DREIFLUSS: A Minimalist Approach for Table Matching

Vishvapalsinhji Parmar*1*, Alsayed Algergawy*1*

1Chair of Data and Knowledge Engineering, University of Passau Passau, Germany

Abstract

This paper introduces DREIFLUSS, an innovative, minimalist approach designed to tackle the Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks in the SemTab challenge. DREIFLUSS efficiently employs semantic information from well-established knowledge graphs, DBpedia, and Schema.org, to improve the annotation process. Experimental evidence illustrates the superior performance of logistic regression models trained via DREIFLUSS, resulting in precise column-type annotations and insightful relationship predictions. The findings substantiate the significance of proper sampling technique while training a model, thereby boosting the accuracy and efficiency of table matching. This research illuminates a promising pathway to enhance table matching techniques, underlining the practical ramifications of DREIFLUSS for data integration and knowledge discovery endeavors.

Keywords

Table matching, Column Type Annotation (CTA), Column Property Annotation (CPA), knowledge discovery, data integration

1. Introduction

Table matching, as a critical part of data integration and knowledge discovery, is gaining increasing attention in this age of digital information proliferation. It harmonizes information across different tables, thereby paving the way for extracting valuable insights. An estimated millions of high-quality tables are currently accessible on the Internet, a figure that continues to rise due to the progression of automated data extraction techniques and an increasing reliance on structured data across various sectors, including business, academia, and government [1].

In this context, the SemTab challenge[1] emerges as a pivotal competition, advancing the frontiers of table understanding and annotation. The challenge emphasizes the importance of Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks, accentuating the significance of accurate column labeling and the identification of inter-column relationships for comprehensive table understanding. Addressing this challenge, we introduce "DREIFLUSS", a minimalist yet effective approach tailored for the tasks presented in the SemTab challenge. DREIFLUSS harnesses the labels defined in two major knowledge graphs, DBpedia and Schema.org, as a guiding force to improve the table matching process. The labels in these knowledge graphs

[1]https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

provide an exhaustive set of semantic details that can improve the accuracy and efficiency of table matching.

DBpedia, a vast knowledge graph sourced from Wikipedia, presents a broad spectrum of labels spanning various domains. Conversely, Schema.org, a collaborative initiative by leading search engines, offers a universally accepted schema vocabulary. By leveraging these labels, DREIFLUSS underscores the potential of using pre-existing semantic resources to enhance the table annotation process. This strategic utilization of label resources simplifies the implementation, while enhancing the system's scalability and adaptability - key traits in the current era of data explosion [2, 3]. The ensuing sections offer a comprehensive overview of the CTA and CPA tasks, data specifics, the experimental design we used, and the outcomes achieved. The detailed evaluation of the DREIFLUSS system clarifies its capabilities in handling table matching tasks within the SemTab challenge and beyond.

The rapid proliferation of structured data on the web opens up vast opportunities for knowledge discovery and data integration but also introduces significant challenges. The data is typically embedded in tables, each with its own unique structure, schema, and notation. Extracting valuable and accessible information from these tables calls for advanced methods of data understanding and harmonization. In this context, competitions such as SemTab take center stage, pushing the boundaries of table understanding and annotation. Of the tasks involved, Column Type Annotation (CTA) and Column Property Annotation (CPA) are pivotal. These tasks entail precise labeling of columns and establishing relationships among them, both crucial for comprehensive table understanding, efficient data integration, and adequate knowledge discovery.

To address these needs, we present an innovative methodology explicitly designed to tackle these tasks. While minimalist in its approach, this system leverages the existing labels in two prominent knowledge graphs - DBpedia and Schema.org. These labels represent a rich, comprehensive, and standardized set of semantic details that can considerably enhance the precision and efficiency of the table matching process. Our solution aims to respond effectively to the SemTab challenge and underlines the value of utilizing pre-existing semantic resources to enhance table annotation. In a time when the need for scalable, adaptable solutions is more pressing than ever, our methodology stands out for its potential to cater to the growing needs of the digital era.

2. Related Work

The SemTab challenge, which began in 2019, has played a pivotal role in pushing the boundaries of semantic table interpretation, a field that aims to understand and annotate tabular data with semantic information. In its inaugural year, 2019, the challenge witnessed some groundbreaking contributions. Oliveira and d'Aquin introduced "ADOG" [4], a system that leverages ontologies for annotating data. This was complemented by the work of Cremaschi et al., who presented "MantisTable" [5], an innovative system designed to automatically interpret tables semantically. Another notable contribution was from Thawani et al., who delved deep into the CTA and CPA tasks, presenting a method for linking entities to knowledge graphs, thereby inferring column types and properties [6]. The subsequent year, 2020, saw the challenge grow in terms

of participation and complexity. Huynh et al. unveiled an enhanced version of "DAGOBAH" [7], which emphasized the importance of scalable annotations for large datasets. Parallelly, Abdelmageed and Schindler introduced "JenTab" [8], a system tailored to match tabular data with knowledge graphs, bridging the gap between structured and unstructured data. By 2021, the challenge had gained significant traction in the research community. Systems that had made their debut in previous years underwent refinements. For instance, "DAGOBAH" [9] was further optimized to provide efficient semantic annotations. Similarly, "MantisTable V" [10] was introduced as a novel and efficient successor to the earlier version, emphasizing innovative methods for table interpretation. The 2022 edition of the challenge was particularly noteworthy for the introduction of specialized datasets. "SOTAB" [11] and "MammoTab" [12] were introduced, both of which align closely with the 2023 round 2 tasks focusing on Schema.org annotations. Additionally, systems like "s-elBat" by Cremaschi et al. [13] underscored the challenges and intricacies of interpreting real-world, messy data. As we approach the 2023 SemTab challenge, the emphasis on CTA and CPA tasks, especially in the context of Schema.org and DBpedia, is more pronounced than ever. The collective works from 2019 to 2022 not only highlight the progress made but also set the stage for future innovations in the domain.

3. Tasks

The second round of the SemTab challenge spotlights two core tasks: Column Type Annotation (CTA) and Column Property Annotation (CPA). These tasks seek to enrich table comprehension by attributing specific labels to columns and establishing inter-column relationships, respectively.

3.1. Column Type Annotation (CTA)

CTA categorizes columns by associating specific labels that signify semantic information about their content. This involves attributing fitting labels to columns based on their purpose or content. In the context of the SemTab challenge, labels used for CTA derive from the DBpedia and Schema.org knowledge graphs. CTA aids efficient data integration and allows downstream applications to understand the structure and semantics of tables, proving pivotal in tasks such as data cleaning, schema matching, and query optimization. The labels assigned to column types offer valuable insights into each column's intended purpose and content, facilitating improved data understanding and analysis.

3.2. Column Property Annotation (CPA)

CPA focuses on establishing relationships between table columns. It involves annotating column pairs with labels indicating their mutual relationship or connection. These relationships could denote concepts such as "startDate," "priceValidUntil," or "recipeIngredient," among others. CPA affords essential context about inter-column relationships, leading to a more comprehensive understanding of the table. The annotations help identify related or interconnected columns, which is beneficial in data integration, schema alignment, and knowledge discovery tasks. By

Table 1
Train CSV file for CTA

	table_name	column_index	label
0	Book_11x17.pt _September2020_CTA.json.gz	3	https://dbpedia.org/ontology/date
1	Book_12min.com _September2020_CTA.json.gz	0	https://dbpedia.org/ontology/Book
2	Book_12min.com _September2020_CTA.json.gz	2	https://dbpedia.org/ontology/Language

Table 2
Example of content in table

	0	1	2	3
0	9789722539739	A Cidade Perdida	728	2020-07-10
1	9789722527118	Misery	480	2013-09-13

capturing the relationships between columns, CPA bolsters the potential to extract meaningful insights from tables and supports precise analysis and decision-making.

CTA and CPA tasks collectively aim to improve table matching and comprehension. These tasks amplify tables' semantic richness by attributing labels to column types and defining inter-column relationships, facilitating effective data integration, knowledge discovery, and other downstream applications. The following sections will delve into the datasets used for CTA and CPA, elucidate the experimental setup, discuss the results achieved using the DREIFLUSS system, and evaluate our approach's effectiveness in addressing these tasks in the SemTab challenge.

4. Dataset

The SemTab 2023 competition page[2] provided the datasets used for the CTA and CPA tasks. Each task has three dataset folders. For CTA, these folders are Round2-SOTAB-CTA-Tables, Round2-SOTAB-CTA-DBP-Datasets, and Round2-SOTAB-CTA-SCH-Datasets. The Round2-SOTAB-CTA-Tables folder contains 44,409 JSON files representing different tables, each file comprising the column index and corresponding value. The remaining two folders, Round2-SOTAB-CTA-DBP-Datasets, and Round2-SOTAB-CTA-SCH-Datasets contain the training, validation, and test datasets in CSV format, along with the appropriate labels (derived from DBpedia or Schema.org) in a TXT file. This file encompasses 46 labels for DBpedia and 80 labels for Schema.org. The CSV files present the data in the format shown in Table 1. The validation dataset shares the same format, whereas the test dataset does not contain labels. The JSON files in the Tables folder can be converted to a table format which can be represented as shown in Table 2

The data folders are similar in structure for the CPA task, encompassing 49 labels for DBpedia

---

Table 3
Train CSV file for CPA

	table_name	main_column_index	column_index	label
0	Book_11x17.pt _September2020_CPA.json.gz	0	3	datePublished
1	Book_1jour-1jeu.com _September2020_CPA.json.gz	0	5	datePublished

Table 4
CTA table after concatenating respective data value column

	table_name	column_index	label	data_value
0	Book_11x17.pt _September2020_CTA.json.gz	3	Date	[2020-07-10, 2016-04-08, 2013-09-13, 2016-08-0...
1	Book_12min.com _September2020_CTA.json.gz	0	Book/name	[The Sleep Revolution, Viva, Ame, Lidere, The ...

and 105 for Schema.org. The Tables folder contains 28,223 JSON files (tables), and the CSV data representation for this task includes an additional column to identify the primary column index, as shown in Table 3. As with the CTA task, the test dataset for the CPA task does not include labels.

The comprehensive dataset provided in the SemTab 2023 competition enhances the complexity and richness of the CTA and CPA tasks, laying the groundwork for evaluating and refining the efficacy of different table matching strategies.

5. Methodology

In order to appraise the efficacy of the DREIFLUSS system, we employed a series of tests using datasets made available by the SemTab challenge. These datasets were specifically designed for Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks. Our feature generation process involved parsing the provided CSV file to extract relevant data points. The features were derived from the semantic information of the columns, the relationships between columns, and the inherent structure of the tables. Initially, when attempting to process the entire dataset on our local machine, we faced computational challenges due to the vastness of the data. To address this, we strategically downsampled the data. We ensured our sample set encompassed representations from each category for both the CTA and CPA tasks. Specifically, we selected 15 samples for each label in the CTA task and 3 samples for each label in the CPA task. This approach allowed us to maintain a balanced representation of each label while optimizing computational efficiency.

Next, we extracted appropriate data values for the respective column indexes from the JSON file in the table dataset folder. This approach allowed us to construct new tables specifically for the CTA and CPA tasks, respectively. For the CTA task, we introduced a single column named "data_value", as illustrated in Table 4, designed to hold specific data values associated with each

Table 5

CPA table after concatenating respective main column and other column

	table_name	main_ column_ index	column_ index	label	main_ column_ value	other_ column_ value
0	Movie_yts-torrent.net _September2020_ CPA.json.gz	0	9	actor	[The Lost Husband (2020), John Henry (2020)...	[[Nora Dunn, Kevin Alejandro, ...
1	Movie_zomraa.com _September2020_ CPA.json.gz	0	9	actor	[Integrity (2019), Gretel & Hansel (2020)...	[Johnny Depp, Johnny Depp, ...
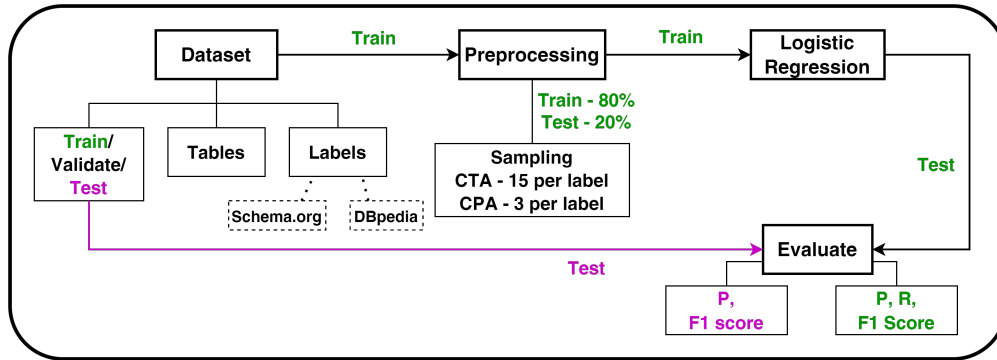
Figure 1: Experiment pipeline

column type. In contrast, for the CPA task, we integrated two columns: "main_column_value" and "other_column_value", as showcased in Table 5. These columns captured the primary data values and their associated or related values, respectively. Given the nature of our data, these columns often contained lists of multiple elements. To streamline our data for modeling, we employed a technique known as "exploding", which transformed each list of values into separate rows. For instance, a row with a list of three elements in the "data_value" column would post-exploding result in three distinct rows, each holding one of those elements. This transformation ensured a singular data point representation for each row, facilitating the subsequent training process.

In our experimental setup, we strategically selected columns from the generated tables to serve as features for our machine-learning tasks. For the Column Type Annotation (CTA) task, our primary features were derived from the label and data_value columns. In contrast, for the Column Property Annotation (CPA) task, we harnessed the information from the main_column_value, other_column_value, and label columns.

Before feeding this data into our machine learning model, we partitioned it into training and testing sets, maintaining an 80-20 split. This division was done with stratification on the label column, ensuring that our test set was representative of the overall distribution of labels. The next crucial step was data vectorization. Raw textual data isn't directly usable in most machine learning algorithms, including logistic regression. Hence, we employed a vectorization technique, specifically the CountVectorizer from scikit-learn, which converts text data into a matrix of token counts. This transformation is pivotal as it translates our textual data into a numerical format that our model can understand and learn from. With our data appropriately vectorized, we proceeded to train our logistic regression model. We utilized scikit-learn's LogisticRegression class. By default, this model applies L2 regularization (ridge regularization) with a penalty hyperparameter set to 'l2'. The strength of this regularization is controlled by the 'C' hyperparameter, which defaults to 1.0, implying a balanced regularization. The solver hyperparameter, set to 'lbfgs', dictates the optimization algorithm used for parameter tuning. Additionally, the model iterates a maximum of 100 times during training, as determined by the 'max_iter' hyperparameter. Post-training, our logistic regression model had learned the intricate mappings between our input features (vectorized data values) and the target outputs, which were either column type labels (for CTA) or column relationship labels (for CPA). To gauge the model's efficacy, we employed evaluation metrics like precision, recall, and F1 scores. These metrics provided insights into how well our model could predict column types and discern inter-column relationships. A comprehensive visual representation of our entire experimental pipeline is depicted in Figure 1. For the broader research community's benefit and to promote reproducibility, we've made our implementation code publicly accessible on GitHub[3]. Through this rigorous methodology, we were able to critically assess the capabilities of the DREIFLUSS system in the context of the SemTab challenge."

6. Results

This section delineates the findings obtained from the experiments carried out utilizing our system for the Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks. The results pertain to the use of both Schema.org and DBpedia labels. The efficacy of the logistic regression models was gauged using evaluation metrics such as precision, recall, and F1 scores. The metrics illustrated in Table 1 present the performance of the models for both CTA and CPA tasks using Schema.org and DBpedia labels. For the CTA task, the metrics indicate the model's effectiveness in precisely predicting column type labels based on the data values provided. In the context of the CPA task, these scores underscore the model's proficiency in identifying relationships between columns based on the main and other column values.

These results accentuate the system's competence in carrying out both the CTA and CPA tasks using Schema.org and DBpedia labels. The precision, recall, and F1 scores achieved exemplify the system's ability to predict column types and delineate relationships between columns accurately. These outcomes pave the way for enhanced table matching, enabling comprehensive data integration and promoting knowledge discovery. Table 6 displays the precision, recall, and F1 scores for the CTA and CPA tasks using both Schema.org and DBpedia

---

[3]https://github.com/vishvapalsinh/cta-cpa-schemaorg-dbpedia

Table 6
Precision, recall, and F1 scores for CTA and CPA tasks with Schema.org and DBpedia labels.

Task	Labels	Our Test Data			SemTab Test Data	
		Precision	**Recall**	**F1 Score**	**Precision**	**F1 Score**
CTA	Schema.org	89.47%	73.50%	77.95%	56.67%	38.04%
CTA	DBpedia	89.55%	70.87%	76.94%	61.14%	40.97%
CPA	Schema.org	91.77%	72.35%	78.01%	31.96%	17.39%
CPA	DBpedia	92.94%	77.73%	83.12%	39.67%	20.69%

labels. It also shows the precision and F1 score generated on the test dataset provided by the SemTab organizers[4]. The data demonstrates the system's competence in executing these tasks effectively.

7. Discussion

The experimental outcomes from the DREIFLUSS system application for the Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks during the SemTab challenge form an insightful basis for subsequent discussion. This segment dives deeper into the repercussions of the achieved results and initiates a dialogue around various facets linked with the system's performance, limitations, and potential enhancements.

7.1. Performance Analysis

The precision, recall, and F1 scores achieved underscore the robustness of the logistic regression models that DREIFLUSS employs for CTA and CPA tasks with both Schema.org and DBpedia labels. Although we used only a fraction of the available samples for training (CTA: 690-1200 out of 85561-115562, CPA: 147-315 out of 62128-97967), the model still yielded convincing results. However, a decrease in scores on the SemTab-2023 challenge test set highlights the need for more effective sample selection strategies.

7.2. Significance of Knowledge Graphs

The DREIFLUSS system benefits significantly from labels provided by Schema.org and DBpedia knowledge graphs, reaffirming the importance of integrating existing semantic resources in table matching tasks. The broad coverage of DBpedia and the standardized schema vocabulary of Schema.org serve as a rich data source, enriching the annotation process and enhancing understanding of column types and relationships.

7.3. Limitations and Challenges

Despite showing promising outcomes, the DREIFLUSS system does have certain limitations. Its heavy reliance on the quality and completeness of labels offered by DBpedia and Schema.org,

---

[4]https://shorturl.at/kFKO9

and the data values used for training can pose a challenge. Only complete or accurate labels or data values can impact the system's performance, resulting in incorrect classifications or relationship annotations. Furthermore, employing a logistic regression model may limit the system's ability to handle complex relationships or certain data variations.

7.4. Future Directions

Future works could incorporate advanced techniques such as deep learning models or ensemble methods to overcome these limitations and augment the results. External knowledge sources beyond DBpedia and Schema.org, like domain-specific ontologies or other domain-specific knowledge graphs, can offer more precise annotations. Additionally, considering more context, such as table structure or content, could further enhance the accuracy of column type annotations and relationship predictions.

7.5. Practical Applications

The DREIFLUSS system holds practical significance in numerous domains and applications revolving around table understanding and integration. Its precise column type annotations and relationship predictions can aid in tasks such as data cleaning, schema matching, query optimization, and knowledge discovery. Its minimalist design and reliance on existing knowledge graphs promise practicality and scalability in real-world applications.

8. Conclusion

This study sheds light on the novel application of the DREIFLUSS system for the Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks as a part of the SemTab challenge. The results obtained through the application of this system underscore its effectiveness in enhancing table matching accuracy and efficiency. Through the use of Schema.org and DBpedia labels, this research highlights the importance of integrating existing semantic resources into the process of table understanding. These knowledge graphs serve as invaluable tools, providing a rich data source that guides the annotation process and enhances the understanding of column types and relationships. However, the journey towards a perfect solution is paved with challenges. Certain limitations, such as the dependence on the quality and completeness of labels and data values used for training, have been identified. The use of logistic regression models may also restrict the system's ability to capture complex relationships or handle data variations. Nevertheless, these challenges present avenues for future research. Prospective advancements in this field could explore the inclusion of more sophisticated techniques such as deep learning models or ensemble methods. Expanding the scope to include domain-specific ontologies or other domain-specific knowledge graphs could provide more precise and specialized annotations. The implications of this research are vast and multi-faceted. Beyond the academic realm, it holds substantial practical value in various domains and applications, such as data cleaning, schema matching, query optimization, and knowledge discovery. The scalability and practicality of the DREIFLUSS system promise its relevance in real-world scenarios.

In conclusion, the DREIFLUSS system has demonstrated promising results in addressing the CTA and CPA tasks within the SemTab challenge, setting a solid foundation for further improvements. The learnings from this research open up exciting possibilities for future endeavors in the realm of table matching and understanding, thereby contributing to the body of knowledge in this ever-evolving field.

References

[1] A. O. Shigarov, Table understanding: Problem overview, WIREs Data Mining Knowl. Discov. 13 (2023). URL: https://doi.org/10.1002/widm.1482.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia - A crystallization point for the web of data, J. Web Semant. 7 (2009) 154–165. URL: https://doi.org/10.1016/j.websem.2009.07.002.

[3] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of structured data on the web, ACM Queue 13 (2015) 10. URL: https://doi.org/10.1145/2857274.2857276.

[4] D. Oliveira, M. d'Aquin, ADOG - annotating data with ontologies and graphs, volume 2553 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 1–6. URL: https://ceur-ws.org/Vol-2553/paper1.pdf.

[5] M. Cremaschi, R. Avogadro, D. Chieregato, Mantistable: an automatic approach for the semantic table interpretation, volume 2553 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 15–24. URL: https://ceur-ws.org/Vol-2553/paper3.pdf.

[6] A. Thawani, M. Hu, E. Hu, H. Zafar, N. T. Divvala, A. Singh, E. Qasemi, P. A. Szekely, J. Pujara, Entity linking to knowledge graphs to infer column types and properties, volume 2553 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 25–32. URL: https://ceur-ws.org/Vol-2553/paper4.pdf.

[7] V. Huynh, J. Liu, Y. Chabot, T. Labbé, P. Monnin, R. Troncy, DAGOBAH: enhanced scoring algorithms for scalable annotations of tabular data, volume 2775 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 27–39. URL: https://ceur-ws.org/Vol-2775/paper3.pdf.

[8] N. Abdelmageed, S. Schindler, Jentab: Matching tabular data to knowledge graphs, volume 2775 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 40–49. URL: https://ceur-ws.org/Vol-2775/paper4.pdf.

[9] V. Huynh, J. Liu, Y. Chabot, F. Deuzé, T. Labbé, P. Monnin, R. Troncy, DAGOBAH: table and graph contexts for efficient semantic annotation of tabular data, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 19–31. URL: https://ceur-ws.org/Vol-3103/paper2.pdf.

[10] R. Avogadro, M. Cremaschi, Mantistable V: A novel and efficient approach to semantic table interpretation, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 79–91. URL: https://ceur-ws.org/Vol-3103/paper7.pdf.

[11] K. Korini, R. Peeters, C. Bizer, SOTAB: the WDC schema.org table annotation benchmark, volume 3320 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 14–19. URL: https://ceur-ws.org/Vol-3320/paper1.pdf.

[12] M. Marzocchi, M. Cremaschi, R. Pozzi, R. Avogadro, M. Palmonari, Mammotab: A giant and

comprehensive dataset for semantic table interpretation, volume 3320 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 28–33. URL: https://ceur-ws.org/Vol-3320/paper3.pdf.

[13] M. Cremaschi, R. Avogadro, D. Chieregato, s-elbat: A semantic interpretation approach for messy table-s, volume 3320 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 59–71. URL: https://ceur-ws.org/Vol-3320/paper7.pdf.