2025 19th International Conference on Semantic Computing (ICSC) | 979-8-3315-2426-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSC64641.2025.00035

Improving Tabular Reusability through Data Dictionary Descriptions

Matthew Johnson
Air Force Research Laboratory
Rome, USA
matthew.johnson.151@us.af.mil
0000-0001-5212-8100

Sabbir M. Rashid
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, USA
rashidsabbir@gmail.com
0000-0002-4162-8334

Deborah L. McGuinness
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, USA
dlm@cs.rpi.edu
0000-0001-7037-4567

***Abstract*—Tables have become a ubiquitous standard for capturing, storing, and sharing data on the web. This is primarily due to the semi-structured nature of tables, where relationships between data are often ambiguously encoded using locality. While this format can be easy for humans to interpret in simple cases, as table complexity increases, so does the difficulty in interpretability. To bridge this context gap, many data publishers provide a data dictionary to capture schema elements' meaning through text descriptions. Existing work compounds the need for data dictionaries to improve tabular interoperability, but few provide detailed requirements for data dictionary descriptions. This paper identifies and defines three common types of data dictionary descriptions in the biomedical domain. We then compare the effectiveness of each description type by normalizing data dictionary descriptions to a single type using large language models and measuring their performance using a semantic tabular interpretation algorithm. Our experiments show that intensional descriptions, which describe the general properties a column member should have, are most effective for tabular alignment and improve the reusability of data dictionaries.**

***Index Terms*—tabular data, data dictionary, semantic data dictionary, semantic annotation, large language model.**

I. Introduction

HTML tables were first introduced to the web in 1997 as a way to capture tabular data or the layout of a web page [1]. In 2008, researchers identified over 14.1 billion HTML tables containing tabular data within English web pages using Google's general-purpose web crawl [2]. Since then, tables have continued to be a common structure on the web and have become a de facto standard for storing and sharing data.

However, there are still issues with sharing data from tables. Tabular data often use locality to capture cell relationships, which can be unclear when the schema is missing or uninformative. Additionally, like any data product, tables are optimized for the data publisher's original task. If the assumptions built into the original task are unclear or uncatalogued, it can cause issues when a third party tries to reuse the data. Earlier work has shown that tables are more challenging for algorithms to interpret as table complexity increases [3]. Factors such as the number of subject columns and named entity cells within a table can significantly impact downstream interpretability.

To improve tabular interpretability, many data publishers include a data dictionary with tabular data to provide additional context around the elements within a table. Data dictionaries are metadata files that can include information such as a description of the data, relationships between data, the origin of the data, data usage, and the format of the data [4]. Unfortunately, there is no consensus on what metadata conventions data dictionaries should adhere to, and as a result, they come in many shapes and sizes. Data dictionaries commonly include text descriptions, data ranges, and units of columns, as well as codebooks [5] that contain templated codes to organize data.

In Figure 1, we provide a segment of a data dictionary from the Centers for Disease Control and Prevention's National Center for Health Statistics' (NCHS) National Health and Nutrition Examination Survey (NHANES) 2017-2018 Demographics dataset [6]. The figure shows the schema element "RIDRETH1," a text description of the element "Recode of reported race and Hispanic origin information," and includes a codebook that maps races to different numbers within the data. This example highlights the need for data dictionaries. If a third party wanted to reuse this data, it would be challenging to determine the data element from the schema label and data entries alone.

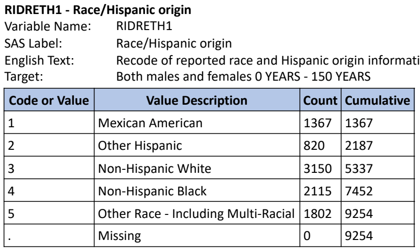
Communities often established best practices to share tabular data using data dictionaries [7]. These guidelines often declare what elements a data dictionary should have, why they are important, and provide some examples. However, few give detailed requirements for data dictionary descriptions. In this paper, we explore the effectiveness of data dictionary description types for tabular interoperability through the lens of semantic tabular interpretation – a problem that looks at mapping tabular data to ontology terms for knowledge graph creation.

We reviewed over 1,600 data dictionary descriptions from several biomedical studies developed by domain experts. From these descriptions, we identified the three most common types of data dictionary descriptions and defined them using inspiration from ontology class design. Finally, we compare the performance of a description type by normalizing descriptions

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited. Case Number: AFRL-2024-5675

2472-9671/25/$31.00 ©2025 IEEE
DOI 10.1109/ICSC64641.2025.00035

to a single type and comparing their performance with a semantic tabular interpretation algorithm. Our experiments show that descriptions that define the conditions needed to be a member of a column and associate the column with a subject are most effective for tabular alignment.



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
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. Adapted from [8]

II. Related Work

One of the most challenging parts of creating data dictionaries is the lack of community consensus on what metadata standards should apply and what format descriptions should be. There are several metadata standards that could apply to data dictionaries in the biomedical domain.

The Data Documentation Initiative (DDI) is a standard for describing the data produced by surveys in the social, behavioral, economic, and health sciences domains [9] using FAIR principles [10]. They establish XML schema standards for capturing the data lifecycle from planning through analysis and the codebooks across these domains [9]. The Dublin Core Metadata Element Set was first developed as an RDF vocabulary to describe web content. However, it has evolved into a general-purpose framework for describing metadata and associating it with resources and agents [11]. ISO 23081 provides general guidance for managing record metadata. It asserts that metadata should capture events that affect the record, enable the record be interpreted in the context of the business, and support integrity, authenticity, reliability, and usability [12]. In addition, it provides a self-assessment to determine the current state of metadata management across an organization. The Visual Resources Association (VRA) Core is a metadata schema for describing images and other works of art [13]. They provide an XML schema for describing the artwork's physical properties and provenance, such as style, period, and techniques. The Investigation, Study, Assay (ISA) framework has established metadata standards to manage life science, environmental, and biomedical experiments [14]. Their ontology is built around three key classes: Investigation, which captures project context; Study, which captures research units; and Assay, which captures analytical measurements [14].

Of these metadata standards, DDI and ISA are most applicable to the biomedical study domain. Both effectively capture the metadata at the variable level and connect them back to the study. However, we have yet to encounter either standard being used in practice in the biomedical domain.

The term data dictionary was coined in the *IBM Dictionary of Computing*; their descriptions are described as definitions that are supposed to capture the meaning of data [4]. The paper *Best Practices for Data Dictionary Definitions and Usage*, advises that a "description of the meaning of the data element" should be included in data dictionaries [7]. In [15], the Smithsonian describes a definition of data elements in a table as a best practice for data management. In [16], Buchanan et al. provide a tutorial on creating a data dictionary to promote sharable data and say that "a detailed description of the information provided for each variable of the data set" is a minimal requirement. While these sources identify a need for a description to capture the meaning and describe a data element, none of them describe what form it should take. This lack of specificity often leads data publishers to adopt descriptions that are not optimized for data reuse, underscoring the need for specific guidelines.

The best data dictionary description definition is from the Center for Open Science, which recommends that data dictionary descriptions: "reflect the way you use the term and intend the term to be used by others;" where possible, be in a genus-differentia form; and avoid circular definitions [17]. A genus-differentia description is an intensional description with a starting definition built upon using a differentiation. For example, a square is a rectangle with four sides of equal length. This recommendation tries to focus descriptions on reuse and builds upon known concepts to improve reusability. However, we have yet to observe data dictionary descriptions like this in practice within the biomedical domain.

III. Approach

To better understand what type of data dictionary descriptions are used in practice, we composed a dataset from thirty-three biomedical studies containing over 1,600 data descriptions. These studies included the National Institute of Environmental Health Sciences' Human Health Exposure Analysis Resource (HHEAR) [18], the NCHS' NHANES [19], and the National Cancer Institute's Cancer Genome Atlas (TCGA) [20]. These data dictionaries are important because they were all developed by domain experts looking to publish real-world study data to a broader audience using data dictionaries, which makes the description types introduced here representative of the broader landscape.

A. Description Types

Surveying this dataset, we identified three common types of data dictionary descriptions: lexical, intensional, and extensional. Lexical descriptions define a column by expanding the column name or acronym and associating the column with a subject. Intensional descriptions define the conditions needed to be a member of that column and associate the column with a subject. Extensional descriptions use codes or data examples to describe the column. Table I shows several examples of each description type for three different table columns. In general, the complexity of the column and the presence of data codes seemed to influence which description types an author would use.

Lexical descriptions are generally short three to five-word sentence fragments and often describe more straightforward columns such as subject identifiers and ages. Their brief and simple syntax should make it easier for humans and machines to interpret tables. However, they have the least context information of all the description types we observed, which can be an issue when processing more complex concepts. For example, in Table I, the "ga" column description doesn't mention the stop date for the age measurement.

For intensional and extensional descriptions, the definitions were influenced by the work done with description logics [21]. Similar to description logic, intensional descriptions should contain necessary and sufficient conditions for membership in the class the column represents. However, they may go beyond a simple class definition of necessary and sufficient conditions for membership to provide information. For example, how something is measured or what kind of units are expected can also be included. Generally, these descriptions are one to two sentences long and are often used to describe more complex concepts, such as sampling a variable under specific circumstances. Intensional descriptions are syntactically complex but contain more contextual data and are potentially more helpful to humans trying to interpret complex tables. In Table I, we see that the intensional description of "ga" includes extra information over the lexical description, including how and when age was measured.

Extensional descriptions define a column of data by enumerating its unique members. These descriptions often expand the column header and include named entities within the column. This description type is most effective when a limited number of named entities compose a class. However, humans and algorithms may struggle with descriptions that don't contain named entities, such as numerical data. In Table I, we can see that the extensional description of "educate" describes the various education levels by showing the actual education categories, which the other description types lack. However, it's unclear how helpful the data in the other extensional descriptions are for tabular interpretation.

B. Evaluation Framework

We explore the effectiveness of data dictionary descriptions through the lens of semantic tabular interpretation. In semantic tabular interpretation, an algorithm is provided with a table and asked to align the tabular data to the terms within a target knowledge graph. These alignments are then used to augment the knowledge graph with data from the table. The problem is often broken down into three subtasks:

- Cell-Entity Annotation (CEA): Matching a cell to an entity within a knowledge graph
- Column-Type Annotation (CTA): Assigning an ontology class to a column
- Column-Property Annotation (CPA): Assigning an ontology property between two columns

For the experiments in this paper, we focus on CTA because our dataset has more ground-truth samples for column types. However, the findings may extend to other tasks as well because CTA has been used to inform CEA and CPA alignments in some solutions [22].

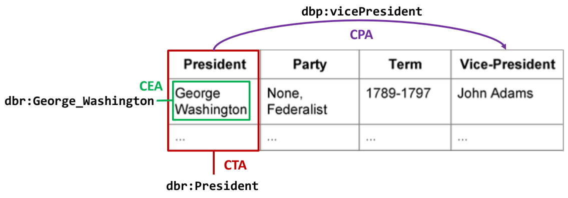


Fig. 2. Example of the three semantic tabular interpretation subtasks over a table. Adapted from [8]

Recent work has begun to utilize data dictionaries to provide additional context and improve alignment accuracy. The Semantic Data Dictionary Generator (SDD-Gen) uses transformer neural networks to align concepts within tables and data dictionaries to ontological terms [3]. The key to semantic tabular interpretation is the algorithm's ability to interpret what the table describes and align with the various ontologies underlying the knowledge graph. Therefore, the better a data dictionary description type performs in semantic tabular interpretation, the more effective it is in tabular interpretation.

To evaluate data dictionary description types, we turn the machine learning evaluation paradigm on its head. Instead of using a dataset to test an algorithm, our framework uses the SDD-Gen algorithm to test three datasets. We normalize the data dictionary descriptions from our dataset to a single description type using Large Language Models (LLM) and repeat this for all three types. The descriptions in these three datasets are then validated and manually reviewed to ensure that they meet the intended description definition. After that, we divide the three new datasets into the same training, validation, and test sets. These are used to train three new SDD-Gen models, one for each description type. Finally, the SDD-Gen models are run on the test set, and the models with higher accuracy statistics indicate the more effective description types for tabular interpretation.

C. Normalizing Descriptions

To normalize a data dictionary description to a specified type, we developed a series of Python scripts that generate an LLM prompt to rewrite the descriptions for a given set of table columns. One of the challenges we had to overcome was the need for more context for ambiguous schema columns. Earlier work on ontology alignment has shown that LLMs such as BERT need a complementary corpus when ontologies are deficient in class labels [23]. Since our goal in this experiment was not to test how well LLMs perform on semantic tabular interpretation but to generate the best possible data dictionaries to determine how effective a description type is

TABLE I
EXAMPLE DATA DICTIONARY DESCRIPTIONS FOR THE PID, GA, AND EDUCAT TABLE COLUMNS AND THE GROUND TRUTH CLASS ALIGNMENT

Column	**PID**	**ga**	**educat**
Lexical	Participant ID identifier	Child gestational age	Maternal education level
Intensional	Participant ID: A unique numerical identifier assigned to each individual enrolled in the study.	Gestational Age: The length of time, typically measured in weeks, from the first day of the mother's last menstrual period to the birth of the child.	Maternal Education: Categorized representation of the mother's educational attainment, simplifying various levels of educational achievement into distinct groups.
Extensional	Participant ID: [61, 41, 94]	Gestational Age: [34, 38, 40]	Maternal Education: 0=illiterate, 1=primary, 2=secondary or higher
Class Alignment	http://semanticscience.org/resource/SIO_000115 (identifier)	http://www.ebi.ac.uk/efo/EFO_0005112 (Gestational Age)	http://purl.obolibrary.org/obo/ExO_0000041 (Education)

for tabular alignment, we provided ground truth alignments to the prompts.

For the lexical and intensional data dictionary descriptions, our prompts use the column name, the ground truth column type alignment, and the existing data dictionary description. Below, we show an example of a full lexical description prompt:

```
Given a list of column headers,
column types, and old descriptions
in the following format:
columHeader [classType] -
oldDescription

Modify the old description for
each column header so that they are
contextually connected to the other
columns. The modified descriptions
should be short three to five words
that expands the column header and
has a subject followed by a type.

The output should have the
following format:
columNumber. columHeader { modified
description

column-list:
PID [identifier] - CHEAR
Participant ID number
ga [gestational age] - Gestational
age at birth of child
educat [education] - Maternal
Education (simplified categories)
```

Running this prompt using GPT-3.5 Turbo[1] generates the lexical description in Table I. The first instruction in the prompt, "Given a list of column headers, column types, and old descriptions in the following format: columHeader [classType] - oldDescription," describes the input data format for the LLM. We modify the first row of the table header to include the original column name in "columHeader" and add class labels from the ground truth column alignment to "classType." We request that descriptions be "contextually connected" to ensure that the column is attributed to several subjects that reoccur throughout the table. We describe the lexical description as "three to five words," that "expands the column header," and "has a subject followed by a type." The final line, "The output should have the following format: columNumber. columHeader – modified description," describes the format we want the LLM to output and allows us to parse and check for hallucinations.

For intensional descriptions, we used the following prompt template:

```
Given a list of column headers,
column types, and old descriptions
in the following format:
columHeader [classType] -
oldDescription

Generate a new description for each
column header that is contextually
connected to the other columns. The
new descriptions should be detailed
and align the column to a subject
and data type. The new descriptions
should not explain the benefits or
why the data is collected.

The output should have the
following format:
columNumber. columHeader {
headerExpansion: new description
```

The intensional prompt reuses the same input instruction as the lexical prompt template but has a different definition and output format. For the definition instructions, we use "descriptions should be detailed and align the column to a subject and data type." This instruction aims to guide LLMs to descriptions that provide column inclusion properties such as data type. We had issues with GPT-3.5 Turbo trying to motivate why the data was collected, which led to the instruction to "not explain the benefits or why the data is collected." The other change was the output format change; we asked it to include an additional field called "headerExpansion." We found that GPT-3.5 Turbo produced better intensional

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

descriptions with the header. We theorize that the LLM can leverage its training better because this format is similar to how humans write dictionary definitions, which are related to intentional descriptions.

When we ran this prompt using GPT-3.5 Turbo, it generated the lexical description in Table I. The results demonstrate how, during the conversion from lexical to intensional description, GPT-3.5 Turbo attempts to provide additional details, such as the "weeks" measurement from column 'ga.' While we cannot guarantee accuracy with the actual dataset, these embellishments should be fine for evaluation.

For extensional description, we used a different approach because of the availability of data and codebooks. First, we reused the header expansion generated from the intensional description to generate an expansion of the column name. Next, we reviewed the data dictionary to determine if a column contained numerical data or had an associated codebook. If the column contained numerical data, we would grab the first three unique data values for class membership. Otherwise, we would add the codebook to the description. We generated the extensional descriptions in Table I using this method.

D. Validating Descriptions

To validate the normalized data dictionary descriptions, we use a three-stage methodology that checks for error messages, detects hallucinations, and reviews a description's type. Using this methodology, we found that only 1.3% of all data dictionary descriptions needed manual changes, demonstrating the effectiveness of the normalization process.

The first step in the validation process was determining whether the LLM provided a valid response by checking for error messages. Every query against GPT-3.5 Turbo has a finished reason indicating why it stopped processing the request. The processing could be stopped early due to message size, content filtering, or insufficient credits. Often, these issues require a developer to address a bug or directly intervene to resolve them.

The next step is to detect hallucinations by ensuring that responses follow the formatting guidelines in the prompt and contain the correct number of column descriptions. In testing, we found that when the GPT-3.5 Turbo would begin to hallucinate, it would no longer follow the formatting guidelines. Often, checking if the column numbering system was off would indicate a hallucination, such as combining semantically similar columns into a single column with one description. Another issue that this step addresses is remapping columns. GPT-3.5 Turbo often renames columns; for example "Year_Collected" becomes "Year Collected." While this change seems reasonable, some very drastic name changes can occur. We developed several automated strategies that automatically remap these columns by examining the name changes, but if the change is drastic, we flag the response for developer disambiguation.

The final step in the validation process is to review the actual content of the new data dictionary description. For this experiment, because the type of the data dictionary description was important, we manually reviewed and modified each description to ensure it met the description type definitions.

The manual review of the new data dictionary descriptions revealed that our methodology was very effective at normalizing data dictionaries to a single type with a better error rate than real-world data dictionaries [3]. The original description type and uncommon abbreviations seem to impact the normalization process, but overall, their effects appear minor.

For lexical description, around 8.7% of errors were caused by GPT-3.5 Turbo not generating high-quality data dictionary descriptions. These descriptions primarily did not link a column type to a subject. Around 62.5% of errors were caused by prompts generating intensional descriptions instead of lexical ones. This happened more often when converting from an intensional to a lexical description. This indicates that the original description type influences the quality of the normalized descriptions. The rest of the lexical description errors were just minor formatting issues.

For intensional descriptions, 53.3% of errors were caused by GPT-3.5 Turbo not generating high-quality data dictionary descriptions. Once again, most of these did not link the column type to a subject. This appeared to happen in descriptions with uncommon abbreviations such as postpartum questionnaire (PPQ) or developmental quotient (DQ). The rest of the errors consisted of descriptions that did not create header expansions. The data did not indicate what caused these issues.

For extensional descriptions, a majority of the errors were caused by the missing header expansions mentioned above. In addition, around 29% of the errors that needed to be addressed were caused by poor codebook formatting. Data publishers include codebooks that occasionally contain unexpected characters like tabs or newlines that prevent the codebooks from being properly parsed. In general, this issue could be handled programmatically in the future to avoid these errors.

IV. Evaluation

To determine which data dictionary description types are most effective for tabular interpretation, we have developed three new datasets, each normalized to a single description type. We then train three new models, one on each dataset, and compare their performance to align a table column to an ontology class with each other and a baseline model [3] trained on the original dataset consisting of all three description types.

For this experiment, we use the SDD-Dataset [3] as the original dataset, which was created by combining biomedical studies data from HHEAR [18], NHANES [19], and TCGA [20]. The SDD-Dataset [3] has 43 tables with 900 column-type alignments aligned against the Clinical Measurement Ontology (CMO) [24], Cognitive Atlas Ontology (COGAT) [25], Human Disease Ontology (DOID) [26], Human-Aware Science Ontology (HAScO) [27], Human Health Exposure Analysis Resource Ontology (HHEAR) [28] (which imports a subset of the Exposure Science Ontology (ExO) [29], Experimental Factor Ontology (EFO) [30] and more), the National Health and Nutrition Examination Surveys ontology (NHANES) [31], the Phenotype And Trait Ontology (PATO) [32] and Semantic science Integrated Ontology (SIO) [33].

This problem is challenging because these ontologies consist of over 40,000 classes, and each column can be aligned with multiple classes. For this experiment, we divided each dataset into the same three training, test, and validation sets, 50% training, 20% validation, and 30% testing, based on the number of column-type alignments made within the ground truth. We were careful not to split tables within a given study across multiple sets to ensure no direct overlap between training and testing. To measure the performance of the models, we use standard performance metrics for precision (1), recall (2), and F1 score (3). We consider the F1 score as the primary metric in this evaluation.

$$Precision = \frac{correctAnnotations}{submittedAnnotations} \tag{1}$$

$$Recall = \frac{correctAnnotations}{groundtruthAnnotations} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

For the SDD-Gen, we trained the transformer only using the column label and description to not disadvantage the extensional description dataset. We tracked the L1 distance between class embeddings on the validation set to ensure our models would not overfit the training set. After that, we calculated the optimal α for each model using the validation set where α is the maximum distance between a description and an ontology class embedding for an alignment to be considered a match. For our final results, we vary k, which is the maximum number of alignments the SDD-Gen will return.

The SDD-Gen model trained for 200 epochs for the lexical description training set, and the optimal α was calculated to be 0.02. For the intensional description, the SDD-Gen model trained for 3,800 epochs, and the optimal α was calculated to be 0.01. Finally, for the extensional description, the SDD-Gen model trained for 1,800 epochs, and the optimal α was calculated to be 0.02.

TABLE II
F1 STATISTICS FOR SDD-GEN MODELS ON DESCRIPTION DATASETS

Description Dataset	F1-Score	Precision	Recall
Baseline [3] (k=1)	0.239	0.525	0.155
Baseline [3] (k=5)	0.212	0.269	0.175
Baseline [3] (k=10)	0.208	0.257	0.175
Lexical (k=1)	0.281	**0.810**	0.170
Lexical (k=5)	0.463	0.802	0.325
Lexical (k=10)	0.463	0.802	0.325
Intensional (k=1)	0.351	0.691	0.235
Intensional (k=5)	**0.555**	0.689	**0.465**
Intensional (k=10)	**0.555**	0.689	**0.465**
Extensional (k=1)	0.355	0.723	0.235
Extensional (k=5)	0.554	0.720	0.450
Extensional (k=10)	0.554	0.720	0.450

We ran each model on its corresponding test set with the final parameters locked in. Figure II shows the final results, with the SDD-Gen achieving an optimal F1 score of 0.406 on lexical descriptions, 0.555 on intensional descriptions, and 0.554 on extensional descriptions.

There are several interesting results from this experiment. First, we observe that as k increases, precision decreases, and recall improves. This makes sense because if you are interested in only the most accurate alignments, you should only include the best alignment. However, when complicated or ambiguous descriptions are involved, providing several good answers may lead to a better alignment among the group. This is reflected in higher k values, maximizing the F1-score across all description types. Another pattern that emerged was the same performance for $k = 5$ and $k = 10$ across all three normalized models. Many factors affect this, including the α range, the semantic overlap of ontologies, and how effective the transformer networks were at embedding data dictionary descriptions. Therefore, it is hard to draw any meaningful insight from this observation. All three normalized models outperformed the baseline model, which is expected. The baseline model [3] had the same amount of training data as the other models but had a more complex learning task. Instead of learning a single description type, it had to learn all three and when to apply them.

Another interesting result from this experiment was that lexical descriptions were the easiest for alignment algorithms to parse. The simpler syntax of the lexical descriptions required significantly fewer training epochs, and lexical description had the highest precision of all three description types. This indicates it provided more accurate alignments within the suggested alignments. However, the recall was significantly worse than the other description models, indicating it had trouble aligning all columns. This supports our hypothesis that complex columns may require additional context to parse.

The other interesting takeaway from this experiment was that both intensional and extensional description models performed equally well regarding F1 score. For $k = 1$, extensional descriptions outperformed intensional models, and for $k = 5$ and $k = 10$, intensional models were slightly more accurate. Both algorithms achieved their highest F1 scores at $k = 5$ and $k = 10$. We had initially hypothesized that intensional descriptions may be harder to parse and that extensional descriptions may suffer from the transformer model's inability to embed numerical data. Both models took several orders of magnitude longer to train. However, the additional context provided by both descriptions gave them an edge over lexical descriptions.

Another interesting result from this experiment was the optimal α parameter learned from each validation set. Lexical and extensional models had an optimal α twice as large as the intensional model. This indicates that the intensional model could align description embeddings closer to the ontology embeddings. Figure 3 highlights this in a toy example. While this resulted in a similar F1 score on the SDD-Dataset, intensional would outperform extensional descriptions on an

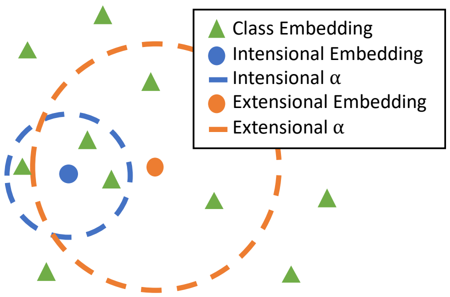


Fig. 3. Toy example showing how α effects alignments. Adapted from [8]

alignment dataset with larger ontology search spaces such as DBpedia [34]. These results also support human intuition. Often, we use intensional descriptions to define objects instead of extensional descriptions because it can be challenging to understand the scope of an extensional-described class with a large number of named entities of varying properties.

V. Conclusion

In this paper, we present our exploration of the effectiveness of different types of data dictionary descriptions for tabular interoperability through the lens of semantic tabular interpretation. Our study provides guidance to data dictionary publishers interested in improving tabular interpretability. In this study, we defined three common types of data dictionary descriptions from a review of biomedical data dictionaries, developed a methodology to normalize descriptions to a single description type, and developed an evaluation framework to compare the effectiveness of each description type. Our experiments show that both intensional and extensional descriptions are effective for tabular alignment. Both types of description models achieved comparable F1 scores. The additional context provided by both descriptions gave them an edge over lexical descriptions. The smaller α value for intensional descriptions indicates that intensional trained models can be more precise. As a result, we recommend that data publishers use intensional data dictionary descriptions if they want to improve tabular interpretability. However, extensional descriptions can be an excellent choice for describing columns with data codes or entity data. This work will allow users to develop better data dictionaries using the guidance laid out, improving the reusability of tabular data.

VI. Future Work

There are two primary future work directions that would be useful to explore. The first is expanding this work outside of the biomedical studies domain. There is a whole world of data dictionaries out there that come in every shape and size. While we did a sufficient job characterizing the biomedical domain, other description types could be common to different domains and combinations of description types. For instance, in [35], we found data dictionary descriptions describing the Lemur population at the Duke Lemur Center that combine intensional and extensional descriptions. It would be nice to see a more comprehensive review of the data dictionary description space to develop community standards.

The other direction for future work is to improve the normalization methodology. While it was accurate overall with fewer errors than data dictionaries in the wild. The observed errors showed that the original description type influenced the quality of the normalized description. It would be interesting to develop a methodology that classifies the original description type and uses a specialized prompt for each conversion, for instance, a specific prompt for lexical to intensional and intensional to lexical.

Acknowledgment

We want to give a special thanks to Dr. Jeanette Stingone, Dr. Sofia Bengoa, Dr. James Masters, and Dr. Paulo Pinheiro, who developed the data dictionaries, semantic data dictionaries, and framework to work with the data in the Health's Human Health Exposure Analysis Resource (HHEAR) project. This work is partially supported by the National Institutes of HHEAR project (NIH U2CES026555) and the Air Force Research Laboratory's Civilian Academic Degree Program. This work would not have been possible without the metadata generated from the HHEAR studies [36]–[57]. This paper has been adapted from a chapter of the author's dissertation [8].

References

[1] D. Raggett, Jan. 1997, "HTML 3.2 Reference Specification," distributed by W3C, Accessed: Sep. 17, 2024. https://www.w3.org/TR/2018/SPSD-html32-20180315/.

[2] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu, "Uncovering the relational web," in *11th Int. Workshop on the Web and Databases (WebDB)*, 2008, pp. 1-6.

[3] M. Johnson, J. A. Stingone, S. Bengoa, J. Masters, and D. L. McGuinness, "Complex semantic tabular interpretation using sdd-gen," in *18th Int. Conf. on Semantic Comput. (ICSC)*, 2024, pp. 317-322.

[4] G. McDaniel, *IBM Dictionary of Computing*. 10th ed., New York, NY, USA: McGraw-Hill, 1994.

[5] B. F. Crabtree and W. F. Miller, "A template approach to text analysis: developing and using codebooks," *Doing Qualitative Res.*, vol. 3, pp. 93–109, Apr. 1992.

[6] National Center for Health Statistics, May 2021, "Demographic Variables and Sample Weights," distributed by Centers for Disease Control and Prevention, Accessed: Aug. 8, 2023. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.htm.

[7] Northwest Environmental Data-Network, "Best practices for data dictionary definitions and usage," pacific Northwest Aquatic Monitoring Partnership. Accessed: April 15, 2024. [Online] Available: https://pnamp.org/cms/assets/d57bac50-7caa-11ed-9fb9-0d1603bac9b4-best_practices_for_data_dictionary_definitions_and_usage_version_1.1_2006-11-14.pdf.

[8] M. Johnson, "A Data Dictionary Based Approach to Semantic Tabular Mapping," Ph.D. dissertation, Dept. Comput. Sci., Rensselaer Polytechnic Inst., Troy, NY, 2024.

[9] M. Vardigan, P. Heus, and W. Thomas, "Data documentation initiative: Toward a standard for the social sciences," *International Journal of Digital Curation*, vol. 3, no. 1, pp. 107–113, 2008.

[10] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, pp. 1–9, Mar. 2016.

[11] J. Kunze and T. Baker, "The dublin core metadata element set," Tech. Rep., 2007.

[12] A. Cunningham, Sep. 2023, "HTML 3.2 Reference Specification," distributed by W3C, Accessed: Sep. 19, 2024. https://committee.iso.org/sites/tc46sc11/home/projects/published/iso-23081-metadata-for-records.html.

[13] V. Core and V. None, "Vra core 4.0 element description." Library of Congress. https://www.loc.gov/standards/vracore/, 2007.

[14] P. Rocca-Serra *et al.*, “Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level,” *Bioinformatics*, vol. 26, no. 18, pp. 2354–2356, 2010.

[15] Smithsonian Institution, “Smithsonian data management best practices,” smithsonian Libraries. Accessed: Sep. 12, 2024. [Online] Available: https://library.si.edu/sites/default/files/tutorial/pdf/datadictionaries20180226.pdf.

[16] E. M. Buchanan *et al.*, “Getting started creating data dictionaries: How to create a shareable data set,” *Adv. in Methods and Pract. in Psychological Sci.*, vol. 4, no. 1, pp. 1–10, Jan. 2021.

[17] Center for Open Science, “How to make a data dictionary,” oSF Support. Accessed: Sep. 12, 2024. [Online] Available: https://help.osf.io/article/217-how-to-make-a-data-dictionary.

[18] S. M. Viet *et al.*, “Human health exposure analysis resource (hhear): A model for incorporating the exposome into health studies,” *Int. J. of Hygiene and Environ. Health*, vol. 235, pp. 14–21, Jun. 2021, doi: 10.1016/j.ijheh.2021.113768.

[19] National Center for Health Statistics, Sep. 2017, “Nhanes - about the National Health and Nutrition Examination Survey,” distributed by Centers for Disease Control and Prevention, Accessed: Sep. 3, 2018. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

[20] National Cancer Institute, Apr. 2022, “The Cancer Genome Atlas Program,” distributed by National Institutes of Health, Accessed: Jun. 29, 2023. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[21] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation, and Applications*. 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2007.

[22] X. Li *et al.*, “Kgcode-tab results for semtab 2022,” in *Semantic Web Challenge on Tabular Data to Knowl. Graph Matching (SemTab)*, 2022, pp. 37-44. [Online]. Available: https://ceur-ws.org/Vol-3320/paper5.pdf.

[23] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, “Bertmap: a bert-based ontology alignment system,” *AAAI Conf. on Artif. Intell.*, vol. 36, no. 5, pp. 5684–5691, Jul. 2022, doi: 110.1609/aaai.v36i5.20510.

[24] J. R. Smith *et al.*, “The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications,” *J. of Biomed. Semantics*, vol. 4, no. 1, pp. 1–12, Dec. 2013.

[25] 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. 1–4, Jun. 2010, doi: 10.1038/npre.2010.4532.1.

[26] L. M. Schriml *et al.*, “Human disease ontology 2018 update: classification, content and workflow expansion,” *Nucleic Acids Res. (NAR)*, vol. 47, no. D1, pp. D955–D962, Nov. 2019, doi: 10.1093/nar/gky1032.

[27] P. Pinheiro *et al.*, “Annotating diverse scientific data with hasco,” in *Seminar on Ontology Res. in Brazil*, 2018, pp. 80-91.

[28] HHEAR, “Hhear ontology,” hADatAc. Accessed: May 12, 2022. [Online] Available: https://www.hadatac.org/hhear-ontology/.

[29] C. J. Mattingly, T. E. McKone, M. A. Callahan, J. A. Blake, and E. A. C. Hubal, “Providing the missing link: the exposure science ontology exo,” *Environmental Sci. and Technol.*, vol. 46, no. 6, pp. 3046–3053, Feb. 2012, doi: 10.1021/es2033857.

[30] J. Malone *et al.*, “Modeling sample variables with an experimental factor ontology,” *Bioinformatics*, vol. 26 no. 8, pp. 1112–1118, Mar. 2010, doi: 10.1093/bioinformatics/btq099.

[31] H. Santos, P. Pinheiro, and D. L. McGuinness, “nhanes-hadatac,” github. Accessed: May 12, 2022. [Online] Available: https://github.com/tetherless-world/nhanes-hadatac.

[32] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, “The anatomy of phenotype ontologies: principles, properties and applications,” *Briefings in Bioinf.*, vol. 19, no. 5, pp. 1008–1021, Apr. 2018, doi: 10.1093/bib/bbx035.

[33] M. Dumontier *et al.*, “The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery,” *J. of Biomed. Semantics*, vol. 5, no. 1, pp. 1–11, Mar. 2014, doi: 10.1186/2041-1480-5-14.

[34] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *international semantic web conference*. Springer, 2007, pp. 722–735.

[35] S. M. Zehr *et al.*, “Life history profiles for 27 strepsirrhine primate taxa generated using captive data from the duke lemur center,” *Sci. Data*, vol. 1, no. 1, pp. 1–11, Jul. 2014, doi: 10.1038/sdata.2014.19.

[36] M. Mahabee-Gittens, May 2020, “Biological Responses to Tobacco Smoke Exposure in Ill Children: Inflammatory Processes and Oral Metabolomic Profiles,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2017-1762_EPI_87.

[37] M. Mahabee-Gittens, May 2020, “Biological Responses to Tobacco Smoke Exposure in Ill Children: Inflammatory Processes and Oral Metabolomic Profiles SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1762/_633/_2022.2.

[38] K. Bendinskas, Feb. 2022, “Childhood Exposures, Epigenetic and Transcriptomic Responses in the Syracuse Lead Study,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2517_399.

[39] K. Bendinskas, Feb 2022, “Childhood Exposures, Epigenetic and Transcriptomic Responses in the Syracuse Lead Study SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2517_639_2022.2.

[40] A. Liu, Sep. 2021, “Denver Asthma Panel Study-CHEAR Ancillary Study,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1450_355.

[41] A. Liu, Sep. 2021, “Denver Asthma Panel Study-CHEAR Ancillary Study SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1450_629_2022.2.

[42] I. Hertz-Picciotto, Sep. 2021, “ECHO ReCHARGE Study - Environmental Exposures,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1461_219.

[43] I. Hertz-Picciotto, May 2020, “ECHO ReCHARGE Study - Environmental Exposures SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1461_630_2022.2.

[44] J. Goodrich, Mar. 2021, “First Trimester Exposures: Influence on Birth Outcomes and Markers of Biological Response,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2273_337.

[45] J. Goodrich, Mar. 2021, “First Trimester Exposures: Influence on Birth Outcomes and Markers of Biological Response SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2273_638_2022.2.

[46] C. Breton, Aug. 2022, “Maternal and Developmental Risks from Environmental and Social Stressors,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1945_177.

[47] C. Breton, Aug. 2022, “Maternal and Developmental Risks from Environmental and Social Stressors SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1945_634_2022.2.

[48] C. C. John, May 2021, “Micronutrient Deficiencies, Environmental Exposures and Severe Malaria: Risk Factors for Adverse Neurodevelopmental Outcomes in Ugandan Children,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1432_403.

[49] C. C. John, May 2021, “Micronutrient Deficiencies, Environmental Exposures and Severe Malaria: Risk Factors for Adverse Neurodevelopmental Outcomes in Ugandan Children SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1432_614_2022.2.

[50] W. Phipatanakul, Sep. 2021, “Pediatric Inner-City Environmental Exposures at School and Home and Asthma Study,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2016-1407_EPI_68.

[51] W. Phipatanakul, Sep. 2021, “Pediatric Inner-City Environmental Exposures at School and Home and Asthma Study SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1407_605_2022.2.

[52] K. Hunt, Sep. 2021, “Phthalates and Childhood Obesity in a Racially, Ethnically and Geographically Diverse Cohort,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2537_286.

[53] K. Hunt, Sep. 2021, “Phthalates and Childhood Obesity in a Racially, Ethnically and Geographically Diverse Cohort SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2537_641_2022.2.

[54] C. Karr, Aug. 2021, “The Dynamics of Exposure, Phthalates and Asthma in a Randomized Trial,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2121_260.

[55] C. Karr, Aug. 2021, “The Dynamics of Exposure, Phthalates and Asthma in a Randomized Trial SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/2121_635_2022.2.

[56] M. Karagas, Dec. 2021, “Validation of Detailed Maternal Cigarette Smoke Exposure Self Reporting by Cotinine Analysis,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1523_291.

[57] M. Karagas, Dec. 2021, “Validation of Detailed Maternal Cigarette Smoke Exposure Self Reporting by Cotinine Analysis SDD,” Human Health Exposure Analysis Resource Data Center, doi: 10.36043/1523_631_2022.2.