2021

Evaluating the Quality of Repurposed Data – The Role of Metadata

Hui Zhou
University of Queensland, hui.zhou1@uq.net.au

Gianluca Demartini
University of Queensland, g.demartini@uq.edu.au

Marta Indulska
University of Queensland, m.indulska@business.uq.edu.au

Shazia Sadiq
University of Queensland, shazia@itee.uq.edu.au

Evaluating the quality of repurposed data – The Role of Metadata

Full research paper

Hui Zhou
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, Queensland, Australia
Email: hui.zhou1@uq.net.au

Gianluca Demartini
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, Queensland, Australia
Email: g.demartini@uq.edu.au

Marta Indulska
School of Business
The University of Queensland
Brisbane, Queensland, Australia
Email: m.indulska@business.uq.edu.au

Shazia Sadiq
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, Queensland, Australia
Email: shazia@itee.uq.edu.au

Abstract

Existing approaches for evaluating data quality were established for settings where user requirements regarding data use can be explicitly gathered. However, users are often faced with new, unfamiliar, and repurposed datasets, where they have not been involved in the data collection and data creation processes. Furthermore, there is evidence that there is typically a lack of supporting information, such as metadata, for such datasets. Yet, users need to evaluate the quality of such data and determine if the data can be used for intended purposes. In this paper, we aim to gain an empirical understanding of the role of metadata in evaluating the quality of repurposed data. Using an interview approach, we collected rich qualitative data that reveals current practices, key challenges, preferences, and approaches for improvement regarding evaluating the quality of repurposed data.

Keywords: data quality, metadata, repurposed data

1 Introduction

Data has long been recognised as a valuable resource and asset for organisations (Fisher 2009). However, poor data quality (DQ) can cause critical issues in data analytics and decision making, also known as the "garbage in garbage out" problem (Redman 2018). Accordingly, DQ has been extensively studied for decades (Sadiq et al. 2011) by various research communities, including Information Systems, Computer Science, Statistics, as well as in various contexts, such as finance, credit, health, etc (Jaya et al. 2019). A large subset of this research is dedicated to studying DQ assessment, which is not surprising considering the widespread need to perform quality evaluation and data cleaning (Krishnan et al. 2016).

In the most general case, the DQ assessment processes follow a top-down approach, starting with determining user requirements, followed by data quality measurement (or profiling) and finally cleaning data as well as ensuring continuous monitoring. To this end, prior research has proposed a variety of methodologies, such as TDQM (Total Data Quality Management Methodology) (Wang 1998), DWQ (Data Warehouse Quality) Methodology (Jeusfeld et al. 1998), DQA (Data Quality Assessment) Methodology (Pipino et al. 2002), and DQAF (Data Quality Assessment Framework) (Sebastian-Coleman 2012). However, the explosive growth of data, both in terms of size and complexity, has challenged these existing methodologies. In particular, this challenge arises due to the rise of so-called data repurposing (Zhang et al. 2019), i.e. using data that was created for a particular purpose by one set of users, for a different purpose, generally by a different user group, one who is unfamiliar with the data characteristics and its original creation processes. A typical example of repurposed data is open data provided through government public portals. Such datasets were created for one purpose e.g., public administration, but are then offered to be used for entirely new purposes by new users. In such cases, users generally have little, if any, involvement in the data creation processes and thus their knowledge of the data is poor. On the other hand, it is difficult, if not impossible, for data owners or custodians releasing these datasets to anticipate the various new uses for the data. For this reason, predetermined user requirements and data quality assessment approaches that have been well developed and utilised in the past are not sufficient to assess such *repurposed* data.

Indeed, as illustrated by Sadiq and Indulska (2017), there is a substantial gap in the body of knowledge in regards to assessing the quality of repurposed data. Current approaches are ad-hoc and manual in nature (Clarke 2016). In recent years, the research community has started to address this challenge. For example, Zhang et al. (2019) developed a "bottom-up" approach, which has the capacity to discover data quality issues for repurposed datasets. However, the approach does not cover the full spectrum of DQ dimensions[1]. Similarly, there has been some progress on advanced machine learning approaches to profile and prepare data for analytical tasks (Stonebraker et al. 2013). However, fully automating DQ assessment has a number of pitfalls, and human judgement is essential to ensure that issues of bias, transparency and fairness can be adequately managed (Cichy and Rass 2019).

A fundamental artifact in supporting the human judgement in the evaluation of the quality and fitness of a dataset for a given purpose is metadata - generally defined as 'data about data'. Over the past 30 years, metadata has been widely used in Information Systems and specifically in data quality management (Lee et al. 2006). However, the study of metadata usage for DQ assessment is lacking (Aljumaili et al. 2016), despite broad recognition of the importance of metadata in this area. For example, Batini et al. (2009) indicate that metadata assists data consumers to understand and evaluate data. Existing data quality assessment approaches also recognise that metadata contains necessary information to evaluate data quality. Existing approaches require high-quality metadata to begin with, for instance, a database schema. In the context of repurposed datasets, however, there is no guarantee of the availability of such a schema, and metadata is often insufficient for these datasets (Belkin and Patil 2018).

As a step towards addressing this challenge, our aim in this paper is to explore the current practices of metadata use in evaluating repurposed datasets and identify the key challenges, approaches and preferences in this regard. Accordingly, we conducted a qualitative study through semi-structured interviews to collect rich data on the experiential knowledge of senior data management practitioners. Our results provide a deep insight into the role of metadata in evaluating the quality of repurposed datasets. In the remaining sections of the paper, we first present related work in terms of data quality assessment and the use of metadata in data quality assessment. Section 3 presents our research method,

[1] Data quality can be measured through so-called dimensions, for instance: Accuracy, Consistency, Completeness, Uniqueness, and Timeless (Batini and Scannapieco 2016).

and Section 4 presents the results. We conclude the paper in Section 5 and provide an outlook of how the findings of our study can guide future research on evaluating the quality of repurposed datasets.

2 Related Work

2.1 Data Quality Assessment

It is widely recognised that data quality is a critical factor in the outcomes of data-driven business decisions (Stvilia et al. 2007). Using poor quality data can cause detrimental impacts for organisations (Wang and Strong 1996). To enable management of data quality[2], substantial effort has gone into classifying data quality dimensions, which can be used as a basis to assess data quality and discover data quality problems. Researchers have proposed various classifications. For example, a basic set of data quality dimensions were defined as accuracy, completeness, consistency, and timeliness (Wang et al. 2002). More recent research has offered a consolidation of the various classifications and identified eight core data quality dimensions, namely: Completeness, Accuracy, Validity, Consistency, Currency, Availability and Accessibility, Reliability and Credibility, and Usability and Interpretability (Jayawardene et al. 2015).

The quality of data is often evaluated on the basis of these various data quality dimensions. Among existing assessment methodologies, Accuracy, Consistency, Completeness, Uniqueness, and Timeless are the most commonly discussed dimensions in literature and in practice (Batini and Scannapieco 2016). Batini et al. (2009) offered a systematic review of existing data quality assessment approaches, indicating that data analysis, DQ requirements analysis, identification of critical areas, process modelling, and measurement of quality are the most common related activities. Their work also emphasises the role of metadata in these activities, as data workers can be assisted by complementary information about the data.

A variety of techniques have been developed to assess and/or improve data quality for various dimensions. Among existing techniques, data profiling is commonly recognised and used in practice to examine datasets and to produce metadata. Data profiling results are utilised by users to evaluate the datasets. For instance, the number of null values, distinct values in a column, the most frequent patterns of data values, cross-column correlations and functional dependencies are some of the results used to make such judgements (Abedjan et al. 2015). A variety of software tools can also be found in the market to conduct or assist data profiling: e.g., InfoSphere Information Analyzer from IBM, Information Steward from SAP, Data Profiler from Talend and Trifacta.

2.2 Metadata

There have been several initiatives towards standardisation of metadata in various contexts (e.g. Dublin Core, Categories for the Description of Works of Art (CDWA), Library of Congress Subject Headings (LCSH) (Gill et al. 2008)). Various communities have explored different typologies of metadata depending on different ways of classification and aggregation. For instance, Lagoze et al. (1996) developed a typology of seven types of metadata: 1) Identification/description metadata; 2) Administrative metadata; 3) Terms and conditions; 4) Content ratings; 5) Provenance; 6) Linkage/relationship; 7) Structural. More recently, Sawadogo and Darmont (2021) classified metadata into two major categories, namely Functional and Structural, and related sub-categories. To date, there is no universal agreement on the classification of metadata. Indeed, different classifications of metadata often overlap. In practice, the most widely used classification considers three types: descriptive, structural, and administrative (Méndez and Hooland 2014). In this paper, we will adopt this classification on the basis of the following understanding:

a) **Descriptive metadata**: describes data related to discovery, retrieval, or identification purposes, such as author, keywords, tags, topics, and titles.

b) **Administrative metadata**: describes the collection, creation, management, licence, access, and use information for data, such as the creation date, method of acquisition, intellectual use,

[2] We note that researchers tend to use the term 'data quality' (DQ) to refer to technical data quality problems, and 'information quality' (IQ) to refer to non-technical quality problems related to effective use of processed data. However, there is no common agreement on the specific distinction between the two (Zhu et al. 2014). We use DQ to refer to both categories of quality issues in this paper to allow as complete as possible coverage of the diversity of data quality dimensions available in literature (Jayawardene et al. 2015).

and contextual information. These metadata elements are used for managing digital objects and are often called metadata about metadata, or "meta-metadata".

c) **Structural metadata**: describes the storage, navigation, presentation, and structure information of data, such as table of contents, entity relationships and Data Type Definition (DTD) for the XML file. They indicate how digital objects are organised and their relationships.

2.2.1 Metadata for data quality assessment

There have been few but notable prior studies on metadata usage in data quality assessment. In the work of Farinha et al. (2009), researchers have developed a metadata-based DQ assessment approach to identify recurring patterns that assist in the inheritance of business rules. Zhang et al. (2009) proposed a method to improve data quality by using structural metadata, such as data structures, integrity constraints, and data atomicity. Aljumaili et al. (2016) developed a metadata-based DQ assessment tool that assesses DQ using both content and database metadata. However, the proposed tool is limited to relational database management systems and requires a well-managed database schema.

Our work aims to provide a foundation for future research on DQ assessment in current settings by investigating the role metadata in assessing the quality and fitness of repurposed datasets, and identifying current challenges, user preferences and potential approaches for improvement.

3 Methodology

We use semi-structured interviews as we aim to gather rich qualitative data regarding participants' experience and knowledge of a topic (Brinkmann and Kvale 2015). In particular, we opt for open-ended questions in alignment with the objective of this study to explore practitioners' hands-on experience with evaluating repurposed data.

Our interview protocol was structured in 4 sections: 1) Background and experience in using repurposed data; 2) Role metadata played in assessing quality for repurposed data; 3) Preferences on the format of metadata to enable DQ assessment; 4) Challenges and suggestions regarding assessing the quality of repurposed data. Questions in section 2 related to the three types of metadata: descriptive, administrative and structural (Méndez and Hooland 2014). We probed on several key aspects for each type of metadata, including availability, format, and usefulness.

We recruited participants from various sectors and roles in their organisation to explore the phenomenon from different perspectives. Participant recruitment was aimed at participants who are senior data managers (engineers and scientists), have over ten years' experience in various data management roles and specific experience in interacting with repurposed data of unknown quality. We followed a purposive approach until data saturation was reached (Palinkas et al. 2015). Seven interviews were conducted, which sits in the range of common interview studies (Richter et al. 2016). We conducted all interviews in English, following the same protocol. All the interviews were recorded, transcribed, and moderated by our research team. The interviews lasted 54 minutes on average (minimum 46 minutes, maximum 61 minutes). A summary of participants can be found in Table 1.

Participant	Sector	Role
P1	Insurance	Data Governance Executive
P2	Telecommunications	Senior Information Quality Analyst
P3	Telecommunications	Data Management Senior Specialist
P4	Government	Manager, Data Services
P5	Higher Education	Research Lead
P6	Higher Education	Senior Strategic Adviser
P7	IT Services	Principal – Analytics Engineering

Table 1. Summary of Interview Participants

To analyse the interview data, we follow the approach described by (Thomas 2006). We build themes from the data using thematic analysis (bottom-up approach) to summarise data into abstract codes, themes, or categories. We used an *a priori* high-level set of codes that aligned with the four sections of the interview. A dual-coder approach was used to reduce any bias in our analysis. Two researchers individually iteratively coded the seven interviews using the *a priori* set of codes and creating new

related codes as necessary. We used NVivo 12[3] to assist in analysing the data and managing the code structure (Miles and Huberman 1994). Each identified code was also annotated as a challenge (C), approach (A), and preference (P). The researchers then compared the coding, accounting for synonyms. The coding was then discussed among all authors and refined until full agreement was reached. Because of the inductive nature of our coding approach, agreement measures, such as Cohen's Kappa, were not feasible in our study. However, all authors reviewed the coding results until 100% agreement was reached. Overall thirty-six codes were identified, including eight related to section 1 (in the interview), fifteen related to section 2, five related to section 3, and eight related to section 4.

4 Results

Table 2 shows a summary of our analysis as it relates the four interview sections and three code types (see Section 3 for interview protocol description and coding process), identifying how many participants raised the observation (NoP in Table 2) and how many times it was mentioned in the seven interviews (NoM in Table 2) collectively. For each observation, we offer a brief explanation in the table. Due to lack of space, we only present and discuss the codes that were identified by at least 3 participants.

Code	CT	IS	NoP	NoM
Lack of adequate metadata: users work with unfamiliar data with inadequate metadata	C	1	7	108
Manual and ad-hoc profiling: the need for manual or semi-automated tasks to evaluate data quality from scratch, for example, through 'eyeballing', or ad-hoc data profiling	A	1	7	44
Waste of resources: a waste of time as users realize the datasets are not fit for purpose after significant investment in the analysis of the data	C	1	3	3
Inconsistent format: lack of format consistency of metadata due to variety of source systems, creators, and projects	C	2	6	29
Personal interpretation: subjective interpretation of datasets due to insufficient metadata	A	2	6	13
Discoverability: difficulties in finding datasets that suit intended use due to lack of descriptive metadata	C	2	4	8
Communication: users know whom they can ask for more information, e.g., senior employees, data creators, or experts	A	2	6	16
Data quality proxy: initial understanding of data quality based on administrative metadata	A	2	4	4
Accessibility: the need for input from data owners/creators due to lack of structural metadata	C	2	4	6
Establishing shared understanding: understanding the structure of the unexplored data is vital for establishing shared understanding between the data analytics team and business stakeholders	A	2	3	6
Interactive dashboards: users prefer an interactive dashboard to view and drill down data if needed	P	3	5	14
Best practice and data governance: keep clear documentation when creating data or repurposing data, and proper data governance	P	4	4	10
Common protocol: common protocol, standards, approaches, methodologies is important for the reuse of unfamiliar data	P	4	3	7
Platform and tools: single platform to access, transform, manage data	P	4	3	5

CT: Code Type; IS: Interview Section; NoP: Number of Participants; NoM: Number of Mentions

Table 2. Summary of the Codes

[3] NVIVO is a qualitative data analysis tool designed for analysing rich text-based and/or multimedia information, where deep levels of analysis of data are required. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home/

4.1 Evaluating Repurposed Data

Lack of adequate metadata (C). All participants in our study advised that they had experience in dealing with unfamiliar data with insufficient information and knowledge about the data, in other words, with missing or inadequate metadata – e.g., "*attribute level, definitions, data type, size, format, carnality and whether it can be allowed to be null or not, whether it's mandatory or not. And I find a lot of data sets in the open-source space don't have that information. And you're not necessarily even provided with publication dates or time.*" (P1). This lack of information led to inefficiency and inaccuracy in their work. As they were not initially involved in the data creation process, they did not have the prior knowledge to make sense of the data. Consequently, this can lead to misinterpretation of fields in structured data and mishandling integrities for relational data.

Manual and ah-hoc profiling (A). In our study, all participants indicated they had to undertake a manual evaluation of data quality from scratch, through eyeballing, looking for missing values in Excel, with Python scripts, or SQL queries, to name a few. As P2 explained, "*Analysts often then do manual profiling of the data attributes that they're interested in, just to have a look at the range of values, whether they look like the consistent, trying to understand fields*". Similar manual work was reported by P6, "*We just do sampling and look at random data outputs, and sort of go through manually to see, you know, how accurately it is filled out, if there are any inconsistencies, it's very much a manual process*". Participants conducted this process on the basis of their personal experience and preference towards getting an understanding of the quality. P1 advised their team used Python and SQL scripts to manually explore datasets or the columns that they were interested in: "*…does it look reliable. Does it look complete, does it look accurate, and then comparing against longitudinal datasets to see does it look consistent across (the whole datasets)*". Additionally, six participants advised they undertake some form of tool-based data profiling as an initial indicator of general quality, especially in large datasets, to develop an initial understanding of DQ issues, e.g., inconsistent formats and missing values. P2 indicated: "*The profiling tools generate … how many characters, how many different formats are in that column, and then you can get an idea of how consistent the data is*", and "*It's your first indication of whether there are data quality issues or not, so it can be an indicator of, you know, if it's all consistent and all looking good, then you start looking at more sophisticated data quality checks. Those basic things are the first checks.*". This behaviour of integrating data profiling to initially evaluate data as the first quality indication is also raised by P1, "*It gives you a starting point. It will give you some indication of whether or not this is a data set you might want to work with […] it might even give you some indication of the kinds of data quality challenges that you might expect when you're trying to repurpose. If you don't have that information, then you're flying blind*".

Waste of resources (C). Three participants highlighted their experiences of wasted time and effort due to insufficient metadata. On the contrary, "*If you have all the information upfront, you can prepare yourself and prepare the data before you ingest it into your pipeline."* (P5). Participants noted a higher difficulty of anticipating data issues, therefore, well-defined data preparation and pre-processing steps do not apply to repurposed datasets. *"Things can happen where you are just waiting [for analysis results] for weeks, and then you find out - oh shoot, we didn't see that - the end of lines were coded slightly differently"* (P5). Similar experience was reported by P2: *"… we produced a thing, went to roll it out and found out it was rubbish, so we actually haven't been on a roll here. So, yeah, it costs business, I don't know, we're probably into the 10,000 dollars so far".*

4.2 Role of Metadata in Evaluating Repurposed Data

Inconsistent format (C). Metadata format inconsistency was raised by 6 participants. Because of the variety of the source systems, creators, projects, and other factors, there is no consistency with regard to the format of the metadata. P1 explained, "*I'd say it was a mixed bag. It wasn't consistent across the different data sets and different source locations*". Similarly, P2 indicated: "*Some of it would have a narrative maybe a paragraph overall explaining the data set and the purpose. Some of them would have a table, for example on the website […] Some of the more advanced ones might tell you the number of records that you can expect in each individual file, but it's very inconsistent across the different data sources. There's no one way that these are all represented to you*". As a result, users need to use various tools to ingest metadata to assist the quality evaluation of the datasets. Among the experience of format inconsistency, supplementary PDF (or Word) documents and Excel spreadsheets are two relatively typical formats, as advised by P4: "*Typically, a PDF. It depends on who does it. But if I look at what our guys do, they will typically have a description of the data what it is as a whole, then they would go to table names, describe the table names.*"; and P5: "*if it's extensive data, it probably will come in something like an Excel spreadsheet*".

Personal interpretation (A). Six participants explained that personal interpretation of the repurposed datasets is necessary when there is insufficient information about the datasets, such as the original use case and purpose, previous transformations that have been conducted, the business logic of creating the metrics and relationships between data objects, to name a few. For example, P3 indicated: "*There was no real descriptive information, say exactly what those fields were. It was more my interpretation of what the information was within the fields"*. This can lead to a misunderstanding of the datasets or incorrect analysis results, which results in a waste of resources: "*The less information provided about the dataset, the more interpretation you have to make which causes a lot more analysis, it's a lot more brain time […] because you have to do all those assessments first, before you can really determine if it's fit for purpose […] you may decide after some time (using that data) that, actually the data wasn't what I was thinking, and it's not giving me the information that I want*." (P1).

4.2.1 Descriptive metadata

Six participants mentioned their experiences of having insufficient descriptive metadata when repurposing datasets. Participants also indicated the importance of descriptive metadata, for instance, the description can give them an indication of whether the dataset is fit for their purpose, "*It will give you some indication of whether or not this is a dataset, you might want to work with, or if it's a very complete description, then it might even give you some indication of the kinds of data quality challenges that you might expect, when you're trying to repurpose"* (P1). Among these participants, discoverability was the key challenge caused by the lack of descriptive metadata.

Discoverability (C). Discovering the right data sources has become increasingly important in the area of big data, as indicated by four participants. P3, for example, indicated they could no longer find the right data by asking colleagues because employees who created the datasets, often years earlier, or who have knowledge of the datasets, may have left the organisation. The complexity of file formats and source systems also makes the discovery task more difficult than before, "*Whereas when they started to move into the big data space, or the data lakes, […] it's up to the data analysts or the data scientists to try and discover that (data) by themselves*."(P3). Under these circumstances, having good descriptive metadata assists users to identify potential datasets for their use. P4 indicated that part of their core responsibility is to assist users to find the right data when requested: "*So, it's typically more about the identification of the data than it is about how we bring together. (The scenario is common where) someone has the idea that this group over here might have something that's useful for that group over there.*". Having descriptive metadata can also allow users to discover related or similar datasets if the current datasets are not fit for intended purposes, "*…having some tags is useful because it allows you potentially to jump between a few data sets with the same subject, or the same topic. […] If there are those tags, being able to click on a tag or having a way to see all the assets that relate to a specific tag (is beneficial to discover datasets)*" (P7).

4.2.2 Administrative metadata

All participants experienced a lack of administrative metadata, which contributed to the complexity of their work. For example, P1 indicated: "*… when I look at the administration data available for datasets, it's pretty limited, you often don't get told what sources it's coming from, what processes were used to generate it. Was it a raw extraction, or was it some kind of interpretation? Is it an aggregation of things, or has it been vetted? Has somebody fiddled or tweaked with the data before they made it available because they wanted the numbers to look nice?*" (P1). P2 also considered administrative metadata being crucial to assess DQ: "*unless you know the business metadata and the context in which the data has been created and for what purpose. That (missing administrative metadata) can often make it difficult to assess the quality*." The most common observations include:

Communication (A). The need to talk to the owner, creator, SME (Subject Matter Expert) or the steward of the data is one of the commonly brought up experience by participants - six participants described such need to gain a better understanding of the datasets before using it for a new purpose. For example, P1 indicated: "*I actually have to arrange a meeting with the person who administered the data and explained what I want to use it for*". The lack of administrative metadata is often the cause of this need. We observed the use of communication platforms or software, e.g., Yammer, to post enquiries regarding datasets and source relevant information about the data. Senior employees or those with experience using the specific datasets are trusted to provide relevant information, which assists the use of such data for a new purpose, "*Without that, users do tend to operate on user groups, and even in that you know how I described, we had that internal check site. There were people that were regarded as SMEs (subject matter expert), and they would be trusted as giving you the correct answer more often than you know*" (P2). The availability of the owner contact information as part of the supplementary

administrative metadata is deemed a key to the success of the value creation from repurposed datasets, "*It makes a huge difference, and if you know who created it as well, then if you have any questions or any uncertainties, you can always reach out to the owner of that data, and potentially get some answers*" (P5).

Data quality proxy (A). Four participants stated that administrative metadata provides some indication of the quality of the datasets, regardless of whether there is a good chance of good quality or not. On the one hand, a lack of administrative metadata can cause data trust issues – e.g. "*There were certain things that we needed to look more carefully at the data if we didn't have good metadata with it. It usually was a little bit of a red flag for us if the metadata wasn't with it."* (P5). On the other hand, having administrative metadata can assist users to develop a better understanding of data. For instance, P7 indicated: "*I will be definitely interested in the lens of the business logic that has been applied or what are the business rules that have been applied to come up to this data set. Because if I have the slight doubt about the business rules, I will not use the dataset because I don't want to show wrong data.*". Two participants (P5 and P7) further indicated that the existence of administrative metadata could raise their trust in the datasets if it enables them to re-create the data or metrics following the information provided, "*Replicability of your study is one key for gaining trust to metadata sharing and repurposing*" (P5).

4.2.3 Structural metadata

All participants experienced a lack of structural metadata in their work with repurposed datasets, with the most common observations outlined in the following.

Accessibility (C). Apart from general metadata insufficiency, four participants stated that lack of access is a specific problem when it comes to structural metadata. Participants reported that typically structural metadata is not shared, resulting in the need to reach out to the team who is responsible for the given datasets to request the data model, which relates to the need to talk to data owners as mentioned in section 4.2.2. P4 indicated: "*It's not that it doesn't exist; in fact, it's probably documented quite well*". Structural metadata often can be retrieved from existing databases or data management platforms, but the lack of active sharing of this metadata causes additional work and communication.

Establishing shared understanding (A). Understanding the structure of unexplored data is essential for data analytics team and for business stakeholders as they are often interested in relationships between multiple tables rather than technical details. Three participants indicated that having good quality structural metadata helps to recreate a trustworthy data model and eliminates subjective interpretation of repurposed data. P6 illustrated their experience of using available structural metadata to create models that are suitable to share with business stakeholders: "*Then we start conceptual modelling. So that is essentially looking very high level, what business entities is what we call them, what data sets, what are the relationships […] And that's how we only talk to business stakeholders*". More specifically, P6 reported "*particularly their data model*" is the key to the communication between the analytics team and the business stakeholders. Presenting data models and the relationships between data objects (attributes, tables, or datasets) in the report enhances the information consumption for business stakeholders.

4.3 User Metadata Preferences

In the following, we describe the participants' preferences towards the format, presentation style or specific elements of metadata that would assist them to assess the data quality of repurposed datasets. We note that two participants advised they do not have a specific preference for the presentation of metadata, indicating that any available metadata would be useful, in whatever form. These two participants shared their experience with having to deal with no or very poor metadata. Specifically, P2 advised: "*there's generally not much coming so there's not much to base it on […] anything is better than nothing"*. P5 advised: *"I don't think there's a specific preference, because it's the content that's more important than the way it's presented […] I don't think there is anything specific that would make it easy or difficult, actually just having information was a happy day"*.

Interactive dashboards (P). When asked for an ideal approach to view metadata and data, five participants indicated a preference for an interactive dashboard to assist their work in evaluating repurposed datasets. P1, who considered themselves a visual learner, indicated their preference of: "*interactive dashboard, that I can narrow down (to) what I'm interested in, I have that as an artifact and persist in a way, that allows me at least to do some discovery and learn about the data through a different mechanism"*. The ability to interactively drill down and drill across datasets is widely preferred in our study, P2 advised: *" being able to find objects, and then drill. So, that's kind of my ideal, I think,*

find an object in the data flow, and then drill down to find, this is the metadata initially, and then you can drill further to find underlying objects". P7 identified a preference in terms of similarity with existing tools in the market, such as Power BI, Tableau and Qlik, "*loading the raw data into a visualisation tool can help. Like, just playing around with the data doing some different features and looking at what that is showing*". However, participants noted that the usability of metadata at present is relatively immature in existing tools, indicating the ability to "*extract metadata*" (P7) from the datasets and "*beautifully visualise*" (P6) supplementary as well as extracted metadata elements in the platform as wish lists for future tool functionality. Additionally, P3 reported, "*a web application*" to run ad-hoc queries, the ability to extract results to Excel, and a "*user-friendly interface*" as the preference.

We note that two participants (P2 and P6) further showed a preference for data models, e.g., Entity Relationship Diagrams, and Logical Data Models.

4.4 Challenges and Suggestions

In addition to existing challenges, participants also made suggestions for improvement to current practice.

Best practice and data governance (P). Four participants discussed the importance of keeping clear documentation when creating data or repurposing data, to ensure others can have good metadata for future uses. P4 advised: "*I can at least produce an audit trail that says to me how we got there, so I can prove some lineage. […] when we do it the lineages, someone can follow that through the steps are taken, when other people do it, they're not necessarily (clearly documenting their transformation steps)"*. P5 illustrated the importance of following best practice and keeping good documentation: *"more and more teams are doing this (good documentation) because of the information development management framework, it's actually pretty clear document in terms of who's responsible for the data"*. Through better data governance, users can reduce data quality issues and preventing "*garbage in, garbage out*" problem (P4).

Common protocol (P). Three participants reported that the lack of common protocols, standards, approaches, methodologies, plays a big part in the difficulties inherent in the reuse of complex, unfamiliar data. "*Everyone thinks in different ways […] everyone has different standards of data preparation, data standardisation*" (P5). Not having protocols or standards to evaluate repurposed data can lead to wrong conclusions and discrepancies, as DQ issues often are not well described in the supplementary information. The evaluation then depends on the reliability of the approach adopted and the personal experience of users. P1 advised that data analysts in their team generally are "*not using any particular module or framework to do that (repurposing). Just more personal experience*". The participants expressed a need for a common protocol for data collection, data preparation and standardisation – e.g., P5 indicated that it is important to have "*clear standards in terms of data collection*", which would improve the quality of data in the initial collection process as well as the quality of its metadata. The experience of creating an industry consortium is also shared by P5, which helped the development of a uniform protocol after they "*put a lot of effort into creating pipelines for data preparation and standardisation in data collection*".

Platform and tools (P). Three participants suggested a single platform to store, share, access and manage data and metadata would be beneficial for repurposing data. Because users adopt different tools or techniques at different stages of evaluating repurposed data, a single platform would increase efficiency by bringing data creators and data consumers together. P2 indicated such a platform would enable *"owners driving producers and capturing metadata at the front end and then making that available throughout the business, throughout the lifecycle. It brings it right down to the users and the entire data management sphere, for one of a better word, governance, quality, and management into a single platform"*. P4 advised that their team is developing such a platform as a key initiative: "*one of the components of that (platform) is a data lake that (stores data) in its purest form. But it's more than that; it's also about creating the data domains that people can use the various lenses, you might want to view something. You might want to look at the same data from a workforce lens versus a capital funding lens, etc"*.

5 Conclusions

With the rise of repurposing data, the evaluation of the quality of repurposed data takes on high importance. This paper is motivated by the lack of empirical evidence on the role of metadata in evaluating repurposed datasets. By collecting rich qualitative data through semi-structured interviews with key stakeholders, we provide principal findings in this regard. These include five challenges, *viz.*

lack of adequate metadata, waste of resources, inconsistent format, discoverability, and accessibility, that arise when assessing the quality of repurposed datasets in practice. Further, our study identified five main approaches, namely manual and ad-hoc profiling, personal interpretation, communication, data quality proxy, and establishing shared understanding; and four preferences or suggestions identified by our study participants, namely interactive dashboards, best practice and data governance, common protocol, and platform and tools.

Although this paper is an initial step in understanding the role of metadata in evaluating repurposed data, where little empirical evidence exists, our study is not without limitations. First, our study involved seven participants. While this is within the range of the number of participants in similar exploratory studies, involving more participants, and from a variety of sectors, might lead to the identification of further insights. Second, as with all qualitative studies, there is a risk of subjective bias in the coding of the rich interview data. To mitigate this risk, we employed a dual coding approach, using a high-level set of *a priori* codes and allowing the code structure to evolve.

6 References

Abedjan, Z., Golab, L., and Naumann, F. 2015. "Profiling Relational Data: A Survey," *The VLDB Journal* (24:4), pp. 557–581.

Aljumaili, M., Karim, R., and Tretten, P. 2016. "Metadata-Based Data Quality Assessment," *VINE Journal of Information and Knowledge Management Systems* (46:2), Emerald Group Publishing Limited, pp. 232–250.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys (CSUR)* (41:3), ACM New York, NY, USA, pp. 1–52.

Batini, C., and Scannapieco, M. 2016. "Data and Information Quality," *Cham, Switzerland: Springer International Publishing*, Springer.

Belkin, R., and Patil, D. 2018. *Everything We Wish We'd Known about Building Data Products*.

Cichy, C., and Rass, S. 2019. "An Overview of Data Quality Frameworks," *IEEE Access* (7), pp. 24634–24648.

Clarke, R. 2016. "Big Data, Big Risks," *Information Systems Journal* (26:1), pp. 77–90.

Farinha, J., Trigueiros, M. J., and Belo, O. 2009. "Using Inheritance in a Metadata Based Approach to Data Quality Assessment," in *Proceedings of the First International Workshop on Model Driven Service Engineering and Data Quality and Security*, New York, NY, USA: Association for Computing Machinery, November 6, pp. 1–8.

Gill, T., Gilliland, A. J., Whalen, M., and Woodley, M. S. 2008. *Introduction to Metadata*, Getty Publications.

Jaya, I., Sidi, F., Affendey, L., Jabar, M., and Ishak, I. 2019. "SYSTEMATIC REVIEW OF DATA QUALITY RESEARCH," *Journal of Theoretical and Applied Information Technology* (97), p. 3043.

Jayawardene, V., Sadiq, S., and Indulska, M. 2015. *An Analysis of Data Quality Dimensions*.

Jeusfeld, M. A., Quix, C., and Jarke, M. 1998. "Design and Analysis of Quality Information for Data Warehouses," in *International Conference on Conceptual Modeling*, Springer, pp. 349–362.

Krishnan, S., Haas, D., Franklin, M. J., and Wu, E. 2016. "Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pp. 1–5.

Lagoze, C., Lynch, C. A., and Daniel Jr, R. 1996. "The Warwick Framework: A Container Architecture for Aggregating Sets OfMetadata," Cornell University.

Lee, Y. W., Pipino, L., Funk, J. D., and Wang, R. Y. 2006. *Journey to Data Quality*, MIT press Cambridge.

Méndez, E., and Hooland, S. van. 2014. "METADATA TYPOLOGY AND METADATA USES" in *Handbook of Metadata, Semantics and Ontologies*, World Scientific, pp. 9–39. (https://doi.org/10.1142/9789812836304_0002).

Miles, M. B., and Huberman, A. M. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*, sage.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. 2015. "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research," *Administration and Policy in Mental Health and Mental Health Services Research* (42:5), Springer, pp. 533–544.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data Quality Assessment," *Communications of the ACM* (45:4), ACM New York, NY, USA, pp. 211–218.

Redman, T. C. 2018. "If Your Data Is Bad, Your Machine Learning Tools Are Useless," *Harvard Business Review* (2).

Richter, I., Raith, F., and Weber, M. 2016. "Problems in Agile Global Software Engineering Projects Especially within Traditionally Organised Corporations: [An Exploratory Semi-Structured Interview Study]," in *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, New York, NY, USA: Association for Computing Machinery, July 20, pp. 33–43. (https://doi.org/10.1145/2948992.2949019).

Sadiq, S., and Indulska, M. 2017. "Open Data: Quality over Quantity," *International Journal of Information Management* (37:3), pp. 150–154.

Sawadogo, P., and Darmont, J. 2021. "On Data Lake Architectures and Metadata Management," *Journal of Intelligent Information Systems* (56:1), pp. 97–120.

Sebastian-Coleman, L. 2012. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*, Newnes.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B., Pagan, A., and Xu, S. 2013. "Data Curation at Scale: The Data Tamer System.," *Cidr* (Vol. 2013).

Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C. 2007. "A Framework for Information Quality Assessment," *Journal of the American Society for Information Science and Technology* (58:12), Wiley Online Library, pp. 1720–1733.

Thomas, D. R. 2006. "A General Inductive Approach for Analyzing Qualitative Evaluation Data," *American Journal of Evaluation* (27:2), pp. 237–246.

Wang, R. Y. 1998. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), ACM New York, NY, USA, pp. 58–65.

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), Taylor & Francis, pp. 5–33.

Wang, R. Y., Ziad, M., and Lee, Y. W. 2002. "Extending the ER Model to Represent Data Quality Requirements," *Data Quality*, Springer, pp. 37–48.

Zhang, J., Wen, Q., and Zhang, H. 2009. "The Research in Improving the Quality of DW Data: The Job-Scheduling and Checking Based Program in Upgrading DW Performance," in *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, pp. 1–4.

Zhang, R., Indulska, M., and Sadiq, S. 2019. "Discovering Data Quality Problems: The Case of Repurposed Data," *Business & Information Systems Engineering* (61:5), pp. 575–593.

Zhu, H., Madnick, S. E., Lee, Y. W., and Wang, R. Y. 2014. *Data and Information Quality Research: Its Evolution and Future.*

Acknowledgements

This study was supported by the Australian Research Council through ARC Discovery Grant DP190102141.

Copyright