2024 IEEE 18th International Conference on Semantic Computing (ICSC) | 979-8-3503-8535-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSC59802.2024.00057

Complex Semantic Tabular Interpretation using SDD-Gen

Matthew Johnson
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, USA
johnsm21@rpi.edu

Jeanette A. Stingone
Department of Epidemiology
Columbia University
New York, USA
ORCID 0000-0003-3508-8260

Sofia Bengoa
Department of Environmental Medicine and Public Health
Icahn School of Medicine at Mount Sinai
New York, USA
sofia.bengoa@mssm.edu

James Masters
Scientific Computing and Data
Icahn School of Medicine at Mount Sinai
New York, USA
james.masters@mssm.edu

Deborah L. McGuinness
Dept. of Computer Science, Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, USA
dlm@cs.rpi.edu

***Abstract*—Knowledge graphs have become an essential technology for both businesses and governments. They enable a wide variety of critical tasks, such as aligning diverse datasets, improving the capabilities of search engines, supporting error checking, and generating explanations using inference engines. However, populating, augmenting, and/or validating a knowledge graph can be challenging because developers need domain knowledge to understand their data and experience in ontology modeling to align concepts properly as well as experience with conflict detection and truth maintenance tools. Previous efforts have explored automatically integrating simple tabular data into knowledge graphs to lower the barrier to entry. These methods heavily rely on named entity overlap and require that tables are similar to relational tables in third normal form. While these methods have been successful under competition, these limitations make them impractical for general usage. In this paper, we introduce the semantic data dictionary generator (SDD-Gen), an algorithm that aligns complex tabular data to ontological terms for knowledge graph generation. Our methodology leverages context information from data dictionaries to make alignments, enabling us to align complex tables with few named entities and multiple subject columns.**

***Index Terms*—Tabular Data, Knowledge Graph, Semantic Annotation**

I. Introduction

The term knowledge graph was first popularized in 2012 by Google, a network of named entities that captures the properties and relationships between named entities [1]. While knowledge graphs have only become popular somewhat recently, their foundations in semantic web technology have been around much longer [2]. This technology has been critical because it enables the integration of diverse datasets, improves the capabilities of search engines [1], powers question-answering services [3], and can be used to generate explanations for inference engines [4]. As a result, knowledge graphs have become important tools for both industry and governments. Unfortunately, creating or adding to an existing knowledge graph can be challenging, and knowledge graphs are most useful when they are current and validated. Semantic tabular interpretation addresses these issues by augmenting knowledge graphs with tabular data. This process can be challenging because knowledge graphs encode meaning through structure and require statements to be unambiguous [5]. In contrast, tabular data uses locality to encode relationships within the data. This encoding is often optimized to the data publisher's original task, requiring additional context for third parties to interpret and reuse a table for subsequent unanticipated tasks. Specifically, context is needed to understand how entities are grouped within a table and how these groups relate to each other [6].

Previous work automating semantic tabular interpretation has focused on developing algorithms that integrate web tables into an existing knowledge graph using data-level alignments. Web tables are similar to relational tables in third normal form, except instead of a primary key they contain a single subject column where every other column is a property of the subject column [7]. These algorithms generally align overlapping entities in the web table and knowledge graph and then use these alignments to vote on a column's upper-level class and the relationships between subject and non-subject columns. The yearly Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) competition at the International Semantic Web Conference (ISWC) aims to solve this problem and has shown that data-level alignment algorithms can be quite effective [8]–[11] however there are several limitations to these solutions. First, there must be an overlap of named entities within the knowledge graph and table. This constraint means that data-level alignment algorithms can not be used to bootstrap new knowledge graphs. In addition, there are open questions on the percentage of entity overlap needed for these algorithms to be effective. Data-level alignment algorithms may be unable to align tables with few or no named entities. Second, the format of the table is a severely limiting factor. In 2008, researchers extracted 14.1 billion HTML tables using Google's

979-8-3503-8535-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ICSC59802.2024.00057

general-purpose web crawl and estimated that only around 1.1% of tables contain high-quality relational data [12]. Some of these tables were rejected because the content contained non-interesting information such as web page layout information, but others were rejected because they did not meet the strict web table formatting requirements. These restrictions have meant that the SemTab competition has often been forced to evaluate algorithms against synthetic datasets [8]–[11].

To address these issues, algorithms need to consider the additional context around the table that asserts groups and relationships among the entities presented in the table. Traditionally, data publishers have attempted to address semantic tabular interpretation by including data dictionaries and code books. These metadata files contain additional context, such as text descriptions, column types, and codes within the data, enabling third parties to understand the data better. While there is no agreed-upon formal specification for a data dictionary, they can include text descriptions of schema elements, codebooks that map data values and data ranges.

In this paper, we introduce the semantic data dictionary generator (SDD-Gen), an algorithm that aligns complex tabular data to ontological terms for knowledge graph generation. Our methodology leverages context information from data dictionaries to make tabular alignments using semantic data dictionaries (SDDs) that formally capture alignments between tabular data and domain ontologies [13]. Unlike data-level alignments, which align named entities and infer class and relations, ontological-level alignments extract class and relations from the data dictionary to interpret tabular data, allowing us to operate on more complex tables with multiple subject columns and few named entities. We compare the column type annotation performance of SDD-Gen against KGCODE-Tab [14], a high-performing algorithm at Semtab 2022 [11], on a complex table dataset composed of actual health study data.

II. Related Work

Semantic tabular interpretation algorithms aim to align tabular data with an existing knowledge graph. The problem is composed of three subtasks: cell-entity annotation (CEA), matching a cell to an entity within the knowledge graph; column-type annotation (CTA), assigning a class type to a column; and column-property annotation (CPA), assigning a property between two columns. Previous papers have categorized semantic tabular interpretation solutions into three main categories [15]: Search-based algorithms that utilize a keyword search against the knowledge graph for alignment, supervised algorithms that utilize machine learning alignment trained on ground truth table data, and unsupervised algorithms that utilize machine learning alignment without ground truth table data and keyword searches at runtime.

The general architecture of many search-based semantic tabular interpretation algorithms is to solve the CEA problem first. Entity alignments are generated by searching the knowledge graph for the terms within cells. Once the entities are aligned, they are used in a voting strategy to determine a common type across a column, which solves the CTA problem. Finally, the CPA problem is solved by searching the knowledge graph for properties from the subject column type to another column type and using matched entities within those columns that observe that property to vote for a property. Variants of this architecture have been successful in the SemTab challenges [8]–[11]. Two of the higher-performing algorithms in the 2022 competition were KGCODE-Tab [14] and DAGOBAH [16]. Both methodologies focused on enhancing the CEA stage. KGCODE-Tab uses Bing searches to help standardize entity labels. Website mentions are collected for a cell entity, and the mention with the closest Levenshtein distance becomes the new label. DAGOBAH relies on an alias table backed by elastic search to link entities to multiple labels. In addition, they have further improved entity disambiguation by parsing column headers using BERT and comparing the semantic correlation with cell entities.

Supervised algorithms are less common in semantic tabular interpretation because datasets with large amounts of labeled data are rare. Rümmele et al. introduced the Data INTegrator (DINT), which takes in a table to solve the CTA subtask [17]. Their methodology trained a random forest model to classify the data type of a column based on hand-crafted features such as whitespace count, class text similarity, string length, and min-max statistics. Wang et al. introduced the Table Convolutional Network (TCN), which uses context within a table and context across similar tables to solve CTA and CPA [18]. They train two neural networks, one for each subtask where the input is an intra-table embedding generated by named entities within the table and an inter-table embedding from similar tables.

In 2016, Neumaier et al. introduced an unsupervised semantic tabular interpretation algorithm that solves the CTA subtask for numerical columns [19]. Their algorithm mines a knowledge graph for examples of numerical data, such as the heights of basketball players and uses K-NN to form clusters. They then cross-validate with the knowledge graph to determine each cluster's most common numeric class types. When processing a numeric column, they find the most likely clusters and use a voting strategy to determine the most likely column type. In 2017 Nguyen et al. extended this methodology to handle numeric data with different units [20].

III. Approach

The SDD-Gen algorithm differs from these existing approaches because it aligns table schema with ontology class and relations using additional context from data dictionaries. This difference means our algorithm relies less on named entities than search-based methods and doesn't need an existing knowledge graph like supervised and unsupervised methods. SDD-Gen takes as input tabular data, a corresponding data dictionary, a set of domain ontologies and generates a semantic data dictionary. We assume the table is row-oriented, where related concepts are in the same row; the first row contains column header information and has no nested cells. For the data dictionary, we assume that it includes a text description and any codes for each column in the table. In the biomedical domain, this metadata is standard [21]–[23].

RIDRETH1 - Race/Hispanic origin

Variable Name: RIDRETH1
SAS Label: Race/Hispanic origin
English Text: Recode of reported race and Hispanic origin information
Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Mexican American	1367	1367	
2	Other Hispanic	820	2187	
3	Non-Hispanic White	3150	5337	
4	Non-Hispanic Black	2115	7452	
5	Other Race - Including Multi-Racial	1802	9254	
.	Missing	0	9254	

Fig. 1. Data dictionary sample from NHNANES 2017-2018 Demographic dataset

In figure 1 we provide a segment of a data dictionary from the National Center for Health Statistics' (NCHS) National Health and Nutrition Examination Survey (NHANES) 2017-2018 Demographics dataset [24] as an example. The figure shows the schema element RIDRETH1, provides a text description of the element and a codebook that maps races to different codes within the table. This example highlights the need for data dictionaries because if a third party wanted to reuse this data and only had the data, it would be very difficult to determine that RIDRETH1 was the reported race of the participants. The domain ontologies contain the class and properties SDD-Gen will map to tabular concepts. We assume there is some overlap of concepts between the domain ontologies and table.

The SDD-Gen generates a semantic data dictionary (SDD), which formally captures the mappings between the table and domain ontology. The SDD specification was developed to provide a machine-readable alternative to data dictionaries that formalize the assignment of a semantic representation of data [13]. Beyond promoting data reuse, semantic data dictionaries can also be used with several tools to generate RDF representations of the tabular data [25]–[27]. Other works have used this approach to align tabular data to existing knowledge graphs allowing them to harmonize diverse datasets [26], [27]. SDDs are composed of rows corresponding to columns within the table and virtual columns representing concepts not explicitly defined within the table. Each row maps a column to properties of several top-level ontologies. These ontologies form the foundation for the resulting model and are used to describe the mappings to domain ontologies. In theory, any ontology can be used as a top-level ontology however, Rashid et al. [25] recommend using well-known ontologies such as Semanticscience Integrated Ontology (SIO) [28] or Basic Formal Ontology (BFO) [29]. To limit the scope of this paper, we will only consider semantic data dictionaries that use the default top-level ontologies, but the methodologies discussed can be adapted to other ontologies.

Generally, every row in an SDD is aligned as an attribute or an entity. Attribute alignments, at minimum, align the table column to an ontological class that represents the attribute and assigns that attribute to a subject. Beyond the minimum, attribute columns can also align units and ascribe the attribute to a time period. Entity alignments, at minimum, assign a class and optionally define a role or a relation between the entity and another column. This representation allows SDDs to align more complex tables where there are multiple subjects or concepts that are only referenced implicitly. Although the SDD-Gen focuses on creating semantic data dictionaries to capture these mappings, it can be applied to other semantic mapping tasks, such as the column-type annotation subtask in semantic tabular interpretation [8].

TABLE I
PARTIAL SEMANTIC DATA DICTIONARY FOR NHNANES 2017-2018 DEMOGRAPHIC DATASET

Column	Attribute	attributeOf	Unit	Time
SEQN	hasco:originalID [original ID]	??participant		
RIDAGEYR	sio:SIO_001013 [age]	??participant	sio:SIO_000428 [year]	??screen
RIDAGEMN	sio:SIO_001013 [age]	??participant	sio:SIO_000429 [month]	??screen
RIDRETH1	hhear:00609 [Race or Ethnicity Combined]	??participant		

TABLE II
PARTIAL SEMANTIC DATA DICTIONARY VIRTUAL FOR NHNANES 2017-2018 DEMOGRAPHIC DATASET

Column	Entity	Role	inRelationTo
??participant	sio:SIO_000485 [human]	sio:SIO_000883 [subject role]	
??screen	nhanes:00020 [Screening time]		??participant

We have included a portion of the SDD for the National Center for Health Statistics' NHANES 2017-2018 Demographics dataset [24] in tables I, II to highlight some of the features introduced. This semantic data dictionary uses SIO [28] as its top-level ontology and the Human Aware Science Ontology (HAScO) [30], SIO, the Human Health Exposure Analysis Resource (HHEAR) ontology [31], and the NHANES ontology [32] for the domain ontologies. To improve the readability, we have separated the semantic data dictionary into two separate tables and have added brackets with the RDF label to all ontology mappings.

In table I, the first column shows all the data column labels from the corresponding data dictionary. It maps their relationship to the top-level ontology terms Attribute, attributeOf, Unit, and Time. The Attribute column maps the class label to a domain ontology attribute type. The attributeOf property assigns that attribute to an entity within the data. The Unit column captures the metric used for the attribute. Finally, the Time column captures when the measurement took place. Throughout table I several virtual columns are referenced with the ?? notation. Table II shows the virtual column declarations which allow users to assign an Entity, Role, Relation, and inReleationTo mapping to a virtual column.

The SDD-Gen generates semantic mapping by evaluating the context information using transformer network encoders and developing an embedding corresponding to the closest ontology terms for a given SDD column. We evaluate each data

dictionary row and generate a text string by concatenating the column header, the data dictionary description, and the named entities within the table. After that, each word in the text string is replaced with a word embedding that captures the semantic similarity between words. For our experiments, we used GloVe vectors because they have been shown to perform better in word similarity and named entity recognition tasks [33]. These vectors are part of a pre-trained word vector space where semantically similar words are closer together (ie ”frog” nearest neighbors would be ”frogs”, ”toad”, ...) and the relationship between words is captured in vector mathematics (ie ”man” - ”woman” = ”king” - ”Queen”) [33]. If an embedding is missing for a particular word, we use spellcheck to replace the word if there is a high-confidence suggestion.

The next step is to take the large data dictionary vector and map it back into the word embedding space. This is accomplished by using the encoder portion of transformer networks, which use self-attention to summarize sequential data [34]. For each column in an SDD, we train the encoders to return the most closely associated word embedding to the ontology class or property from the list of target ontologies. After generating our SDD embedding, we search for the nearest class or property neighbors depending on the SDD column. The distance between the SDD embedding and potential matches is used to measure confidence in the alignment. We parametrically set a minimum confidence and return the top matches above that threshold. The SDD-Gen has an additional mode that takes in a ranked list of ontologies. This list encodes the priority of an ontology so that if we have a close match in two ontologies, we will prefer the higher-ranked ontology. This is accomplished by adding a bias to the confidence score based on the priority of the ontology.

IV. Evaluation

To evaluate the performance of SDD-Gen on complex tables with multiple subject columns and a few named entities we created a complex table dataset and compared it to the current state of the art in semantic tabular interpretation. For the complex table dataset we created and gathered manually aligned semantic data dictionaries from several biomedical studies. The dataset is composed of complex tables, data dictionaries, and semantic data dictionaries from a variety of sources, including the National Institute of Environmental Health Sciences' (NIEHS) Human Health Exposure Analysis Resource (HHEAR) [21], the National Center for Health Statistics' (HCHS) National Health and Nutrition Examination Survey (NHANES) [35], and the National Cancer Institute's (NCI) The Cancer Genome Atlas (TCGA) [36]. Altogether we have collected 39 semantic data dictionaries containing over 600 ontology mappings.

To ensure the data quality of the complex table dataset, we developed a set of rules to validate both data dictionaries and semantic data dictionaries. For data dictionaries, we check that every column has a unique name that's not empty and that we do not have multiple data dictionary entries for the column. Next, we check that we have good column descriptions. We decided that descriptions should describe the attribute being measured and link the variable to a subject. One common issue we encountered was descriptions filled out incorrectly. Some data publishers would use the description almost as a tag field, ”raw data” or ”source variable,” and others would expand the column name ”AdmHeight” to ”adm height.” To flag poor descriptions for review, we created a system that measured the length and Shannon entropy of the description. Any description with less than three words or less than 3.5 Shannon entropy would be flagged and rewritten. The three-word minimum helped identify tags, and the Shannon entropy test ensured more complexity in the descriptions. After that, we check the spelling and flag any potentially misspelled words for review. Finally, we check to ensure that columns are properly linked to code books by confirming that every column has a defined type and, if that type is categorical or enumerated, checking that a code is defined correctly.

For semantic data dictionaries, we check that every column has been defined and all columns have the correct data types. Next, we check that a variable has been defined as an attribute or entity and that all required columns for that designation have been filled out. For instance, if we designate a variable as an attribute, it's recommended that it be associated with another variable using the attributeOf column. After that, we check that all ontology IRIs exist and that the SDD imports the corresponding ontology. Finally, we check that all implicit variables are referenced or reference another variable connected to the data dictionary. This ensures that all RDF generated from the SDD will be linked.

Using these rulesets, we parsed each study's data dictionaries and SDDs and corrected any issues we encountered. After that, we created a target list of ontologies by parsing the prefix sheet within each semantic data dictionary. The domain ontologies include the Clinical Measurement Ontology (CMO) [37], Cognitive Atlas Ontology (COGAT) [38], Human Disease Ontology (DOID) [39], Human-Aware Science Ontology (HAScO) [30], Human Health Exposure Analysis Resource Ontology (HHEAR) [31], the National Health and Nutrition Examination Surveys ontology (NHANES) [32], the Phenotype And Trait Ontology (PATO) [40] and Semantic science Integrated Ontology (SIO) [28].

To train the SDD-Gen, we split the dataset into 50% training, 20% validation, and 30% testing based on the number of alignments made in the SDDs, ensuring that no SDDs were split between sets. We trained our encoder for 5,000 epochs to minimize the L1 distance between ontology class alignments and data dictionary descriptions for column attributes. We achieved a validation accuracy of 23.57% for column-type annotation on the complex table dataset. To understand what this accuracy means in terms of actual capabilities, we evaluated what general patterns caused issues for the SDD-Gen through the lens of high and low-confidence misalignments. Confidence is a measure from [0, 1] that represents how close a match is between a data dictionary embedding and the ontology class embedding. If confidence is lower than 0.85, we consider it a low-confidence alignment, and do not report a match.

When evaluating the performance of SDD-Gen we see several general trends. For high-confidence incorrect answers, 18.6% of the correct answers were within the top ten. This implies that the SDD-Gen did extract the correct attribute from the data dictionary description, but there were multiple suitable ontology class matches. In practice, we addressed this issue by establishing a priority system for the ontologies and weighting results from higher priority ontologies as better. For the other high-confidence answers, there were various issues, such as not choosing a more general ontology class (year vs. time instant) or struggling to align ontology classes with large labels (Race or Ethnicity Combined). To address issues like these, we need a better class embedding strategy that considers the ontology's structure. For low-confidence answers, the most common issue was poor data dictionary descriptions. The SDD-Gen struggled with descriptions that contained multiple sentences and descriptions posed as questions (In what country were you/was sample person born?). We need to develop larger SDD-Gen models that can reason over more complex sentence patterns to address these issues.

TABLE III
F1 STATISTICS ON THE COMPLEX TABLE TEST SET

Algorithm	Precision	Recall	F1-Score
KGCODE-Tab	0.025	0.015	0.019
SDD-Gen	0.525	0.155	0.239

Next, we wanted to compare the SDD-Gen column-type annotation accuracy on the test set to the current state of practice for semantic tabular interpretation. We compared SDD-Gen against KGCODE-Tab, a data-level alignment algorithm that performed well in the 2022 SemTab competition. To make this a fair comparison, we aligned our test set to the DBpedia ontology, which KGCODE-Tab was developed for, and populated codes within the tables for more named entities. We then ran both algorithms on the test set and calculated the F1 statistics shown in III.

In this experiment, SDD-Gen outperformed KGCODE-Tab by 12x in terms of accuracy, but the highlight of this experiment is how challenging semantic tabular interpretation can be on complex tables. Reviewing the results, KGCODE-Tab succeeded when column names were clear and several named entities were present. For example, columns like Race with entries like Asian, Hispanic, and White were easy for KGCODE-Tab to align. However, the lack of named entities within study data highlights the need to leverage additional context to make alignments.

CONCLUSION AND FUTURE WORK

Creating and augmenting knowledge graphs has long been the domain of data scientists and ontology modelers. But to continue growing this technology, data publishers need to be able to contribute content easily. Previous efforts to automate the semantic tabular alignment process have centered around algorithms that rely on data-level alignments, which are only effective on simple tables. The SDD-Gen leverages context information for data dictionaries, enabling it to align more complex tabular data.

The next steps for the SDD-Gen are exploring additional embedding strategies and artificial expanding the training set. Word and class embeddings do a lot of heavy lifting in this task; therefore, it is crucial to improve them. We are interested in class embedding strategies that encode the relationship between classes. One potential limitation of our methodology is insufficient data to train the transformer encoders properly. To mitigate this risk, we want to leverage existing semantic tabular interpretation datasets to train our models. These datasets contain tables and the semantic mappings needed to generate SDDs but do not provide data dictionaries. We are exploring using large language models (LLMs) to generate data dictionaries for these datasets. Recent work has shown that LLMs can be effective for similar zero-shot learning tasks [41] and would enable us to train deeper networks to improve performance further.

ACKNOWLEDGMENT

This work is partially supported through the National Institutes of Health's Human Health Exposure Analysis Resource (HHEAR) project (NIH U2CES026555) and the Air Force Research Laboratory's Civilian Academic Degree Program. This work could not have been possible without the metadata generated from the HHEAR studies [42]–[59].

REFERENCES

[1] A. Singhal. Introducing the knowledge graph: things, not strings. [Online; accessed 12-May-2022]. [Online]. Available: https://blog.google/products/search/introducing-knowledge-graph-things-not/

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys (Csur)*, vol. 54, no. 4, pp. 1–37, 2021.

[3] How alexa keeps getting smarter. [Online; accessed 12-May-2022]. [Online]. Available: https://www.aboutamazon.com/news/devices/how-alexa-keeps-getting-smarter

[4] F. Lecue, "On the role of knowledge graphs in explainable ai," *Semantic Web*, vol. 11, no. 1, pp. 41–51, 2020.

[5] J. P. McCusker, D. L. McGuinness, J. S. Erickson, and K. Chastain. What is a knowledge graph? [Online; accessed 12-May-2022]. [Online]. Available: https://www.authorea.com/users/6341/articles/107281-what-is-a-knowledge-graph/_show_article

[6] A. K. Dey, "Understanding and using context," *Personal and ubiquitous computing*, vol. 5, pp. 4–7, 2001.

[7] S. Balakrishnan, A. Halevy, B. Harb, H. Lee, J. Madhavan, A. Rostamizadeh, W. Shen, K. Wilder, F. Wu, and C. Yu, "Applying webtables in practice," *Proceedings of the CIDR*, 2015.

[8] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, and K. Srinivas, "Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems," in *European Semantic Web Conference*. Springer, 2020, pp. 514–530.

[9] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, and V. Cutrona, "Results of semtab 2020," in *CEUR Workshop Proceedings*, vol. 2775, 2020, pp. 1–8.

[10] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira *et al.*, "Results of semtab 2021," *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching*, vol. 3103, pp. 1–12, 2022.

[11] N. Abdelmageed, J. Chen, V. Cutrona, V. Efthymiou, O. Hassanzadeh, M. Hulsebos, E. Jiménez-Ruiz, J. Sequeda, and K. Srinivas, "Results of semtab 2022," *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching*, 2022.

[12] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "Webtables: exploring the power of tables on the web," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 538–549, 2008.

[13] S. M. Rashid, K. Chastain, J. A. Stingone, D. L. McGuinness, and J. McCusker, "The semantic data dictionary approach to data annotation & integration." *SemSci@ ISWC*, vol. 2017, 2017.

[14] X. Li, S. Wang, W. Zhou, G. Zhang, C. Jiang, T. Hong, and P. Wang, "Kgcode-tab results for semtab 2022," *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org*, 2022.

[15] L. de Alwis, A. Dissanayake, M. Pallewatte, K. Silva, and U. Thayasivam, "Survey on semantic table interpretation," *Semantic Web Journal*, Jul 2018. [Online]. Available: http://www.semantic-web-journal.net/content/survey-semantic-table-interpretation

[16] V.-P. Huynh, Y. Chabot, T. Labbé, J. Liu, and R. Troncy, "From heuristics to language models: A journey through the universe of semantic table interpretation with dagobah," *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2022.

[17] N. Rümmele, Y. Tyshetskiy, and A. Collins, "Evaluating approaches for supervised semantic labeling," *arXiv preprint arXiv:1801.09788*, 2018.

[18] D. Wang, P. Shiralkar, C. Lockard, B. Huang, X. L. Dong, and M. Jiang, "Tcn: Table convolutional network for web table interpretation," in *Proceedings of the Web Conference 2021*, 2021, pp. 4020–4032.

[19] S. Neumaier, J. Umbrich, J. X. Parreira, and A. Polleres, "Multi-level semantic labelling of numerical values," in *International Semantic Web Conference*. Springer, 2016, pp. 428–445.

[20] P. T. Nguyen and H. Takeda, "Semantic labeling for quantitative data using wikidata," in *Extended Semantic Web Conference*, 2018.

[21] S. M. Viet, J. C. Falman, L. S. Merrill, E. M. Faustman, D. A. Savitz, N. Mervish, D. B. Barr, L. A. Peterson, R. Wright, D. Balshaw *et al.*, "Human health exposure analysis resource (hhear): A model for incorporating the exposome into health studies," *International journal of hygiene and environmental health*, vol. 235, p. 113768, 2021.

[22] (2017, Sep) Nhanes - about the national health and nutrition examination survey. [Online; accessed 30-April-2022]. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about_nhanes.html

[23] The cancer genome atlas program. [Online; accessed 30-April-2022]. [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[24] "Demographic variables and sample weights," May 2021, [Online; accessed 01-May-2022]. [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.htm

[25] S. M. Rashid, J. P. McCusker, P. Pinheiro, M. P. Bax, H. O. Santos, J. A. Stingone, A. K. Das, and D. L. McGuinness, "The semantic data dictionary–an approach for describing and annotating data," *Data intelligence*, vol. 2, no. 4, pp. 443–486, 2020.

[26] P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, and M. P. Bax, "Hadatac: A framework for scientific data integration using ontologies." in *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

[27] J. McCusker, S. M. Rashid, N. Agu, K. P. Bennett, and D. L. McGuinness, "The whyis knowledge graph framework in action." in *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

[28] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath *et al.*, "The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery," *Journal of biomedical semantics*, vol. 5, no. 1, pp. 1–11, 2014.

[29] B. Smith, A. Kumar, and T. Bittner, "Basic formal ontology for bioinformatics," *IFOMIS Reports*, 2005.

[30] P. Pinheiro, M. P. Bax, H. Santos, S. M. Rashid, Z. Liang, Y. Liu, Y. Ne'eman, J. P. McCusker, and D. L. McGuinness, "Annotating diverse scientific data with hasco." in *ONTOBRAS*, 2018, pp. 80–91.

[31] "Hhear ontology," Dec 2019. [Online]. Available: https://www.hadatac.org/hhear-ontology/

[32] H. Santos, P. Pinheiro, and D. L. McGuinness, "Nhanes knowledge graph," *ISWC*, vol. 2022, 2022.

[33] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] "Nhanes - about the national health and nutrition examination survey," Sep 2017, [Online; accessed 30-April-2022]. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about_nhanes.html

[36] "The cancer genome atlas program," https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, [Online; accessed 30-April-2022].

[37] J. R. Smith, C. A. Park, R. Nigam, S. J. Laulederkind, G. T. Hayman, S.-J. Wang, T. F. Lowry, V. Petri, J. D. Pons, M. Tutaj *et al.*, "The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications," *Journal of biomedical semantics*, vol. 4, no. 1, pp. 1–12, 2013.

[38] E. Miller, C. Seppa, A. Kittur, F. Sabb, and R. Poldrack, "The cognitive atlas: employing interaction design processes to facilitate collaborative ontology creation," *Nature Precedings*, pp. 1–1, 2010.

[39] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein *et al.*, "Human disease ontology 2018 update: classification, content and workflow expansion," *Nucleic acids research*, vol. 47, no. D1, pp. D955–D962, 2019.

[40] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "The anatomy of phenotype ontologies: principles, properties and applications," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 1008–1021, 2018.

[41] S. Ubani, S. O. Polat, and R. Nielsen, "Zeroshotdataaug: Generating and augmenting training data with chatgpt," *arXiv preprint arXiv:2304.14334*, 2023.

[42] "Biological responses to tobacco smoke exposure in ill children: Inflammatory processes and oral metabolomic profiles," doi: 10.36043/2017-1762_EPI_87 and 10.36043/1762_633_2022.2, 2020.

[43] "Childhood exposures, epigenetic and transcriptomic responses in the syracuse lead study," doi: 10.36043/2517_399, 10.36043/2517_639_2022.2 and 10.36043/2517_640_2022.2, 2022.

[44] "Denver asthma panel study-chear ancillary study," doi: 10.36043/1450_355 and 10.36043/1450_629_2022.2, 2021.

[45] "Echo recharge study - environmental exposures," doi: 10.36043/1461_219 and 10.36043/1461_630_2022.2, 2021.

[46] "Environmental phenols and pesticide levels in relationship to autism," doi: 10.36043/CHEAR-2016-1449-ADOS and 10.36043/1449_628_2022.2, 2020.

[47] "First trimester exposures: Influence on birth outcomes and markers of biological response," doi: 10.36043/2273_337 and 10.36043/2273_638_2022.2, 2021.

[48] "Internal metabolomic biomarker exposome and developmental disorders," doi: 10.36043/1438_297 and 10.36043/1438_620_2022.2, 2021.

[49] "Maternal and developmental risks from environmental and social stressors," doi: 10.36043/1945_177 and 10.36043/1945_634_2022.2, 2021.

[50] "Metabolomics linking air pollution, obesity and type 2 diabetes," doi: 10.36043/1448_333 and 10.36043/1448_621_2022.2, 2020.

[51] "Micronutrient deficiencies, environmental exposures and severe malaria: Risk factors for adverse neurodevelopmental outcomes in ugandan children," doi: 10.36043/1432_403 and 10.36043/1432_614_2022.2, 2021.

[52] "Mitochondrial dna biomarkers of prenatal metal mixture exposure: intergenerational inheritance and infant growth," doi: 10.36043/2017-1740_EPI_58 and 10.36043/1740_632_2022.2, 2020.

[53] "Pediatric inner-city environmental exposures at school and home and asthma study," doi: 10.36043/2016-1407_EPI_68 and 10.36043/1407_605_2022.2, 2021.

[54] "Perfluoroalkyl substances and lipid composition in human milk," doi: 10.36043/2539_513, 2022.

[55] "Phthalates and childhood obesity in a racially, ethnically and geographically diverse cohort," doi: 10.36043/2537_286 and 10.36043/2537_641_2022.2, 2022.

[56] "Relating metals exposure to birth and early childhood outcomes via the metabotype of cord blood," 2021.

[57] "The dynamics of exposure, phthalates and asthma in a randomized trial," doi: 10.36043/2121_260 and 10.36043/2121_635_2022.2, 2021.

[58] "The impact of tobacco smoke exposure and environmental exposures on the pulmonary microbiome and outcomes of critically ill children," doi: 10.36043/2120_229 and 10.36043/2120_637_2022.2, 2020.

[59] "Validation of detailed maternal cigarette smoke exposure self reporting by cotinine analysis," doi: 10.36043/1523_291 and 10.36043/1523_631_2022.2, 2021.