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. However, extracting and presenting this statistical information in an easily consumable format typically demands significant effort and specialized skills. This research introduces an automated methodology for constructing and visualizing an ontology enriched with relational Database Statistics (DBS). By integrating collected metadata and statistics into a visually accessible database ontology graph, this approach enhances the ontology with the ability of supporing learning integration tasks such as ontology simplification, decision-making for database design and optimization, and ingestion by other artificial intelligence (AI) processes. The outcomes of this research include a methodology for augmenting database ontology graphs with statistical details, a proposed model for a DBS ontology, and use cases demonstrating 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, 4 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 [17]. One of the major objectives of ontology is to make knowledge in a domain computationally useful [15]. For heterogeneous systems, it provides a common language by which disparate systems can communicate [5, 17]. 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 [7, 18].

A multitude of resources (vendor documentation, domain experts' articles, and posts [10–12]) address RDBMS statistics maintenance (in this research, 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 DBS 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 DBS—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 DBS 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 DBS 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.

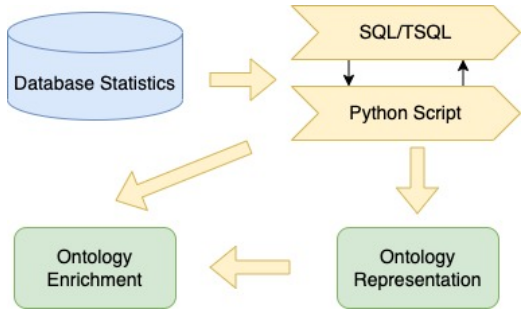

Figure 1: An Overview of the Study.

This work contributes to DBS 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 DBS 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 (using DBPedia as external source) [1] and Rajput and Gurulingappa (using a unified medical language system) [14].

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. [9] 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.

DBS, 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. [8] 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

This research aims to achieve three primary objectives: producing a DBS ontology, generating statistics-centric ontology graphs enriched with data resources, and preparing the groundwork for data mining opportunities. Each of these motivations is elaborated upon below, with use cases embedded to demonstrate their practical relevance.

3.1 Motivation 1: Produce a DBS Ontology

Ontology, by definition, specifies a common vocabulary for a domain, facilitating information exchange between heterogeneous systems through semantic-level representation. This research aims to create a DBS ontology in XML format, visualized as an ontology graph via WebVOWL, as well as a regular diagram. The ontology will enhance understanding of DBS and their significance in evaluating database ontologies.

Use Case: An ontology defines a domain with common terms used by heterogeneous systems, and the creation of a DBS ontology aids in understanding what statistical information is available at different levels. For instance, a junior database administrator (DBA) tasked with comparing schemas from different databases may lack familiarity with DBS. Without the appropriate knowledge, they might resort to inefficient methods such as manually selecting counts from tables or researching distinct column values. By referencing the DBS ontology, the DBA can quickly comprehend what statistical data is available and at what object level, streamlining their tasks and enhancing their understanding of the database's statistical landscape.

3.2 Motivation 2: Produce an DB Ontology Graph With Enriched Data from DBS

While analyzing data is essential for improving the accuracy and understanding of an ontology, collecting and formatting this information manually is often cumbersome and time-consuming. This research proposes an alternative approach: automatically and semi-automatically collecting and formatting enrichment data directly from the RDBMS. The resulting ontology graph will display meaningful statistical characteristics inline, providing insights that can be assessed either manually or automatically.

Use Case: Ontology creation tools often "collapse" or prune attributes and relationships that are deemed unimportant based on processing rules. For example, an ontology generated from a Human Resources schema might prune a relationship represented by a foreign key if the connection is not recognized as significant. However, if this research's enriched ontology shows that the associated index is used 98% of the time when the table is accessed, the relationship would be preserved in the ontology. This highlights the importance of integrating statistical data to prevent the omission of critical relationships, thus ensuring that the ontology accurately represents the underlying database structure. Furthermore, the enriched ontology can assist in evaluating the relative importance of indexes and tables, providing more nuanced insights into database performance.

3.3 Motivation 3: Prepare for Data Mining Opportunities

Although general statistics like "Number of Tables" or "Is Used in Index" are useful for research and design, the primary purpose of detailed statistics like column statistics and histograms is for internal use by query optimizers in estimating query plans. These statistics, which can be accessed using TSQL in SSMS or through built-in stored procedures, are typically underutilized outside of query optimization. This research explores how these detailed statistics can be organized within an ontology, allowing the data to be prepared for ingestion by systems—potentially AI-based—that are trained to extract meaningful insights through data mining.

Use Case: Poor maintenance of statistics can lead to inaccurate execution plans, resulting in inefficient query performance. Key factors such as sample size, update frequency, and row modification thresholds complicate the maintenance of accurate statistics. The traditional methods for gathering and analyzing statistics data, including histograms, are often cumbersome and prone to errors. By organizing this information within an ontology, the research prepares it for efficient analysis and ingestion by data mining systems. For example, a system trained to recognize patterns in histogram data could use the structured ontology to identify data skews or other noteworthy distributions, thereby contributing to more accurate query planning and improved database performance.

4 METHODOLOGY

4.1 Modeling

Ontology and class diagram can be transformed to each other [13, 16]. We are using a class diagram to represent the static designing aspects of ontology and use it to guide the ontology generation. Figure 2 presents a class diagram of the db statistics that we are concerning with. The class diagram can serve as a model for DBS ontology. Based on the concepts and relationships expressed in this model, the code for generating the infrastructure of the DBS ontology concepts and relationships can be generated using an OO language or a scripting language.

There are mainly four levels/types of statistics, namely:

- DBS or schema statistics:
- Table statistics:
- Column statistics:
- Index statistics:

Thus, as one direction of our future work, for each type of the databases, we can explore various ways and concerns to enrich those data, because for different domain of knowledge, it might need different business logic to concern with so that different aspects in the statistics will be emphasized.

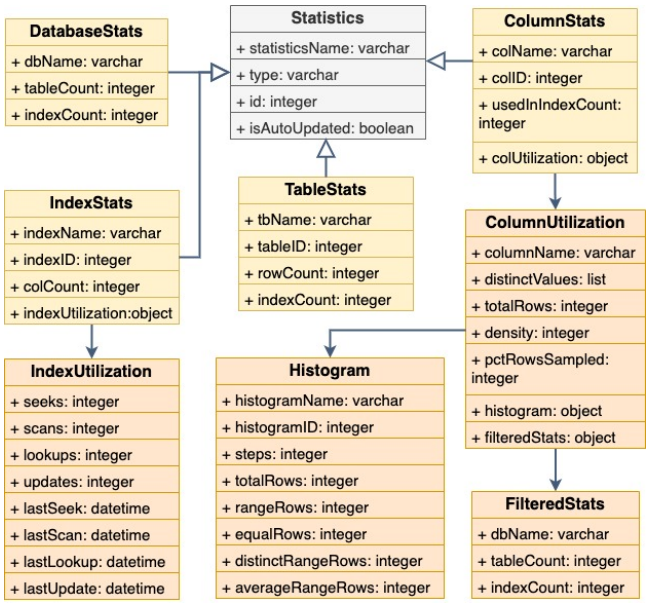

Figure 2: A UML Class Diagram for building DBS Ontology.

4.2 Data Processing

4.2.1 Creation of DBS Ontology. The initial step involves creating an ontology for DBS, focusing on columns and indexes. This is achieved through the following process:

UML to XML Conversion: Based on expertise in SQL Server, the ontology is first designed using UML diagrams. This model is then manually converted into XML format and represented using OWL (Web Ontology Language). Visualization and Validation: The OWL and XML code are composed using a text editor, and WebVOWL is utilized to validate and visualize the ontology graph. This ontology serves as the framework for developing ontology graphs tailored to specific target schemas.

4.2.2 Generation of Enriched DB Ontology Graphs. Following Noy and McGuinness's guideline that "there is no one correct way to model a domain," our methodology includes:

Data Collection and Transformation: Scripts are employed to collect and transform data from schema tables into RDF OWL formatted files. Python scripts play a crucial role in this process, including generating an HTML file for visualization. Visual Representation: The final ontology graphs, enriched with data, are visualized and analyzed using WebVOWL, as depicted in Figure 6, which outlines the high-level workflow.

4.2.3 Extraction of Metadata to OWL/XML. Metadata, including tables, columns, relationships, and enriching data, are extracted and

transformed into OWL format using Python scripts and JDBC connections to the SQL Server instance. This process involves wrapping and writing the data into XML files, with additional data mining logic applied for deeper insights as needed.

4.2.4 Visualization and Evaluation. The ontology graph is further evaluated using WebVOWL to ensure accuracy and completeness. The visualization process includes:

Accuracy Check: Verifying the ontology graph for correctness. Enrichment Assessment: Assessing the added data to confirm its relevance and utility. Function Utilization: Employing features such as collapse and re-evaluate as necessary to refine the ontology.

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. Thus, we will not only develop specific applications according to the use cases in section 3 in order to demonstrate the neccesity of it, but also evaluate the overall workflow, which will at least contain the following parts.

- Evaluation/Review of the statistics ontology setup and execution
- Collect base data and no enriching data for column statistics simulated.
- Evaluation of ontology graph produced in the last steps, allowing enriched data.
- Collection of index utilization statistics, and generation of HTML enriching data.
- Generation of index-oriented statistics ontology on target schema.
- Evaluation of the accuracy and completeness of the generated DBS ontology.

5 CONCLUSION / FUTURE WORK

This research produce the intended ontology for DBS 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] 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).

[8] 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.

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

[10] 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

[11] 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

[12] 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

[13] Meriem Mejhed Mkhinini, Ouassila Labbani-Narsis, and Christophe Nicolle. 2020. Combining UML and ontology: An exploratory survey. *Computer Science Review* 35 (2020), 100223.

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

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

[16] Minh Hoang Lien Vo and Quang Hoang. 2020. Transformation of uml class diagram into owl ontology. *Journal of Information and Telecommunication* 4, 1 (2020), 1–16.

[17] Songhui Yue, Xiaoyan Hong, and Randy K Smith. 2024. CSM-HR: A Context Modeling Framework in Supporting Reasoning Automation for Interoperable Intelligent Systems and Privacy Protection. *IEEE Access* (2024).

[18] 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.