



School of Electrical Engineering, Computational
and Mathematical Sciences

PhD. Candidacy Proposal – Milestone 1

**Attribute-Centric Approach for Data Quality Assessment in
Diverse Datasets and Domains**

Marcelo Valentim Silva

20882194

Primary supervisor – Dr Johannes Herrmann

Co-supervisor – Dr Valerie Maxville

Committee Chairperson – Prof Andrew Rohl

"Simplicity is about subtracting the obvious and adding the meaningful." [Maeda, 2006]

Attribute-Centric Approach for Data Quality Assessment in Diverse Datasets and Domains

Marcelo Valentim Silva

Supervisors: Dr Johannes Herrmann, Dr Valerie Maxville

July 2023

Abstract

In an era where data-driven decision-making is increasingly prevalent across various sectors, ensuring data quality has become a top priority. However, data quality problems persist, leading to inaccuracies and errors that can significantly impact outcomes, such as misinformed business strategies or flawed scientific research. While there are existing methods for data quality assessment, there is potential for further exploration and improvement in leveraging attribute labels (or column names in tables), their metadata, and content across diverse datasets and domains. This research addresses this gap by developing a comprehensive Attribute-Based Data Quality Assessment Framework (A-DQAF). A key feature of A-DQAF is its emphasis on well-known data quality issues and their corresponding data quality dimension violations, making the framework comprehensive and adaptable. The proposed framework will employ advanced natural language understanding techniques and machine learning to analyse attribute information and content, identify data quality issues and recommend improvements. Furthermore, this research will develop domain-specific heuristics, informed by a comprehensive Data Error Catalogue, to address unique challenges in data quality assessment across various fields. Preliminary results, detailed in the appendices, indicate the successful identification and mitigation of several common data quality issues, demonstrating the potential of this approach. This research is expected to make a significant contribution to the field of data quality assessment, enhancing the accuracy and effectiveness of data-driven decision-making.

1 Introduction

1.1 Rationale

The growing importance of data in decision-making has placed increasing emphasis on the need for high-quality data. Attribute information, such as labels, names, or descriptions, is often neglected in data quality assessments. However, these elements harbour valuable semantic information that can be instrumental in identifying data quality issues related to their content, such as ID, Name, Percentage, and Date. This research is dedicated to creating a robust, adaptable framework for data quality assessment to harness the power of attribute information, their associated metadata, and content across a broad spectrum of datasets and domains. The goal is to enhance the reliability and effectiveness of data-driven decisions by ensuring the quality of the underlying data.

1.2 Problem Statement

While data quality is critical for effective decision-making, existing methodologies for data quality assessment need to utilise the potential of attribute names, metadata, and content in improving data quality. Furthermore, their adaptability and scope could handle diverse datasets and domain variability better. There is a clear gap in the field: we need a robust and adaptable framework that utilises attribute information to detect data quality issues such as incorrect data types, missing values, and inconsistent formatting but also recommends suitable attribute names based on content analysis. Addressing this gap will improve confidence in given datasets, enhance the accuracy of insights, improve decision-making, and allow data scientists and analysts to focus on more strategic, value-adding tasks.

1.3 Aim

This research aims to develop an efficient, effective, and adaptable framework for data quality assessment using attribute names and their related metadata and content in diverse datasets and data domains.

1.4 Research Question

This is the research question that guides this study:

“How can attribute names, associated metadata, and content analysis be leveraged to identify data quality issues across various datasets and domains and suggest suitable refinements to give greater data confidence?”

1.5 Objectives/Research Elements

Unpacking the Research Question, we have defined the following Research Elements:

- **Research Element 1 (RE1):** Development of Framework
 - Develop a comprehensive tool for data quality assessment that leverages attribute labels, metadata, and content across diverse datasets and domains.
 - Collect and standardise datasets.
 - Compile and analyse a list of domain-specific attribute words.
- **Research Element 2 (RE2):** Development of Data Error Catalogue and Heuristics
 - Develop a Data Error Catalogue that will inform the development of heuristics and rules for the framework, enhancing its ability to identify and address data quality issues.
 - Investigate methods to suggest attribute names based on content analysis.
- **Research Element 3 (RE3):** Documentation, Validation and Dissemination of Research Findings and Impact Assessment
 - Document the development process, results, and tools associated with the framework.
 - Validate the effectiveness and robustness of the framework through rigorous testing and evaluation.
 - Share the findings and tools with the research community and relevant stakeholders through platforms like GitHub and academic conferences or journals.
 - Investigate the impact of improved data quality on the interpretability and usability of datasets.

1.6 Significance

This research holds the potential to make a significant contribution to the field of data quality assessment. It will address a critical gap in the field by developing an innovative framework that leverages a deeper understanding of attribute information for data quality assessment. The anticipated outcomes will influence the data science community by enhancing data quality and enriching the academic discourse around the role of attributes and associated metadata in data quality management.

1.7 Delimitation and Limitations

This research focuses on using attribute names, metadata, and content for data quality assessment across various datasets. The limitations include data files that can't be analysed automatically, such as complex files like documents, images, or video files. Furthermore, there may be limitations in applying natural language and machine learning techniques to attribute names and metadata due to the complexity of language and semantics. Besides that, the dataset collection process may also face challenges like access restrictions and data privacy concerns.

1.8 Theoretical Framework

The theoretical framework of this research integrates concepts from data quality, NLP, and machine learning to detect and address data quality issues at an individual or collective attribute level in datasets. It will use attribute names, which contain semantic information related to their content, and NLP and machine learning to analyse attribute names, related metadata, and content. This will be a foundation for a novel approach to data quality assessment in diverse datasets and domains.

1.9 Assumptions

- Attribute labels in datasets, or column names in tables, associated descriptions, or metadata, often carry meaningful semantic information related to their content, including many data quality issues that their analysis can obtain.
- The developed framework can be adapted to various domains and use cases.

1.10 Research Approach

This research develops a data quality assessment and improvement framework using a pragmatic approach incorporating quantitative and qualitative methods, including heuristics, advanced NLP, and machine learning

techniques. It mirrors the Unified Software Development Process methodology with an iterative approach for continuous refinement based on feedback and results.

Key phases include Inception, Elaboration, Construction, and Transition. Adopted methodologies include correlational and experimental quantitative methods and content analysis, grounded theory, and systematic review qualitative methods.

1.11 Research Contributions

These are the contributions that this research will produce:

- The Framework for data quality assessment,
- The domain-specific attribute words compilation,
- The Data Error Catalogue,
- The heuristics and rules development,
- The actual implementation of the developed methods and heuristics to pre-process datasets, analyse attribute labels and content, suggest attribute labels based on content, and identify data quality issues,
- The documentation and dissemination of the Framework, developed heuristics, rules, and any created software, along with an explanation of the Data Error Catalogue's contents and usage.

2 Literature Review

Issues with data quality are not new. In 1992, Computerworld magazine reported that over 60% of the companies had data quality problems that year [Wang et al., 1992]. We have moved from human knowledge doubling every century to doubling every year in 1982 [Fuller & Kuromiya, 1982], and storage of new data in the period 2021-2025 is estimated to be greater than twice the amount stored since data storage began [IDC, 2021]. This emphasises the critical need for robust data quality management practices and the demand for comprehensive strategies and tools to address these prevalent issues.

2.1 Data quality assessment

Data Quality is defined as '*data that are fit for use by data consumer*' [Wang & Strong, 1996]. It implies that the data should meet the requirements of its users. It should be accurate, reliable, timely, relevant, and complete, which needs to be assessed.

Data quality assessments can be produced subjectively by interviews and questionnaires with the stakeholders or objectively by measurements of the data, such as in the measure of the completeness dimension by obtaining the missing values in a column of a table [Pipino et al., 2002]. They also present data quality dimensions that can be obtained subjectively, such as Believability, Interpretability and Relevancy or objectively, such as Completeness, Free-of-Error (also called, Accuracy), and Timeliness.

In their review of data quality methodologies, [Batini et al., 2009] define a data quality methodology as "*a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data*". Thus, the application context and a rational (repeatable) process must be included. Key stages in data analysis are obtaining, describing, and cleaning unfamiliar datasets [Dasu & Johnson, 2003].

Processes such as Exploratory Data Analysis and Data Profiling help obtain metadata. However, data descriptions and business rules are extremely helpful if available. Frequently, the documentation of metadata could be more thorough and accurate. When the time comes to scrutinise the data, doubts arise about its integrity and reliability [Dasu & Johnson, 2003]. If problems in data are not discovered, the misleading results could result in financial losses, upset customers and potentially the loss of property or lives. This reinforces the importance of good data quality.

2.2 Data quality dimensions and attribute analysis

Data is usually available as datasets, text files or tables in databases. Attributes, also known as columns or fields in tables, are usually related to the features or characteristics of the associated data. Attributes or columns usually have a label or a name, which may be a word, phrase or some defined text that should infer meaning to the characteristic.

Unfortunately, it is common to encounter tables where the names of attributes need to provide meaningful information, hindering a clear comprehension of their content. Examples of unhelpful attribute names are “V1”, “V2”, or “AQ0023”. Often this is due to older application development tools and Data Base Management Systems (DBMS) that restricted the length of a column name, making it difficult to use descriptive and easy-to-understand names. Data may also have descriptions or metadata, providing information about the data. This may also happen at the column level, including their name, data type or domain, constraints, and comments about the meaning of the column. Data Profiles may also provide metadata for columns, enabling better data understanding.

Some data quality dimensions associated with attributes include [Batini & Scannapieco, 2016]:

- **Accuracy** is the closeness between a recorded value and the represented real-world item or state. Some instances, like numbers or incorrect values for a ‘State’ attribute, can be easily discovered with access to the data domain (like a list of all states).
- **Completeness** measures the incidence of missing values for an attribute. For this dimension, attribute names may be analysed if they contain the words “ID” or “name”.
- **Consistency** “captures the violation of semantic rules defined over (a set of) data items”. Some of these semantic rules are Integrity constraints. Among integrity constraints, these are the main types of dependencies:
 - Key dependency. This is related to the Primary Key (PK) in databases. Usually, PKs should uniquely identify the contents of the tables/files. When key dependency constraints are enforced, the tables will not contain duplicated data, enabling the best possible data quality.
 - Inclusion dependency. It is also known as a referential constraint. “A foreign key (FK) constraint is an example of inclusion dependency”, meaning that the information in a table’s Primary Key columns can refer to the columns in a Foreign Key in another related table.
 - Functional dependency. It allows dependencies such as ZIP and City to be related to a State.

In addition, [Dasu and Johnson, 2003] identified another data quality dimension related to attribute names:

- **Timeliness:** the currency of the data is critical for attributes that change over time but not for those that do not change. Classifying the names of attributes that change over time, such as “weight” or “age”, can be easy. “Species type” is a column name that does not change over time.

In the data cleaning stage, we detect and remove errors from the data to improve its quality [Rahm & Do, 2000].

Among Single-Source Data Problems, when a file or table is analysed alone, [Rahm & Do, 2000] show possible problems/ “dirty data” related to column names, including illegal values, value violations, missing values, duplications, contradictions, and wrong references (see Table 1).

Data Problem	Example
Illegal values	date with an illegal value of 30.13.70
Violated attribute dependencies	age different than the value of (current date – birth date)
Uniqueness violation	two different names for a single ID value
Referential integrity violation	department value non existing in the department dimension table
Missing value	a possible default value in the telephone column
Misspelling	city content
Abbreviated values	content of columns
Embedded values in a field	when name and birth date are found
Misfielded values	when a country name is found in the city column
Violated attribute dependencies	when the ZIP code is inconsistent with the city name
Word transpositions	when names sometimes include the first name followed by the last name, and other times the opposite
Duplicated records	when the same employee appears twice on the same table
Contradicting records	when different values describe the same real value
Wrong references	when the department number is not the correct one

Table 1 - Single-Source Data Problems [Rahm & Do, 2000]

When considering Multi-Source Problems, when more than one file or table is being analysed, [Rahm & Do, 2000] state that the main problems for columns are naming and structural conflicts. Naming conflicts occur when different names are used for the same case (synonyms), e.g., tables Customer and Client, columns Sex and Gender. Structural conflicts occur in cases when the name of the client is full in one table and separated as FirstName and

LastName in another. Other common problematic cases exist when the representation varies between tables, such as gender (0 and 1 vs “F” and “M”). Heuristics for evaluating all the cases described have been published in Rahm & Do [2000].

A summary of 22 well-known Data Quality Issues, including Missing data, Incorrect data, and Misspellings, has been published by [Visengeriyeva and Abedjan, 2020]. This is augmented with the associated Data Quality Dimension violations in Appendix 1. Along with considering single and multi-source problems, we have a wide lens for assessing data quality.

2.3 Natural Language Processing (NLP) and Machine Learning

Addressing data quality necessitates the examination of both coded and textual data. Natural Language Processing (NLP) techniques such as tokenisation, stemming, and lemmatisation are crucial for pre-processing and normalising text, facilitating data cleaning. Tokenisation isolates data by eliminating punctuation and whitespace while stemming reduces words to their root form, stripping off variations that do not impact meaning. Lemmatisation then connects these processed words to known dictionary words, enhancing the understanding of the data [Bird et al., 2009]. NLP techniques such as topic modelling (e.g., Latent Dirichlet Allocation) can also identify common themes in textual data, indicating data quality issues [Blei et al., 2003].

In addition to NLP, ontologies can provide a structured representation of knowledge, which can be particularly useful in understanding and categorising attribute information [On Behavioral et al., 2022].

Once the data has been prepared, machine learning techniques can be applied to identify and take advantage of patterns in the data. There are supervised and unsupervised machine learning techniques. The choice between them depends on the nature of the data and the specific research objectives.

Unsupervised learning techniques, such as K-means and DBSCAN clustering algorithms, can group similar attribute labels, aiding in identifying data quality issues [Ester et al., 1996]. These techniques are particularly useful when exploring the data and looking for patterns without a specific prediction task.

On the other hand, supervised learning techniques, including classification algorithms like logistic regression, decision trees, and support vector machines, can predict the quality of new, unseen data based on the patterns learned from the training data [De Cock et al., 2019]. These techniques are suitable when we have labelled data and a clear prediction task, such as identifying specific data quality issues.

Pre-trained language models like BERT and GPT series (including GPT-2, GPT-3, GPT-4, and ChatGPT) can be fine-tuned for various tasks related to data quality assessment. BERT, for instance, can identify semantic relationships between attribute labels and the data they contain, enabling the discovery of data quality issues [Devlin et al., 2019]. Similarly, GPT and its successors can assess semantic similarity, helping to identify patterns and relationships between attribute labels and potential data quality issues [Radford & Narasimhan, 2018].

This research aims to develop a comprehensive and adaptable framework for data quality assessment across diverse datasets and domains by leveraging these NLP and machine learning techniques maximises. The choice of these techniques is influenced by the nature of the data and the specific objectives of the research, and they will be applied in a way that maximises their potential for identifying and addressing data quality issues.

3 Methodology

This research involves the development of a framework for assessing and improving data quality. Case study data will be used in developing the techniques and again in evaluating their effectiveness.

3.1 Research Philosophies

There are five major research philosophies: Positivism, Critical Realism, Interpretivism, Postmodernism, and Pragmatism [Saunders et al., 2016]. Each philosophy offers a different lens through which to view and interpret reality, and thus, influences the research approach.

Positivism, emphasising observable and measurable facts, was deemed less suitable for this research due to its limited scope for interpreting attribute information's complex and nuanced nature and data quality issues. Critical Realism, while acknowledging a reality independent of human perceptions, often requires a level of abstraction that may not align with this research's practical, solution-oriented nature.

Interpretivism, which prioritises subjective interpretation and understanding, could offer valuable insights into interpreting attribute information. However, it might not fully support the quantitative aspects of data quality assessment and the use of machine learning methods in this research.

Postmodernism, with its doubt towards grand theories and ideologies, could challenge the notion of a 'one-size-fits-all' framework for data quality assessment. However, this research operates on the premise that while data and contexts may vary, common data quality issues can be addressed through a comprehensive framework.

Therefore, Pragmatism was chosen as the most suitable philosophy for this research. It allows for integrating multiple methods and perspectives to address the research problem. In analysing attribute information and detecting data quality issues, pragmatism permits employing quantitative and qualitative approaches in developing and evaluating techniques, including heuristics, advanced NLP, and machine learning methods. This flexibility and focus on practical outcomes can help develop a comprehensive and effective solution for identifying data quality issues in diverse datasets and domains while remaining open to incorporating new ideas and methods as the research progresses.

3.2 Research Methodology

This research involves a framework's scoping, planning, implementation, and evaluation. In this, there are similarities and synergies with software development. The commonly used approaches to developing software are Waterfall; Agile Family such as Rapid Application Development (RAD) and Extreme Programming (XP); and Unified Process Family (UP) [Young, 2013]. An iterative approach is preferred to formalise the cycle of plan-implement-test-evaluate to have reflection and improvement throughout the project. As an individual project, XP and Agile's structure for teams is unnecessary, and client interaction is not required. The approach chosen for this research is Unified Software Development Process or Unified Process (UP) [Jacobson et al., 1999] to allow for continuous improvement, refinement, and adaptation of the methods based on feedback and results obtained during an iterative process. One of the more popular versions of UP is the Rational Unified Process (RUP). Figure 1 illustrates how the various phases of a software project overlap when using UP, allowing flexibility and feedback.

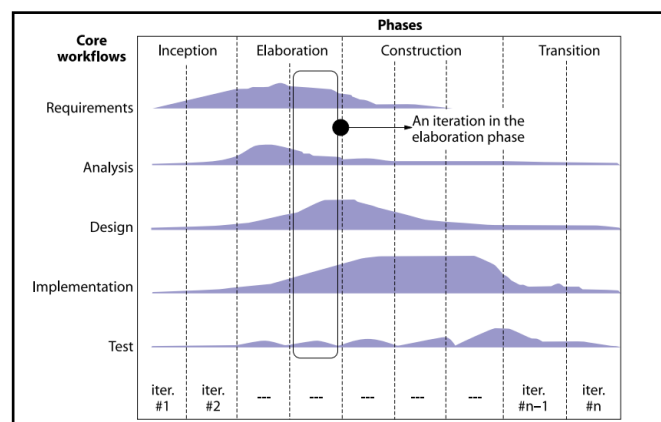


Figure 1 – Unified Process [Jacobson et al., 1999]

The Unified Software Development Process or Unified Process (UP) (Figure 1) contains the following phases:

- **Inception** sets the project parameters, delineating the project scope, identifying requirements (both functional and non-functional), and assessing potential risks. It provides a high-level overview but with sufficient detail to facilitate work estimation.
- **Elaboration** produces a functional architecture that effectively addresses key risks and satisfies the non-functional requirements, providing a robust framework for the project.
- **Construction** methodically builds upon the established architecture, creating production-ready code. This is achieved through analysis, design, implementation, and rigorous testing of the functional requirements. There might be some iterations in this phase.
- **Transition** ensures smooth system delivery, ensuring optimal functionality and integration.

Each phase may be divided into one or more iterations, normally time boxed. UP breaks the system functionality into increments. Functionalities are delivered in each increment.

3.3 Research Methodologies adopted in the Research

1. Quantitative (Correlational): This methodology involves determining the relationship between two (or more) variables.
2. Quantitative (Experimentation): This methodology systematically investigates phenomena by gathering quantifiable data and performing statistical, mathematical, or computational techniques.
3. Qualitative (Content Analysis): This method systematically and objectively examines language to classify and describe phenomena.
4. Qualitative (Grounded Theory): This methodology involves discovering emerging data patterns to create a theory.
5. Qualitative (Systematic Review): This methodology involves a comprehensive literature review focused on a research question that tries to identify, appraise, select, and synthesise all high-quality research evidence relevant to that question.

3.4 Application of Unified Process (UP) to the Research Elements

This is how the UP methodology [Figure 1] relates to the Research Elements from Section 1.5:

UP Phases	Research Elements
Inception	RE1
Elaboration	RE1 and RE2
Construction	RE1, RE2 and RE3
Transition	RE3

Table 2 – UP Phases and Research Elements

RE1: Development of Framework
 RE2: Development of Data Error Catalogue and Heuristics
 RE3: Documentation, Validation and Dissemination of Research Findings and Impact Assessment

3.5 Ethical Issues

The Research follows the Australian Code for the Responsible Conduct of Research, and all Research Integrity Training modules have been completed, as seen in Appendix 2. Besides that, the Research Initiation Guide RIG – HIT2023-0388's first page is available in Appendix 3.

3.6 Data Management

Data Sets, Codes and Documentation produced will be stored on the Research Drive provided by Curtin University and (while the work is ongoing) on Microsoft Teams and OneDrive, with security and backup enabled by the University. They will also be provided in GitHub for easy availability for the research community. Data will be retained for at least seven years. The Research Data Management Plan is annexed in Appendix 4.

4 Preliminary work

Many steps have already been produced according to the UP methodology:

4.1 Inception Phase

Following the phases of the UP methodology, the first phase is the Inception (Research Element 1 – RE1) with the following items:

1. Sets the project parameters, delineating the project scope. This step has already been accomplished with the information described in the next section, which describes a first draft of a Data Quality Assessment Framework: The Attribute-Based Data Quality Assessment Framework (A-DQAF).
2. Identifies requirements (both functional and non-functional); this corresponds to the Requirements step. The Research Elements presented in Section 1.5 cover the functional requirements - each represents a functional requirement of the framework being developed. The non-functional requirements, such as the framework's efficiency, reliability, and scalability, are implied in the research elements and will be kept in mind throughout the research.
3. Assesses potential risks. This is integral to the Inception phase. It has already been assessed, and the potential risks for this research and suggestions for mitigating them are listed in Appendix 5.

4.1.1 Attribute-Based Data Quality Assessment Framework (A-DQAF)

These are the steps for an Attribute-Based Data Quality Assessment Framework (A-DQAF) for this PhD Research. Figure 2 shows its diagram. It considers the summary of well-known Data Quality Issues and the Data Quality Dimension violations [Visengeriyeva and Abedjan, 2020] presented in Appendix 1:

1. Data collection: Gather diverse datasets from different domains to test the generalizability of the proposed methods to obtain potential data quality issues. Information from 622 datasets available in the UCI Machine Learning Repository Catalogue has been obtained. The Catalogue is a well-known source of research datasets encompassing various domains (called areas in the catalogue) and data types. The code used to download this information and the dataset created is available on GitHub [Silva, 2023a]. This allowed an in-depth Data Quality Assessment [Silva, 2023b], which is about to be sent to UCI, the University of California, Irvine, to improve their excellent and extremely important Catalogue of research data sets that has had millions of accesses over time.

For the first part of this research, ten datasets from five different areas were chosen from the Data Quality assessment produced on the 622 datasets from the UCI Catalogue of Datasets, and their attribute labels were obtained. The decision to choose these ten datasets and the spreadsheet with the top 30 datasets containing the ten chosen is presented in Appendix 6. The columns from the ten datasets chosen are presented in Appendix 7.

Appendix 8 lists other catalogues of datasets besides UCI that may be used during this PhD Research.

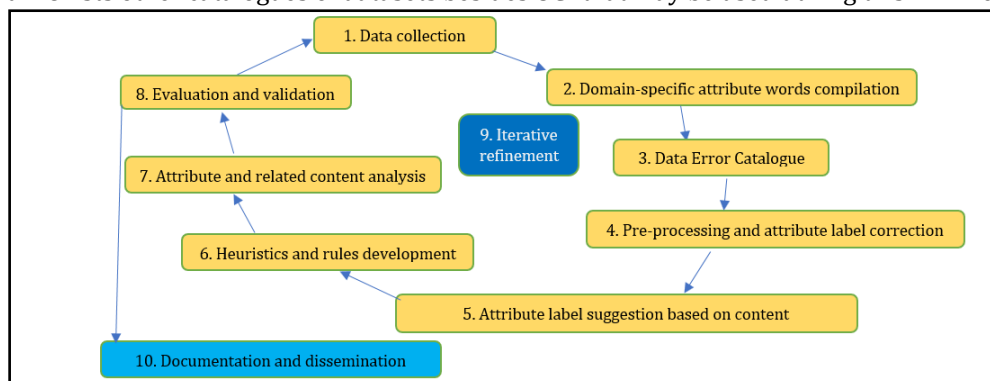


Figure 2 - Attribute-Based Data Quality Assessment Framework (A-DQAF)

2. Domain-specific attribute words compilation: Consolidate the attribute names/words extracted from the datasets already selected across different domains. Conduct a preliminary analysis of these attribute names to understand their characteristics, usage, and common patterns or inconsistencies. A first attempt is shown in Appendix 9, with the first ten datasets described previously. Alternatively, consult domain experts or conduct a focused literature review to complement the findings, especially if unfamiliar patterns emerge or further domain understanding is needed. The primary methodology adopted for this phase is Qualitative (Content Analysis).

3. **Data Error Catalogue:** Develop a Data Error Catalogue that contains all kinds of obtained data errors associated with their Data Quality Issues and Dimensions. This catalogue will serve as a comprehensive resource for understanding and categorising data errors in the datasets under examination. Appendices 10, 11 and 12 exhibit Data Errors from different sources and have all been correlated with the summary of well-known Data Quality Issues and the Data Quality Dimension violations presented in Appendix 1. They are all grouped in Appendix 13 for a first draft of the Data Error Catalogue. The methodology adopted for this step is Qualitative (Content Analysis).
4. **Pre-processing and attribute label correction:** Apply methods to pre-process the datasets to standardise and clean the attribute labels, making them more interpretable and meaningful, addressing misspellings, ambiguous data, extraneous data, and misfielded values. Apply NLP techniques like tokenisation, stemming, and lemmatisation to pre-process and normalise attribute labels; or unsupervised learning techniques, such as clustering algorithms (e.g., K-means, DBSCAN), to group similar attributes. Obtain associated metadata through data profiling. The methodology adopted here is Qualitative (Content Analysis).
5. **Attribute label suggestion based on content:** Utilize advanced NLP techniques and machine learning algorithms to analyse the content of the attributes and suggest appropriate attribute labels. This approach will be instrumental in addressing issues like incorrect references, different word orderings, and the use of synonyms. The methodology adopted here is Quantitative (Experimentation).
6. **Heuristics and rules development:** Formulate domain-specific heuristics and rules to enhance the framework, making it more adaptable to different domains and use cases. This process should use the Data Error Catalogue developed to consider data quality issues like different units/ representations, domain violation, FD violation, and wrong data type. An initial list of rules according to some attributes is informed in Appendix 14, and a list of expected formats across multiple domains is presented in Appendix 15. The methodology adopted here is Qualitative (Ground Theory).
7. **Attribute and related content analysis:** Apply advanced NLP techniques and machine learning algorithms to analyse the relationships between attributes and their content, identifying potential data quality issues. Various unsupervised and supervised learning techniques and state-of-the-art NLP methods will be employed to perform this analysis. An initial list of items to be followed for this content analysis is shown in Appendix 16. The methodology adopted here is Quantitative (Correlational).
8. **Evaluation and validation:** Evaluate and validate the developed methods using a test set of datasets, measuring their effectiveness in identifying and addressing data quality issues. Incorporate a detailed error reporting system that maps detected data quality issues directly to the corresponding entries in the developed Data Error Catalogue. Establish metrics that would effectively measure the success of handling issues like structural conflicts, temporal mismatch, and different encoding formats. The methodology here is Quantitative (Experimentation).
9. **Iterative refinement:** Continuously refine and improve the methodology based on the evaluation results and feedback from domain experts.
10. **Documentation and dissemination:** Document the Attribute-Based Data Quality Assessment Framework results, and any developed tools, sharing the findings with the research community and relevant stakeholders at GitHub.

Currently, items 1, 2, 3, 6, and 7 have gone through some of the first iteration, and part of the results of item 10 are on GitHub and Teams.

4.2 Elaboration Phase

- **Planning:** Detail the approach for building a list of domain-specific attribute words (RE1).
- **Requirements:** Identify the additional requirements for building the Data Error Catalogue and developing heuristics for suggesting suitable attribute labels (RE2).
- **Analysis and Design:** Conduct a preliminary analysis of the datasets to understand their characteristics, usage, common patterns, or inconsistencies (RE1). Compile a list of domain-specific attribute words (RE1).
- **Implementation:** Begin the development of the data quality assessment framework (RE1). Implement the initial data error catalogue version and develop heuristics for suggesting suitable attribute labels (RE2).
- **Testing and Evaluation:** Evaluate the initial implementation of the framework using a small subset of datasets. Test the functional architecture and refine it based on the results.

- **Deployment:** Apply the framework to the gathered datasets for a preliminary validation (RE1). Make necessary adjustments and refinements based on the deployment results.

4.3 Construction Phase

Usually, there are more than two iterations in this Phase. One will be for the first set of datasets analysed; then, at least two more sets of datasets will be analysed in two different iterations of this phase.

- **Planning and Requirements:** Re-evaluate the datasets and incrementally update the pre-processing methods as more datasets are gathered (RE1).
- **Analysis and Design:** Continuously analyse the attribute labels and incrementally add to the list as more datasets are reviewed (RE1). Refine the data error catalogue to include the data quality issues found (RE2).
- **Implementation:** Develop strategies and techniques for detecting and correcting attribute labels (RE1). Implement the suggestion of attribute labels based on content analysis (RE2). Develop and update the Data Error Catalogue as new errors are discovered (RE2).
- **Testing and Evaluation:** Validate the developed methods using a test set of datasets (RE1). Evaluate the impact of improved data quality on the interpretability and usability of datasets (RE3).
- **Deployment:** Apply the refined framework to the collected datasets (RE1). Iteratively improve the methods for suggesting attribute labels and developing heuristics (RE2).

4.4 Transition Phase:

This phase corresponds to the developed methodologies' testing, validation, and deployment. It involves RE3, where the methodologies, results, and tools developed from the previous phases are documented and shared with the research community and stakeholders.

5 Timeline

	2022			2023				2024				2025				2026
	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
Literature Review																
Milestone 1 Preparation																
Inception																
Elaboration and Construction of first set of datasets																
Construction of second set of datasets																
Construction of third set of datasets																
Transition																
Milestone 2 Preparation																
Paper Writing																
Milestone 3 Preparation																

Table 3 - Timeline

6 Budget

Expense	Item	Year 1	Year 2	Year 3 / Year 4	Total
Conference	Travel, Accommodation, Registration – National Conference		\$1,600		\$1,600
Conference	Travel, Accommodation, Registration – International Conference			\$4,000	\$4,000
Thesis	Editing and Binding			\$300	\$300
Total					\$5,900

Table 4 - Budget

7 References

- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). [Methodologies for data quality assessment and improvement | ACM Computing Surveys](#)
- Batini, C., & Scannapieco, M. (2016). [Data and Information Quality: Dimensions, Principles and Techniques | SpringerLink](#)
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media. Available at <https://www.nltk.org/book/ch03.html>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022. Available at [blei03a.dvi \(jmlr.org\)](http://blei03a.dvi.jmlr.org)
- Dasu, T., Johnson, T., (2003). [Exploratory Data Mining and Data Cleaning \(wiley.com\)](#)
- De Cock, M., Dowsley, R., Horst, C., Katti, R., Nascimento, A. C. A., Poon, W-S., & Truex, S. (2019). [Efficient and Private Scoring of Decision Trees, Support Vector Machines and Logistic Regression Models Based on Pre-Computation | IEEE Journals & Magazine | IEEE Xplore](#). *IEEE Transactions on Dependable and Secure Computing*, 16(2), 217-230. <https://doi.org/10.1109/TDSC.2017.2679189>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). [\[PDF\] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Semantic Scholar](#)
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). [A density-based algorithm for discovering clusters in large spatial databases with noise | Proceedings of the Second International Conference on Knowledge Discovery and Data Mining \(acm.org\)](#)
- Fuller, R.B., Kuromiya, K. (1982). [Critical Path - R. Buckminster Fuller - Google Books](#)
- IDC (2021). [Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All? \(marketresearch.com\)](#)
- Jacobson, I., Booch, G., Rumbaugh, J. (1999). [The unified software development process: Jacobson, Ivar : Free Download, Borrow, and Streaming : Internet Archive](#) and [rup-unified-process.pdf \(ku.edu\)](#)
- Maeda, J., (2006). The Laws of Simplicity: Design, Technology, Business, Life available at [The Laws of Simplicity: Design, Technology, Business, Life : Maeda, John: Amazon.com.au: Books](#)
- On Behavioral, B., Beatty, A. S., Kaplan, R. M., & National Academies of Sciences, Engineering, and Medicine. (2022). [Understanding Ontologies - Ontologies in the Behavioral Sciences - NCBI Bookshelf \(nih.gov\)](#)
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). [Data quality assessment | Communications of the ACM](#)
- Radford, A., & Narasimhan, K. (2018). [\[PDF\] Improving Language Understanding by Generative Pre-Training | Semantic Scholar](#)
- Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4), 3-13. Available at [\(4\) \(PDF\) Data Cleaning: Problems and Current Approaches \(researchgate.net\)](#)
- Saunders, M., Lewis, P., & Thornhill, A. (2016). Research Methods for Business Students (7th ed.). Pearson Education Limited. Available at [\(7\) \(PDF\) Research Methods for Business Studies \(researchgate.net\)](#)
- Silva, M. V. (2023a). Code and UCI Dataset Catalogue obtained. Available at [marcelovalentimsilva/UCI-Catalog-with-all-622-Datasets: UCI Catalog with all 622 Datasets \(github.com\)](https://marcelovalentimsilva.github.io/UCI-Catalog-with-all-622-Datasets/)
- Silva, M. V. (2023b). [Data Quality Assessment of UCI Datasets Catalog | Kaggle](#)
- Visengeriyeva, L. and Abedjan, Z. (2020). [Anatomy of Metadata for Data Curation \(acm.org\)](#). *ACM Journal of Data and Information Quality (JDIQ)*.
- Wang, R. Y., Reddy, M. P., Kon, H. B. (1992). [Toward quality data: an attribute-based approach \(mit.edu\)](#)
- Wang, R. Y., Strong, D. M. (1996). [Beyond Accuracy: What Data Quality Means to Data Consumers \(jstor.org\)](#)
- Young, D. (2013). [\(1\) \(PDF\) Software Development Methodologies \(researchgate.net\)](#)

8 Appendices

8.1 Appendix 1 - Summary of Well-known Data Quality Issues and the Data Quality Dimension Violations Associated

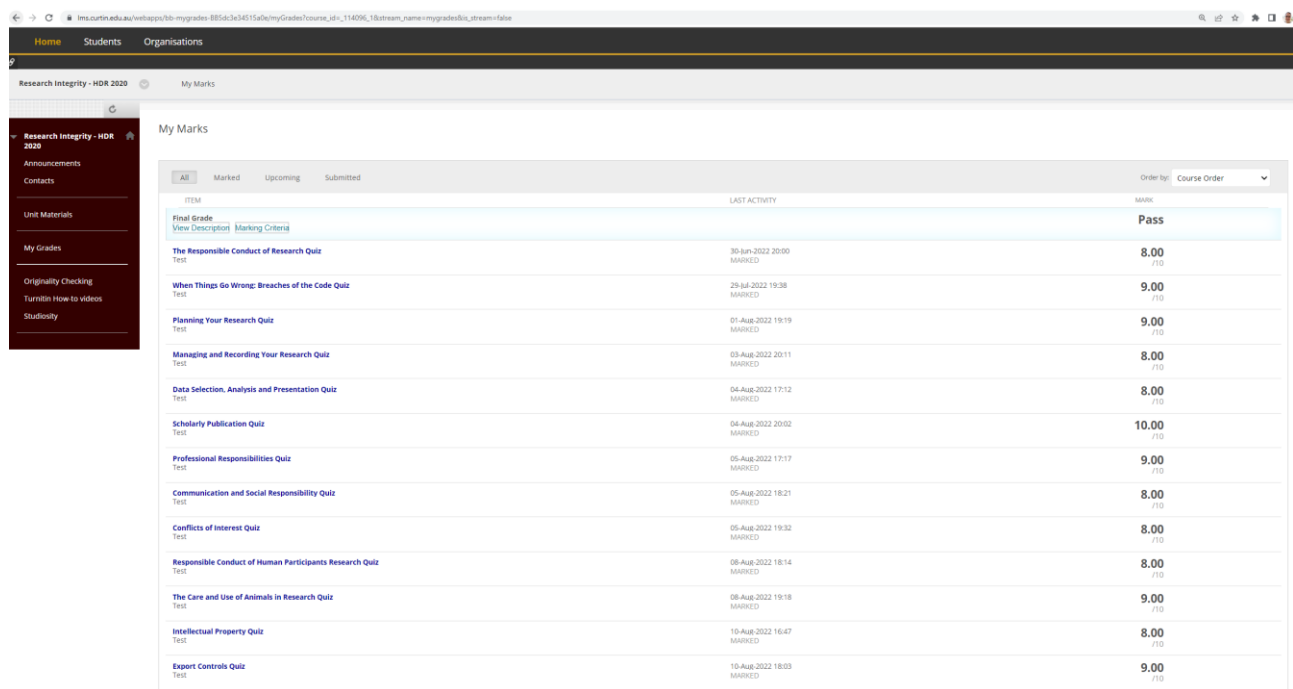
#	Data Quality Issue	Description	Data Quality Dimensions
1	Missing data	Comprises missing tuples and missing values. Tuple completeness requires that all tuples are present in the table. Missing value issue consists of either null values or disguised values. Value completeness requires that all values are present in the table, while null values indicate that the value is unknown or non-existent.	Accuracy, Completeness
2	Incorrect data	Data that differ from the values of the real-world entity (e.g., wrong date of birth).	Accuracy
3	Misspellings	Syntactic deviation of the data value from its ground truth (e.g., "Smiht" instead of "Smith").	Accuracy
4	Ambiguous data	Data values which might be interpreted in several ways (e.g., abbreviations or cryptic values).	Accuracy, Consistency
5	Extraneous data	Presence of additional data in the attribute value (e.g., the address column contains a person's name in addition to the address).	Consistency, Uniqueness
6	Outdated temporal data	Values that are obsolete or outdated.	Timeliness
7	Misfielded values	Values that are placed inside the wrong attribute.	Accuracy, Consistency, Completeness
8	Incorrect references	Entities that contain wrong information concerning the reference relation (e.g., the employee is associated with a wrong department).	Accuracy
9	Duplicates	Tuples/values that represent the same real-world entity.	Uniqueness
10	Structural conflicts	Conflicting duplicates in different sources.	Consistency, Uniqueness
11	Different word orderings	Values that violate the expected word order (e.g., first name precedes last name).	Consistency, Uniqueness
12	Different aggregation levels	Entities produced by applying different aggregation methods (e.g., entries per quartal vs. entries per year).	Accuracy, Consistency
13	Temporal mismatch	Refers to erroneous data that arise due to non-enforcement of integrity constraints for temporal data.	Accuracy, Timeliness
14	Different units/representations	Occurrence of multiple representations for the same concept (e.g., Price in different currencies).	Consistency
15	Domain violation	Values that violate semantic rules defined on the specific attribute.	Accuracy
16	FD violation	Values that violate previously specified functional dependencies.	Accuracy, Consistency
17	Wrong data type	Values that violate the data type specification of the corresponding attribute, i.e., data type constraint violation.	Consistency
18	Referential integrity violation	Tuples that violate the referential integrity constraints defined on multiple relations (e.g., missing foreign key).	Accuracy, Consistency, Completeness
19	Uniqueness violation	Duplication of values under the uniqueness constraint.	Uniqueness
20	Use of synonyms	Occurrence of synonymous representations for the same concept inside the same column (e.g., "lecturer" and "professor").	Uniqueness
21	Use of special characters (space, no space, dash, parentheses)	Refers to different representations of compound data, such as Social Security Number or phone number.	Consistency
22	Different encoding formats	Inconsistent usage of encodings for values within a dataset (e.g., ASCII or EBCDIC).	Consistency

Figure 3 - Summary of Well-known Data Quality Issues and the Data Quality Dimension Violations
[Adapted from Visengeriyeva and Abedjan, 2020]

Appendices 10, 11 and 12 show these cases associated with Data Quality Issues in attribute labels and descriptions.

From these 3 Appendices, it was obtained Appendix 14, which contains the first draft of a Data Error Catalogue with the distribution of all data errors in this document by Data Quality Dimension and Data Quality Issue from Figure 3 above.

8.2 Appendix 2 – Research Integrity courses which were taken and passed in full



ITEM	LAST ACTIVITY	MARK
Final Grade View Description Marking Criteria		Pass
The Responsible Conduct of Research Quiz Test	30-Jun-2022 20:00 MARKED	8.00 /10
When Things Go Wrong: Breaches of the Code Quiz Test	29-Jul-2022 19:38 MARKED	9.00 /10
Planning Your Research Quiz Test	01-Aug-2022 19:19 MARKED	9.00 /10
Managing and Recording Your Research Quiz Test	03-Aug-2022 20:11 MARKED	8.00 /10
Data Selection, Analysis and Presentation Quiz Test	04-Aug-2022 17:12 MARKED	8.00 /10
Scholarly Publication Quiz Test	04-Aug-2022 20:02 MARKED	10.00 /10
Professional Responsibilities Quiz Test	05-Aug-2022 17:17 MARKED	9.00 /10
Communication and Social Responsibility Quiz Test	05-Aug-2022 18:21 MARKED	8.00 /10
Conflicts of Interest Quiz Test	05-Aug-2022 19:32 MARKED	8.00 /10
Responsible Conduct of Human Participants Research Quiz Test	08-Aug-2022 18:14 MARKED	8.00 /10
The Care and Use of Animals in Research Quiz Test	08-Aug-2022 19:18 MARKED	9.00 /10
Intellectual Property Quiz Test	10-Aug-2022 16:47 MARKED	8.00 /10
Export Controls Quiz Test	10-Aug-2022 18:03 MARKED	9.00 /10

Figure 4 - Research Integrity courses

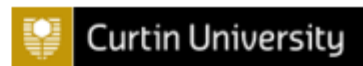
Figure 3

8.3 Appendix 3 – RIG

RIG Submission HIT2023-0388

8.4 Appendix 4 – Research Data Management Plan

Document generated: 08/07/2023



Research Data Management Plan

Attribute-Centric Approach for Data Quality Assessment in Diverse Datasets and Domains

Supervisor	Valerie Maxville
Data Management Plan Edited by	Marcelo Valentim Silva
Modified Date	8/07/2023
Data Management Plan ID	MAXVIV-SE22024
Faculty	Science and Engineering

1 Research Project Details

1.1 Research project title

Attribute-Centric Approach for Data Quality Assessment in Diverse Datasets and Domains

1.2 Research project summary

In an era where data-driven decision-making is increasingly prevalent across various sectors, ensuring data quality has become a top priority. However, data quality problems persist, leading to inaccuracies and errors that can significantly impact outcomes, such as misinformed business strategies or flawed scientific research. While there are existing methods for data quality assessment, there is potential for further exploration and improvement in leveraging attribute labels (or column names in tables), their metadata, and content across diverse datasets and domains. This research addresses this gap by developing a comprehensive Attribute-Based Data Quality Assessment Framework (A-DQAF). A key feature of A-DQAF is its emphasis on well-known data quality issues and their corresponding data quality dimension violations, making the framework comprehensive and adaptable. The proposed framework will employ advanced natural language understanding techniques and machine learning to analyse attribute information and content, identify data quality issues and recommend improvements. Furthermore, this research will develop domain-specific heuristics, informed by a comprehensive Data Error Catalogue, to address unique challenges in data quality assessment across various fields. Preliminary results, detailed in the appendices, indicate the successful identification and mitigation of several common data quality issues, demonstrating the potential of this approach. This research is expected to contribute to the field of data quality assessment significantly, enhancing the accuracy and effectiveness of data-driven decision-making.

1.3 Keywords

Data, Data Quality, Data Quality Assessment, Attribute, Column, Metadata, Data Problems, Data Errors

2 Research Project Data Details

2.1 Research project data summary

Information from 622 datasets available in the UCI Machine Learning Repository Catalogue has been obtained. The Catalogue is a well-known source of research datasets encompassing various domains (called areas in the catalogue) and data types. The code used to download this information and the dataset created is available on GitHub at <https://github.com/marcelovalentimsilva/UCI-Catalog-with-all-622-Datasets>. This allowed an in-depth Data Quality Assessment (available at <https://www.kaggle.com/code/marcelovalentimsilva/data-quality-assessment-of-uci-datasets-catalog?scriptVersionId=127674648>), which is about to be sent to UCI, the University of California, Irvine, to improve their excellent and extremely important Catalogue of research data sets that has had millions of

accesses over time. For the first part of this research, ten datasets from five different areas were chosen from the Data Quality assessment produced on the 622 datasets from the UCI Catalogue of Datasets, and their attribute labels were obtained. There are other catalogues of datasets besides UCI that may be used during this PhD Research.

2.2 Will the data be identifiable

- Not applicable — no human data used

2.3 Will human participant information be sent overseas? Will biospecimens be sent overseas?

No

2.4 Will novel information about controlled goods or technologies on the Defence and Strategic Goods List (DSGL) be sent overseas?

No

2.5 How is the data being organised and structured?

This research aims to pursue Data Quality Assessments of at least three sets of different datasets. The first one is just a set of 10 different data sets from the UCI Catalogue of Data Sets. The second set will have at least 100 datasets, and the Third one will have at least 200 datasets. All Assessments will evaluate the Attribute labels or Column names in order to obtain data quality issues. The analysis results will be inserted in an Open Source Data Base, such as MySQL. The data will have a column with the date and time of each analysis result to keep track of the process.

3 Research Project Data Storage, Retention and Dissemination Details

3.1 Where will the data be stored?

The data will be stored in the Research Drive at Curtin. The access will be controlled through the Curtin network. There will be no physical data stored.

3.2 How much storage space will the data need?

The space is not going to be large.

3.3 How will the data be kept safe from human or technical failure?

The backup will follow the rules in Curtin network. There will be data saved in .csv files as well as in an Open Data Base such as MySQL.

3.4 What is the minimum retention period for the data?

7 years (All other research with outcomes that are classed as Minor)

3.5 Who will have access to the data and how will they access it?

The author, and the supervisors, Valerie Maxville and Johannes Herrmann.

3.6 How will the data be published?

Besides Curtin infrastructure, the data is being shared in Github, at
<https://github.com/marcelovalentimsilva/>

3.7 Will the data be embargoed prior to publication?

It will not be embargoed.

8.5 Appendix 5 - List of some potential risks for this research and suggestions on how to mitigate them:

1. **Data Availability and Quality:** There's a risk that the datasets needed for testing this framework may not be available or may not be of the quality needed. A solution for this risk should be to access as many datasets as possible. This risk is mitigated with over 600 datasets in the UCI Catalogue and other possible dataset catalogues described in the appendices.
2. **Technological Challenges:** Implementing NLP and machine learning techniques can bring unexpected difficulties. A mitigation strategy could be to stay updated with the latest research and technologies in the field and to be prepared to pivot if a certain method proves unworkable.
3. **Time Constraints:** The research might take longer than expected, causing delays. It is necessary to maintain a flexible timeline, with allowances for unexpected delays, to mitigate this.
4. **Changes in Technology/Field:** Given the rapid pace of advancement in fields like NLP and machine learning, new technologies or methods may emerge that make the work obsolete. Regular literature and technology reviews can help us stay updated.
5. **Dissemination Challenges:** There could be challenges in making the findings accessible and understandable to others in the field. Clear documentation, an explanation of methodologies, and a user-friendly interface can mitigate this risk.

8.6 Appendix 6 – Spreadsheet with top two datasets in five areas from UCI Catalogue

The decision to choose these ten datasets was the following:

- The top two datasets from each of the areas: Life, Social, Physical, Computer and Financial.
- This is how the top two datasets were chosen for each area:
 - Sorting from top to bottom of the 'webhits' column that showed how many web hits each dataset obtained. The top ones have over a million web hits.
 - Sorting from top to bottom of the 'num_papers' column, which showed the number of research papers that cited each dataset. The top ones have many dozen citations.
 - Ranking from 1 to the lowest amount for the 'webhits' column. The top one is number 1
 - Ranking from 1 to the lowest amount for the 'num_papers' column. The top one is number 1
 - Addition of the two rankings.
 - Ranking from 1 to the lowest amount for the sum of rankings.
 - Only the first 2 top in each area were chosen, and they are presented in green below:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	index	name	url	instances	attributes	year	area	date_donat	web_hits	#webhits	dataset	attribute_info	papers_that_cite_this_data	num_papers	#numpapers	2#	FinalRank						
2	52	Iris	http	150	4	1988	Life	1/07/1988	5319836	1	data	1. sepal length in	11	Sotiris B. Kotsiantis and Pan	100	1	2	1					
3	45	Heart Disease	http	303	75	1988	Life	1/07/1988	2184270	4	data	1. Only 14 attributes	11	Jeroen Eggermont and Joos	58	3	7	2					
4	2	Adult	http	48842	14	1996	Social	1/05/1996	2769887	2	data	1. Listing of	11	Rakesh Agrawal and Ramaki	51	8	10	3					
5	107	Wine	http	178	13	1991	Physical	1/07/1991	2167533	5	data	1. All attributes are	11	Ping Zhong and Masao Fuku	40	13	18	4					
6	17	Breast Cancer Wisconsin (Diagn	http	569	32	1995	Life	1/11/1995	1972842	8	data	1. ID number	11	Gavin Brown. Diversity in N	40	14	22	5					
7	14	Breast Cancer	http	286	9	1988	Life	11/07/1988	705622	29	data	1. Class: no-	11	Igor Fischer and Jan Poland.	91	2	31	6					
8	34	Diabetes	https://archive	20		Life	N/A		741148	26	Z	Diabetes files	11	Prem Melville and Raymond	53	5	31	7					
9	15	Breast Cancer Wisconsin (Origir	http	699	10	1992	Life	15/07/1992	875007	17	data	1. Sample code	11	Gavin Brown. Diversity in N	40	15	32	8					
10	72	Mushroom	http	8124	22	1987	Life	27/04/1987	821573	21	data	1. cap-shape:	11	Manuel Oliveira. Library Rel	46	12	33	9					
11	1	Abalone	http	4177	8	1995	Life	1/12/1995	1431163	12	data	1. Given is the	11	Ilhan Uysal and H. Altay Guv	29	22	34	10					
12	42	Glass Identification	http	214	10	1987	Physical	1/09/1987	471902	36	data	1. Id number: 1 to	11	Ping Zhong and Masao Fuku	52	7	43	11					
13	19	Car Evaluation	http	1728	6	1997	N/A	1/06/1997	1748706	10	data	1. Class Values:	11	Qingping Tao Ph. D. MAKIN	16	38	48	12					
14	20	Census Income	http	48842	14	1996	Social	1/05/1996	768711	22	data	1. Listing of	11	Aristides Gionis and Heikki	24	29	51	13					
15	51	Ionosphere	http	351	34	1989	Physical	1/01/1989	311140	62	data	1. All 34 are	11	Jennifer G. Dy and Carla Bro	55	4	66	14					
16	9	Auto MPG	http	398	8	1993	N/A	7/07/1993	856989	19	data	1. mpg:	11	Dan Pelleg. Scalable and Pri	12	48	67	15					
17	58	Letter Recognition	http	20000	16	1991	Computer	1/01/1991	489555	33	data	1. Lettcr capital	11	Jaakko Peltonen and Arto K	16	39	72	16					
18	46	Hepatitis	http	155	19	1988	Life	1/11/1988	359983	56	data	1. Class: DIE, LIVE	11	Amaury Habrard and Marc B	33	20	76	17					
19	142	Statlog (German Credit Data)	http	1000	20	1994	Financial	17/11/1994	890119	16	data	1. Attribute 1:	11	Jeroen Eggermont and Joos	7	63	79	18					
20	109	Zoo	http	101	17	1990	Life	15/05/1990	448279	40	data	1. animal name:	11	Eibe Frank and Stefan Kram	16	40	80	19					
21	100	Thyroid Disease	http	7200	21	1987	Life	1/01/1987	351663	57	data	1. N/A	11	Ken Tang and Ponnuthurai f	26	24	81	20					
22	149	Connectionist Bench (Sonar, Mi	http	208	60		Physical	N/A	263556	82	https://	1. N/A	11	Zhi-Hua Zhou and Yuan Jian	53	6	88	21					
23	16	Breast Cancer Wisconsin (Progn	http	198	34	1995	Life	1/12/1995	276394	76	data	1. ID number	11	Gavin Brown. Diversity in N	40	16	92	22					
24	99	Tic-Tac-Toe Endgame	http	958	9	1991	Game	19/08/1991	310422	63	data	1. top-left-square:	11	Saher Esmeir and Shaul Mar	19	33	96	23					
25	92	Spambase	http	4601	57	1999	Computer	1/07/1999	744313	25	data	1. The last column of	11	Don R. Hush and Clint Scove	4	75	100	24					
26	10	Automobile	http	205	26	1987	N/A	19/05/1987	873273	18	data	1. Attribute:	11	Geraldine E. Rosario and Elk	3	84	102	25					
27	103	Congressional Voting Records	http	435	16	1987	Social	27/04/1987	290644	71	data	1. Class Name: 2	11	Aristides Gionis and Heikki	20	32	103	26					
28	108	Yeast	http	1484	8	1996	Life	1/09/1996	374003	54	data	1. Sequence	11	Vassilis Athitsos and Stan Sc	11	52	106	27					
29	31	Covertypes	http	581012	54	1998	Life	1/08/1998	444144	41	data	1. Given is the	11	Joao Gama and Ricardo Rod	6	66	107	28					
30	27	Credit Approval	http	690	15		Financial	N/A	666861	31	data	1. A1:b, a,	11	Xiaoming Huo. FBP: A Front	4	76	107	29					

Figure 5 – Definition of ten datasets chosen from 5 areas

8.7 Appendix 7 - Columns from the top two datasets in five areas from ICU Catalogue

8.7.1 Iris Dataset – Area Life

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm

8.7.2 Heart Disease Dataset – Area Life

- age
- sex
- cp
- trestbps
- chol
- fb
- restecg
- thalach

9. exang
10. oldpeak
11. slope
12. ca
13. thal
14. num (the predicted attribute)

8.7.3 Adult Dataset – Area Social

1. age
2. workclass
3. fnlwgt
4. education
5. education-num
6. marital-status
7. occupation
8. relationship
9. race
10. sex
11. capital-gain
12. capital-loss
13. hours-per-week
14. native-country

Observation. The second dataset for Area Social should be Index 20 – “Census Income”. But it was discarded because it is another version with the same columns as the one above.

8.7.4 Wine Dataset – Area Physical

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

8.7.5 Glass Identification Dataset – Area Physical

1. Id number
2. RI: refractive index
3. Na: Sodium
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass (class attribute)

8.7.6 Letter Recognition – Area Computer

1. lettr: capital letter (26 values from A to Z)
2. x-box: horizontal position of the box (integer)
3. y-box: vertical position of the box (integer)
4. width: width of the box (integer)
5. high: height of the box (integer)
6. onpix: total number of on pixels (integer)
7. x-bar: mean x of on pixels in the box (integer)
8. y-bar: mean y of on pixels in the box (integer)
9. x2bar: mean x variance (integer)
10. y2bar: mean y variance (integer)
11. xybar: mean x y correlation (integer)
12. x2ybr: mean of $x * x * y$ (integer)
13. xy2br: mean of $x * y * y$ (integer)
14. x-ege: mean edge count left to right (integer)
15. xegvy: correlation of x-ege with y (integer)
16. y-ege: mean edge count bottom to top (integer)
17. yegvx: correlation of y-ege with x (integer)

8.7.7 Statlog (German Credit Data) – Area Financial

1. Status of existing checking account (qualitative)
2. Duration in months (numerical)
3. Credit history (qualitative)
4. Purpose (qualitative)
5. Credit amount (numerical)
6. Savings account/bonds (qualitative)
7. Present employment since (qualitative)
8. Installment rate in percentage of disposable income (numerical)
9. Personal status and sex (qualitative)
10. Other debtors/guarantors (qualitative)
11. Present residence since (numerical)
12. Property (qualitative)
13. Age in years (numerical)
14. Other instalment plans (qualitative)
15. Housing (qualitative)
16. Number of existing credits at this bank (numerical)
17. Job (qualitative)
18. Number of people being liable to provide maintenance for (numerical)
19. Telephone (qualitative)
20. Foreign worker (qualitative)

8.7.8 Spambase – Area Computer

1-48. word_freq_WORD (48 continuous real attributes): Percentage of words in the e-mail that match a specific word

49-54. char_freq_CHAR (6 continuous real attributes): Percentage of characters in the e-mail that match a specific character

55. capital_run_length_average (1 continuous real attribute): Average length of uninterrupted sequences of capital letters
56. Capital_run_length_longest (1 continuous integer attribute): Length of the longest uninterrupted sequence of capital letters
57. capital_run_length_total (1 continuous integer attribute): Sum of the length of uninterrupted sequences of capital letters or total number of capital letters in the e-mail
58. spam (1 nominal class attribute): Denotes whether the e-mail was considered spam (1) or not (0)

The WORDS are as follows:

1. make
2. address
3. all
4. 3d
5. our
6. over
7. remove
8. internet
9. order
10. mail
11. receive
12. will
13. people
14. report
15. addresses
16. free
17. business
18. email
19. you
20. credit
21. your
22. font
23. 000
24. money
25. hp
26. hpl
27. george
28. 650
29. lab
30. labs
31. telnet
32. 857
33. data
34. 415
35. 85
36. technology
37. 1999
38. parts
39. pm
40. direct
41. cs
42. meeting
43. original
44. project
45. re
46. edu
47. table
48. conference

And the six characters they analysed are:

1. ;
2. (
3. [
4. !
5. \$
- 6.

8.7.9 Congressional Voting Records – Area Social

1. Class Name (party affiliation): democrat, republican
2. handicapped-infants: yes, no
3. water-project-cost-sharing: yes, no
4. adoption-of-the-budget-resolution: yes, no
5. physician-fee-freeze: yes, no
6. el-salvador-aid: yes, no
7. religious-groups-in-schools: yes, no
8. anti-satellite-test-ban: yes, no
9. aid-to-nicaraguan-contras: yes, no
10. mx-missile: yes, no
11. immigration: yes, no
12. synfuels-corporation-cutback: yes, no
13. education-spending: yes, no
14. superfund-right-to-sue: yes, no
15. crime: yes, no
16. duty-free-exports: yes, no
17. export-administration-act-south-africa: yes, no

8.7.10 Credit Approval – Area Financial

1. A1: Categorical with values: b, a
2. A2: Continuous
3. A3: Continuous
4. A4: Categorical with values: u, y, l, t
5. A5: Categorical with values: g, p, gg
6. A6: Categorical with values: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff
7. A7: Categorical with values: v, h, bb, j, n, z, dd, ff, o
8. A8: Continuous
9. A9: Categorical with values: t, f
10. A10: Categorical with values: t, f
11. A11: Continuous
12. A12: Categorical with values: t, f
13. A13: Categorical with values: g, p, s
14. A14: Continuous
15. A15: Continuous
16. A16: Categorical with values: +, - (class attribute)

8.8 Appendix 8 - List of other catalogues of datasets besides UCI:

1. Kaggle Datasets: A platform offering a variety of datasets contributed by the Kaggle community for data science and machine learning competitions, as well as for general research purposes. URL: <https://www.kaggle.com/datasets>
2. Google Dataset Search: A search engine developed by Google that allows users to find datasets hosted across various repositories on the web. URL: <https://datasetsearch.research.google.com/>
3. Data.gov: The United States government's open data portal provides access to thousands of datasets on various topics, such as agriculture, climate, education, and health. URL: <https://www.data.gov/>
4. Eurostat: The statistical office of the European Union, offering a wide range of datasets on various topics related to the EU member countries. URL: <https://ec.europa.eu/eurostat/>
5. World Bank Open Data: A repository of global development data, providing datasets on various topics such as economics, education, health, and the environment. URL: <https://data.worldbank.org/>
6. AWS Public Datasets: Amazon Web Services provides a collection of public datasets that can be analysed and used with their cloud-based tools and services. URL: <https://aws.amazon.com/opendata/>
7. FiveThirtyEight: A data journalism platform that publishes datasets used in their articles, covering topics such as politics, sports, and economics. URL: <https://data.fivethirtyeight.com/>
8. OpenML: An open-source platform for machine learning researchers to share, organise, and collaborate on datasets, code, and experiments. URL: <https://www.openml.org/>
9. UNdata: A world of information from the United Nations: URL: <https://data.un.org/>

8.9 Appendix 9 – Compilation of attribute labels from first ten datasets

Appendices 6 and 7 show the ten datasets chosen from the UCI Catalogue analysis and the columns obtained, with descriptions when available.

The ten datasets were copied to a new spreadsheet with all its columns. A second tab brought only some columns:

1	index	name	url	instances	attributes	year	area	date_donated	web_hits	#webhits	data_fold/dataset_file_url	num_papers	#numpapers	2#	FinalRank
2	52	Iris	https://archive.ics.uci.edu/ml/datas	150	4	1988	Life	1/07/1988	5319836	1	https://ar https://archive.ic	100		1	2
3	45	Heart Disease	https://archive.ics.uci.edu/ml/datas	303	75	1988	Life	1/07/1988	2184270	4	https://ar https://archive.ic	58		3	7
4	2	Adult	https://archive.ics.uci.edu/ml/datas	48842	14	1996	Social	1/05/1996	2769887	2	https://ar https://archive.ic	51		8	10
5	107	Wine	https://archive.ics.uci.edu/ml/datas	178	13	1991	Physical	1/07/1991	2167533	5	https://ar https://archive.ic	40		13	18
6	42	Glass Identification	https://archive.ics.uci.edu/ml/datas	214	10	1987	Physical	1/09/1987	471902	36	https://ar https://archive.ic	52		7	43
7	58	Letter Recognition	https://archive.ics.uci.edu/ml/datas	20000	16	1991	Computer	1/01/1991	489555	33	https://ar https://archive.ic	16		39	72
8	142	Statlog (German Credit Data)	https://archive.ics.uci.edu/ml/datas	1000	20	1994	Financial	17/11/1994	890119	16	https://ar https://archive.ic	7		63	79
9	92	Spambase	https://archive.ics.uci.edu/ml/datas	4601	57	1999	Computer	1/07/1999	744313	25	https://ar https://archive.ic	4		75	100
10	103	Congressional Voting Records	https://archive.ics.uci.edu/ml/datas	435	16	1987	Social	27/04/1987	290644	71	https://ar https://archive.ic	20		32	103
11	27	Credit Approval	https://archive.ics.uci.edu/ml/datas	690	15		Financial	N/A	666861	31	https://ar https://archive.ic	4		76	107

Another tab containing each dataset's 'index', 'name', and 'area', and the 'Original Column' was created. Then the ID was extracted as well as the 'Column', which was divided into 'Separated Column Name' and 'Description', with 'Description' created from what came after the characters "(" or ":" in 'Column'. Here is a sample:

1	A	B	C	D	E	F	G	H
1	index	name	area	Original Column	ID Column	Separated Column Name	Description	
5	45	Heart Disease	Life	1. age	1 age	age		
6	2	Adult	Social	1. age	1 age	age		
7	142	Statlog (German Credit Data)	Financial	13. Age in years (numerical)	13 Age in years (numerical)	Age in years	(numerical)	
8	142	Statlog (German Credit Data)	Financial	5. Credit amount (numerical)	5 Credit amount (numerical)	Credit amount	(numerical)	
9	92	Spambase	Computer	55. capital_run_length_average (1 c	55 capital_run_length_average (1 c	capital_run_length_average	1 continuous real attribute): Average length of uninterrupted sequences of capital letters	
10	52	Iris	Life	1. sepal length in cm	1 sepal length in cm	sepal length in cm		
11	52	Iris	Life	2. sepal width in cm	2 sepal width in cm	sepal width in cm		
12	52	Iris	Life	3. petal length in cm	3 petal length in cm	petal length in cm		
13	52	Iris	Life	4. petal width in cm	4 petal width in cm	petal width in cm		
14	142	Statlog (German Credit Data)	Financial	2. Duration in months (numerical)	2 Duration in months (numerical)	Duration in months	(numerical)	
15	92	Spambase	Computer	1. ;	49 ;	char_freq_;		

Then the pre-processing and analysis of each column name were realised. First, the "Separated column name" above had all its characters turned to lowercase. It became the Pre-Standardized Column Name. Then the characters '-' and '_' were removed, and "Standardized Column Name" was created. Later, the author visually analysed to find 'target words' in the attribute labels that suggested something. The words "age, amount, average, cm, duration and freq" suggested that each column's content was 'numeric >= 0':

Original Column	ID Column	Separated Column Name	Pre Standardized Column Name	Standardized Column Name	target_word_found	analysis of Clean Column
1. age	1 age	age	age	age	age	numeric >= 0
1. age	1 age	age	age	age	age	numeric >= 0
13. Age in years (numerical)	13 Age in years (numerical)	Age in years	age in years	age in years	age	numeric >= 0
5. Credit amount (numerical)	5 Credit amount (numerical)	Credit amount	credit amount	credit amount	amount	numeric >= 0
55. capital_run_length_average (1 c	55 capital_run_length_average (1 c	capital_run_length_average	capital_run_length_average	capital run length average	average	numeric >= 0
1. sepal length in cm	1 sepal length in cm	sepal length in cm	sepal length in cm	sepal length in cm	cm	numeric >= 0
2. sepal width in cm	2 sepal width in cm	sepal width in cm	sepal width in cm	sepal width in cm	cm	numeric >= 0
3. petal length in cm	3 petal length in cm	petal length in cm	petal length in cm	petal length in cm	cm	numeric >= 0
4. petal width in cm	4 petal width in cm	petal width in cm	petal width in cm	petal width in cm	cm	numeric >= 0
2. Duration in months (numerical)	2 Duration in months (numerical)	Duration in months	duration in months	duration in months	duration	numeric >= 0
1. ;	49 ;	char_freq_;	char_freq_;	char freq ;	freq	numeric >= 0
1. make	1 make	word_freq_make	word_freq_make	word freq make	freq	numeric >= 0

Other analyses were made, where the 'target word found' was telephone, name, status, class, country & education.

Each one resulted in a separate analysis:

Original Column	ID Column	Separated Column Name	Pre Standardized Column Name	Standardized Column Name	target_word_found	analysis of Clean Column
19. Telephone (qualitative)	19 Telephone (qualitative)	Telephone	telephone	telephone	telephone	telephone format
1. Class Name (party affiliation): d	1 Class Name (party affiliation): d	Class Name (party affiliation)	class name (party affiliation)	class name (party affiliation)	name	text
6. marital-status	6 marital-status	marital-status	marital status	marital status	status	text
2. workclass	2 workclass	workclass	workclass	workclass	class	text with domain
14. native-country	14 native-country	native-country	native-country	native country	country	text with domain
4. education	4 education	education	education	education	education	text with domain

Similar analyses were made when no 'target word' was found, but description words were found:

Original Column	ID	Separated Column Name	Description	description_word_found	analysis of description_found
2. A2: Continuous	2 A2	Continuous	Continuous	continuous	numeric >= 0
3. A3: Continuous	3 A3	Continuous	Continuous	continuous	numeric >= 0
8. A8: Continuous	8 A8	Continuous	Continuous	continuous	numeric >= 0
15. xegvy: correlation of x-egw with y	15 xegvy	correlation of x-egw with y (integer)	correlation	correlation	numeric >= 0
17. yegvx: correlation of y-egw with x	17 yegvx	correlation of y-egw with x (integer)	correlation	correlation	numeric >= 0
5. high: height of the box (integer)	5 high	height of the box (integer)	height	height	numeric >= 0
10. y2bar: mean y variance (integer)	10 y2bar	mean y variance (integer)	mean	mean	numeric >= 0
11. xybar: mean x y correlation (inte	11 xybar	mean x y correlation (integer)	mean	mean	numeric >= 0
4. Purpose (qualitative)	4 Purpose	qualitative	qualitative	qualitative	text
6. Savings account/bonds (qualita	6 Savings account/bonds	qualitative	qualitative	qualitative	text
1. A1: Categorical with values: b, a	1 A1	Categorical with values	categorical	categorical	text with domain
10. A10: Categorical with values: t, f	10 A10	Categorical with values	categorical	categorical	text with domain
12. A12: Categorical with values: t, f	12 A12	Categorical with values	categorical	categorical	text with domain

8.10 Appendix 10 - Real problems in datasets related to columns

This is a list of real problems in datasets related to columns that are going to be researched in this PhD, with their association with the summary of well-known Data Quality Issues and the Data Quality Dimension violations presented in Appendix 1:

The first dataset is the Online Retail Dataset from the UCI Catalogue

(<https://archive.ics.uci.edu/ml/datasets/Online+Retail>), which contains negative values in the columns

Minimum 10 values

Value
-80995

Minimum 10 values

Value
-11062.06

Quantity -74215 and UnitPrice -11062.06. It is easy to see that the names Quantity and UnitPrice are supposed to be positive only. So, two problems were found related to columns' names and contents.

These cases fall under the category of "Domain violation" (Issue 15) from Appendix 1, as these columns should not contain negative values according to their semantics (quantity and price should be positive). And this is a violation of the "Accuracy" dimension.

The second dataset with real problems related to columns found is Books, from GitHub:

<https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.csv>. The column original_publication_year also contains negative values. Years should not contain negative values in this case.

Minimum 10 values

Value

-1750

-762

In a similar case to the first dataset, this is also a "Domain violation" (Issue 15). Therefore, also a violation of the "Accuracy" dimension.

A third dataset with problems related to the attribute labels is Bike Sharing Dataset, also from the UCI Catalogue:

<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>. This dataset contains information about bike-sharing. Some attribute labels, such as "dteday", "yr", "mnth", "hr", "weathersit", "temp", "hum" and "cnt" are abbreviated and need to be corrected. This issue could be seen as a type of "Misspellings" (Issue 3) and "Ambiguous data" (Issue 4), as the attribute labels could be open to multiple interpretations due to their abbreviated forms. This violates the "Consistency" and "Accuracy" dimensions.

The column "yr" brings the values 0 and 1 instead of years with the expected format YYYY. This falls under "Different units/representations" (Issue 14) as the same concept (year) is represented differently. It also could be considered a "Domain violation" (Issue 15) as the values do not meet the expected domain for years. This violates the "Consistency" and "Accuracy" dimensions.

8.11 Appendix 11 - Some examples of errors related to attribute labels in datasets

These examples illustrate some of the potential errors that can be found in datasets based on attribute labels and their expected content. Analysing attribute labels and developing techniques to detect and correct these errors can improve data quality and ensure more accurate data analysis. Each item is connected to the well-known Data Quality Issues summary and the Data Quality dimension violations shown in Appendix 1.

1. Abbreviation pattern errors: A column named "pctComplete" is expected to contain percentage values, but it contains decimal values instead (e.g., 0.85 instead of 85%). This falls under "Different units/representations" (Issue 14) as the same concept is represented differently. This violates the "Consistency" dimension.
2. Categorical column errors: A column named "Gender" is expected to contain only "Male" or "Female" values, but it contains other unexpected values or misspellings (e.g., "Mael", "Femael"). This can be categorised as "Incorrect data" (Issue 2) and "Misspellings" (Issue 3). It violates the "Accuracy" dimension.

3. Data entry errors: A column named "ZIPCode" contains various formats of postal codes, making it difficult to validate and analyse the data. This could be seen as "Ambiguous data" (Issue 4) due to inconsistent formatting. It violates the "Accuracy" and the "Consistency" dimensions.
4. Composite column name errors: A column named "FirstNameLastName" contains concatenated first and last names without a clear delimiter, making it challenging to separate the individual components. This is an instance of "Extraneous data" (Issue 5), as the column contains more data than expected. It violates the "Consistency" and "Uniqueness" dimensions.
5. Outliers and anomalies: A column named "Age" contains implausible values, such as negative numbers or ages exceeding 150 years. This issue can be categorised as "Domain violation" (Issue 15). It violates the "Accuracy" dimension.
6. Data hierarchy errors: A dataset contains columns "Country", "State", and "City", but some rows have inconsistent hierarchies, such as a city located in the wrong state or country. This falls under "Incorrect references" (Issue 8). It violates the "Accuracy" dimension.
7. Column name similarity errors: A dataset contains columns "Income" and "AnnualIncome", or "Gross_Profit", and "GrossProfit", causing confusion and potential data integration issues due to duplicated or closely related information. This might be seen as "Structural conflicts" (Issue 10) and possibly "Duplicates" (Issue 9). It violates the "Consistency" and "Uniqueness" dimensions.
8. Data sparsity and missingness errors: A column named "OptionalNotes" is expected to have sparse data, but it contains important information that should not be missing for certain records. This refers to "Missing data" (Issue 1). It violates the "Accuracy" and "Completeness" dimensions.
9. Column name standardisation errors: A dataset has columns with different naming conventions, such as "BirthDate" and "Date_of_Birth", which may lead to data quality issues when integrating or comparing datasets. This falls under "Different encoding formats" (Issue 22) and "Structural conflicts" (Issue 10). It violates the "Consistency" and "Uniqueness" dimensions.
10. Domain-specific column name errors: In a healthcare dataset, a column named "BP" is expected to contain blood pressure values, but it contains unrelated values or data entry errors (e.g., incorrect units or implausible values). This is a "Domain violation" (Issue 15) and "Incorrect data" (Issue 2). It violates the "Accuracy" dimension.
11. Column name evolution errors: A longitudinal dataset has a column named "MaritalStatus" in earlier records but changes to "RelationshipStatus" in later records, causing confusion and potential data integration issues. This is related to "Structural conflicts" (Issue 10). It violates the "Consistency" and "Uniqueness" dimensions.
12. Data provenance errors: A column named "Source" is expected to contain information about the data's origin but has inconsistent or missing values, making it difficult to assess data quality and reliability. This issue can be seen as "Missing data" (Issue 1) and "Incorrect data" (Issue 2). It violates the "Accuracy" and "Completeness" dimensions.
13. Column name inconsistency errors: A dataset has columns with inconsistent naming conventions, such as mixed capitalisation ("FirstName" vs "lastname") or different delimiter styles ("Date_of_Birth" vs "DateOfBirth"). This refers to "Different encoding formats" (Issue 22). It violates the "Consistency" dimension.
14. Column name clustering errors: A dataset has columns named "CustomerID" and "CustID", which are semantically similar but not identified, leading to incorrect data validation and analysis processes. This is like "Structural conflicts" (Issue 10) and possibly "Duplicates" (Issue 9). It violates the "Consistency" and "Uniqueness" dimensions.

8.12 Appendix 12 – Data problems from [Rahm & Do, 2000] associated with the summary of well-known Data Quality Issues and the Data Quality dimension violations shown in Appendix 1

8.12.1 Single-Source Problems

1. Date with illegal value (e.g., 30.13.70): This falls under the "Incorrect data" (Issue 2) category, violating the "Accuracy" dimension.
2. Violated attribute dependencies (e.g., age different than the value of current date - birth date): This is an example of "FD violation" (Issue 16), violating both "Accuracy" and "Consistency" dimensions.
3. Uniqueness violation (e.g., two different names for a single ID value): This is a "Uniqueness violation" (Issue 19), violating the "Uniqueness" dimension.
4. Referential integrity violation (e.g., department value non-existing in department dimension table): This is a "Referential integrity violation" (Issue 18), violating "Accuracy", "Consistency", and "Completeness" dimensions.
5. Missing value (e.g., a possible default value in the telephone column): This falls under "Missing data" (Issue 1), violating both "Accuracy" and "Completeness" dimensions.
6. Misspelling in city content: This is an example of "Misspellings" (Issue 3), violating the "Accuracy" dimension.
7. Abbreviated values in the content of columns: This falls under "Ambiguous data" (Issue 4), violating both "Accuracy" and "Consistency" dimensions.
8. Embedded values in a field where name and birth date are found: This is an example of "Extraneous data" (Issue 5), violating "Consistency" and "Uniqueness" dimensions.
9. Misfielded values (e.g., the country name is found in the city column): This is an example of "Misfielded values" (Issue 7), violating "Accuracy", "Consistency", and "Completeness" dimensions.
10. Violated attribute dependencies when the ZIP code is inconsistent with the city name: This is another example of "FD violation" (Issue 16), violating both "Accuracy" and "Consistency" dimensions.
11. Word transpositions (e.g., names include sometimes first name followed by last name, and other times the opposite): This falls under "Different word orderings" (Issue 11), violating both "Consistency" and "Uniqueness" dimensions.
12. Duplicated records (e.g., the same employee appears twice in the same table): This is an example of "Duplicates" (Issue 9), violating the "Uniqueness" dimension.
13. Contradicting records (e.g., the same real value is described by different values): This could be an example of "Domain violation" (Issue 15), violating the "Accuracy" dimension.
14. Wrong references (e.g., the department number is not the correct one): This falls under "Incorrect references" (Issue 8), violating the "Accuracy" dimension.

8.12.2 Multi-Source Problems:

15. Naming conflicts (e.g., tables Customer and Client, columns Sex and Gender, etc.): This falls under "Use of synonyms" (Issue 20), violating the "Uniqueness" dimension.
16. Structural conflicts (e.g., the name of the client is full in one table and classified as FirstName and LastName in another table): This is a type of "Structural conflicts" (Issue 10), violating both "Consistency" and "Uniqueness" dimensions.
17. Different gender representations in different tables (e.g., gender 0 and 1 vs 'F' and 'M'): This falls under "Different units/representations" (Issue 14), violating the "Consistency" dimension.

8.13 Appendix 13 – Data Error Catalogue - Distribution of all data errors in this document by Data Quality Dimension and Data Quality Issue from Appendix 1

Dimension	DQI #	Data Quality Issue (DQI)	Description of Data Error	Source
Accuracy	2	Incorrect data	Column "Gender"	Appendix 11 - 02. Categorical column errors
Accuracy	2	Incorrect data	Column "BP"	Appendix 11 - 10. Domain-specific column name errors
Accuracy	2	Incorrect data	Column "Source"	Appendix 11 - 12. Data provenance errors
Accuracy	2	Incorrect data	Column Date e.g., 30.13.70	Appendix 12 - 01. Date with an illegal value
Accuracy	3	Misspellings	Column "Gender"	Appendix 11 - 02. Categorical column errors
Accuracy	3	Misspellings	Column City	Appendix 12 - 06. Misspelling in city content
Accuracy	3	Misspellings	Columns "dteday", "yr", "mnth", "hr", "weathersit", "temp", "hum" and "cnt"	Appendix 10 – Bike Sharing Dataset
Accuracy	8	Incorrect references	Columns "Country", "State", and "City"	Appendix 11 - 06. Data hierarchy errors
Accuracy	8	Incorrect references	The department number is not the correct one	Appendix 12 - 14. Wrong references
Accuracy	15	Domain violation	Column "Age"	Appendix 11 - 05. Outliers and anomalies
Accuracy	15	Domain violation	Column "BP"	Appendix 11 - 10. Domain-specific column name errors
Accuracy	15	Domain violation	Different values describe the same real value	Appendix 12 - 13. Contradicting records
Accuracy	15	Domain violation	Column "yr"	Appendix 10 – Bike Sharing Dataset
Accuracy	15	Domain violation	Original_publication_year column	Appendix 10 – Books dataset
Accuracy	15	Domain violation	Quantity and UnitPrice columns	Appendix 10 - Online Retail Dataset
Accuracy/ Completeness	1	Missing data	Column "OptionalNotes"	Appendix 11 - 08. Data sparsity and missingness errors
Accuracy/ Completeness	1	Missing data	Column "Source"	Appendix 11 - 12. Data provenance errors
Accuracy/ Completeness	1	Missing data	A possible default value in the telephone column	Appendix 12 - 05. Missing value
Accuracy/ Consistency	4	Ambiguous data	Column "ZIPCode"	Appendix 11 - 03. Data entry errors
Accuracy/ Consistency	4	Ambiguous data	Columns	Appendix 12 - 07. Abbreviated values in the content of columns
Accuracy/ Consistency	4	Ambiguous data	Columns "dteday", "yr", "mnth", "hr", "weathersit", "temp", "hum" and "cnt"	Appendix 10 – Bike Sharing Dataset
Accuracy/ Consistency	16	FD violation	Column age different than the value of current date - birth date	Appendix 12 - 02. Violated attribute dependencies
Accuracy/ Consistency	16	FD violation	Columns ZIP/City	Appendix 12 - 10. Violated attribute dependencies
Accuracy/ Consistency/ Completeness	7	Misfielded values	Columns Country/City	Appendix 12 - 09. Misfielded values
Accuracy/ Consistency/ Completeness	18	Referential integrity violation	Department value does not exist in the department dimension table	Appendix 12 - 04. Referential integrity violation
Consistency	14	Different units/representations	Column "pctComplete"	Appendix 11 - 01. Abbreviation pattern errors
Consistency	14	Different units/representations	Column Gender	Appendix 12 - 17. Different representations in different tables
Consistency	14	Different units/representations	Column "yr"	Appendix 10 – Bike Sharing Dataset
Consistency	22	Different encoding formats	Columns "BirthDate" and "Date_of_Birth"	Appendix 11 - 09. Column name standardisation errors
Consistency	22	Different encoding formats	Columns ("FirstName" vs "lastname") and ("Date_of_Birth" vs "DateOfBirth")	Appendix 11 - 13. Column name inconsistency errors
Consistency/ Uniqueness	5	Extraneous data	Column "FirstNameLastName"	Appendix 11 - 04. Composite column name errors
Consistency/ Uniqueness	5	Extraneous data	Column Name/Birth	Appendix 12 - 08. Embedded values in a field
Consistency/ Uniqueness	10	Structural conflicts	Columns "Income" and "AnnualIncome", or "Gross_Profit" and "GrossProfit"	Appendix 11 - 07. Column name similarity errors
Consistency/ Uniqueness	10	Structural conflicts	Columns "BirthDate" and "Date_of_Birth"	Appendix 11 - 09. Column name standardisation errors
Consistency/ Uniqueness	10	Structural conflicts	Column "MaritalStatus"/"RelationshipStatus"	Appendix 11 - 11. Column name evolution errors

Consistency/ Uniqueness	10	Structural conflicts	Columns "CustomerID" and "CustID"	Appendix 11 - 14. Column name clustering errors
Consistency/ Uniqueness	10	Structural conflicts	The name of the client is full in one table and classified as FirstName and LastName in another table	Appendix 12 - 16. Structural conflicts
Consistency/ Uniqueness	11	Different word orderings	Columns name, first name, last name	Appendix 12 - 11. Word transpositions
Uniqueness	9	Duplicates	Columns "Income" and "AnnualIncome", or "Gross_Profit" and "GrossProfit"	Appendix 11 - 07. Column name similarity errors
Uniqueness	9	Duplicates	Columns "CustomerID" and "CustID"	Appendix 11 - 14. Column name clustering errors
Uniqueness	9	Duplicates	Column employee appears twice in the same table	Appendix 12 - 12. Duplicated records
Uniqueness	19	Uniqueness	Two different names for a single ID value	Appendix 12 - 03. Uniqueness violation
Uniqueness	20	Use of synonyms	Tables Customer and Client, columns Sex and Gender	Appendix 12 - 15. Naming conflicts

8.14 Appendix 14 - List of rules/heuristics according to some attributes/columns

These are some rules to be checked:

- Percentage attributes: Columns with names previously cleaned containing 'percentage' are expected to contain numerical values ranging from 0 to 100. Content issues may include values outside this range, non-numeric values, or inconsistent formatting (e.g., using decimals or fractions).
- Date attributes: Columns with names containing words like 'date', 'month', 'day', 'week', or 'year' are expected to have date-related values. Content issues may involve incorrect or ambiguous date formats, missing or inconsistent delimiters, invalid dates (e.g., 30 February, years well in the future or past), or mixing different date formats within the same column.
- ID attributes: Columns with names containing the word 'ID' or similar terms are expected to contain unique identifiers, usually integers or alphanumeric strings. Content issues can include duplicate IDs, missing IDs, inconsistent formatting (e.g., leading zeros, variable-length strings), or non-unique values that should be unique.
- Name attributes: Columns with pre-cleaned names containing words like 'name' or similar terms are expected to contain textual data representing names. Content issues may involve misspellings, inconsistent capitalisation, abbreviations, or mixing different name formats (e.g., first and last names in the same column).
- Address attributes: Columns with names related to addresses, such as 'street', 'city', 'state', or 'country', are expected to contain location-related information. Content issues may include missing data, misspellings, abbreviations, inconsistent formatting (e.g., using different delimiters or ordering of address components), or mixing different address formats within the same column.

All attribute content must be validated against the rules:

- For each attribute in the dataset, identify the most relevant rule from the rule-based system based on the pre-processed column name.
- Check whether the actual content of the attribute matches the expected content as per the rule. This may involve verifying data types, value ranges, or specific constraints.

8.15 Appendix 15 - List of expected formats according to some labels of attributes.

Here's a list of some common attribute labels with their expected formats across multiple domains:

1. Personal information:
 - a. Name: "FirstName", "LastName", "FullName" - String (text)
 - b. DateOfBirth: "DOB", "BirthDate" - Date (YYYY-MM-DD) or (DD-MM-YYYY)
 - c. Gender: "Sex" - String (text), usually "Male", "Female", or abbreviations like "M", "F"
 - d. Email: "EmailAddress" - String (text) following the email format (e.g., example@example.com)
2. Geographical information:
 - a. Address: "StreetAddress", "City", "State", "PostalCode", "Country" - String (text)
 - b. Latitude: "Lat" - Decimal number (floating-point)
 - c. Longitude: "Lon" - Decimal number (floating-point)
3. Temporal information:
 - a. Date: "Date", "DateTime" - Date (YYYY-MM-DD) or Date & Time (YYYY-MM-DD HH:MM:SS)
 - b. Year: "Year" - Integer (YYYY)
 - c. Month: "Month" - Integer (MM) or String (text) (e.g., "January", "Feb")
4. Financial information:
 - a. Price: "Price", "Cost" - Decimal number (floating-point)
 - b. Currency: "Currency" - String (text) (e.g., "USD", "EUR", "AUD")
5. Healthcare information:
 - a. PatientID: "Patient_ID" - Integer or String (text)
 - b. Diagnosis: "Diagnosis", "Condition" - String (text)
 - c. Medication: "Drug", "Medicine" - String (text)
6. Educational information:
 - a. StudentID: "Student_ID" - Integer or String (text)
 - b. Course: "CourseName", "Subject" - String (text)
 - c. Grade: "Grade", "Score", "Mark" - Integer or Decimal number (floating-point)

8.16 Appendix 16 - List of items to be followed for content analysis

- a. For each dataset attribute, apply the rules or heuristics identified in Appendix 14.
- b. Check whether the actual content of the column matches the expected content based on the column name features. Use supervised learning techniques, such as classification algorithms (e.g., logistic regression, decision trees, support vector machines), or NLP techniques, like topic modelling (e.g., Latent Dirichlet Allocation) or transformer-based models (e.g., BERT, GPT), to predict the type of data quality issue based on the column name features and content.
- c. If a column's content does not match the expected content based on the rule, flag the column as having a potential data quality issue.
- d. Identify any data quality issues detected, such as inconsistencies, missing values, incorrect data types, value range violations, constraint violations or outliers.
- e. Generate a report summarising the detected data quality issues, categorised by the type of issue and the corresponding rule.
- f. Suggest how to correct the identified issues, including data transformation, data imputation, or domain-specific validation. Data transformation techniques may involve standardising or normalising data, applying log transformations, or converting categorical variables into numerical ones. Data imputation methods may include mean, median or mode imputation, regression imputation, or more advanced techniques like k-Nearest Neighbours (k-NN) or multiple imputations by chained equations (MICE). Domain-specific validation may involve external knowledge sources like ontologies, knowledge graphs, or expert input to verify and correct the data.