Ontology Construction for Database Statistics in Preparation for Learning Integration

Leo Scott Fitzsimmons
Charleston Southern University
North Charleston, SC, USA
lsfitzsimmons@csustudent.net

Songhui Yue
Charleston Southern University
North Charleston, SC, USA
syue@csuniv.edu

Sean Hayes
Charleston Southern University
North Charleston, SC, USA
shayes@csuniv.edu

Valerie Session
Charleston Southern University
North Charleston, SC, USA
vsessions@csuniv.edu

ABSTRACT

Ideally, data-focused researchers and database administrators should easily reference statistics and other relevant metadata about their databases, but statistical information often requires additional skills and effort to be manually extracted into an easily consumable, understandable format. This research aims at an automatic methodology for constructing and visualizing an ontology for database statistics. The collected metadata and statistics are transformed and presented visually alongside corresponding objects in a database ontology graph. Statistics are included to enrich the ontology graph for increasing researcher knowledge and may assist in learning integration tasks such as ontology simplification, decision-making for database design and optimization, and ingestion by other artificial intelligence (AI) processes. The results of this research consist of a methodology for complementing database ontology graphs with statistics details, a candidate model for database statistics ontology, and use cases of potential learning integrations.

CCS CONCEPTS

• **Information systems → Extraction, transformation and loading**; **Ontologies**; *Business intelligence.*

KEYWORDS

Ontology, Database, Statistics, Interoperability, Visualization

ACM Reference Format:
Leo Scott Fitzsimmons, Songhui Yue, Sean Hayes, and Valerie Session. 2024. Ontology Construction for Database Statistics in Preparation for Learning Integration. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

An ontology is a set of concepts and categories within a subject area or domain that defines their properties and the relationships between them. It forms the scaffolding that establishes standard terms, labels, and relationships, creating a common language for both structured and unstructured data [3, 6]. Ontology allows for the processing and understanding of data upon the defined, common categorization of data and relationships. [18]. One of the major objectives of ontology is to make knowledge in a domain computationally useful [17]. For heterogeneous systems, it provides a common language by which disparate systems can communicate [5, 18]. For AI systems, it gives context to data and enables AI to "decide" based on content and relationship. Ontology graphs are a way to visually represent an ontology and increase learning and understanding of a domain. They provide a visual representation of relationships between classes, attributes, and properties [8, 19].

A multitude of resources (vendor documentation, subject matter expert (SME) articles, and posts [11–13]) address RDBMS statistics maintenance (specifically Microsoft SQL Server© –or the rest of the document, referred to as "SQL Server") and how the query optimizer uses statistics and histograms. The vast majority of the writings are functional and pragmatic, addressing how to solve issues related to performance. Academic work related to database statistics and histograms is sparse; the limited information is curious because there are intricacies and subtleties to statistics updates and histograms, which make the query optimizer choose questionable plans that lead to poor and often unpredictable performance. Most efforts to resolve these issues are labeled and resolved as "statistics maintenance" problems and do not go much further. Furthermore, histogram data, primarily used by the query optimizer, are available to the analyst and have the potential for analysis and mining, but querying and presenting these data can be somewhat cumbersome and have a less-than-desirable default format.

The scarcity of information/research regarding database statistics—specifically index utilization, column statistics, and histograms—and lack of deep study/mining into the same motivates this research. This effort first produces a statistics ontology to establish a standard ontology for database statistics based on the subject matter knowledge of the RDBMS. It is intended for the ontology to define a domain and assist in making the statistics data "computationally useful." The ontology shows database statistics

as they exist at different levels, taking opportunities to aggregate characteristics or flag important aspects. Next, with the resultant ontology in mind, a semi-automated process has been created and executed. The process produces two ontology graphs: an ontology graph centered on column statistics and an ontology graph based on index utilization. This transformation effort also produces enriching data in a supporting Hypertext Markup Language (HTML) document. Figure 1 below summarizes the full objective. The resultant ontology helps promote learning and establish a framework from which data can potentially be formatted for use by machine learning/AI.

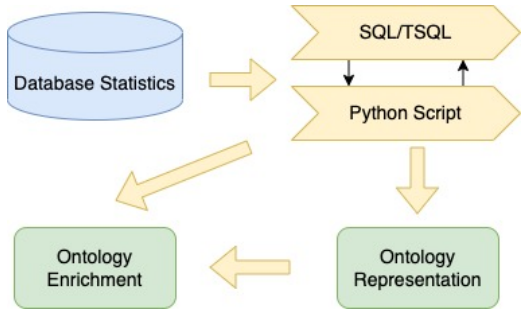

Figure 1: An Overview of the Study.

This work contributes to database statistics learning, reinforcing metadata concepts by visually organizing and representing statistics on appropriate object (table, index, column) levels. This work also contributes to statistics ontology (and, more specifically, database-related statistics) as it can be a starting point for future work regarding database statistics modeling and mining opportunities using machine learning or other AI methods.

2 RELATED WORK

There is a significant amount of research regarding ontology (in the scope of computer science) and the automatic/semiautomatic creation of ontology, especially as extracted from a relational database; the related works highlight many of these efforts, to include Protégé, RDB2OWL, and Mapping of Relational Schema [4]. Deeper or specific understanding and learning are achieved by incorporating "enriching" data into the ontology, added in from an external source. Examples of automatic/semiautomatic ontology enrichment are the works by Booshehri and Luksch [1] and Rajput and Gurulingappa [15].

Creating an ontology can be very manually intensive and often requires domain expert involvement, so a great deal of research and effort has been given to automate the task of ontology creation. There are several research documents that summarize ontology approaches; for example, the approach in [4] is very similar to this research, with at least the major difference of the RDBMS (Oracle) platform. The research conducted in the work of Loudi et al. [10] is also similar, using MySQL RDBMS. Būmans and Čerāns [2] outline a process to map the database entities to Web Ontology Language (OWL) and includes a schema that may be used and/or improved. In these cases, the ontology was focused on a user schema/data, not statistics, and not toward a standardization or extraction of statistical metadata. Issues common to building an ontology are the following:

- Domain experts needed for validation.
- Objects such as views may not translate into ontology, or important facts or relationships are sometimes missed.
- Ontology creation is dependent on the domain and the purpose of the ontology.

Database statistics, particularly in SQL Server, have been the subject of numerous white papers, conference sessions, blogs, and "best practices" lists. More specifically, the attention is more so on the maintenance of statistics due to their importance in accurately representing the actual data. Likewise, the primary purpose of the histogram is internal use; the query optimizer uses a column's histogram to help estimate the number of rows a query will return. It is speculated that due to their functional nature and common perception of "internal use only" by the RDBMS, there is very little if any, academic-level research on column statistics and histograms. The work by Lee et al. [9] addressed the operation of SQL Server Live Query Statistics feature, but no other scholarly works have addressed or capitalized on the internal collections or analysis of statistics metadata.

3 MOTIVATION AND USE CASES

Numerous ontology-from-relational-database generation efforts are successful in converting data definition language (DDL) output into OWL format. It is common to have external enriching data to help with context and understanding, and ontology is a "knowledge scaffolding" upon which AI can be built. A relational database, due to its structure and data organization, provides an effective platform not only to extract the metadata but to work with business data as well. With these characteristics and ideas in mind, the intended objectives of the research effort are listed below and explained with use cases in the subsections:

- Produce a database statistics ontology
- Produce statistics-centric ontology graphs with enriched data resources
- Prepare for data mining opportunities

3.1 Motivation I: Produce a Database Statistics Ontology

Ontology, by definition, specifies a common vocabulary for a domain or subject, and from this common vocabulary, information can be exchanged between heterogeneous systems. When using ontology, the goal is to use semantic-level representation, meaning it is formal and independent of modeling or implementation strategies [7]. In order to better understand database statistics and their importance when evaluating a database ontology, this research is to produce a database statistics ontology in the XML format and displayed as an ontology graph (via WebVOWL) and regular diagram.

3.2 Motivation 2: Produce an Ontology Graph With Enriched Data

An analysis of the data themselves is typically used to improve the accuracy and understanding of an ontology, but collecting information about the data in a useful manner can be cumbersome and time consuming. There may not be a full replacement for independently evaluating column data and table usage manually, and it is a time and resource-intensive step. Furthermore, the ontology creation process may collapse or process out important data features. Whereas enrichment is commonly from outside source—such as the Wikipedia [16] and DBpedia [1] examples—this research proposes the source of enrichment to be sourced from the data automatically and semi-automatically collected and formatted by the RDBMS. It is intended to identify and then effectively display meaningful statistical characteristics in the ontology graph for manual or automatic assessment.

3.3 Motivation 3: Prepare for Data Mining Opportunities

Although some rollup or generic statistics such as "Number of Tables" or "Is Used in Index" are helpful for research and design purposes, the primary purpose of column statistics and the column histogram is for internal use by the query optimizer to estimate query plans. Column statistics can be viewed with TSQL in SSMS, and histogram data can be viewed via SSMS (see Figure 3) as well as by utilizing TSQL and built-in stored procedures. Histogram formatting is tedious to work with, and the gathering of its data is multi-step. Manual or programmatic evaluation of a histogram may show data skew or other noteworthy distributions of data, but this method of use is minor at best. Database statistics and histograms are not commonly used outside of the query optimizer, so it is rather novel territory to utilize the information in this way. With the overwhelming amount of information and disparate statistics formats, an ontology would provide the structure needed to best organize the data and prepare it in a format that could be consumed by a processing program (perhaps AI), that could mine the information.

3.4 Use Case 1: Establish and increase knowledge of a domain

As previously described, an ontology defines a domain with common terms to be used by heterogeneous systems. Whether ontologies are created manually or automatically generated, they are refined by subject matter experts (SME) (The author does not claim to be an SME for database statistics, but based on real-world experience). Creating a database statistics ontology aids in the understanding of statistics (what information is available at what level) and establishes the framework for creating database statistics-centric ontologies that may be used to describe a database on a statistical level or be used to compare statistics between systems. Example: A junior database administrator (DBA) or analyst is tasked with comparing schemas from different databases. They haven't worked with database statistics and therefore unfamiliar with what is collected or even available. Without the statistics knowledge, they resort to selecting counts from every table, researching columns for distinct values. And maybe even attempting to collect information with custom processes. If the analyst referenced the database statistics ontology, they would better understand what is already available and at what object level.

3.5 Use Case 2: Importance of relationships, tables, and indexes

Ontology visual tools or generating tools have the ability to "collapse", essentially pruning off attributes or combining classes. The pruning/collapsing may be based on degree or "connectedness" to other classes or attributes that may be pruned off and deemed unnecessary to the ontology. Manual intervention can correct such exclusions, but it requires expert knowledge of the subject matter. There are also techniques to analyze the data programmatically to find possible missing elements, which may require additional passes for the needed analysis. The proposed work relies on statistical data already collected by the RDBMS. It will display index and table usage statistics inline in the ontology graph and amplify information in HTML that can assist in the evaluation of the importance of the relationship.

Example: An ontology is created from a Human Resources-related schema. Based on its processing rules, it prunes off a relationship between two tables that was represented by a foreign key (which also has an index). The analysis of this research shows that the index was used 98% of the time when the table was accessed. In this case, the ontology creation process deems the relationship unimportant, whereas the enriched data may indicate otherwise. Similarly, the proposed study will produce data that will help compare the "importance" of indexes and tables to each other overall.

3.6 Use Case 3: Preparation for Mining

Poor maintenance of statistics leads to inaccurate statistics, and inaccurate statistics lead to bad execution plans and poor and often inefficient query performance. Statistics sample size and frequency of statistics updates play major roles in statistics accuracy. A sample size of statistics updates can be especially inaccurate due to RDBMS row count thresholds and whether or not automatic statistics updates are enabled. Likewise, the update frequency is troublesome because the internal thresholds set for automatic statistics updates are often not triggered early enough to avoid bad execution plan estimates. There is a multitude of strategies and scripts to get statistics to behave consistently, and their use is viewed as functional, not complementary to database/data work. Statistics information is used mainly for troubleshooting when statistics are suspected to be contributing to poor query performance. With multiple variables such as formatting, timing of automatic and manual statistics updates, percent of rows sampled for manual statistics updates, poor percentage of rows sampled on automatic updates, and varying triggering levels of row modifications based on row counts, how can all the relevant data be collected and formatted for analysis? How can information from a histogram, in its current format, be analyzed? By using an ontology, the data can be formatted and prepared for ingestion by a system trained to know what to do with the data.

4 METHODOLOGY

4.1 Modeling

Figure 2 presents a class diagram of the db statistics that we are concerning with. We would like to embed those data into the generated ontology, as an enrichment. as a meaningful enrichment. To achieve this, we will build an ontology based on the model in the figure itself. Thus, for each type of the databases, we can explore various ways and concerns to enrich those data, becuase for different domain of knowledge, it might need different business logic to concern with so that different aspects in the statistics will be emphasized. That is for future work.

4.2 Data Processing

4.2.1 Part I—Create a Database Statistics Ontology (Column and Index). Based on subject matter expert experience and knowledge of the RDBMS (specifically SQL Server), an ontology for database statistics was created and revised first via UML diagram and then converted manually to XML. The OWL language was used to represent the database statistics ontology as an ontology graph. A text editor was first used to compose OWL and XML code. Next, the web-based utility WebVOWL was utilized to validate and visually display the resultant ontology graph. The ontology created is the framework for the creation of ontology graphs for a given target schema.

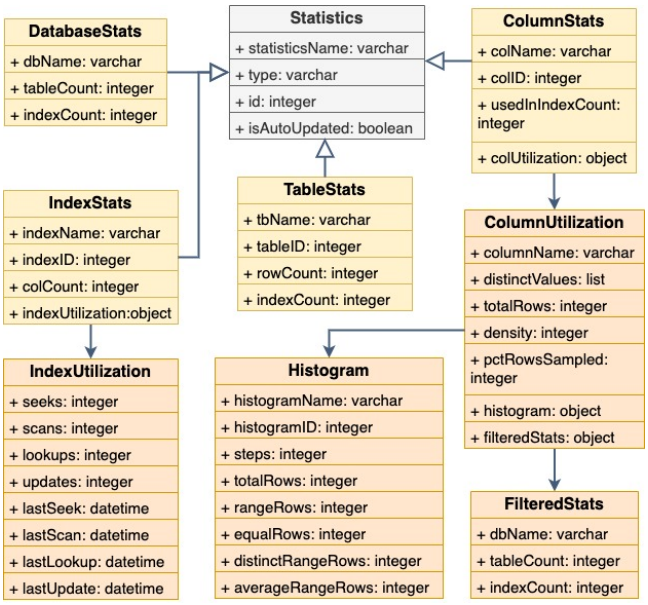

Figure 2: A UML Class Diagram of Database Statistics Ontology.

4.2.2 Part II—Generate Column and Index Ontology Graphs With Enriched Data. Noy and McGuinness [14] suggested three fundamental rules for ontology creation, one which is "There is no one correct way to model a domain— there are always viable alternatives." As such, the proposed methodology is presented. It consists of a collection of tables in a schema isolated from the target schema in the RDBMS, scripts to collect and transform the data, and a Python script to construct a Resource Description Framework Schema (RDF) OWL formatted file and an HTML file. The final step is visual representation and analysis using WebVOWL. Figure 6 represents the high-level workflow.

4.2.3 Part III—Extract Metadata to OWL/XML for Overall/Column Statistics Ontology. Using Python scripts and JDBC connections to the SQL Server instance, the OntologyEngine tables were sourced to wrap/transform data into OWL format and write to XML file. This extraction included all tables, columns, relationships, and enriching data at the appropriate object level. Additional logic for data mining may take place for additional insights.

4.2.4 Part IV—Visualize and Evaluate. The WebVOWL utility at https://service.tib.eu/webvowl/ is sourced, and the created XML is used as input. The ontology graph is checked for accuracy, and the enriching data noted. Functions such as collapse and re-evaluate will be used as needed.

4.3 Plan for Experimentation and Evaluation

The objective is to build a semi-automated way to construct ontology for DB statistics, as a priliminary work for a full automated method. Also, according to the use cases, we will develop detailed use cases in order to demonstrate the neccesity of it, to demonstrate the learning integration with AI and business intelligence. For builidng the overall workflow, the test will contain the following parts.

- Evaluation/Review of the Statistics Ontology Setup and Execution
- Base Data Collected, No Enriching Data for Column Statistics Simulated.
- Evaluation of Ontology Graph Produced in Test 2, Allowing Enriched Data.
- Collection of Index Utilization Statistics, Generation of HTML Enriching Data.
- Generation of Index-Oriented Statistics Ontology on Target Schema.
- Evaluation of the Accuracy and Completeness.

5 CONCLUSION / FUTURE WORK

This research produce the intended ontology for database statistics in XML format, using OWL, and display via WebVOWL. The ontology is used as a framework to creating a process that collect, transform, and extract data to XML files, producing statistics-centric ontologies. The generated ontologies included graphs of column-based statistics, index utilization statistics, and supporting HTML documents with enriching data. Further efforts may work toward improving and formalizing the ontology such that it can be used practically as a standard resource ontology. Other ontologies exist as standard and are referenced (there were several used by the generated ontologies in this research). After refinement and formalization, the ontology could be an official reference; including additional subject matter experts to review may give additional insight and refinement. Next, continuing to pursue efforts with machine learning/AI by creating an application to extract/consume/use

the data is highly recommended, possibly in artificial intelligence or other technology for decisioning.

Use of AI/ML technologies to analyze the statistics and identify data skew, troublesome statistics, or inefficient indexes may lead to better insights and database/query engine improvement. AI/ML most likely has more logic, processing, and learning capabilities than current methods and techniques; processes could be developed for deeper analysis and mining of the histograms and index utilization. Third, advancing research with current RDBMS metadata for additional opportunities and information may enhance the ontology or offer opportunities to roll up information. Finally, research may move one step further with standardization of ontology by expanding to other types of database management system (DBMS) platforms, such as NoSQL, or Graph DB.

REFERENCES

[1] Meisam Booshehri and Peter Luksch. 2015. An ontology enrichment approach by using DBpedia. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*. 1–11.

[2] Guntars Būmans and Kārlis Čerāns. 2018. RDB2OWL: A Language for Database to OWL Mapping and its Implementation. *INNOVATIONS AND CREATIVITY* (2018), 9.

[3] Shruthi Chari, Oshani Seneviratne, Daniel M Gruen, Morgan A Foreman, Amar K Das, and Deborah L McGuinness. 2020. Explanation ontology: a model of explanations for user-centered AI. In *International Semantic Web Conference*. Springer, 228–243.

[4] Aniagu Ugochukwu Christian. 2017. Mapping of relational schema to ontology model. In *The Fourth International Conference on Artificial Intelligence and Pattern Recognition (AIPR2017)*, Vol. 1.

[5] Alvaro Luis Fraga, Marcela Vegetti, and Horacio Pascual Leone. 2020. Ontology-based solutions for interoperability among product lifecycle management systems: A systematic literature review. *Journal of Industrial Information Integration* 20 (2020), 100176.

[6] Tom Gruber. 1993. What is an Ontology.

[7] Tom Gruber. 2008. Ontology. Retrieved 2024-05 from http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/ontology-definition-2007.htm#:~:text=An%20ontology%20specifies%20a%20vocabulary

[8] Victor Gutierrez Basulto and Steven Schockaert. 2018. From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. (2018).

[9] Kukjin Lee, Arnd Christian König, Vivek Narasayya, Bolin Ding, Surajit Chaudhuri, Brent Ellwein, Alexey Eksarevskiy, Manbeen Kohli, Jacob Wyant, Praneeta Prakash, et al. 2016. Operator and query progress estimation in microsoft SQL server live query statistics. In *Proceedings of the 2016 International Conference on Management of Data*. 1753–1764.

[10] MR Chbihi Louhdi, H Behja, and S Ouatik El Alaoui. 2013. Hybrid Method for Automatic Ontology Building. (2013).

[11] Microsoft.com. 2023. DBCC SHOW_STATISTICS (Transact-SQL). Retrieved 2024-05 from https://learn.microsoft.com/en-us/sql/t-sql/database-console-commands/dbcc-show-statistics-transact-sql?view=sql-server-ver16

[12] Microsoft.com. 2023. sys.dm_db_index_usage_stats (Transact-SQL). Retrieved 2024-05 from https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-index-usage-stats-transact-sql?view=sql-server-ver16

[13] Microsoft.com. 2023. sys.dm_db_stats_histogram (Transact-SQL). Retrieved 2024-05 from https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-stats-histogram-transact-sql?view=sql-server-ver16&viewFallbackFrom=sql-server-%20ver16

[14] Natalya F Noy, Deborah L McGuinness, et al. 2001. Ontology development 101: A guide to creating your first ontology.

[15] Abdul Mateen Rajput and Harsha Gurulingappa. 2013. Semi-automatic approach for ontology enrichment using umls. *Procedia Computer Science* 23 (2013), 78–83.

[16] Emiliano Tramontana and Gabriella Verga. 2022. Ontology Enrichment with Text Extracted from Wikipedia. In *Proceedings of the 2022 5th International Conference on Software Engineering and Information Management*. 113–117.

[17] Tania Tudorache. 2020. Ontology engineering: Current state, challenges, and future directions. *Semantic Web* 11, 1 (2020), 125–138.

[18] Songhui Yue, Xiaoyan Hong, and Randy K Smith. 2023. CSM-HR: A Context Modeling Framework in Supporting Reasoning Automation for Interoperable Intelligent Systems and Privacy Protection. *arXiv e-prints* (2023), arXiv–2308.

[19] Songhui Yue and Randy K Smith. 2021. Applying context state machines to smart elevators: Design, implementation and evaluation. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–9.