developed criteria are theoretically sound and shown to be internally coherent (Price and Shanks 2005) they lack an explicit pathway to identification of data quality requirements that are widely defined using the notion of data quality dimensions and associated measures. Without the clearly defined dimensions, it is not possible to perform a data driven discovery. We rely on the recent consolidation of data quality dimensions by Jayawardene et al. (2013a, 2013b) and the resulting definitions, and map them to corresponding criteria above. The mapping was conducted independently by two researchers who each mapped, based on definitions, every information quality criterion derived from the respective semiotic level Price and Shanks (2004, 2005) to one or more data quality dimensions (Jayawardene et al. 2013a; 2013b). The results of the mapping are presented in Table 1.

<table>
<thead>
<tr>
<th>Syntactic Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency</td>
</tr>
<tr>
<td>1. Uniqueness: Data stores should not have duplicate data records within a dataset. (Based on the definitions provided in (Byrne et al. 2008; English 2009; G. Gatling 2007; Loshin 2006; McGilvray 2008))</td>
</tr>
<tr>
<td>2. Structural consistency: The extent to which similar attributes or elements of an information object are consistently represented using the same structure, format and precision. (Based on the definitions provided in (HIQA 2011; McGilvray 2008; Stvilia et al. 2007))</td>
</tr>
<tr>
<td>3. Referential Integrity: Data instances have to be in accordance with database referential integrity constraint. (Based on the definitions provided in (English 2009; Loshin 2006))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
</tr>
<tr>
<td>4. Null Values: A missing value, because it is unavailable or inapplicable, might or might not be valid for a certain attribute. (Based on the definitions provided in (Loshin 2006; Redman 1997))</td>
</tr>
<tr>
<td>5. Record existence: System is not missing any entity which are relevant to the organization (Based on the definitions provided in (English 2009; Price and Shanks 2005))</td>
</tr>
<tr>
<td>6. Fact completeness: Certain attributes of a record should consist of values to make the record usable. (Based on the definitions provided in (Byrne et al. 2008; English 2009; G. Gatling 2007; HIQA 2011; Loshin 2006))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Mapped consistency: Every data unit has only one corresponding real world object. (Based on the definitions provided in (McGilvray 2008; Price and Shanks 2004))</td>
</tr>
<tr>
<td>8. Semantic consistency: Meanings of the data are consistent across data sets and applications. (Based on the definitions provided in (Byrne et al. 2008; English 2009; G. Gatling 2007; Loshin 2006; Redman 1997))</td>
</tr>
</tbody>
</table>
Value consistency: Data values are consistent between records and between different data sets. (Based on the definitions provided in (Byrne et al. 2008; G. Gatling 2007; Loshin 2006; Redman 1997))

Equivalence of Redundancy: Data sets should not have redundant data. (Based on the definitions provided in (English 2009; McGilvray 2008))

<table>
<thead>
<tr>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy to reference source: The data agrees with a reliable source of information. (Based on the definitions provided in (HIQA 2011; Loshin 2006; Lyon 2008; McGilvray 2008; Stvilia et al. 2007))</td>
</tr>
<tr>
<td>Conciseness: The data values should be accurate lexically, syntactically, semantically and to the right level of detail or granularity</td>
</tr>
<tr>
<td>Accuracy to reality: Every entity instance in the database has one and only one corresponding real world object which can be identified meaningfully. (Based on the definitions provided in (English 2009; Price and Shanks 2005))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformance to Metadata: Data values are legitimate according to their defined domain. (Based on the definitions provided in (Byrne et al. 2008; English 2009; Loshin 2001; Price and Shanks 2005; Redman 1997; Stvilia et al. 2007))</td>
</tr>
<tr>
<td>Rules validity: Attribute values should be compatible with business rules (Based on the definitions provided in (Byrne et al. 2008; English 2009; HIQA 2011))</td>
</tr>
<tr>
<td>Signage accuracy and clarity: Signs should be standardized and universally used across the broadest audience possible (Based on the definitions provided in (English 2009))</td>
</tr>
<tr>
<td>Conformance to definition: calculations/estimates/forecasts have been done by conformance to the policies, procedures and standards</td>
</tr>
</tbody>
</table>

**Table 1. DQ dimensions relating to Syntactic and Semantic Levels.**

Note that ‘Record existence’, ‘Accuracy to reality’ and ‘Mapped consistency’ all involve comparison between data and its external referent (or surrogate source), demanding user assessment (or available surrogate source). These three characteristics are thus beyond the scope of our approach.

**Approach for Discovering Data Quality Requirements**

The approach has four stages, *viz.* preparation stage, syntactic stage and semantics stage and requirements identification stage. We make some basic assumptions regarding the setting in which this approach can be feasibly applied. First, metadata or minimal information relating to primary keys, foreign keys, and attributes (clear label of attributes) is preferred to be available. Second, data format should be simple, or, alternatively, semantic meanings for each component of a complicated data format should be specified in the documentation e.g. the data format for Alighting_stop values such as ‘A B Paterson Dr : A B Patterson College [54496]’, and ‘Coorparoo Primary – 24 [BT0002566]’ was briefly explained. Lastly, an external reference source for checking the accuracy and validity of attribute values should be available, e.g. for stops information, a mapping between stop identifier and stop names would have facilitated the approach better.

**Preparation Stage**

The preparation stage is based on the format and available documentation of the data set. The goCard dataset is contained in csv files, and has attributes: *operator* (Name of the Operator running the service), *operations_date* (date of the trip), *run, route, service* (these three attributes together identifies a specific operation on a route), *direction* (inbound/outbound), *smartcard id* (A unique identifier of a go card), *Passengers* (number of passengers), *boarding_stop, alighting_stop, boarding_time, alighting_time, journey_id* (unique journey identifier), *trip_id* ( trips between transfers).
In this phase, assuming a csv file format, tables with generic columns (strings) are created, and the data is loaded into a relational database. Simple visual analysis of the loaded data provides further information such as some columns should be loaded as special types, such as Number or Date. Loading also helps discovering textual issues in csv files such as non-default delimiters, as is the case with the goCard data.

**Syntactic Stage**

Investigation in this stage focuses on detection of the following problems:

1. **Uniqueness**: concerns duplicate records within a dataset. Comparison between the count of all records resulting from the `select COUNT [column name]` and `select COUNT (distinct [column name])` can determine whether the single attribute value is unique among the records. Whereas when primary keys are specified in the documentation, we can use SQL query `select key1, key2, ..., count(*)` from goCard_all group by key1, key2, ... to examine uniqueness of record in the dataset.

2. **Structural Consistency**: examines consistency of format within a column. We can manually examine the sample subset of data (first 200 records) for data type and format. Then use SQL queries to verify our findings. Once the format pattern is known, SQL can be used with regex expressions or enumeration of legally formatted values to identify structural consistency problems. Verification is important especially when complex SQL regex expression is used for format patterns, e.g. `regexp_like(boarding_stop, '([[:alnum:]]+)\+([BT][0-9] ([A-Z][0-9]))')`. If the resultset of query is not null, either the summarized format does not cover all patterns, then iterative verification continues i.e. the regex expression is revised and verified again, or there are discrepancies in the column i.e. identification of discrepancy. For instance, for ‘direction’ column, we use observed value range in SQL condition ‘IN (inbound’, ‘outbound’, ‘unknown’)’, and resultset contains values ‘0’ and ‘1’, which are potential discrepancies and awaits further investigation regarding attribute domain in the next stage.

3. **Referential Integrity**: pertains to Referential Integrity constraints that involve cross-table investigation. This requires knowledge of primary key and foreign key, and can be examined using SQL queries, that is, examine foreign key column, select keys in the column that are not in the primary key column of the referenced table. If the resultset is not null, referential integrity is violated. The goCard data only contains one table, therefore, referential integrity problems are not relevant.

**Semantic Stage**

Documentation needs to be consulted for semantic information of the data, e.g. in the goCard data, ‘Service’, representing the Route’s stopping pattern for buses, have the value ‘unknown’ for trains. At this stage, problems concerning the following dimensions are inspected:

1. **Null value**: pertains to missing values, which may or may not be valid for an attribute. Based on the understanding of the metadata, we can determine which columns have null value problems. Most null value existence in columns can be detected by manual work in the sample records. SQL queries with ‘is not null’ key words can also be applied to columns. In the goCard data, several columns have null value problems, e.g. ‘journey_id’ which uniquely identifies a journey repeatedly has null values.

2. **Fact Completeness**: underlines that facts are available i.e. attributes has value, to make a record usable in terms it is complete. Based on ‘null value’ investigation result and documentation, we can determine if some null values contribute to the fact completeness problem. Frequent co-occurrence of null values in certain columns can be used as clues. In our investigation, ‘boarding_stop’ and ‘boarding_time’, which constitutes the facts of boarding event of a passenger, and similarly ‘alighting_stop’ and ‘alighting_time’ for alighting event, constantly have ‘null value’ problem, leading to fact incompleteness.

3. **Semantic Consistency**: specifies that meanings of the data are consistent across data sets and applications. To investigate, we employ single-column and cross-column examinations. SQL queries needs to be conducted on entire dataset for problem detection. The goCard data is semantically inconsistent in that trip duration calculated by boarding_time minus alighting_time spans up to 203 hours. Another example is trips recorded within a journey starts from trip 2, with no trip 1 record. Note that one journey can consist of several trips.
4. Value Consistency: requires values to be consistent between records and between data sets. It involves manual observation and SQL queries to detect contradictions between data values. We use a reference source to verify values if possible. Based on format (observed under 'Structural Consistency') and semantic information, columns belonging to the same semantic domain can be detected and can be used for cross-reference i.e. use regex expression in SQL to verify that attributes of the same semantic domain are consistent in values. In the goCard data, a same bus stop, e.g. 'Adelaide St @ City Hall frnt (Stop 20) [BT000020]' and 'Adelaide St @ City Hall frnt (Stop 20) [BT0020_1]', constituting value inconsistency.

5. Equivalence of Redundancy: redundant data can be observed by investigation of overloaded columns. If two column are of an identical type and range, especially when the value are the same in every record, then either there is a overloading (redundancy) or they have different semantic meaning, which will depend on the explanatory documentation of the data.

6. Accuracy to Reference Source: specifies that the data agrees with a reliable source of information. To investigate, we can compare reference source and dataset using SQL. In the goCard data set, no reference source regarding service or customer information was provided. Problems regarding this dimension are undetectable.

7. Conciseness: inspects granularity and lexical, syntactical and semantic accuracy of attribute value. To investigate, single column inspection is involved. We can check sample (first 200 records of the dataset) for lexically, syntactically and semantically incorrect data. We look for values that are not populated to the right level of granularity as specified by the documentation. When problems are detected in the sample, we use SQL on the whole dataset to verify whether the discrepancy represents individual anomalies or inconsistency of all the column values with the metadata.

8. Conformance to Metadata: pertains to Domain Constraint. It is examined mainly within single columns based on domain knowledge provided by the documentation. Problems can be detected by manual or SQL query inspection. Value range of the column needs to be inspected by using SQL to select all distinct attribute values within a column for conformance check with the domain knowledge. An example in the goCard data is 'Route' taking values of 'Wait' and 'Hold', with the metadata specifying routes as numbers. Possible causes of these values include their being default representation of routes undergoing changes or incomplete metadata documentation of certain routes.

9. Rules Validity: examines compatibility of attribute value with business rules taking the form of metadata e.g. one passenger travels on one Go Card. The inspection by SQL query involves single-column and cross-column investigations according to different rules. Rules validity problem within goCard data lies in the value range of 'Passenger' being (-1,0,1,2). Observation of the records with unusual values suggests several possible causes of the problem: refund, repeated touch-offs, separated records of one trip. Particularly interesting is the potentially ‘separated records of one trip’, inferred from three records within one day by the same passenger, with one complete record from train station A to B, and two incomplete records appearing to be of consecutive trips within one journey from B to A. Travelling time of first record approximately equals the added travelling time of latter records, with the second and third record alighting and boarding information respectively.

10. Signage Accuracy and Clarity: signs should be standardized and universally used (e.g. using red to indicate low efficiency and green to indicate the contrary). Detection of this type of problem mainly involves manual inspection.

11. Conformance to Definition: derivation procedures e.g. formulas should be repeated if possible, to reproduce the value for verification. In the Go Card data example, this criterion is not applicable.

12. Random error detection: besides the dimensions used as criteria for data quality problem detection, random error detection is used throughout the data quality assessment process and can happen at any stage of the approach. The detected problems involving single or multiple columns should be noted and investigated. For example, in investigations of the metadata and semantic layer, queries regarding domain range of ‘direction’ result in the discovery that records with ‘go card reload’ value for ‘Ticket Number’ always have ‘null’ for ‘Alighting_Time’, ‘Alighting_Stop’, ‘Journey_id’, ‘Trip_id’.
Thus we suspect potential functional dependency relationship to exist between these attributes. This relationship is confirmed by verification with SQL over the entire dataset.

**Requirements Identification Stage**

The identified problems using the proposed approach provide a concrete set of data quality requirements that can be used to gain data quality improvements or at the very least, data quality aware data usage. In the goCard data we found data quality problems under ‘Completeness’, ‘Consistency’ and ‘Validity’ pertaining to dimension definition 2,4,6,8,9,14,15 (see table 1), and the above 11 requirements were identified. The requirements can be stated as ‘at most x (e.g. four) trip records can be linked with one journey_id’, and ‘time lapse between boarding_time of first trip and alighting_time of last trip in one journey must be within y (e.g. five and a half) hours (x, y to be advised by user), etc. Thus although the user can be provided with the set of requirements relating to the syntactic and semantic aspects of the dataset, to implement data cleansing and/or enforcement mechanisms for the requirement(s), specific metrics, thresholds and priorities (especially for large datasets) need to be agreed upon by the users based on data usage and value.

**Conclusion and future work**

In this paper we have presented the development of an approach for data driven discovery of data quality requirements. The approach has been developed on the basis of a public transport data set. The four stages of the approach provide guidelines on undertaking a systematic and comprehensive assessment of the syntactic and semantic data quality of an unfamiliar data set. By following these guidelines, the user can arrive at a concrete set of data quality requirements that can be used to perform data cleansing and data quality enforcement tasks.

In our immediate future work, we will conduct an empirical study in the form of a focus group, to evaluate the validity of the results (data quality requirements) generated from the approach. The study will reveal the extent to which the generated requirements align with actual requirements and we hope to be able to identify both false positives (identified but not considered as genuine requirements) and false negatives (genuine requirements the approach failed to identify) from this study. Results from the focus group will allow us to authentically assess the validity of the requirements identified through our approach, and accordingly refine the approach as needed and better understand its limitations.

**References**


