The University of Bradford Institutional Repository

Link to conference webpage: *http://cse.stfx.ca/~CPSCom2017/acceptedlist.php*

Citation: Micic N, Neagu D, Campean F and Habib Zadeh E (2017) Towards a Data Quality Framework for Heterogeneous Data. Presented at: The International Workshop on Engineering Data- & Model-driven Applications (EDMA-2017) within the IEEE International Conference on Cyber, Physical and Social Computing (CPSCom) June 21-23, 2017, Exeter UK.

Towards a Data Quality Framework for Heterogeneous Data

Natasha Micic*, Daniel Neagu†, Felician Campean‡ and Esmaeil Habib Zadeh§
Advanced Automotive Analytics Research Institute
Faculty of Engineering and Informatics
University of Bradford
Email: *N.Micic@bradford.ac.uk, †D.Neagu@bradford.ac.uk, ‡F.Campean@bradford.ac.uk, §E.Habibzadeh@bradford.ac.uk

***Abstract*—Every industry has significant data output as a product of their working process, and with the recent advent of big data mining and integrated data warehousing it is the case for a robust methodology for assessing the quality for sustainable and consistent processing. In this paper a review is conducted on Data Quality (DQ) in multiple domains in order to propose connections between their methodologies. This critical review suggests that within the process of DQ assessment of heterogeneous data sets, not often are they treated as separate types of data in need of an alternate data quality assessment framework. We discuss the need for such a directed DQ framework and the opportunities that are foreseen in this research area and propose to address it through degrees of heterogeneity.**

***Keywords*—Heterogeneous Data Sets; Data Quality; Metadata; Data Cleaning; Data Quality Assessment.**

I. Introduction

Latest developments across engineering industries, including automotive, have been dominated by the proliferation of sensors and data-driven technologies connected through the Internet of Things (IoT). This enables a global feedback loop throughout the product life cycle, including product development (from R&D testing to design verification and production validation) and data collected from units (e.g. vehicles) in the field. However, owing to historic development and evolution of systems, in many companies, the exploitation of data collected from the product, is often limited to the purpose for which the data was originally collected.

For example, an engine test during a vehicle development programme might be focused on collecting information about the durability of a particular component or subsystem. Other engine tests are carried out to support characterisation of the engine performance in terms of fuel efficiency and environmental emissions. In any such system level test, all components are effectively tested, with significant amounts of data collected for relevant functional or operational parameters. Yet such data, while catalogued and stored, is seldom used beyond the immediately intended purpose of the test. Such data stores useful information about the behaviour of components and subsystems, which could be used, for example, to predict important performance properties of the system in operation. A specific example would be the reliability performance of the system. If information collected from tests could effectively be combined with data from system operation, including for example field failure data available from warranty databases, prognostics for components or subsystems could be produced, and effective pro-active action could be taken to manage the health of the system in operation. Data quality and data / model governance are essential enablers for the fusion of data from heterogeneous engineering sources, such as various product development tests and system operations, including condition monitoring, diagnostics data, and warranty information.

This paper specifically focuses on the data quality problem. The final aim of the research is to develop a rigorous data quality framework that can be uniformly applied across the range of heterogeneous data, to give an unambiguous indication of the quality of the data, regardless of specificity of data (e.g. continuous data, discrete event data, categorical / qualitative / descriptive type data). The work presented in this paper summarises our initial steps towards the development of this data quality framework. The methodology adopted was to first characterise the heterogeneous nature of engineering data, based on references to multiple domains concerned with data quality. A review of the relevant literature was then carried out to identify the techniques used for describing and quantifying data quality (category of assessment). The deficiencies of such domain data that affect their quality and impact their quality ranking and may have further effects on business intelligence processes, including modelling, knowledge discovery and decision making are also discussed. A list of DQ assessment goals/focuses are derived from these.

In section II we discuss the properties of heterogeneous engineering data sets, and decompose their degrees of heterogeneity. Section III introduces the conceptual overview of data quality governance in particular assessment methodologies. Section IV provides conclusions for the importance of data models and frameworks in information systems when developing DQ assessment processes. Then in section V we provide a review and conclusions of there assessment metrics in literature.

II. Engineering Data Context

The paper focuses on engineering data quality assessment although most techniques have been applied in other domains such as financial, social and bioinformatics. There are multiple types of data stored in engineering environments. To name a few we have, supplier detail information, commodity in product data, information about routine tests, often warranty and claims data, even diagnostics from larger products. Much of this information is disparate and not always handled with efficient data management governance. This causes issues especially for consistent maintenance and re-usage, and for meaningful knowledge discovery. Such a motivation highlights the current need for data quality assessment techniques.

For engineering data there are increasing expectations that data driven modelling is an important predictive analytics tools. But the model will then only be as good as the quality of data used. One such example of engineering application is knowledge discovery from warranty records and failure information [45]. Big data management has shown significant benefits [10] such as process improvements and costs saved from integrating data in such a way to make the most of its content.

The number and variety of data resources (one of the 4V's of big data [46]) generate one of the big data challenges: the presence and need to consistently process heterogeneous data. The various available definitions of data heterogeneity are different depending on disciplinary viewpoint, and we present a definition bounded by the composition of data being evaluated.

In real world data-driven applications, homogeneity exists by exception. However it is also widely used for benchmarking problems in data mining, and for many real world condition monitoring applications where decisions are based on just one time signal (e.g. vibration or heart rate / ECG / EKG etc).

From a database point of view a heterogeneous data system offers a unified query interface for various data resources. From a statistical point of view, heterogeneity refers to different populations, samples or results. Within the paper context mixed data types such as (1) time-continuous data – on variable timescale/frequency; (2) event based data (e.g. Diagnostic Trouble Codes); and (3) categorical/text based data (such as comments) or any of their combinations motivate our use of the term heterogeneous data.

We can however consider the case where heterogeneity occurs at the attribute level (see for similar work [15], discussing levels of heterogeneity). Take, for instance, the field recording the date of claims in a warranty data base, these can contain correct but differently formatted dates, causing heterogeneity. Consider instead no formating inconsistencies in any fields of the data, taking one field for analysis is then considered homogeneous, however taking two fields like claims date and part repaired would in turn be considered a subset of various formats that is a heterogeneous case. Similarly any analysis considering data from separate tables will also be considered heterogeneous as long as they are not representing identical fields. So effectively we consider the data involved in an analysis or modelling process to be heterogeneous or not based on the criteria described below. The "analysis" in the context of this paper is in reference to the DQ assessment methodologies (subsetted in section III), and consequently we categorise the types of heterogeneity explored in the literature in the following degrees:

- **Attribute Homogeneity**: assessment occurs at attribute level on identical data formatting;
- **Attribute Heterogeneity**: assessment on attributes with formatting inconsistencies;
- **Intra-Data Set Homogeneity**: assessment combining multiple similar data attributes;
- **Intra-Data Set Heterogeneity**: assessment combining multiple different data attributes;
- **Inter-Data Set Homogeneity**: assessment considering multiple similar data sets (e.g. identical in schema or context);
- **Inter-Data Set Heterogeneity**: assessment considering multiple different data sets[1] (e.g. dissimilar to in schema or context).

We identified in our research the need to measure heterogeneity as part of data quality assessment; the current approach is categorical and this should be addressed through a more continuous way of measuring.

There are, in fact, methodologies in place like the Heterogeneous Data Quality Management (HDQM) methodology that deals with the assessment of Heterogeneous data [7]. In HDQM the authors describe heterogeneity in the context of different data sets contained inside an organisation. The authors emphasise heterogeneity being the combination of structured, semi-structured and unstructured data sets. This approach is one considering data composition on a data set wide basis. Our breakdown, which is related to data used for type of data quality analysis, provides for a more distinct deeper understanding of relationships between data, while in [7] the methodology links into developing a unified conceptual representation of all the data considered then assessing quality based on the homogeneous unified representation.

III. Concepts of Data Quality Assessment

Data quality is important for businesses for many reasons. Data quality foundations have drawn analogies from the product quality domain [40], where quality checks and standards are applied. The argument is therefore made that data too should have these rigorous quality checks as nowadays data is considered also a valuable business product. It is argued that data of good quality provide many benefits to various data consumers [41], such as data analysists, business intelligence and decision makers.

DQ Methodologies and DQ Frameworks are terms that are often used interchangeably in literature. DQ methodologies are described in [7] as providing "a set of guidelines and techniques that, starting from information that describes a given application context, define a process to assess and improve the quality of data". The view of a framework is usually more of a conceptual one that is used as a helpful map of the process to provide a structure to the use of quality assessment methods, theories and approaches [12][26].

Data quality dimensions are the types of data quality issues that arise in information systems; metrics are the tangible

[1] Note that this category does not consider metadata as a separate data set since it is more related to ideas of record linkage for data cleaning methodologies.

measurements for data dimensions. There is a vast amount of research that attempts to summarise data quality dimensions and their descriptions [13][26][31][38][35][5] in the context of DQ frameworks. In [5] the authors offer a comprehensive analysis of data quality dimensions and DQ methodologies including a list of metrics used to describe dimensions and how they have been calculated. These metrics allow for a quantification of the dimensions that can be measured and compared where needed, and is an important step in the DQ assessment and management process.

Many existing frameworks are structured so the dimensions are subdivided into logical categories; for example in [41] authors use intrinsic and contextual dimensions. Intrinsic dimensions are often discussed as being an objective view of the data quality. Other examples of subgrouping DQ dimensions are found in the Hyperdimensionality framework proposed in [21] (sectioning into Data, Process and User related dimension groups), as well as a categorisation described in [36] (Intrinsic, Accessibility, Contextual and Representational groupings). The paper [25] provides a summary of the most well documented frameworks and their dimensional groupings.

The methodologies discussed so far [41][21][36] produce DQ dimensions to define and capture the fundamental properties of quality, and such a dimension-based analysis is regularly applied successfully in literature [32][18][42].

Other research publications take a less abstract interpretation and view data quality assessment from a user-centred perspective. Namely in [3][33] the authors categorise the issues related to single-source and multi-source contexts to map the data deficiencies based on schema and instance level problems. This work emphasises the importance of knowing the domain and knowing the data. Only then one can bring meaning to metrics being grounded in the context and logical deduction with minimal knowledge e.g. in the case of poor metadata. Authors in [44] suggest that there is no better alternative to understanding the issues in the data than human interaction and exploration (such as the rules of data quality assessment exemplified below); this causes difficulty for automating tasks for discovering data deficiencies. Automating these tasks is commonly known as Data Quality Mining (DQM), as first explained in [44].

Existing research has broken down data quality assessment and improvement activities into (1) proactive and (2) reactive assessments [27]. Both begin with a process of applying some assessment criteria, but reactive strategies simply alter values demonstrating deficiency where proactive assessment attempts to analyse and produce reasoning for defects. This is implemented through an environment that helps the user assess issues at their source [2] [34], but is usually heavily reliant on user input based and potentially only need to happen in cases of severe deficiencies where very little knowledge can be gained because of it. Reactive methodologies are summarised as **Cleaning Deficiencies** in the literature summary Table I. Although proactive strategies involve data cleaning they also allow the user to make decisions, thus when the focus of an assessment is on decision making we use the classifiers **Aid User Decisions**, see Table I. There are cases, usually with no user input, where previously undefined deficiencies are found. We express these cases as (automated) **Deficiency Discovery**. And finally if any data quality assessment is used for purposes of purely observing the quality, with the reason of improving data gathering processes or observing the quality propagated through models as to the use of the assessment, we identify these as **Deficiencies Observed**.

The importance of data cleaning has also been stated in numerous literature resources [8] [11] [29], and there are many different automated methods of data cleaning based on the domain of assessment. The paper [20] actually divides the techniques of data cleaning into Statistical Methods (**SM**), Machine Learning (**ML**) and knowledge-based methods. Recent literature suggests that also some other important subcategorisations of techniques that extend to all data quality assessment methods should be introduced for this analysis. We can split knowledge-based methods based on Rules and Metadata representation. **Rules** describe the cases where a user, based on expert knowledge, defines queries, data constraints or thresholds. **Metadata** provides annotations in other static files that add information to aid in producing DQ rules or queries for the data. We also identify pattern recognition (**PR**) techniques as a point of interest in assessment methodologies. In the summary table we want to make it clear that some methodologies do not define metrics, and so the **Metrics** classification identifies methodologies that utilise classical DQ dimension associations. For instance, in data quality mining **DQM** techniques like in [1], there is no mention of any data quality dimensions but still we are able to evaluate data quality. Other research is focused on the impact of unbalanced data sets [9][43] rather than data quality metrics on the classifiers performance.

As a clarification of some of these methodologies presented above, let us take a durability test for an engine on a test bed as an implementation example. Data collected from here would include mostly sensor recordings of certain frequencies, in the form of numeric values. Some fields may contain imputed or missing values for when a sensor is broken or malfunctioning. We have an accompanied flat static file that describes the ranges of each sensor and details about other fields the software produces (through aggregation, calculation or other means). We also have metadata of each durability test file with some specific details about the test itself (such as which engineer ran the test and on what piece of machinery, the expected frequency and test specific input variables). Assuming that we want to qualify our assessment into DQ dimensions, we would classify it with the identifier "Metric". Choosing dimensions Completeness, Uniqueness, Semantic Consistency and Structural Consistency might be done in the following way:

- **Completeness**: describes if all required rows, based on frequency and start and end time, are accounted for in the output file (a definition of population completeness is provided in [31]). Because we require knowledge on the frequency based on metadata as well as user knowledge of what values are incomplete, it is both classified as a "Metadata" and a "Rule" methodology. A metric to calculate completeness could be defined as:

$$\frac{\text{Total Non Missing Values}}{\text{Number of Expected Values}} \tag{1}$$

  - This form of assessment is classified as an Intra-Data Set Heterogeneity, or Attribute Heterogeneity if performed on one field. [30] describes the IPMAP (Information Product map) methodology and in it the context-independent completeness metric similarly to above but without considering the expected number, it is instead just the total number of stored values.

- **Structural Consistency**: describes the structure of the values in the data. Because in this case all the values are numeric, structural consistency is defined by a value falling within the expressed ranges stated in the accompanying static metadata file. This is classified as a "Metadata" assessment:

$$\frac{\text{Total Structurally Consistent Values}}{\text{Total Values}} \tag{2}$$

  - This consolidates two separate files and so can be classified as Inter-Data Set Heterogeneity

- **Semantic Consistency**: describes rules that explain mandatory relationships between fields. This is classified as a "Rule" that the knowledge of relationships are designed by the user:

$$\frac{\text{Total Semantically Consistent Rows}}{\text{Total Rows}} \tag{3}$$

  - This is an assessment that makes use of Intra-Data Heterogeneous data.

- **Uniqueness**: is defined by time records not being duplicated. Again this is a "Rule" and is based on the users knowledge of the constraint that is required from the output:

$$\frac{\text{Total Unique Rows}}{\text{Total Rows}} \tag{4}$$

  - This use case is utilising Attribute Homogeneity as the analysis is done on one correctly formatted field of data.

The example above gives us a more intuitive feel for the categorisations described in this paper and are mainly based on ideas of metric development suggested in [31]. We do not address some assessment methodologies in the example, but logically SM techniques might be used for detecting anomalous data, and so are highly applicable for this kind of numerical homogeneous attribute data. As well as this many data cleaning techniques employ methods of automated ML techniques. PR is widely used also in general time series analysis and so this context might definitely feel the benefit of assessment with user defined patterns or even learned patterns (as an example of PR tasks on time series outside of the DQ domain see [39]). Pattern recognition then would be an attribute homogeneous task in this case if we were using it to detect errors when expected patterns did not occur.

IV. The Role of Data Models, Frameworks and DBMS

In [28] the authors assert that in specific cases "[the] data quality problems could be avoided by defining convenient constructs at the model level". The idea of the cost of poorly structured data produced in vast amounts, requests a much needed data quality assessment framework and potential data quality metrics [31]. However when we decide to link disparate data sources for mining purposes it is often the case that the sources were never intended to collaborate with each other in the first place and that the data being mined only ever had one initial purpose or is the byproduct of a process [19]. Usually the linking of data sets is done through a process named ETL (Extract Transform Load) with the intention to create a data warehousing framework for efficient data mining processes to be implemented [22].

The problem faced with many ETL tools, as described in [33], is that there is a lack of transformations (or cleaning processes) that are specific for the domain and so have to be done by the user while still being affected by a pitfall in such tools - and this motivates more research needed in this area especially because of current data-nami applications.

We have discussed the operation of some of the existing DQ frameworks generally follow the three distinct steps: state reconstruction, assessing/measuring the quality and then improving the quality [7]. This is the general form that most methodologies take, including that of the Heterogeneous Data Quality Methodology (HDQM) [7]. HDQM focuses the heterogeneity as defined by a combination of structured, semi-structured and unstructured data sources. They analyse the quality of "Conceptual Entities" that combine quality measures from all different sources relating to an entity to produce a quality of that conceptual entity. They discuss that one issue with most standing frameworks is their lack of qualitative assessment that can be used in practice. This may be the cause of the observation made in paper [25] that there is often little suggestion as to how we might assess the "quality score" of an information product. This then leads to the fact that a lot of DQ assessment methodologies that are implemented in practice are better described as "lists of criteria" [12]. This becomes especially true the more specialised the domain data and DQ deficiencies become.

Our degrees of heterogeneity break down the types of heterogeneity at a more granular level than HDQM does and are more concerned with the heterogeneity of the data that we subset for a particular analysis. With this and the fact we have in mind both the assessment categorisations as well as the different potential focuses of an assessment strategies, we believe there to be space for a new framework. This framework would connect more closely with the data analyst and potentially give them suggestions as to which DQ assessment category is the most appropriate for the degree of heterogeneity present in their analysis.

V. Data Quality Metrics as Applied in Literature

We now introduce a small survey of resources reporting the use of data quality assessment processes, comparing the previously defined degrees of heterogeneity, focus and assessment methods. As reviewed by Gschwanter [16] there are numerous taxonomies for data quality problems. Which taxonomy will give the best foundation for the problem is not always clear. Often case studies for DQ will not follow precisely a formalised framework for assessing dimensions as discussed previously and it exemplified by the literature in this review. We are attempting to build a foundation of similarities between data quality in different domains to get a sense of generalisations from useful metadata formats that can be utilised in the future.

Pastorello [30] makes the observation that there are three main types of data quality assessment tools used during the data governance in the context of data pertaining to some time series domain.

Relevant research reported in recent literature that applies some form of metrics will describe them as one of the following: rule, check, test or query. We will however use the definitions provided in section III to give an overview of their assessment methodologies.

Dimensions and metrics are not always described in terms of the fundamental definitions in the established texts in this data quality assessment domain like [5],[31] or [38]. Suggesting a disparity between theoretical and practical application based on the domain of interest. This could be caused by the fact that a data issue may be thought of in a specialised domain way and not under the DQ field of study. This kind of specialised data quality mapping provides a more labouring process of cross domain application of metrics. Conversely there is much ambiguity in choice of metric described in the establishing texts that metric defining inconsistent throughout literature.

A large focus of some literature is exactly observing inconsistencies from data linkage techniques which are then resolved using learning algorithms to impute the errors in some meaningful way[29][11]. The automation of such imputations is important in the field of big data and we have seen that initial quality assessment rules and queries will give us the landscape of quality but only based on the initial knowledge of domain DQ issues. Not much work has gone into the automated discovery of data quality issue when there is very little domain knowledge. For data representing a time series we can however use the well founded ideas of outlier and anomaly detection to produce somewhat meaningful DQ metrics[4].

Even more interesting topics can be discussed in the multivariate time series data quality domain, where fitting a model to a particular set of variables will introduce calculated data metrics into the models to provide a more realistic model. In [21] the authors question the effect poor data quality might have on a data fitted model and discuss the lack of metrics related to some data based inference. Similarly as we have discussed they mention the subjectivity of the data metrics and the lack of comparability the metrics actually give up eluding to the need for more inference based metrics. While inference based data cleaning has been highlighted in the literature, there is little reported yet on metrics based on prior knowledge of the system. We may be able to think about the system "behaving as it should" based on previous behaviours, then flags in quality could be made if the behaviours deviated too far from that we would infer. These are just very surface level observations which again bring us back to the problem of domain-specific metrics and the amount of subjectivity needed for initial DQ assessment, though in some specific case we may be able to generalise.

For multivariate time series, DQ assessment happens on a whole data set, but also relies on inference taken from many other data sets of this type. The nature of this allows for poor data quality to propagate if the data we make inferences with are incorrect. This topic is of high interest to researchers and could potentially lead to metrics often designed by experts, [5] to be automated effectively replacing the long subjective analysis processes to develop assessment tools. These processes have been described in [14] as "no easy task" and being the cause for information loss by missing small details in defined metrics. Ideally a framework for DQ assessment would work together with the expert to improve the metrics and have them adapt to the purpose of use. And equally it would, independently of the expert, automatically make changes that are trivial but time consuming for a user to do manually.

The work in [23] attempts to produce reasoning on metrics for the dimensions of accuracy, confidence and completeness in the domain of data streams. Having a stream with a set size sampled at a constant rate the methodology introduces a windowing approach applying the DQ metrics reasoned in the paper to sections of streams rather that an approach of assessing individual sample points. The article also mentions how the DQ metrics are affected or should take into account altered data in the way of sampling, aggregation, joining, algebraic operators or threshold controls. It explains that there is room to define a good window size and aggregation function to minimise the resources whilst maximising quality, which is another aspect of data quality to consider; the resources that are taken up producing measures and also storage required in long streams of data. In [34] there is also a focus on sensor networks and they too discus the layers of data quality assessment necessary at which levels of the processing from raw data accusation to the user level (the authors call the dimensions data quality criteria). This paper also has a consideration of avoiding "information overload" by the way of windowing the streams and updating metadata when needed in such a way that they are also integrating DQ criteria, or when a specified updating criteria is triggered.

In [34] authors discuss a "global vision" of a dataset, referred to as an object, that is a weighted sum of all assigned data quality criteria of that object.

In many of these papers the process of acquisition and

TABLE I: Data quality assessment summary

Citation	Domain	Category of Assessment	Focus	Degree of Heterogeneity
[21]	Event data	SM, Metadata, Rules	Observe Deficiencies	Intra-Data Heterogeneity, Attribute Homogeneity, Attribute Heterogeneity
[6]	Sensor Data	Rules	Observe Deficiencies, Aid User Decisions	Attribute Homogeneity, Intra-Data Homogeneity
[4]	Sensor Data/Simulated experiment	SM	Cleaning Deficiencies, Aid User Decisions	Attribute Homogeneity
[14]	Machine/asset data	Rules	Observe Deficiencies, Aid User Decisions	Attribute Homogeneity
[2]	Sensor Data	Metadata, Rules, SM	Observe Deficiencies, Aid User Decisions	Attribute Homogeneity
[42]	Clinical Research Records	Rules	Deficiencies Observed, Aid User Decisions	Inter-Data Heterogeneity, Attribute Homogeneity
[18]	Supply chain data	Rules	Deficiencies Observed, Aid User Decisions	Attribute Homogeneity, Attribute Heterogeneity, Inter-Data Heterogeneity
[29]	Longitudinal data	DQM, Rules	Data Cleaning, Aid User Decisions, Discover Deficiencies	Intra-Data Heterogeneity
[34]	Geo sensor data	Rules, Metadata, SM	Deficiencies Observed, Aid User Decisions	Inter-Data Heterogeneity, Attribute Homogeneity
[37]	Geographical data	Rules, Metadata	Deficiencies Observed, Aid User Decisions	Attribute Homogeneity, Inter-Data Heterogeneity, Inter-Data Homogeneity
[30][2]	Sensor data	SM	Deficiencies Observed	Attribute Homogeneity,
[17]	Sensor data	Rules, Metadata, SM	Deficiencies Observed, Aid User Decisions	Attribute Homogeneity, Inter-Data Heterogeneity
[20]	Sensor event data	SM, PR	Deficiencies Observed, Aid User Decisions	Attribute Homogeneity, Inter-Data Heterogeneity
[24]	Sensor data streams	Metrics, ML, Metadata, SM	Deficiencies Observed, Aid user decisions,	Attribute Homogeneity
[1]	Event data	DQM	Deficiency Discovery, Aid user decisions	Inter-Data Homogeneity
[11]	Event data	DQM	Cleaning Deficiencies , Discover Deficiency	Attribute Heterogeneity, Intra-Data Heterogeneity

[2] This paper is eluding to the use of metrics to solve the DQ issues raised, no formal assessments are actually implemented or described in detail. It describes motivation for PR DQ assessment methodologies.

processing of the data is thought of as points where there is a possibility for data deficiencies to manifest themselves. Also ideas of developing and incorporating metadata or usually other linked data into data quality checks and rules is demonstrated. These metrics can be incorporated into the metadata and thought of as a valuable asset in the data recording process.

A strong theme through the literature is that knowing and categorising DQ will affect the decisions being made surrounding the use of the data by the user. With this in mind we need a strong validation and reasoning process as to why a particular metric is a good fit for a dimension assessment, which is often not discussed in literature. This is probably because many assessment methods are derived from classifications of data quality deficiencies based on known issues with the data, especially metrics that are dependent on the data breaking expected rules. This process is highly intuitive and logical and supports the need for data cleansing in cases where the quality can be improved by a particular method.

Other literature has a focus on the need for data metrics; for example in [6] the authors describe the need for assessing the completeness of the data given in a particular "smart home" application. They focus on time series data streams discussing the system of data accumulation rates in sensors and by querying the sensors, describing both system and query completeness. The authors use their query completeness to identify streams where sampling rates are unnecessarily high based on run time, and so using the metrics as an indicator when changing other variables, are able to make informed optimisation decision.

VI. Conclusions and Future Work

Selection of data quality metrics rely heavily on the users' requirements of the data and often the process by which the data is collected, stored and aggregated. A working knowledge of the data domain and context helps to produce clear metrics but is also time consuming when relying on knowledge and human inspection of data to produce representative metrics. There have been suggestions for data cleaning solutions in structured data scenarios that use Bayesian networks and other model-based approaches to impute data and take those decisions out of the users hands: this will often identify data deficiencies. These also have seen application in more text based structured data storage systems containing address, names or dates where record linkage and probabilistic record linkage become a helpful pre-processing activity to support smarter cleaning methodologies and are more conducive of machine learning techniques.

A frequent problem of interest is mostly in numeric data: time series analysis approach in such directions as outlier detection, distribution modelling and trend analysis on homogeneous data sets. Where sensor data networks are concerned accuracy and reliability of the data is related to linked meta data as well as user expectation. If we begin to map these expectations in terms of rules, in large networks of streaming data, the process starts to become a large undertaking. There are limitations in this rule based methodology. An open question is if we can provide some issue identification methodology to alert the users of machine discovered issues in time series data. Research on this topic of automated issue discovery in a more text based information system has already made significant progress. The challenge we are proposing may be thought of as an outlier detection of a higher level than analysis of one time series but a system of sensor network outputs (or the realm of inter- and intra-heterogeneous time series data).

Another issue highlighted by this review is the divide between the conceptual frameworks and the criteria list approach of DQ assessment. We begin to acknowledge the benefits of both and think toward developing a framework that amalgamates the conceptual view with the tangible measurement process, to create a more accessible DQ framework to engineering domain applications.

ACKNOWLEDGEMENTS

We acknowledge and greatly appreciate that this work was partly funded by the Soroptimists Bradford charity scholarship.

REFERENCES

[1] Nawaf Alkharboush and Yuefeng Li. A decision rule method for data quality assessment. In *Proceedings of the 15th International Conference on Information*, volume 3, pages 84–95, 2010.

[2] Clemens Arbesser, Florian Spechtenhauser, Thomas Mühlbacher, and Harald Piringer. Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):641–650, 2017.

[3] José Barateiro and Helena Galhardas. A survey of data quality tools. *Datenbank-Spektrum*, 14(15-21):48, 2005.

[4] Sabyasachi Basu and Martin Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11(2):137–154, 2007.

[5] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16, 2009.

[6] Jit Biswas, Felix Naumann, and Qiang Qiu. Assessing the completeness of sensor data. In *International Conference on Database Systems for Advanced Applications*, pages 717–732. Springer, 2006.

[7] Batini Carlo, Barone Daniele, Cabitza Federico, and Grega Simone. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems*, 3(1):60–79, 2011.

[8] Arthur D Chapman. *Principles and methods of data cleaning*. GBIF, 2005.

[9] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[10] Thomas H Davenport and Jill Dyché. Big data in big companies. *International Institute for Analytics*, page 3, 2013.

[11] Sushovan De, Yuheng Hu, Venkata Vamsikrishna Meduri, Yi Chen, and Subbarao Kambhampati. Bayeswipe: A scalable probabilistic framework for improving data quality. *Journal of Data and Information Quality (JDIQ)*, 8(1):5, 2016.

[12] Martin J Eppler and Dörte Wittig. Conceptualizing information quality: A review of information quality frameworks from the last ten years. In *IQ*, pages 83–96, 2000.

[13] Christopher Fox, Anany Levitin, and Thomas Redman. The notion of data and its quality dimensions. *Information processing & management*, 30(1):9–19, 1994.

[14] Ralf Gitzel. Data quality in time series data an experience report. *Proceeding of CBI 2016 Industrial Track*, 2016.

[15] Aubrey Gress and Ian Davidson. A flexible framework for projecting heterogeneous data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1169–1178. ACM, 2014.

[16] Theresia Gschwandtner, Johannes Gärtner, Wolfgang Aigner, and Silvia Miksch. A taxonomy of dirty time-oriented data. In *International Conference on Availability, Reliability, and Security*, pages 58–72. Springer, 2012.

[17] Jianwen Guo and Feng Liu. Automatic data quality control of observations in wireless sensor network. *IEEE Geoscience and Remote Sensing Letters*, 12(4):716–720, 2015.

[18] Benjamin T Hazen, Christopher A Boone, Jeremy D Ezell, and L Allison Jones-Farmer. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72–80, 2014.

[19] Jochen Hipp, Ulrich Güntzer, and Udo Grimmer. Data quality mining-making a virute of necessity. In *DMKD*, 2001.

[20] Thomas Hubauer, Steffen Lamparter, Mikhail Roshchin, Nina Solomakhina, and Stuart Watson. Analysis of data quality issues in real-world industrial data. In *Poster Presentation at the 2013 Annual Conference of the Prognostics and Health Management Society*, 2013.

[21] Alan F Karr, Ashish P Sanil, and David L Banks. Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173, 2006.

[22] Ralph Kimball and Joe Caserta. *The Data Warehouse? ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. John Wiley & Sons, 2011.

[23] Anja Klein. Incorporating quality aspects in sensor data streams. In *Proceedings of the ACM first Ph. D. workshop in CIKM*, pages 77–84. ACM, 2007.

[24] Anja Klein and Wolfgang Lehner. Representing data quality in sensor data streaming environments. *Journal of Data and Information Quality (JDIQ)*, 1(2):10, 2009.

[25] Shirlee-ann Knight and Janice Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8, 2005.

[26] Anany Levitin and Thomas Redman. Quality dimensions of a conceptual view. *Information Processing & Management*, 31(1):81–88, 1995.

[27] David Loshin. *The practitioner's guide to data quality improvement*. Elsevier, 2010.

[28] Kashif Mehmood, Samira Si-Said Cherfi, and Isabelle Comyn-Wattiau. Data quality through conceptual model quality-reconciling researchers and practitioners through a customizable quality model. *ICIQ*, 2009:61–74, 2009.

[29] Mario Mezzanzanica, Roberto Boselli, Mirko Cesarini, and Fabio Mercorio. A model-based approach for developing data cleansing solutions. *Journal of Data and Information Quality (JDIQ)*, 5(4):13, 2015.

[30] Gilberto Pastorello, Deb Agarwal, Dario Papale, Taghrid Samak, Carlo Trotta, Alessio Ribeca, Cristina Poindexter, Boris Faybishenko, Dan Gunter, Rachel Hollowgrass, et al. Observational data patterns for time series data quality assessment. In *e-Science (e-Science), 2014 IEEE 10th International Conference on*, volume 1, pages 271–278. IEEE, 2014.

[31] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[32] Zhijing Qin, Qi Han, Sharad Mehrotra, and Nalini Venkatasubramanian. Quality-aware sensor data management. In *The Art of Wireless Sensor Networks*, pages 429–464. Springer, 2014.

[33] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[34] Claudia C Gutiérrez Rodríguez and Sylvie Servigne. Managing sensor data uncertainty: a data quality approach. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, 4(1):35–54, 2013.

[35] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A Jabar, Hamidah Ibrahim, and Aida Mustapha. Data quality: A survey of data quality dimensions. In *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*, pages 300–304. IEEE, 2012.

[36] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[37] Howard Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.

[38] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[39] Peng Wang, Haixun Wang, and Wei Wang. Finding semantics in time series. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 385–396. ACM, 2011.

[40] Richard Y Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.

[41] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[42] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

[43] Karl R Weiss and Taghi M Khoshgoftaar. Investigating transfer learners for robustness to domain class imbalance. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 207–213. IEEE, 2016.

[44] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[45] Huaiqing Wu and William Q Meeker. Early detection of reliability problems using information from warranty databases. *Technometrics*, 44(2):120–133, 2002.

[46] Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.