Attribute-Based Semantic Type Detection and Data Quality Assessment

Marcelo Valentim Silva
Curtin University
Perth, WA, Australia
marcelo.valentimsilva@postgrad.
curtin.edu.au

Hannes Herrmann
Curtin University
Perth, WA, Australia
hannes.herrmann@curtin.edu.au

Valerie Maxville
Curtin University
Perth, WA, Australia
v.maxville@curtin.edu.au

***Abstract* - The increasing reliance on data-driven decision-making highlights the critical need for high-quality data. Despite advancements, data quality issues continue to impact both business strategies and scientific research. Current methods often fail to utilize the semantic richness embedded in attribute labels (or column names/headers in tables), leading to a crucial gap in comprehensive data quality evaluation.**

This research addresses this gap by introducing an innovative methodology focused on Attribute-Based Semantic Type Detection and Data Quality Assessment. By leveraging semantic information in attribute labels, combined with rule-based analysis and comprehensive dictionaries, our approach effectively addresses four key Big Data challenges: variety, veracity, volume and value.

Our method provides a practical classification system of 23 semantic types, including numerical non-negative, categorical, ID, names, strings, geographical, temporal, and complex formats like URLs, IP addresses, email, and binary values plus several numerical bounded types, such as age and percentage (variety). The approach was validated across fifty diverse datasets from the UCI Machine Learning Repository, covering multiple domains, further highlighting its adaptability (variety). We also compared our types with the ones from Sherlock, a renowned method for Semantic Type Detection.

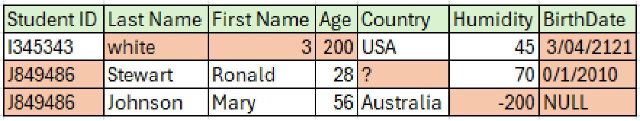
Our evaluation showcases our method's proficiency in identifying data quality issues, detecting 81 missing values out of 922 attributes, compared to only one detected by YData Profiling (veracity). One dataset, containing over 2 million records, was processed efficiently, demonstrating the scalability of our approach (volume). These results underscore the enhanced capabilities of our method in streamlining data cleaning processes, ultimately improving the efficiency and effectiveness of data-driven decision-making across various domains (value).

Keywords - Data quality, Semantic Type Detection, Big Data

I. 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 should pay more attention to the semantic richness of attribute information, such as words in attribute labels, names or headers in table columns, and/or related descriptions/metadata. This oversight leads to a significant 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.

TABLE I. EXAMPLE OF A DIRTY DATASET



Student ID	Last Name	First Name	Age	Country	Humidity	BirthDate
I345343	white	3	200	USA	45	3/04/2121
J849486	Stewart	Ronald	28	?	70	0/1/2010
J849486	Johnson	Mary	56	Australia	-200	NULL

As illustrated in Table I, common data quality issues can be detected by analyzing the words in the attribute labels/column names and their formats. Issues include duplicated IDs in the 'Student **ID**' column, incorrect capitalization in 'Last **Name'**, numerical values in 'First **Name'**, out-of-range values in '**Age**', missing values represented by '?' in **'Country'**, negative values in **'Humidity'**, and invalid dates in 'Birth**Date**'. The **bold words** in the column names provide an example of how we might find useful semantic information to support quality assessment. Our method detects these problems simply by analysing the words in the attribute labels, inferring probable data formats, and examining the content accordingly. Importantly, it links them with Data Quality Issues and their associated Dimensions as defined in respected research by Visengeriyeva and Abedjan [19].

This research introduces a novel approach to enhance data quality assessment by harnessing the untapped semantic information within attribute labels. Our method identifies and categorises potential data quality issues before traditional data cleaning processes begin by systematically analyzing these labels to derive knowledge about the expected format and data content.

Given the current challenges, we formulated and addressed the following research questions:

A. Research Questions

1) Can attribute labels be effectively used for semantic type detection and subsequent data quality assessment?

2) How does our attribute-based approach identify data quality issues across diverse datasets?

3) What types of data quality issues can our method detect that might be overlooked by traditional data profiling tools?

B. *Key Contributions aligned with four Big Data Vs*

1) **Quality Assessment with Semantic Labels**: Enhances **Variety** by leveraging semantic attribute labels to classify around 30 types covering diverse domains such as Life, Social, Physical, Computer, and Financial.

2) **Improved Issue Detection**: Supports **Veracity** by providing superior detection of key data quality issues, such as missing values and domain violations, validated across diverse UCI datasets and against YData Profiling.

3) **Scalable**: Handles **Volume** by assessing the data quality of datasets exceeding 2 million records efficiently.

4) **Valuable Data Insights**: Enhances **Value** by improving data usability and ensuring more accurate decision-making for practical applications.

C. *Approach and Significance*

To address these questions, we developed a methodology for Attribute-Based Semantic Type Detection and Data Quality Assessment by using semantic cues within attribute labels. Our method combines rule-based analysis with comprehensive dictionaries, focusing on 23 semantic types, including numerical non-negative, categorical, names, geographical, temporal, and complex formats like URLs, IP addresses, email, and binary values plus several numerical bounded types, such as age and percentage.

Our methodology addresses four key Big Data characteristics: **variety**, **veracity**, **volume**, and **value**. We address **variety** through multi-domain semantic classification, **veracity** with rigorous quality checks, **volume** with efficient processing of large datasets, and **value** by enhancing data reliability and usability for better decision-making

A comparative analysis with Sherlock, a state-of-the-art semantic type detection system, demonstrated our approach's effectiveness, particularly in addressing practical data science types and bounded numerical values.

Our evaluation of fifty datasets from UCI uncovered 106 data quality issues, far surpassing traditional profiling tools like YData Profiling. This showcases our approach's potential to address data quality needs effectively across multiple domains.

D. *Delimitation*

This research only works where there are words in English in the label of the attributes in datasets or the names or headers of columns in tables in databases and where content can be easily evaluated against the formats determined.

It evaluates file formats, including .txt, .csv, .data, .xls and .xlsx. Compacted files such as .zip, .tar.gz, and .gz. that contain the previous file formats can also be analyzed. When the initial file is a .zip or a .tar.gz file, the user chooses the file to be analyzed. Besides that, .gz files contain only one file inside, so it is automatically analyzed.

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).

II. Related Work

Historical and ongoing challenges in data quality significantly affect both business strategies and scientific research, highlighting the importance of robust data quality management [7], [11], [21]. Data cleaning inefficiencies, as noted in [4], can consume up to 80% of analysis time. Automated solutions for detecting data quality issues through semantic analysis of attribute labels aim to streamline this process.

A. *Data quality assessment*

Data quality is crucial for data usability [22]. Data profiling enhances data quality findings and improves user understanding [13], [14]. Traditional methods often overlook semantic richness in attribute labels, which is a significant area of opportunity. Similar needs are reflected in tools like Data X-Ray [23], which emphasize automation in profiling, and ISO/IEC standards [6], which stress the importance of syntactic, semantic, and pragmatic dimensions in data quality.

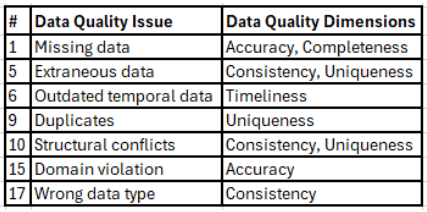
B. *Data quality dimensions and attribute analysis*

Data attributes (columns or fields) contain labels (names or headers) that can offer valuable semantic cues. Metadata at the column level often detail data types and constraints. Key quality dimensions for attribute analysis include accuracy, completeness, consistency, and timeliness [2, 4]. Our research builds on the limited literature around column names, such as [18], by focusing on semantic signals in attribute labels to enhance data quality assessment, addressing gaps not widely covered in current literature.

C. *Data quality issues*

Loshin [12] demonstrated a business rules approach for identifying the underlying causes of poor data quality by transforming declarative data quality rules into actionable code. A study of 22 well-known data quality issues, including missing data and associated data quality dimension violations, is detailed in [19]. Table II presents a summary of the main issues and their associated dimensions which are better discussed in Appendix 1 available in our GitHub page [17].

TABLE II. The Main Data Quality Issues and their Associated Dimensions [19]



#	Data Quality Issue	Data Quality Dimensions
1	Missing data	Accuracy, Completeness
5	Extraneous data	Consistency, Uniqueness
6	Outdated temporal data	Timeliness
9	Duplicates	Uniqueness
10	Structural conflicts	Consistency, Uniqueness
15	Domain violation	Accuracy
17	Wrong data type	Consistency

Our method generates alerts similar to those from YData Profiling (previously Pandas Profiling) [3], such as Missing and Unique Values. Importantly, our alerts are mapped to 11 of the 22 Data Quality Issues from [19], anchoring our study in existing academic knowledge. A detailed comparison with YData Profiling illustrates our approach's effectiveness.

D. *Semantic Type Detection*

Semantic type detection is crucial for data cleaning as it identifies data types/formats based on semantic meaning. One notable approach is **Sherlock** by Hulsebos et al. [10], which uses deep learning trained on over 680,000 data columns from the VizNet corpus [9] to identify 78 semantic types. It also leverages DBpedia [1] for mapping column headers to a knowledge base, providing a rich semantic context. However, the high number of types identified by Sherlock can introduce unnecessary complexity for practical data quality tasks. Our method contrasts this by initially analyzing attribute labels, followed by content verification, to maintain a focused and practical set of approximately 30 semantic types. This approach aims to balance comprehensiveness and practical application while avoiding excessive complexity, maintaining high accuracy while being more directly applicable to common data quality assessment tasks.

III. METHODOLOGY

This section details our approach for semantic type detection, data quality assessment, and identifying overlooked issues in traditional tools.

A. *Dataset Selection and Preparation*

- Selection of diverse datasets to test semantic type detection and data quality methods.
- Ensure dataset diversity and preprocessing for various file formats and structures.

B. *Semantic Type Detection Using Attribute Labels*

RQ 1: *Can attribute labels be effectively used for semantic type detection and subsequent data quality assessment?*

- Development of formats and abbreviations dictionaries based on Web Data Commons [1, 5, 15].
- Development of the rule-based classification algorithm for attribute labels/ column names/headers.
- Definition of distinct semantic types, including common and edge cases like URLs and IP addresses.

C. *Data Quality Assessment Across Diverse Datasets*

RQ 2*: How does our attribute-based approach identify data quality issues across diverse datasets?*

- Validation of content against expected formats and identification of potential data quality issues.
- Mapping identified issues to data quality dimensions.

D. *Comparative Analysis with Traditional Tools*

RQ 3: *What types of data quality issues can our method detect that might be overlooked by traditional data profiling tools?*

- Check issues detected compared to YData Profiling.
- Employment of quantitative and qualitative measures.

IV. RESULTS

A. *Dataset Selection and Preparation*

We collected information from all 622 datasets available in the well-known UCI Machine Learning Repository (https://archive.ics.uci.edu/) covering many domains to ensure generalizability and robustness in our research. Using Python libraries, like Pandas and Beautiful Soup, we performed web scraping to gather dataset attributes and metadata. The code for downloading dataset information ('UCICatalog-622DataSets.ipynb') and the resulting dataset ('622_Full_UCI_datasets.csv') are available in our GitHub [17].

Using iterative selection criteria, we focused on two groups. The first selection included ten datasets based on popularity and academic relevance, across the domains of Life, Social, Physical, Computer, and Financial (Appendix 2 [17]). The second selection added forty additional datasets for broader representation (Appendix 3 [17]). Ranking was based on web hits and citation counts, ensuring robust foundations for semantic type detection in real-world contexts.

The final selection ('Fiftydatasets.xlsx' [17]) represents diverse data quality challenges with 922 attributes, providing a strong base for testing our semantic type detection. All attributes from the selected datasets are compiled in the file 'AllColumnsFromFiftyDatasets.xlsx' [17], which served as the basis for our Semantic Type Detection iteration.

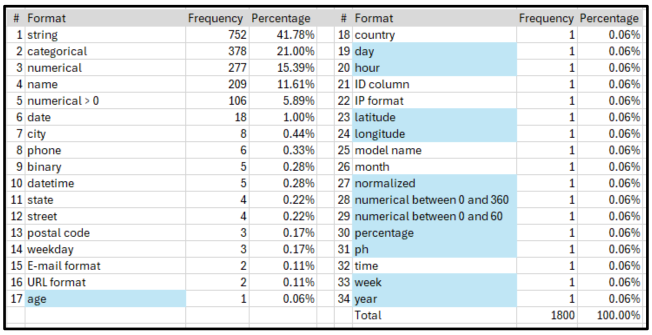
One of the datasets, ID 235 - 'Individual household electric power consumption', contained over 2 million records, and the analysis lasted only 20 minutes using a 16 GB RAM notebook, demonstrating scalability.

B. *Semantic Type Detection Using Attribute Labels*

1) *Development of Semantic Analysis Tools*

We created the Formats Dictionary ('formats_dictionary.txt' [17]) with insights from over 90 million tables from the overall Web Table Corpus - Web Data Commons [15] and a list with 1100 of the words most commonly found in column names [1, 5] and expanded it to over 1800 words connected to their respective formats, using ChatGPT – GPT 4. We also created the Abbreviations Dictionary ('abbreviations_dictionary.txt' [17]) which translated over 300 common abbreviations into full terms in the other dictionary (e.g., 'pct' to 'percentage').

TABLE III. FREQUENCY DISTRIBUTION OF THE 1800 WORDS REGARDING THE FORMATS ASSOCIATED IN THE DICTIONARY.



#	Format	Frequency	Percentage	#	Format	Frequency	Percentage
1	string	752	41.78%	18	country	1	0.06%
2	categorical	378	21.00%	19	day	1	0.06%
3	numerical	277	15.39%	20	hour	1	0.06%
4	name	209	11.61%	21	ID column	1	0.06%
5	numerical > 0	106	5.89%	22	IP format	1	0.06%
6	date	18	1.00%	23	latitude	1	0.06%
7	city	8	0.44%	24	longitude	1	0.06%
8	phone	6	0.33%	25	model name	1	0.06%
9	binary	5	0.28%	26	month	1	0.06%
10	datetime	5	0.28%	27	normalized	1	0.06%
11	state	4	0.22%	28	numerical between 0 and 360	1	0.06%
12	street	4	0.22%	29	numerical between 0 and 60	1	0.06%
13	postal code	3	0.17%	30	percentage	1	0.06%
14	weekday	3	0.17%	31	ph	1	0.06%
15	E-mail format	2	0.11%	32	time	1	0.06%
16	URL format	2	0.11%	33	week	1	0.06%
17	age	1	0.06%	34	year	1	0.06%
					Total	1800	100.00%

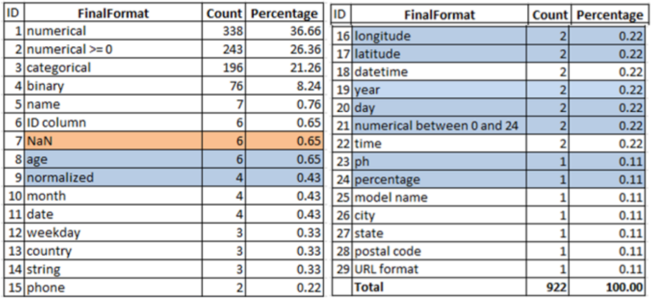
Table III shows the frequency distribution of the formats in the dictionary, with 106 words related to non-negative numbers and numerically bounded cases, such as 'age' and 'latitude' (marked in blue).

2) *Rule-Based Classification Results*

We developed a Rule-Based Classification Approach in the Python notebook 'Attribute-BasedSemanticType Detection.ipynb' [17], which automated the analysis of attribute labels by cross-referencing them with our Formats and Abbreviations Dictionaries.

The high-level algorithm, described in Appendix 4 [17], details the steps for extracting, transforming, and analyzing data from attribute labels.

TABLE IV. FREQUENCY DISTRIBUTION OF ALL 922 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

3) Semantic Type Classification System

Our semantic type classification system was designed to balance comprehensiveness with practicality. We identified 19 distinct formats/types and nine numerical bounded cases such as age or year (in blue in Table IV) across the 922 columns analyzed from the fifty datasets.

- **Numerical Format**: The most common type, found in 36.66% of columns.
- **Non-Negative Numbers**: Represented 26.36%, allowing unique insights not commonly seen in other data quality assessments.
- **Categorical Data**: Present in 21.26% of columns.
- **Binary Data**: Appeared in 8.24% of columns.
- **Bounded Numerical Types**: Included specific cases like 'age' and 'normalized' values, enhancing data quality checks, allowing new discoveries.
- **Rare Formats**: Identification (e.g., 'name', 'ID'), geographical (e.g., 'country', 'city'), and temporal (e.g., 'date') formats were less common but still relevant.
- **Unclassified Data (NaN)**: A small portion (0.65%) remained unclassified.
- **Other Formats**: Formats like 'Street', 'IP', and 'Email' were not found in these 50 datasets, highlighting areas for future data expansion.

Process for mapping attribute labels to semantic types:

The mapping process matched words in attribute labels against the Formats Dictionary. We first analyzed column names and, if unsuccessful, examined column descriptions. Conflicts between names and descriptions were resolved, with special handling for 'ID' columns and specific patterns. When no format was determined, we referenced the Abbreviations Dictionary or assigned a default 'NaN'. Matches were assigned a semantic type based on a priority system considering specificity and position in the Formats Dictionary.

4) Addressing Research Question 1:

'Can attribute labels be effectively used for semantic type detection and subsequent data quality assessment?'

Our results strongly affirm the effectiveness of using attribute labels. We successfully classified 99.35% of the 922 columns analyzed, with only 0.65% remaining unclassified. This demonstrates that attribute labels can be effectively leveraged for semantic type detection, identifying 19 distinct formats, including specialized types like non-negative numbers and bounded numerical types. This provides a solid basis for subsequent data quality assessments.

5) Comparative analysis with Sherlock

We compared our semantic types with Sherlock's types (Appendix 8 [17]). This comparison helps evaluate the effectiveness of our classification system against state-of-the-art methods. Here are the key differences from Sherlock:

- **Granularity:** Sherlock offers 78 types, while our method focuses on 23 types plus some bounded numerical types.
- **Practicality:** Our types are directly applicable to common data quality tasks.
- **Handling of Bounded Types:** We explicitly handle bounded types (e.g., age, percentage), whereas Sherlock does not.
- **Ease of Classification:** Fewer classes reduce the risk of misclassification.
- **Grouping and Extensibility:** We aggregate Sherlock's specific types into broader categories (e.g., 'Artist', 'Creator', 'Director' under the 'name' type in the Formats Dictionary) and include types like URL, latitude, and phone, enhancing our classification for geospatial, contact, and web data**.**

C. Data Quality Assessment Findings

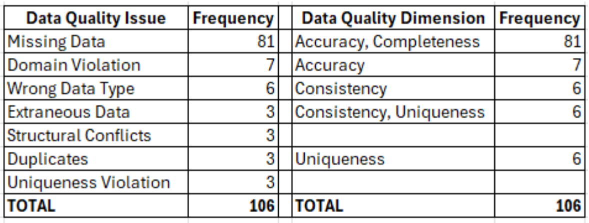
1) Data Quality Issue Identification Results

We developed a Data Quality Assessment approach in the Python notebook 'Attribute-BasedDataQualityAssessment.ipynb' [17], aligning with what is in Appendix 1 [17]. This approach leveraged previous semantic type identifications to analyze dataset attributes and content. The detailed algorithm and results are provided in Appendix 5 [17].

2) Data Quality Issues and Dimensions Analysis

Below are quantitative measures obtained from the analysis of 106 columns (11,5% from the 922 columns in total) with at least one Data Quality Issue found:

TABLE V. DATA QUALITY ISSUES AND DIMENSIONS FREQUENCIES

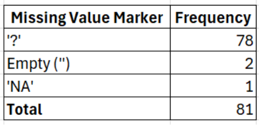


Data Quality Issue	Frequency	Data Quality Dimension	Frequency
Missing Data	81	Accuracy, Completeness	81
Domain Violation	7	Accuracy	7
Wrong Data Type	6	Consistency	6
Extraneous Data	3	Consistency, Uniqueness	6
Structural Conflicts	3		
Duplicates	3	Uniqueness	6
Uniqueness Violation	3		
TOTAL	**106**	**TOTAL**	**106**

3) Key Findings and Missing Value Analysis

- **Missing data identification:** A key result was the precise identification of missing data markers, notably '?', across 14 of the 18 datasets:

TABLE VI. MISSING DATA FREQUENCY



Missing Value Marker	Frequency
'?'	78
Empty ('')	2
'NA'	1
Total	81

- **Format-specific errors:** Identified implausible negative values, e.g., humidity levels in dataset 360 (Air quality).

- **Geographical and temporal format-specific analysis**: Identified capitalization issues in geographical formats, e.g., city and state in dataset 225 (Restaurant & consumer data) and validated complex formats like URLs and postal codes.
- **Advanced error detection**: Detected non-numerical values in numerical fields ('InvoiceNo' in dataset 352 (Online Retail) and instances with over 4000 categories in the 'Description' column of the same dataset 352.

4) Data Quality Issues Summary

Among the results are two 'Discoveries' documents that present the complete outputs from the analysis of all 50 datasets (initial 10 plus the additional 40 datasets) [17]. These documents provide detailed explainability, offering thorough insights into each output.

Additionally, two sheets in the spreadsheet, 'Summaryof Discoveries.xlsx' [17] were created to analyze the outcomes from the two 'Discoveries' documents:

- First sheet: Summarizes Data Quality Issues and their respective Data Quality Dimensions, explaining scenarios where different formats were tested with created Bad Data. The detailed results of testing all 23+ formats with Bad Data are presented in Appendix 6 [17].
- Second sheet: Provides a summary of all Data Quality Issues and their associated dimensions found (in 18 datasets out of the 50) that showed data quality issues. The detailed results are presented in Appendix 7 [17].

5) Addressing Research Question 2:

'How does our attribute-based approach identify data quality issues across diverse datasets?'

Our attribute-based approach identified data quality issues in 106 columns (11.5% of the total 922 columns), spanning 18 of the 50 datasets (36%). Missing values accounted for 76.4% of identified issues. The robust performance across different datasets and domains underscores the approach's versatility in detecting a wide range of data quality issues.

D. Comparative Analysis with Traditional Data Profiling Tools

1) Comparison with YData Profiling

To validate our approach, we conducted a comparative analysis with YData Profiling, a widely-used data profiling tool downloaded over 50,000 times per month [8, 20]. Both YData Profiling and our methods were applied to the same 50 UCI datasets and 922 attributes. The 'Discoveries' documents were updated with findings from each YData alert analysis.

Alerts

CustomerID has 135080 (24.9%) missing values Missing

YData identifies missing values, and **it found only one single case** of empty strings, while our approach found the same case, plus 78 cases of '?', another empty string, and one 'NA', yielding more comprehensive results in identifying missing data.

YData also identifies various other types of alerts, but these are not comparable to our findings, as they do not align with the 22 Data Quality Issues defined in Appendix 1 [17].

Our research strictly focuses on these specific issues, providing a targeted and consistent framework for data quality evaluation.

2) Addressing Research Question 3:

'What types of data quality issues can our method detect that might be overlooked by traditional data profiling tools?'

TABLE VII. QUALITATIVE RESULTS OF COMPARISON WITH YDATA PROFILING

Situation	Our research	YData Profiling Alerts
Content '?'	78	0
Missing Values	81	1
Negative Values	2 (e.g.' Humidity')	0
Geographical and Temporal	5	0
Non-String and Non-Numerical	4	0

Table VII presents the qualitative results of our comparison with YData Profiling. This comparative analysis reveals that our attribute-based method detects several data quality issues that YData Profiling overlooks:

1. Identification of '?' as missing values, which YData Profiling consistently missed. It only identified one case of missing data with an empty value.

2. Detection of domain-specific errors, such as negative values in non-negative fields (e.g., humidity levels).

3. Ability to analyze and validate a broader range of data types, including geographical and temporal data and other cases such as Non-string and Non-numerical data.

These findings clearly demonstrate the enhanced detection capabilities of our attribute-based approach compared to traditional data profiling tools, addressing complex data quality issues that might otherwise go unnoticed.

V. CONCLUSION

This research introduced an **Attribute-Based Semantic Type Detection** and **Data Quality Assessment** approach that leverages semantic information from attribute labels to improve data quality management.

A. Key Findings and Significance:

Effectiveness: Successfully classified 99.35% of the 922 columns across 50 datasets, demonstrating the effectiveness of using semantic analysis of attribute labels/ column names/headers for data quality assessment.

Improved Detection: Identified 81 instances of missing values compared to YData Profiling's one instance, highlighting enhanced detection capabilities.

B. Impact and Implications:

Big Data Relevance: Our method effectively addresses four key Big Data characteristics: **variety** (through diverse semantic classifications across domains), **veracity** (with comprehensive data quality checks), **volume** (via scalable processing of large datasets), and **value** (by improving usability for informed decision-making).

Cost Efficiency: Streamlines data cleaning processes, reducing the time and resources required for data preparation.

C. Limitations:

The approach was tested on datasets from the UCI Machine Learning Repository. Future testing across broader data sources will help verify generalizability.

Summary: This research advances data quality management by leveraging semantic information in attribute labels. The approach significantly enhances both efficiency and effectiveness in data quality assessment, addressing **Big Data** challenges in **variety, veracity, volume**, and **value**.

VI. FUTURE WORK

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

- **Machine Learning and Large Language Models Integration**: We plan to incorporate machine learning for more efficient, automated semantic type detection, reducing manual effort and improving adaptability to diverse data, following what was done in the Sherlock paper. We also intend to use Large Language Models API to improve the maintenance of the dictionaries.
- **Expanding Dataset Analysis**: We plan to broaden our analysis beyond the UCI Repository to include fifty additional database tables from an Open-Source Database (db.relational-data.org) enhancing our framework's generalizability refining its detection capabilities.
- **Expanding Comparison with other Data Profiling Alert systems**: Besides YData Profiling, we intend to analyze other Open-Source libraries that provide similar alert results, such as DataPrep.EDA and Autoviz.
- **Adhering to ISO and Industry Standards**: We plan to align with international data quality standards like ISO and Industry Standards such as HL7, SNOMED CT (Health), FIBO and XBRL (Finance), and GS1 and EDI (Commerce) which will ensure that our methodology meets global data quality benchmarks, increasing its applicability and credibility.

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.

For a more comprehensive version of this research, with 10 pages, please refer to our ArXiv version [16].

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

[3] 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. ydata-profiling: Accelerating data-centric AI with high-quality data - ScienceDirect

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

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

[6] 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. https://doi.org/10.3389/fdata.2022.850611. PMID: 35434611; PMCID: PMC9009315.

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

[8] 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 Open12, e054186. https://doi.org/10.1136/bmjopen-2021-054186

[9] 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. [1905.04616] 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)

[10] 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

[11] IDC. 2023. Worldwide IDC Global DataSphere Forecast, 2023-2027: It's a Distributed, Diverse, and Dynamic (3D) DataSphere International Data Corporation

[12] David Loshin. 2002. Rule-Based Data Quality. https://doi.org/10.1145/584792.584894

[13] Felix Naumann. 2014. Data profiling revisited | ACM SIGMOD Record. ACM SIGMOD Record 42, 40–49. https://doi.org/10.1145/2590989.2590995

[14] Jack E. Olson. 2003. Chapter 7 - Data Profiling Overview - ScienceDirect, in Data Quality – The Accuracy Dimension, The Morgan Kaufmann Series in Data Management Systems, Pages 121-142. https://doi.org/10.1016/B978-155860891-7/50011-1

[15] Petar Ristoski, Oliver Lehmberg, Heiko Paulheim and Christian Bizer. 2012. WDC - Web Table Corpus 2012 - Statistics about English-language Relational Subset http://madata.bib.uni-mannheim.de/212/

[16] Marcelo V. Silva, Hannes Herrmann, and Valerie Maxville. 2024. [2410.14692] Attribute-Based Semantic Type Detection and Data Quality Assessment

[17] Marcelo V. Silva. 2024. All codes and files are available at https://github.com/marcelovalentimsilva/Attribute-Based-Semantic-Type-Detection-and-Data-Quality-Assessment

[18] 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

[19] Larysa Visengeriyeva and Ziawasch Abedjan. 2020. Anatomy of Metadata for Data Curation | Journal of Data and Information Quality

[20] Pepy.tech. 2024. ydata-profiling download stats (pepy.tech). https://pepy.tech/project/ydata-profiling

[21] Richard Y. Wang, M. P. Reddy and Henry B. Kon. 1992. Toward quality data: An attribute-based approach - ScienceDirect

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

[23] 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