Deriving Knowledge from Attribute Labels for Data Quality Assessment

Short Title for ACM: Deriving knowledge from attribute labels

Subtitle: Deriving knowledge from attribute labels for Semantic Type Detection and Data Quality Assessment

MARCELO V SILVA

Curtin University, marcelo.valentimsilva@student.curtin.edu.au

HANNES HERRMANN

Curtin University, hannes.herrmann@curtin.edu.au

VALERIE MAXVILLE

Curtin University, valerie.maxville@curtin.edu.au

The increasing reliance on data-driven decision-making across various sectors underscores the critical need for high-quality data. Despite advancements, data quality issues persist, leading to significant consequences for business strategies and scientific research alike. Current data quality assessment methods fail to leverage the semantic richness embedded in attribute labels (or column names/headers in tables), their metadata, and content across diverse datasets and domains, leaving a crucial gap in comprehensive data quality evaluation.

This research addresses this gap by introducing a methodology centred around Attribute-Based Semantic Type Detection and Data Quality Assessment. By leveraging the semantic information within attribute labels, along with rule-based analysis and comprehensive Formats and Abbreviations dictionaries, our approach skilfully navigates a wide array of data types - including numerical (also non-negative and bounded), categorical, ID, names, strings, geographical, temporal, and complex formats like URLs, IP addresses, email, and binary values - offering insights that go beyond conventional data quality assessment methods.

A key feature of our research is a focus on 22 well-known data quality issues and their corresponding data quality dimension violations, grounding our methodology in a comprehensive and academic framework. Insights from analysing 50 datasets from the UCI Machine Learning Repository are included in the 'Discoveries' documents, which detail this extensive data quality assessment, underscore the research's proficiency in semantic type detection and its unmatched clarity in identifying potential data quality issues based on attributes. Compared to tools like YData Profiling, our method demonstrates superior accuracy in identifying missing values and nuanced data quality issues, as highlighted by our findings of 81 missing values in 922 attributes across fifty datasets, where YData identified only one.

By filling the existing gap with our innovative use of attribute labels for semantic analysis, our methodology advances data quality assessment, aiming to streamline the traditionally time-consuming process of data cleaning and boosting our ability to effectively use data in the digital age.

CCS CONCEPTS • Information systems → Data management systems → Information integration → Data cleaning

Additional Keywords and Phrases: Data quality, data quality assessment, semantic type detection, missing data.

1 INTRODUCTION

The increasing reliance on data across various sectors underscores the critical importance of ensuring high-quality data before it enters the data cleaning phase. Traditional data quality assessment methods overlook the semantic richness of attribute information, such as words in labels, names or headers in table columns, or descriptions, leading to a gap in our ability to identify and understand data quality issues at their source. However, these elements contain valuable semantic information that can be instrumental in identifying data quality issues related to their content, such as ID, name, temporal and geographical information.

This research introduces a novel approach focused on harnessing the untapped semantic information within attribute labels to enhance data quality assessment. By systematically analysing attribute labels to derive knowledge about the expected data content and format, this method aims to identify and categorize potential data quality issues before traditional data cleaning processes begin. This shift towards utilizing attribute labels as a preliminary step in data quality assessment promises to streamline data management practices by providing early insights into data quality, potentially reducing the time and resources dedicated to subsequent data cleaning and processing.

The primary aim of this research is to demonstrate the feasibility and effectiveness of using attribute labels for semantic type detection and data quality assessment. By developing and evaluating a method that interprets the semantic richness of attribute labels, we seek to fill a critical gap in the literature and offer a new pathway for enhancing data quality assessment practices across diverse datasets and domains.

Through a comprehensive evaluation across various datasets, our methodology identified 81 instances of missing values across 922 columns, markedly surpassing traditional tools like YData Profiling, which detected a single instance. This stark contrast not only showcases our approach's superior ability to uncover and address data quality issues effectively but also highlights its transformative potential for enhancing data quality assessment across the digital landscape.

1.1 Research Question

This is the research question that guides this study:

> *"How does semantic analysis of attribute labels, using comprehensive dictionaries, enhance early data type/ format detection, and improve data quality assessment compared to existing tools across various datasets and domains?"*

1.2 Contributions

The core contributions of this study focus on the innovative use of attribute labels for data quality assessment through the development of a method and an accompanying software package. These contributions include:

- **Comprehensive Dictionaries**: Creation and refinement of comprehensive dictionaries for semantic translation of attribute labels into specific data formats, enhancing the precision of data type detection and issue identification.
- **Issue Identification and Categorization**: Advanced techniques for identifying and categorizing potential data quality issues and respective data quality dimensions based on an in-depth analysis of attribute labels and content.
- **Empirical Validation**: Application of the methodology to diverse datasets for empirical validation of its effectiveness in identifying data quality issues, supported by detailed documentation of findings.
- **Dissemination and Impact**: Comprehensive documentation and planned dissemination of the research findings and software package, aimed at fostering adoption and application within the data quality research community and industry practices.

These contributions collectively underscore the potential of attribute label analysis to enhance the early stages of data quality assessment, providing foundational insights that support more informed data management strategies.

1.3 Delimitation

This research analyses fifty of the over six hundred datasets that exist in the UCI Machine Learning Catalogue in many areas/domains. The core analysis revolves around attribute names, metadata, and content across these diversified datasets.

It evaluates file formats including .txt, .csv, .data.xls and .xlsx. Compacted files such as .zip, .tar.gz, .gz. that contain the previous file formats can also be analysed. When the initial file is a .zip or a .tar.gz file, the user chooses the file to be analysed. Besides that, .gz files contain only one file inside, so it is automatically analysed. Two datasets were obtained from extractions of the @DATA part of .arff files inside a zip file. Separate .txt files were created for both analyses.

Besides that, it always analyses the first line to check if it is a header or not and excludes the analysis of the first line when it is a header line. It also always evaluates the symbol that separates the data items, allowing ';', ',' and ' ' (blank).

2 RELATED WORK

Issues with data quality have been recognized for decades, with significant implications for business strategies and scientific research [Wang et al., 1992]. The exponential growth of data underscores the urgency for effective data quality management practices [Fuller & Kuromiya, 1982], [IDC, 2021]. A critical inefficiency highlighted by [Dasu and Johnson, 2003] is the disproportionate amount of time spent on data cleaning, which can consume up to 80% of the total analysis time. Our research introduces an innovative solution by automating the detection of data quality issues through the semantic analysis of attribute labels, aiming to streamline the data quality assessment process.

2.1 Data quality assessment

Data quality is paramount for data to be considered fit for use by consumers [Wang & Strong, 1996]. Traditional assessments often concentrate on quantifiable dimensions such as accuracy and completeness but typically overlook the semantic richness inherent in attribute labels or column names/headers. This oversight represents a significant gap in data quality assessment methodologies, which our research seeks to address.

The literature showcases various data quality assessments, each with its own focus but often missing the opportunity to utilize semantic information from attribute labels. For instance, Data X-Ray [Wang et al., 2015] emphasizes data diagnosis, while Ehrlinger & Wöß (2022) review over 660 tools for data quality measurement, noting the need for enhanced automation and clearer methodologies in data profiling. These studies underscore the existing gaps and the potential for innovative approaches like ours, which prioritizes semantic analysis to improve data quality insights.

Furthermore, ISO/IEC standards highlight the importance of syntactic, semantic, and pragmatic dimensions in data quality [Ehrlinger & Wöß, 2022], supporting the need for a comprehensive approach that includes semantic analysis. By focusing on attribute labels, our research contributes a novel perspective to data quality assessment, aligning with these standards to offer a more nuanced and efficient evaluation method.

2.2 Data quality dimensions and attribute analysis

Data in datasets, text files, or database tables consist of attributes (columns or fields) describing the data's features or characteristics. These attributes typically have labels, names, or headers, potentially providing semantic cues about the data they represent. Additionally, metadata or descriptions at the column level can further detail the nature of the data, including data type and constraints.

There may be descriptions or metadata, providing information about the data. This may happen at the column level, including their name, data type, constraints, and comments about the meaning of the column.

Key data quality dimensions critical to attribute analysis include [Batini & Scannapieco, 2016], [Dasu and Johnson, 2003]:

- **Accuracy** is the closeness between a recorded value and the represented real-world item or state.

- **Completeness** measures the incidence of missing values for an attribute.
- **Consistency** identifies breaches of semantic rules established over (a group of) data items.
- **Timeliness**: the currency of the data.

Literature on directly leveraging attribute labels or column names for data quality is sparse, with notable exceptions like [Trummer, 2023], highlighting the novelty of our approach. In contrast to [Trummer, 2023], which utilizes Kaggle datasets for predicting data correlations through column names, our research distinctively targets data quality issues via semantic examination of attribute labels. Our method is not limited to .csv formats but extends to various file types enhancing our research's applicability. Crucially, our focus is direct engagement with data quality, identifying issues through the inherent semantic signals in column names - a strategy not widely addressed in current literature.

To pursue attribute analysis, we employed formats and abbreviations dictionaries. The definition of words to be considered followed a study from the Web Data Commons [Ristoski et al., 2012] which analysed over 90 million tables from the overall Web Table Corpus, and obtained a table that connected their work with the cross-domain knowledge base DBpedia [Auer et al., 2007] resulting in a list with over 1100 of the words commonly found in column names [DBHeaders file]. Our formats dictionary ('formats_dictionary.txt' [Silva, 2024]) expands upon this, associating these common words with probable data formats, containing 1800 words, providing a foundational tool for our attribute analysis.

Table 1, derived from the work of Ristoski et al. (2012), illustrates the prevalence of certain words in column names across a vast number of web tables, indicating the potential for these labels to convey meaningful insights into data quality.

Table 1 - Most commonly found words in column names/attribute labels, with the respective amount from [DBHeaders file]:

DBpediaProperty	#tables	DBpediaProperty	#tables	DBpediaProperty	#tables
name	3645104	model	843211	city	484069
date	2168479	age	805379	day	438331
title	1538149	rating	591323	distance	426854
description	1510237	artist	579331	team	406040
size	1225375	rank	578530	author	369922
location	1032050	album	557798	result	369457
type	955059	address	507539	length	357595
year	853473	category	489397	id	354066

This focus on attribute labels represents a significant shift towards utilizing semantic analysis for enhancing data quality assessment, setting our research apart from existing methodologies.

2.3 Data quality issues

[Loshin, 2002] demonstrated a business rules approach that provided a framework for identifying the underlying causes of poor data quality. By transforming declarative data quality rules into actionable code, his approach aligns with our aim to develop comprehensive strategies that address real data quality issues across diverse datasets and domains.

A summary of 22 well-known Data Quality Issues, including Missing data, Ambiguous Data, and Extraneous, has been published by [Visengeriyeva and Abedjan, 2020]. This is augmented with their associated Data Quality Dimension violations in Appendix 1. The table presented there forms the grounding of this research into academic knowledge. All data quality problems found in our research are linked to these Data Quality Issues, and the ones below are the most common in our research:

Table 2 – Most common Data Quality Issues according to our understanding

#	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.	Accuracy, Completeness
4	Ambiguous data	Data values which might be interpreted in several ways (e.g., abbreviations).	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
9	Duplicates	Tuples/values that represent the same real-world entity.	Uniqueness
15	Domain violation	Values that violate semantic rules defined on the specific attribute.	Accuracy
17	Wrong data type	Values that violate the data type specification of the corresponding attribute, i.e., data type constraint violation.	Consistency

Ydata-Profiling, previously known as Pandas Profiling, is a very powerful and popular open-source Python library for Data Profiling, but one of the main features is the automatic generation of data quality alerts [Clemente et al., 2023]. Our research also produces some of their alerts, such as Missing Values, and Unique Values.

By analysing attribute labels or column names, our approach not only identifies standard data quality issues flagged by tools like Ydata-Profiling but also uncovers additional concerns not readily detectable through conventional means. A detailed comparison of our findings with those from their tool further illustrates the effectiveness and innovation of our approach and is described later.

2.4 Semantic Type Detection

Semantic type detection, crucial for data cleaning, involves identifying the type/format of columns in tables or attributes in datasets based on their semantic meaning. [Hulsebos et al., 2019] introduced a model, Sherlock, that employs a deep neural network for semantic type detection. Trained on over 680,000 data columns from the web sourced VizNet corpus [Hu et al., 2019], Sherlock successfully matches 78 semantic types to column headers from DBPedia [Auer et al., 2007]. Similarly, other systems utilize matching-based techniques, like dictionary look-up tables and regular expression matching, for semantic detection from column headers. Our methodology aligns with matching-based techniques for semantic detection from column headers. By significantly expanding the scope from Sherlock's 78 types to include 1800 words, including the 1100 words from [Ristoski et al., 2012] this extensive increase enables us to cover a much wider spectrum of insights for format identification, thus greatly enhancing our semantic detection capabilities.

[Yan and He, 2018] developed AUTOTYPE, which can synthesize type detection logic from data types, by using code from open-source repositories like GitHub and detect 84 semantic types with high precision. Their semantic types were developed in a different way, and our future research will observe all of them.

3 METHODOLOGY

This section outlines the systematic approach employed to leverage semantic information from attribute labels/ table columns or headers for data quality assessment, emphasizing the development of comprehensive dictionaries, the semantic type detection and analysis process, and the heuristic rule development for data quality issue identification.

By articulating the methodology in terms of objectives, methods, and expected outcomes, we establish a clear and focused research design that sets the stage for the subsequent practical application and results of the study.

3.1 Development of Semantic Analysis Tools and Initial Data Collection

- **Objective**: Develop comprehensive dictionaries for semantic type detection and collect a diverse set of datasets.
- **Methodology**: Creation of the first Formats Dictionary and Abbreviations Dictionary, followed by the collection of datasets from the UCI Machine Learning Repository using Python libraries.
- **Expected Outcome**: Foundational tools for semantic analysis and a comprehensive dataset collection.

3.2 Dataset Selection for Semantic Type Detection and Data Quality Assessment

- **Objective**: To select a first and second focused set of datasets from the comprehensive collection for an in-depth semantic type detection, testing the general applicability of the developed semantic analysis tools.
- **Methodology**: Application of the developed dictionaries to the collected datasets to identify a range of semantic types and potential data quality issues. Selection of datasets is based on diversity and relevance.
- **Expected Outcome**: A refined selection of sets of datasets for semantic type detection and data quality assessment, ensuring a broad representation of potential data quality challenges.

3.3 Attribute-Based Semantic Type Detection

- **Objective**: To conduct a semantic evaluation of attributes labels/column names or headers, and descriptions, to identify potential data formats/types for each attribute, utilizing the Formats and Abbreviations dictionaries.
- **Methodology**: This involves descriptive analysis, text mining, and rule-based classification to semantic evaluate attribute labels, standardize, and clean them across datasets. This step is crucial for determining the semantic type of each attribute based on the dictionaries, which are iteratively refined and enhanced based on the outcome.
- **Expected Outcome**: Standardized attribute labels associated with probable formats/types.

3.4 Attribute-Based Data Quality Assessment

- **Objective**: Conduct a thorough attribute-based data quality assessment of datasets, leveraging format identifications from the semantic type detection. This step entails validating attribute content against expected formats to ensure data integrity and identifying potential data quality issues where discrepancies occur. The analysis also extends to categorizing these issues according to their associated data quality dimensions, as in Appendix 1.
- **Methodology**: This phase capitalizes on insights from the semantic type detection to conduct attribute-based data quality assessment. It meticulously evaluates dataset attributes for conformity to determined formats, identifying potential data quality issues akin to alerts seen in tools like YData Profiling. When data adhere to expected formats, the process not only confirms data integrity but also details compliance with ranges and criteria. This examination covers many formats/data types, providing a detailed assessment of data quality and integrity.
- **Expected Outcome**: An attribute-based data quality assessment output that both affirms data quality when compliant with expected formats and reveals challenges where deviations occur. This includes validating data integrity for attributes matching expected formats and identifying specific data quality issues, along with their associated dimensions as detailed in Appendix 1.

3.5 Validation and Comparative Analysis

- **Objective**: Validate the effectiveness of "Attribute-Based Semantic Type Detection" and "Attribute-Based Data Quality Assessment" methodologies by assessing their performance in identifying and categorizing data quality issues compared to tools like YData Profiling.
- **Methodology**:
  - **Evaluation and Refinement in Action**: Outline criteria and metrics for comparison with YData Profiling, such as the range of detected data quality issues and depth of analysis.
  - **Validation Through Comparative Analysis:** Detail the framework comparing the developed methodology against tools like YData Profiling, specifying evaluation criteria to demonstrate effectiveness and advancements.
  - **Key Findings and Outcomes:** Expected enhancements in data quality assessment and methodology contributions, including anticipated results from the application.

- **Expected Outcome**: A comprehensive evaluation demonstrating the methodologies' effectiveness in advancing data quality assessment practices. This includes:
  - Enhanced detection of data quality issues not identified by existing tools.
  - A validated and refined framework emphasizing the innovative use of attribute labels.
  - Comparative analysis underscoring the methodologies' advantages over YData Profiling

4 RESULTS

4.1 Development of Semantic Analysis Tools and Initial Data Collection

The initial phase of our research involved the comprehensive development of semantic analysis tools and the collection of a diverse dataset set for testing. This section outlines the practical achievements related to these objectives:

4.1.1 Data Collection

- **Objective Achievement**: Successfully collected information from all 622 datasets available in the UCI Machine Learning Repository. This catalogue is a well-known source of research datasets encompassing many domains (called areas in the catalogue) and some of these datasets have had millions of accesses over time and were analysed in dozens of research papers. This rich catalogue was crucial for the generalizability and robustness of our semantic analysis approach.
- **Methodology Application**: Utilized Python libraries, such as Pandas and Beautiful Soup, for web scraping and data gathering. This process enabled the systematic collection of dataset attributes, metadata, and other relevant information necessary for our analysis.
- **Outcome**: The code used to download information from all 622 datasets and the resulting dataset created are available on GitHub ('UCICatalog-622DataSets.ipynb' and '622_Full_UCI_datasets.csv' [Silva, 2024]).

4.1.2 Development of Dictionaries

- **Objective Achievement**: Developed two critical resources for semantic analysis: the Formats Dictionary and the Abbreviations Dictionary, which are pivotal for detecting semantic types from attribute labels.
- **Methodology Application**: The Formats Dictionary was initially assembled with 1000 words associated with specific formats and expanded with insights from the Web Data Commons and Sherlock's semantic types, resulting in 1800 words/formats. Similarly, the Abbreviations Dictionary was compiled, translating more than 300 common abbreviations to their full expressions in words of the Formats Dictionary.
- **Outcome**: 'formats_dictionary.txt' and 'abbreviations_dictionary.txt ' [Silva, 2024].

Table 2 – Frequency distribution of the 1800 words regarding the Formats associated in the Formats Dictionary

#	Format	Frequency	Percentage	#	Format	Frequency	Percentage	#	Format	Frequenc	Percentage
1	string	752	41.78	13	postal code	3	0.17	25	month	1	0.06
2	categorical	378	21.00	14	weekday	3	0.17	26	normalized	1	0.06
3	numerical	277	15.39	15	E-mail format	2	0.11	27	numerical between 0 and 24	1	0.06
4	name	209	11.61	16	URL format	2	0.11	28	numerical between 0 and 360	1	0.06
5	numerical >= 0	106	5.89	17	age	1	0.06	29	numerical between 0 and 60	1	0.06
6	date	18	1.00	18	country	1	0.06	30	percentage	1	0.06
7	city	8	0.44	19	day	1	0.06	31	ph	1	0.06
8	phone	6	0.33	20	ID column	1	0.06	33	time	1	0.06
9	binary	5	0.28	21	IP format	1	0.06	34	week	1	0.06
10	datetime	5	0.28	22	latitude	1	0.06	35	year	1	0.06
11	state	4	0.22	23	longitude	1	0.06	Semi total		10	0.55556
12	street	4	0.22	24	model name	1	0.06	**Total**		**1800**	**100**
	Semi total	1772	98.44444		Semi total	18	1				

Table 2 shows the frequency distribution of the 1800 words regarding the Formats associated. Notice the cases in blue that are all 'Numerical Bounded', using the function 'Numerical Between'. Age is in range [0, 130] for example.

4.2 Dataset Selection for Semantic Type Detection and Data Quality Assessment

To ensure the developed semantic analysis tools were applied to a diverse and representative sample of data, a first targeted selection of ten datasets from the UCI Machine Learning Repository was undertaken. This process aimed to test the tools' applicability across a range of semantic types and potential data quality issues inherent in different domains.

4.2.1 First Dataset Selection

The first selection criteria prioritized ten datasets ensuring both popularity and academic relevance. We sought to cover a wide spectrum of data challenges by choosing datasets from five key areas: Life, Social, Physical, Computer, and Financial.

A systematic approach was employed to rank the datasets within each area, combining their web hits and the number of citations each dataset has received in research papers to identify the most significant datasets. This methodological approach ensured that the selected datasets would provide a robust foundation for our semantic type detection, illustrating the effectiveness of our semantic analysis tools in real-world contexts. For a detailed explanation of the dataset selection criteria and process, please refer to Appendix 2. Additionally, Figure 1 in Appendix 2 visually summarizes the selection of these ten datasets, highlighting the diversity and relevance of the chosen data.

Outcome: This refined selection process culminated in identifying ten datasets that offer a broad representation of potential data quality challenges across various domains ('TenDatasets.xlsx' [Silva, 2024]). Another important outcome is the file 'AllColumnsFromTenDatasets.xlsx' [Silva, 2024] which is the source for the next step. It contains information from the one of the ten datasets (index, name, and area) as well as the content of all the Columns, obtained from the column 'Attribute_info' from the '622_Full_UCI_datasets.csv' file. See Table 3 for all columns for the 'Heart Disease' dataset.

Table 3 – All columns from the 'Heart Disease' dataset

index	name	area	Original Column
45	Heart Disease	Life	1. age
45	Heart Disease	Life	2. sex
45	Heart Disease	Life	3. cp (chest pain type)
45	Heart Disease	Life	4. trestbps (resting blood pressure Integer)
45	Heart Disease	Life	5. chol (serum cholestoral in mg/dl Integer)
45	Heart Disease	Life	6. fbs (fasting blood sugar > 120 mg/dl Categorical) (1 = true; 0 = false)
45	Heart Disease	Life	7. restecg (resting electrocardiographic results Categorical)
45	Heart Disease	Life	8. thalach (maximum heart rate achieved Integer)
45	Heart Disease	Life	9. exang (exercise induced angina (1 = yes; 0 = no) Categorical)
45	Heart Disease	Life	10. oldpeak (ST depression induced by exercise relative to rest Integer)
45	Heart Disease	Life	11. slope (the slope of the peak exercise ST segment Categorical)
45	Heart Disease	Life	12. ca (number of major vessels (Categorical 0-3) colored by flourosopy)
45	Heart Disease	Life	13. thal (3 = normal; 6 = fixed defect; 7 = reversable defect Categorical)
45	Heart Disease	Life	14. num (the predicted attribute)

4.2.2 Second Dataset Selection

After the initial analysis of the first ten datasets, we expanded our focus to include a second set of forty datasets. The criteria for selecting these datasets are detailed in Appendix 3. The list of the forty chosen datasets can be found in 'FortyDatasets.xlsx' [Silva, 2024]. Another key document produced from this selection is 'AllColumnsFrom FortyDatasets.xlsx' [Silva, 2024]. The final number of attributes/columns analysed from the 50 datasets were 922.

4.3 Attribute-Based Semantic Type Detection

In the initial phase of our data quality assessment, we have automated the analysis of attribute labels through a Python notebook titled 'Attribute-BasedSemanticTypeDetection.ipynb' ([Silva, 2024]). This notebook facilitates a structured approach to identifying potential data formats for each attribute by analyzing target words and abbreviations found in attribute labels, cross-referencing them with our meticulously developed Formats and Abbreviations Dictionaries.

The process is encapsulated in a high-level algorithm, outlined in Appendix 4, which describes the systematic steps taken to extract, clean, and analyze the data from the attribute labels. This Appendix also shows the detailed results associated.

Some metrics were obtained. Table 4 below shows the frequency distribution of the formats found in the 922 columns.

Table 4 – Frequency Distribution of all final formats

ID	FinalFormat	Count	Percentage
1	numerical	338	36.66
2	numerical >= 0	243	26.36
3	categorical	196	21.26
4	binary	76	8.24
5	name	7	0.76
6	ID column	6	0.65
7	NaN	6	0.65
8	age	6	0.65
9	normalized	4	0.43
10	month	4	0.43
11	date	4	0.43
12	weekday	3	0.33
13	country	3	0.33
14	string	3	0.33
15	phone	2	0.22
16	longitude	2	0.22
17	latitude	2	0.22
18	datetime	2	0.22
19	year	2	0.22
20	day	2	0.22
21	numerical between 0 and 24	2	0.22
22	time	2	0.22
23	ph	1	0.11
24	percentage	1	0.11
25	model name	1	0.11
26	city	1	0.11
27	state	1	0.11
28	postal code	1	0.11
29	URL format	1	0.11
	Total	**922**	**100.00**

Here are some insights:

Numerical: The most common format, appearing 338 times, which is about 36.66% of the total.

Numerical >= 0: The second most common, with 243 instances, making up 26.36%.

Categorical: This format was identified 196 times, accounting for 21.26%.

Binary: Appears 76 times, or 8.24%.

NaN (Not a Number): There are 6 instances where the final format wasn't identified, which is about 0.65%. This means that for these columns, the format could not be determined based on the available information, or they did not match any criteria set in the dictionaries used for semantic type detection.

Other interesting analyses can be made regarding the formats in blue. They are all numerical bounded, which means that they have a minimum and a maximum value. Age is an example, which has the range [0, 130].

Formats related to identification ('name', 'ID column') are relatively uncommon.

Geographical data formats ('country', 'state', 'city') and temporal formats ('date', 'datetime', 'time') are also less common, with only a few occurrences.

Interestingly, there are 22 different formats being measured in our code, and 19 appeared here. The only ones that did not appear were 'Street', 'IP Format' and 'Email Format'.

Table 5 – 26 most prevalent words in the 50 datasets

SourceKeyword	Frequency	Amount	SourceKeyword	Frequency	Amount
0	number	61	13	boolean	15
1	binary	58	14	yes	14
2	freq	54	15	class	14
3	categorical	35	16	continuous	12
4	length	30	17	mean	11
5	take	28	18	count	10
6	nominal	22	19	burstiness	8
7	integer	18	20	contribution	8
8	area	18	21	qualitative	8
9	rate	17	22	interaction	8
10	numeric	16	23	increase	8
11	temperature	16	24	num	8
12	humidity	15	25	weight	7

Regarding the words that defined formats and were more prevalent, Table 5 shows the 26 most prevalent. 'Number', 'binary' and 'categorical' are in the top 4. 'Freq' appeared 54 times in just one dataset. Another similar word was 'take', appearing 28 times in just one dataset. 'Length', 'nominal', 'integer' and 'area' are some interesting words that appeared at least 18 times.

4.4 Attribute-Based Data Quality Assessment

During this stage, a Python notebook titled 'Attribute-BasedDataQualityAssessment.ipynb' [Silva, 2024] was developed leveraging format identifications from the previous semantic type detection part, critically analysing the attributes and contents of the selected datasets, aligning with what is outlined in Appendix 1.

The process is encapsulated in a high-level algorithm, outlined in Appendix 5, which describes the systematic steps taken to produce the data quality assessment. This Appendix also shows the detailed results associated.

4.4.1 Interesting results

Some of the interesting results from this analysis are pointed here:

- **Ambiguous Data Markers Identification**: A key result was the precise identification of ambiguous data markers, notably '?', across multiple datasets (e.g., datasets 45, 2, 103, 27 and many others).
- **Format-Specific Errors**: Our methodology adeptly identified format-specific errors, including implausible negative values (e.g., humidity levels in dataset 360).
- **Geographical and Temporal Format-Specific Analysis**: Detailed geographical (such as formats city, street, state, and country in dataset 225) and temporal analyses were conducted, revealing capitalization errors in geographical columns, and validating complex formats such as URLs, IP addresses, postal codes, and binary formats.
- **Advanced Error Detection**: Advanced error detection capabilities were showcased by identifying non-numerical values in numerical columns (e.g., 'InvoiceNo' in dataset 352) and uncovering instances with over 4000 categories in the 'StockCode' column in the same dataset 352. In the same dataset 352 it also found empty values in the 'Description' column, and also found a 'Non-String' value of 20713 in what is considered a String column.

Among the results presented are the Discoveries documents that exhibit all the outputs provided when each one of the 50 datasets had their Data Quality Assessment executed. Analysing the output from all Discoveries, two spreadsheets were created. The first one has the summary of all Data Quality Issues associated with their Data Quality Dimensions and explanations for situations when the formats were tested with created Bad Data. All 23 different formats have been tested with many different Bad Data. These are presented in Appendix 6.

Besides that, the second spreadsheet with the summary of all Data Quality Issues associated with their Data Quality Dimensions and explanations for the 18 datasets among the 50 datasets analysed which had Data Quality Issues, is presented in Appendix 7. The sheet 'SummaryofDiscoveries.xlsx' [Silva, 2024] contains all the results.

Below are general statistics obtained from the analysis of 106 columns (from the 922 columns in total) that had some Data Quality Issue associated (some had more than one):

1. **Data Quality Issue (DQI) Types Count:**
   - Ambiguous Data: 42 instances
   - Wrong Data Type: 33 instances
   - Domain Violation: 9 instances
   - Extraneous Data: 8 instances
   - Structural Conflicts: 3 instances
   - Duplicates: 3 instances
   - Uniqueness Violation: 3 instances
   - Missing Data: 3 instances
   - Non-String Data Type: 2 instances
2. **Columns With Issues Count:**

The columns 'Sample code number', 'CustomerID', and 'OSM_ID ' each have 3 instances of data quality issues.
Description, Name, Address, City, State, and Country each have 2 instances.
Other individual columns have 1 instance each. A total of 94 out of the 922 columns had Data Quality Issues: 10%.

3. **Data Quality Dimension Frequency:**
    - Accuracy, Consistency: 42 instances
    - Consistency: 35 instances
    - Consistency, Uniqueness: 11 instances
    - Accuracy: 9 instances
    - Uniqueness: 6 instances
    - Completeness: 3 instances
4. **Error Explanations Frequency:**
    - Unacceptable content: 42 instances
    - Non-numeric values: 29 instances
    - Extraneous data: 8 instances
    - Capitalization/Format issues: 5 instances
    - Data seems not categorical or has too many categories: 3 instances
    - Duplicate values: 3 instances
    - Uniqueness violation: 3 instances
    - Blank/Empty/Null/NaN values: 3 instances
    - Non-string values: 2 instances
    - Inconsistent length in alphanumeric values: 2 instances
    - Values outside range [1800, 2100]: 2 instances
    - Negative values: 2 instances
    - Incorrect telephone number format: 1 instance
    - Non-alphanumeric values: 1 instance
    - Invalid URL format: 1 instance

It is easy to see that the results presented are quite thorough.

4.4.2 'Missing Value' analysis

A final analysis can be made regarding the Data Issues themselves that were found. Although the content '?' was found in many different Data Quality Issues and Error Explanations, it appeared in the staggering amount of 77 out of the 106 total Data Quality Issues found. If we consider it as a 'Missing Value', which seems to be the case as many of the Attributes Descriptions detailed, and if we add the three cases where " was found, and one 'NA', we arrive to the total of 81 columns where Missing Values were found: 76,5%. Another interesting fact is that Missing Values were found in 14 out of the 18 datasets with Data Quality Issues. In other words, 28% of all the 50 datasets analysed had Missing Values found.

4.5 Validation and Comparative Analysis

It is crucial to contextualize its development through a comparative analysis with existing tools, such as YData Profiling (previously Pandas Profiling), the best performing Data Profiling library in Python [Gordon et al., 2022]. This section aims to highlight the limitations of current data quality assessment tools and underscore the necessity for our research.

4.5.1 Evaluation and Refinement in Action

Following the methodologies outlined in "Attribute-Based Semantic Type Detection" and "Attribute-Based Data Quality Assessment," applied across fifty diverse datasets from the UCI Machine Learning Repository, the process led to significant findings:

- **Iterative Refinement**: Adjustments in the Formats and Abbreviations Dictionaries enhanced our semantic analysis capabilities, showcasing the depth of our approach in identifying nuanced data quality issues beyond the reach of traditional tools.
- **Detected Data Quality Issues**: Our methodologies flagged a range of data quality issues, notably ambiguous data markers and wrong-data type errors, not typically detected by conventional methods.

4.5.2 Validation Through Comparative Analysis with YData Profiling

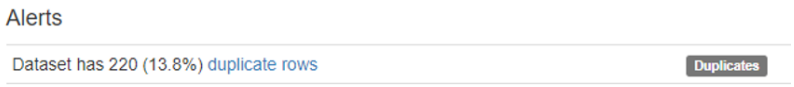
The comparative analysis was structured around applying both YData Profiling and our research methodologies to all the same set of fifty datasets. The two Discoveries documents were improved with each YData Alerts Analysis in all fifty datasets explanations.

4.5.3 Key Findings and Outcomes

Our analysis revealed several critical areas where YData Profiling Alerts falls short in addressing the nuanced needs of comprehensive data quality assessment:

- **Ambiguous Data Markers and Wrong Data Type Issues:** YData Profiling always missed the content '?' in datasets where they were prevalent, indicating a gap in its semantic analysis capabilities. Depending on the format of the attribute, the error was mostly classified as Ambiguous Data for Categorical formats or Wrong Data Type for Numerical formats.
- **Domain-Specific Error Detection**: The tool did not identify negative values as problematic in contexts where they are implausible, such as humidity levels in dataset 360, highlighting a lack of format-specific checks.
- **Unsupported Data Types**: YData Profiling showed limitations in supporting and analysing certain column titles and data formats, such as 'Time', which our method not only supported but validated for expected ranges.
- **Temporal and Geographical Data Analysis**: We provided detailed analysis for temporal and geographical data, as well as specific format checks like URL, IP, Postal Code, and Binary formats, while YData Profiling did not.

One of each of the Alerts found from all YData Analysis on the fifty datasets of this research are shown below:

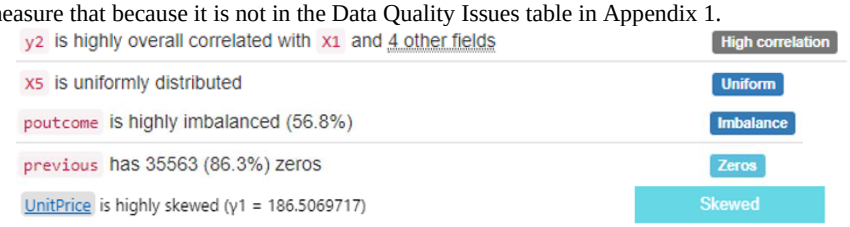


Observe that they do analyse Duplicates, but their duplicates relate to the whole record, with all columns. Our research analyses duplicate values only at the column level themselves. So, we measure on a different level.

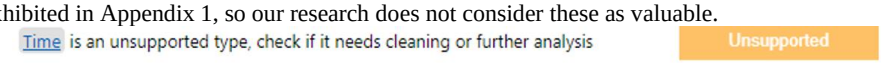
YData checks for Missing values, as can be seen in the case above. We also check for that, but in a broader way. We found cases where '?' or 'NA' or '' can qualify the column as Missing.

YData measures Uniqueness when there is no repetition in all their data for a column. We only measure this kind of uniqueness when we are evaluating probable ID columns, that need to have Unique values so that they can be considered apt to become a Primary Key.

YData measures when the content for a column has only one value, which can be considered a Constant. We do not measure that because it is not in the Data Quality Issues table in Appendix 1.



Above are the other 5 analysis YData produces. They do not relate to anything according to the Data Quality Issues exhibited in Appendix 1, so our research does not consider these as valuable.



And finally, to our surprise, YData has been producing this Unsupported output for some columns being analysed. We observed 11 different cases when they could not evaluate the column, and in all these cases we not only could evaluate, but we have found Data Quality Issues in almost all of them.

4.5.4 Precision, Recall, and F1 Score Calculation

Concluding, the only Data Quality Issue that can be compared regarding our research method and YData Profiling's method must be the Missing Data Issue.

Precision (P) measures the accuracy of the positive predictions (i.e., the proportion of positive identifications that were actually correct).

$$P = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Positives (FP)}}$$

For our method:

- TP for Missing Values = 81 (assuming all Missing Values we identified are correct)
- FP for Missing Values = 0 (assuming no incorrect Missing Values were identified, as we don't have evidence to the contrary)
- Therefore, P = 100%

Recall (R) measures the proportion of actual positives that were identified correctly (i.e., the ability of the model to find all the relevant cases).

$$R = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Negatives (FN)}}$$

For our method:

- TP for Missing Values = 81
- FN for Missing Values = 0 (since YData identified only one case that we also identified; we can assume there are no FNs for Missing Values)
- Therefore, R = 81/(81 + 0) = 100%

For YData Profiling:

- TP for Missing Values = 1 (since they identified one case that we also agree with)
- FN for Missing Values = 81 - 1 = 80 (all the cases they missed)
- Therefore, R = 1/(1 + 80) = 1.23%

F1 Score is the harmonic mean of Precision and Recall, a measure that takes both FP and FN into account.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

For our method, since P and R are both 100%, the F1 score is also 100%.

For YData Profiling:

- Given P is 100% (as they made one correct prediction and no wrong predictions for missing data)
- And R is 1.23%
- The F1 score for YData would be: F1 = 2 x (1 x 0.0123)/(1 + 0.0123) = 2.44%.

5 DISCUSSION

This research introduces the Attribute-Based Semantic Type Detection and Data Quality Assessment. Unlike traditional methods that primarily focus on structural and syntactic data issues, ours leverages semantic analysis to detect and categorize data quality issues. It produces semantic type detection, using words that exist in column names/headers or attribute labels, rule-based logic, setting new benchmarks for identifying and addressing complex data quality issues.

5.1 Semantic Rule-Based Analysis and Format-Specific Checks

At the heart of our research is a sophisticated rule-based analysis that translates column names/headers or attribute labels into standardized formats, guided by the Formats and Abbreviations dictionaries, which together contain 1800 words with the associated proposed formats, and 300 abbreviations with an associated word that is used in the previous dictionary. These dictionaries are pivotal in our semantic type detection, enabling it to discern and categorize data with a precision that outstrips conventional data quality tools. For example, our research effectively recognized '?'—often used in older datasets to denote 'unknown' or what could now be considered a 'Null' value—as an indicator of 'Ambiguous Data' in various datasets (e.g., datasets 45, 2, 103, 27). Furthermore, '?' was flagged under 'Wrong Data Type', 'Extraneous Data', and 'Domain Violation' categories. In contrast, YData Profiling failed to detect '?' as a data quality issue in 80 instances, marking it only once as 'Missing Data', a classification also used by our methodology.

Moreover, our method uniquely identifies format-specific errors, such as negative values in columns where they are implausible (e.g., humidity levels in dataset 360). It also recognizes 11 columns, such as 'Time', that YData Profiling deems unsupported. Our methodology not only supports these columns but also validates their data against expected ranges.

5.2 Geographical, Temporal, and other Format-Specific Insights

Another area where our research excels is in its geographical, temporal, and other format-specific analyses. For example, in dataset 225, we detected capitalization errors in geographical columns and performed specific analyses for different geographical formats (e.g., city, street, state, country). This capability extends to the analysis of URLs, IP addresses, postal codes, and binary formats, highlighting our research 's versatility and depth of analysis.

5.3 Advanced Error Detection

Our research’s advanced error detection capabilities are exemplified by its identification of non-numerical values in numerical columns (e.g. 'InvoiceNo' column in dataset 352), and the discovery of over 4000 categories in the 'StockCode' column in the same dataset 352. It may be correct to have over 4000 different Stock Codes, but at least the program exhibits this as a possible Data Quality Issue #10, related to Structural Conflicts. In the same dataset 352 it also found empty values in the 'Description' column, and besides that it also found a 'Non-String' value of 20713 in what is considered a String Format column. All these three columns are considered Unsupported in YData Profiling. Such findings are pivotal, demonstrating our ability to uncover a breadth of data quality issues, from completeness and consistency to accuracy and uniqueness violations.

5.4 Comparative Advantages

Our research methodology demonstrates a nuanced approach to data quality assessment, significantly surpassing tools like YData Profiling in detecting a broader spectrum of data quality issues. In our comprehensive analysis of 922 columns across 50 datasets, our method identified 81 instances of missing values—markedly more than the single instance YData Profiling recognized. This stark contrast underscores our method's superior ability to identify critical data quality issues, including ambiguous markers like '?' and 'NA', which YData frequently overlooked.

The precision and recall metrics further accentuate our method's effectiveness. With a precision of 100% and a recall of 100%, leading to an F1 score of 100%, our approach demonstrates flawless accuracy in identifying missing data, a fundamental aspect of data quality. In contrast, YData's performance on the same dataset yielded a precision of 2.44%, recall of 1.23%, and an F1 score of 2.44% in detecting missing data, highlighting significant gaps in its capability to identify such crucial data quality issues.

This comparative analysis not only highlights our methodology's comprehensive and precise nature but also its practical implications for enhancing data-driven decision-making processes. By ensuring analyses and conclusions are based on

accurate and reliable data, our approach promises to significantly improve data quality management practices, making it an invaluable tool for researchers, data scientists, and organizations aiming for high data integrity.

6 CONCLUSION

The Attribute-Based Semantic Type Detection and Data Quality Assessment, and its software package, significantly advance data quality assessment, highlighting the importance of attribute labels or column names/headers. This research integrates semantic rule-based analysis with format-specific examination, identifying a wide spectrum of data quality issues and enhancing data understanding. Using comprehensive Formats and Abbreviations dictionaries, it excels in semantic type detection and analysis, providing insights beyond traditional methods. It adeptly handles various data types - from numerical, including not negative and limited numerical, passing through categorical, ID, name, string, geographical and temporal data to diverse formats like URLs, IP addresses, E-mail, and binary values, showcasing its versatility.

Central to our methodology is a structured approach that includes data collection, semantic type detection, and the development of heuristic rules. Appendix 1 anchors our work in academic rigor, detailing data quality dimensions and issues. This precision, coupled with insights from analysing 50 datasets from the UCI repository, demonstrates our research's broad applicability and adaptability.

Particularly noteworthy is our method's capability to identify 81 instances of missing values, out of 922, significantly outperforming tools like YData Profiling, which detected only one. This finding, among others, highlights our approach's nuanced understanding and comprehensive coverage in detecting data quality issues.

The 'Discoveries' documents illustrate our research's practical impact and its capability in semantic type detection. By leveraging attribute label analysis for data quality issue detection—a feature often overlooked in traditional data quality assessments - our approach could streamline data cleaning processes, improving data management efficiency.

In conclusion, our methodology significantly outperforms existing tools like YData Profiling, as evidenced by an F1 score of 100% in detecting missing values—demonstrating a stark contrast to YData's 2.44%. Validated across 50 datasets, this research underscores the effectiveness of combining semantic analysis with format-specific examination, highlighting its potential to revolutionize data quality assessment and management practices.

7 FUTURE WORK

Our research is set to evolve further, focusing on key advancements:

- **Machine Learning Integration**: We plan to incorporate machine learning for more efficient, automated semantic type detection, reducing manual effort and improving adaptability to diverse data.
- **Expanding Dataset Analysis**: Broadening our analysis beyond the UCI Repository to include 50 additional datasets from varied sources will enhance our framework's generalizability and refine its detection capabilities.
- **Adhering to ISO Standards**: Aligning with international data quality standards like ISO will ensure our methodology meets global data quality benchmarks, increasing its applicability and credibility.
- **Enhancing Resources**: By expanding our dictionaries and resources, we aim to address a broader range of data types and formats, improving our framework's comprehensiveness.
- **Impact Evaluation**: Future work includes rigorous evaluations of the framework's effectiveness in real-world scenarios, demonstrating its value and potential for adoption.

These directions aim to refine our approach, ensuring it remains at the forefront of data quality assessment through innovative techniques and adherence to global standards.

8 REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. 722–735. dbpedia.pdf (upenn.edu)

[2] Anna Barberio. 2023. Large Language Models in Data Preparation: Opportunities and Challenges. 2023-12_Barberio_Tesi_01_.pdf (polimi.it)

[3] Carlo Batini and Monica Scannapieco. 2016. Data and Information Quality: Dimensions, Principles and Techniques | SpringerLink

[4] Fabiana Clemente, Gonçalo Martins Ribeiro, Alexandre Quemy, Miriam Seoane Santos, Ricardo Cardoso Pereira and Alex Barros. 2023, ydata-profiling: Accelerating data-centric AI with high-quality data - ScienceDirect, Neurocomputing, Volume 554, , 126585, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.126585.

[5] Tamraparni Dasu and Theodore Johnson. 2003. Exploratory Data Mining and Data Cleaning (wiley.com)

[6] DBheaders file. 2014. data.dws.informatik.uni-mannheim.de/webtables/2014-02/statistics/DBheaders.txt

[7] Lisa Ehrlinger and Wolfram Wöß. 2022. A Survey of Data Quality Measurement and Monitoring Tools - PMC (nih.gov). Front Big Data. 2022 Mar 31;5:850611. doi: 10.3389/fdata.2022.850611. PMID: 35434611; PMCID: PMC9009315.

[8] R. Buckminster Fuller. 1982. Critical Path - R. Buckminster Fuller - Google Books

[9] Ben Gordon, Clara Fennessy, Susheel Varma, Jake Barrett, Enez McCondochie, Trevor Heritage, Oenone Duroe, Richard Jeffery, Vishnu Rajamani, Kieran Earlam, Victor Banda and Neil Sebire. 2022. Evaluation of freely available data profiling tools for health data research application: a functional evaluation review | BMJ Open 12, e054186.. doi:10.1136/bmjopen-2021-054186

[10] Kevin Hu, Snehalkumar 'Neil' S. Gaikwad, Madelon Hulsebos, Michiel A. Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan and Çağatay Demiralp. 2019. Viznet: Towards a large-scale visualization learning and benchmarking repository. In CHI. ACM,. VizNet | Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (acm.org)

[11] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çağatay Demiralp, and César Hidalgo. 2019. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330993

[12] IDC. 2021. Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All? (marketresearch.com)

[13] David Loshin. 2002. Rule-Based Data Quality. https://dl.acm.org/doi/pdf/10.1145/584792.584894

[14] Petar Ristoski, Oliver Lehmberg, Heiko Paulheim and Christian Bizer. 2012 WDC - Web Table Corpus 2012 - Statistics about English-language Relational Subset (webdatacommons.org)

[15] Marcelo V. Silva. 2024. All codes and files developed for this research. Available at https://github.com/marcelovalentimsilva/Deriving-Knowledge-from-Attribute-Labels-for-Data-Quality-Assessment

[16] Immanuel Trummer. 2023. Can Large Language Models Predict Data Correlations from Column Names? | Proceedings of the VLDB Endowment (acm.org) 16, 13 (September 2023), 4310–4323. https://doi.org/10.14778/3625054.3625066

[17] Larysa Visengeriyeva and Ziawasch Abedjan. 2020. Anatomy of Metadata for Data Curation (acm.org). ACM Journal of Data and Information Quality (JDIQ).

[18] Richard Y. Wang, M. P. Reddy and Henry B. Kon. 1992. Toward quality data: an attribute-based approach (mit.edu)

[19] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers (jstor.org)

[20] Xiaolan Wang, Xin Luna Dong, and Alexandra Meliou. 2015. Data X-Ray | Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). Association for Computing Machinery, New York, NY, USA, 1231–1245. https://doi.org/10.1145/2723372.2750549

[21] Cong Yan and Yeye He. 2018. Synthesizing Type-Detection Logic for Rich Semantic Data Types using Open-source Code | Proceedings of the 2018 International Conference on Management of Data (acm.org) In SIGMOD'18: 2018 International Conference on Management of Data, June 10–15, 2018, Houston, TX, USA. ACM, NewYork, NY, USA,16 pages. https://doi.org/10.1145/3183713.3196888

9 HISTORY DATES

In case of submissions being prepared for Journals or PACMs, please add history dates after References as (*please note revised date is optional*):

Received; revised; accepted