Every Data Lake Has a Past: Analytical Exploration of Wikipedia History as a Temporal Data Lake

Mahdi Esmailoghli[1], Steven Purtzel[1], Roee Shraga[2], Renée J. Miller[3] and Matthias Weidlich[1]

[1]*Humboldt-Universität zu Berlin, Berlin, Germany*
[2]*Worcester Polytechnic Institute (WPI), Worcester, USA*
[3]*University of Waterloo, Waterloo, Canada*

Abstract

The rise of data lakes has created complex and temporally rich repositories in which tabular data exists in multiple versions. Current data discovery methods fail to utilize this crucial temporal dimension, treating each version individually. This limits the effectiveness of data discovery and integration for downstream tasks, e.g., machine learning model training. To address this, we conduct an analytical study focusing on the Wikipedia table history data lake to characterize its temporal dimensions. Our work provides essential statistics on table evolution and revision types. This foundational understanding of table evolution can help guide the development of future data lake and data management systems capable of leveraging temporal table properties.

Our exploratory analyses reveal distinct patterns in how table versions transform over time, identifying specific update frequencies, editor behaviors, and rollback prevalence that characterize the lake's lifecycle.

Keywords

Temporal Data Lakes, Data Discovery, Data Analysis, Wikipedia History

1. Introduction

The proliferation of data lakes over the past decade has established them as a primary resource for data scientists. Data lakes are widely capitalized to enrich existing datasets for various downstream tasks, including training data augmentation for machine learning (ML) models [1]. Data search or discovery within this context typically uses several established operators: join discovery [2, 3], where a query table is horizontally extended by joining with tables from the lake; union discovery [4], which involves the vertical aggregation of rows from tables sharing similar schemata; correlation discovery [5, 6] aims at finding informative features to improve a model's predictive accuracy; and integrated discovery that combines a variety of these operators [7, 8].

The evolution of data lakes has inherently resulted in temporally-rich corpora of data, particularly tabular data, where tables often exist in multiple versions [9]. These table corpora are referred to as *temporal data lakes* [10]. Examples of such data lakes include: tables encapsulated within Wikipedia pages, where each change to the page generates a new version of its corresponding tables, GitHub repositories, where updates to the codebase lead to new versions of accompanying data, and open data portals, where organizations and governments frequently provide updated reports [10].

However, current data discovery approaches are rendered impractical when applied to temporal data lakes because they treat each table version as an isolated entity. However, the state of the art has shown that utilizing the temporal dimension of data lakes can significantly improve the efficacy of data integration solutions. As one example, Bornemann et al. [11, 12] showed that inclusion dependencies that are consistently valid across different versions of a Wikipedia table tend to be more semantically correct than dependencies observed in a single instance of a table.

Wikipedia history is a publicly available source of versioned data. Tables within Wikipedia pages have been widely used before to enrich data at hand [13], such as in the WikiTables [14] project. However,

DOLAP 2026: 28th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2026, March 24, 2026, Tampere, Finland
mahdi.esmailoghli@hu-berlin.de (M. Esmailoghli); purtzesc@hu-berlin.de (S. Purtzel); rshraga@wpi.edu (R. Shraga); rjmiller@uwaterloo.ca (R. J. Miller); matthias.weidlich@hu-berlin.de (M. Weidlich)

the evolution of these tables is not well studied, which is a vital prerequisite for effectively leveraging such rich corpora. For instance, the bottom layer in Figure 1 demonstrates an example of a simple table evolution, which goes through schema modification, row addition, content edits, and structure re-ordering. Without comprehending and being able to systematically track these evolutions, one cannot leverage temporal data lakes to build a reliable data discovery system.

In this paper, we take an exploratory approach to analyzing the temporal dimensions of Wikipedia tables (Wiki Lake), validated and curated in previous research [15], which represent a prominent and accessible example of a large-scale temporal data lake. This corpus provides necessary labels, allowing us to accurately track the lineage of individual table versions.

We aim to move beyond the analysis of static data snapshots by characterizing the fundamental mechanics of table evolution across the entire data lake. Specifically, we focus on three interrelated temporal dimensions: the frequency and distribution of versions, the semantic nature of changes, e.g., schema modifications vs. content updates, and the lifecycle of updates, e.g., persistent changes vs. the ones rolled back. By quantifying these dynamics, we seek to identify evolutionary patterns and editor behaviors that distinguish stable data from volatile revisions.

We acknowledge that this work represents only the initial step toward understanding the complex concept of temporally-rich tabular data lakes. Private data lakes may have different temporal patterns than the Wiki Lake. Nonetheless, our study provides a first step in characterizing the temporal evolution of tables. Further research and exploration are essential to develop robust data management systems tailored for modern and temporal data lakes.

In the remainder of this paper, we first review related work and then provide a brief overview of the Wiki Lake and how it organizes table versions. Subsequently, we define the temporal dimensions we aim to understand and outline the exploratory actions taken to investigate them. Following each dimension, we conduct an analysis, extracting statistics from the data to characterize the defined data lake attributes. Finally, we discuss the lessons learned as well as future directions.

2. Related Work

A large body of prior work has analyzed Wikipedia through its revision histories, treating article evolution as temporal events, from which high-level semantic behaviors can be inferred. For instance, a body of research has explored conflicts such as edit wars through analyzing revision metadata [16, 17]. Another line of research focuses on rapid rollbacks, whether they help discover vandalism [18, 19] or recurring maintenance [20]. In contrast, we do not target behaviors at the article level, but study table evolution at scale, treating tables as first-class temporal citizens in the data lake.

While the domain of temporal data lakes remains significantly under-explored, recent work has begun to address the analysis and utilization of these repositories, with a particular emphasis on the evolving histories of the Wiki Lake. Notably, Bleifuß et al. [15] conducted an analytical characterization of the Wiki Lake. Although their work acknowledges the temporal aspect of the data lake, their primary investigation targets a broader spectrum of properties, such as the lifecycle of tables, including creation and deletion, and provenance regarding contributor activity. In contrast, this paper is dedicated to a more fine-granular examination of a data lake's temporal dimensions. We prioritize a detailed analysis of schema evolution dynamics, the correlation between distinct modification types, and the semantic significance of change intervals within this specific context. Moreover, Bleifuß et al. [21, 22] and Esmailoghli and Weidlich [10] introduce specific systems and corresponding data structures, including *Change-cube* and a hierarchical index. In this paper, we merely focus on an analytical approach to explore the temporal aspect of the data lake to understand the data before building such data structures.

In data management, the efficient storage of multiple table versions has received considerable attention in the literature [23, 24, 25, 26, 27, 28]. However, characterizations of the evolution itself have only recently been studied [9], and this work is limited to two versions, not the broad characterization of evolution in a data lake that we are undertaking.

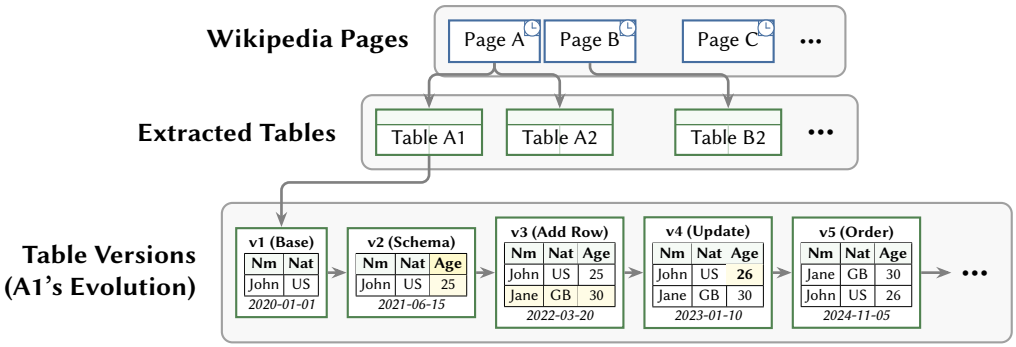

Figure 1: Structured view showing pages, tables, and versions from the Wikipedia Temporal Data Lake, illustrating the temporal evolution of Table A1 through schema changes, row additions, value updates, and reordering.

3. Wikipedia, a Temporal Data Lake

We study the Wiki Lake, which is derived from Wikipedia revision history dumps that track all modifications to Wikipedia pages. This corpus encapsulates the tables embedded within these pages and stores the discrete changes applied to each.[1] We utilize the corpus prepared by Bornemann et al. [11, 12], in which columns across versions are aligned,[2] enabling the analysis of schema evolution.

The data lake comprises 512 JSON files with a collective size of 809GB, containing 2.8M tables. The data is structured hierarchically around pages, including `PageID`, `PageTitle`, and their corresponding tables. Each table entity contains a `TableId` and a chronological sequence of table versions. Note that versions of a table are contained in the same page's history. Each version includes its unique identifier, a timestamp, contributor metadata, e.g., username or IP address, column identifiers, and the table content represented as a list of row values. Figure 1 illustrates an abstract representation of this data model.

4. Dimensions and Analysis

Traditional data lake research has focused primarily on structural and relational characteristics. These include volumetric measures, such as the number of tables [3, 29], dimensionality of tables, including the average number of rows and columns, and how tables are relevant to either each other [30] or to a given input table [8]. Similarly, Wikipedia has been examined through various lenses, ranging from the topological connectivity of its pages and entities [31] to the semantic relevance of WikiTables [32]. Extensive research has also leveraged the Wikipedia revision history as a corpus for supervised learning, utilizing human-annotated updates to improve data integration tasks [33, 34, 35].

Despite these advancements, the temporal dynamics of Wiki Lake remain largely underexplored. While current literature acknowledges Wikipedia as a repository of interconnected tables, there is a lack of comprehensive analysis regarding the evolution of these tables throughout their lifecycle [36, 15].

To address this, we categorize and analyze the temporal dimensions of the Wiki Lake by focusing on *Lake Overview*, *Change Interval*, and *Rollback*.

4.1. Lake Overview

First, we characterize the aggregate statistics of the Wiki Lake. It comprises 2.8M tables. Each table contains 14 versions, averaging 5 columns and 11 rows per version. The distribution of version counts conforms to a power-law distribution illustrated in Figure 2, characterized by a substantial number of sparsely updated tables and a small subset of highly volatile ones. Notably, the tables exhibiting the highest revision frequency are found within the pages *List of social networking websites*, *America's Next Top Model*, and *List of the verified oldest people*, recording 10k, 7k, and 6k changes, respectively.

[1] https://dumps.wikimedia.org/
[2] https://github.com/HPI-Information-Systems/tindResources

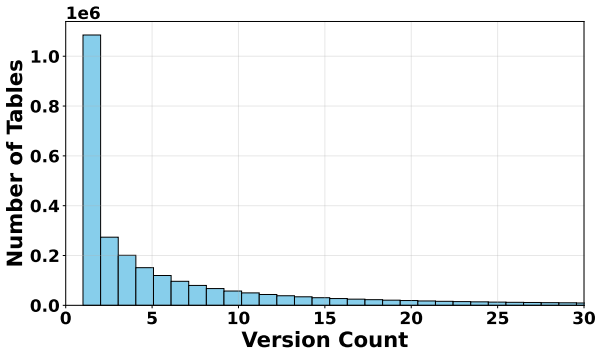

Figure 2: Version count histogram.

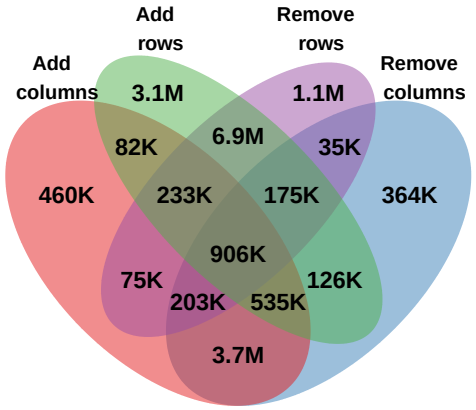

Figure 3: Venn diagram illustrating changes.

To further investigate these changes, we track schema and row-level evolution across 37M consecutive table versions. To maintain column provenance across revisions, we utilize generated unique column identifiers [11]. For row tracking, we employ a heuristic to identify a surrogate Primary Key (PK). Specifically, we find the first attribute where the uniqueness ratio, i.e., the cardinality of unique values relative to the total row count, exceeds 90%. We begin with the leftmost column, based on our observation that the first column is typically the most identifying one. Figure 3 presents a Venn diagram characterizing four distinct classes of structural and content modifications: *column addition*, *column deletion*, *row insertion*, and *row deletion*. The distribution indicates that the most frequent transformation between successive versions is the concurrent insertion and deletion of rows. This pattern can also be interpreted as modifying existing row values, including updates to the identified PK.

4.2. Change Interval

Change interval is defined as the time difference between two consecutive revisions. Analyzing this provides insights into the content-agnostic behavioral dynamics of table evolution. Specifically, change intervals can reveal the underlying intent of modifications, such as anomalous or fraudulent edits, and unveil the inherent semantics of the data. For instance, tables representing national GDP are not expected to be updated daily. We explore the distribution of change intervals across the lake to establish expected temporal baselines within various contexts. For instance, in the context of edit wars,[3] continuous edits are performed by different editors due to the inherent controversy of the Wikipedia article and conflicting views. One expects shorter change intervals on such tables [37] as opposed to more fact-driven content, e.g., historical records, which typically remain static for extended periods and undergo updates only upon the release of official reports.

Figure 4 illustrates the global distribution of change intervals, revealing a bimodal distribution: a short-term peak concentrated around one minute and a long-term peak spanning weekly to monthly intervals. This characterizes two distinct behavioral patterns within the Wiki Lake. The first represents high-frequency revisions, often attributable to iterative editing by a single contributor. In practice, substantial modifications to Wikipedia pages are non-atomic, requiring a sequence of successive commits. Additionally, short-term intervals often result from the correction of erroneous or malicious modifications by other users, moderators, or automated bots. These entities identify inappropriate edits and execute rollbacks to restore the table to its prior state.

Figure 5 depicts the distribution of change intervals aggregated at the table level. In this analysis, we utilize the median change interval per table to mitigate the influence of temporal outliers. The figure demonstrates that the high frequency of near-instantaneous changes diminishes when aggregated, suggesting that while rapid edits are numerous in total volume, they do not constitute the primary mode of long-term temporal evolution for the majority of tables.

To further understand the characteristics of temporal intervals, we analyze their variance using the Coefficient of Variation (CV). The CV is defined as the ratio of the standard deviation of change

[3] https://en.wikipedia.org/wiki/List_of_edit_wars_on_Wikipedia

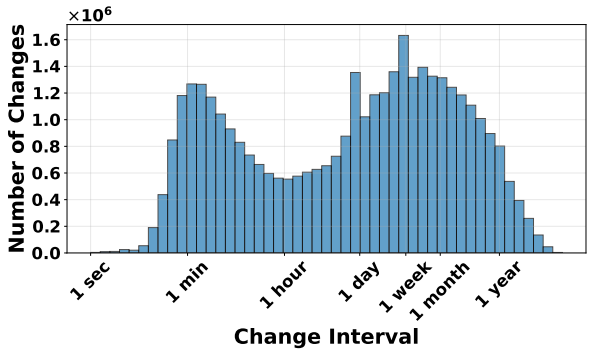

Figure 4: Distribution of change intervals.

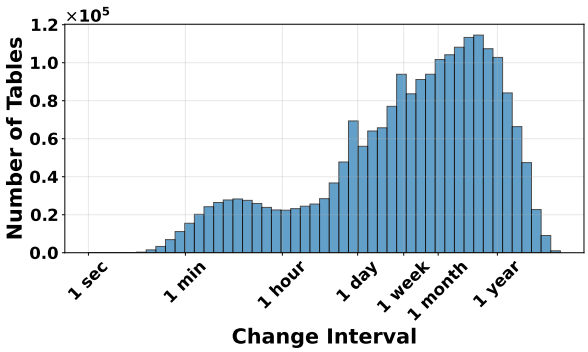

Figure 5: Median change interval distribution.

intervals (σ) to their mean (μ), $CV = \frac{\sigma}{\mu}$. It serves as a normalized measure of interval dispersion. Intuitively, tables updated at regular frequencies, such as those documenting the *Grammy Award for Song of the Year*, should exhibit a low CV, with a perfectly periodic interval yielding $CV = 0$. Conversely, tables subject to random update patterns result in higher CV values, typically $CV > 1$.

Figure 6 illustrates the distribution of CV across the Wiki Lake. Tables with $CV < 1$ are characterized as having relatively deterministic update schedules, whereas those with $CV > 1$ exhibit higher temporal dynamicity. Our analysis reveals a significant population of tables with near-static update patterns, followed by a broad distribution of tables displaying varying degrees of temporal irregularity across a wide spectrum of CV values.

4.3. Rollback

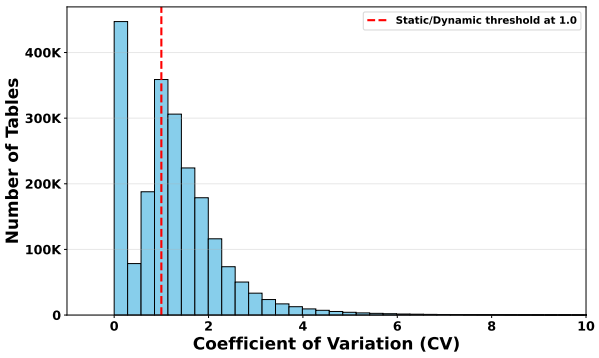

Figure 6: Coefficient of Variation (CV) distribution.

A rollback is the restoration of a table, column, or row to a previous state. Within the Wikipedia ecosystem, rollbacks constitute a critical mechanism for maintaining data integrity and trustworthiness, typically executed by moderators or automated systems[4]. Rollbacks are recorded as independent modifications on tables, therefore, identification is necessary to detect incorrect information and also avoid duplicate data. We analyze rollbacks to distinguish between reversionary actions and evolutionary content updates. Specifically, we investigate the relevant dimensions to temporal rollbacks, the profiles of the contributors involved, and the extent to which a table's revision history can serve as a proxy for its immunity to fraudulent data.

Figure 7 (left) illustrates the distribution of temporal intervals between consecutive revisions across the lake, categorized by contributor consistency, i.e., whether or not the same user initiated the successive changes. The experiment reveals a significant correlation between temporal proximity and contributor identity. High-frequency revisions, i.e., those occurring within a very short temporal window, are predominantly executed by the same user. This behavior is largely attributed to the non-atomic nature of complex updates and subsequent self-corrections.

In contrast, limiting the analysis to rollbacks $T_A \rightarrow T_B \rightarrow T_A$, where T_A and T_B represent distinct table versions, results in a different distribution. Figure 7 (right) presents the temporal distribution restricted exclusively to these reversionary cycles. Several key observations emerge: (i) rollbacks account for about 7% of the 37M consecutive changes, (ii) the distribution deviates significantly from the global baseline, showing a disproportionately high density of near-instantaneous changes, and (iii) unlike general edits, rollbacks are significantly more likely to be performed by a different user.

[4]https://en.wikipedia.org/wiki/Wikipedia:Bots/Requests_for_approval/ClueBot_NG

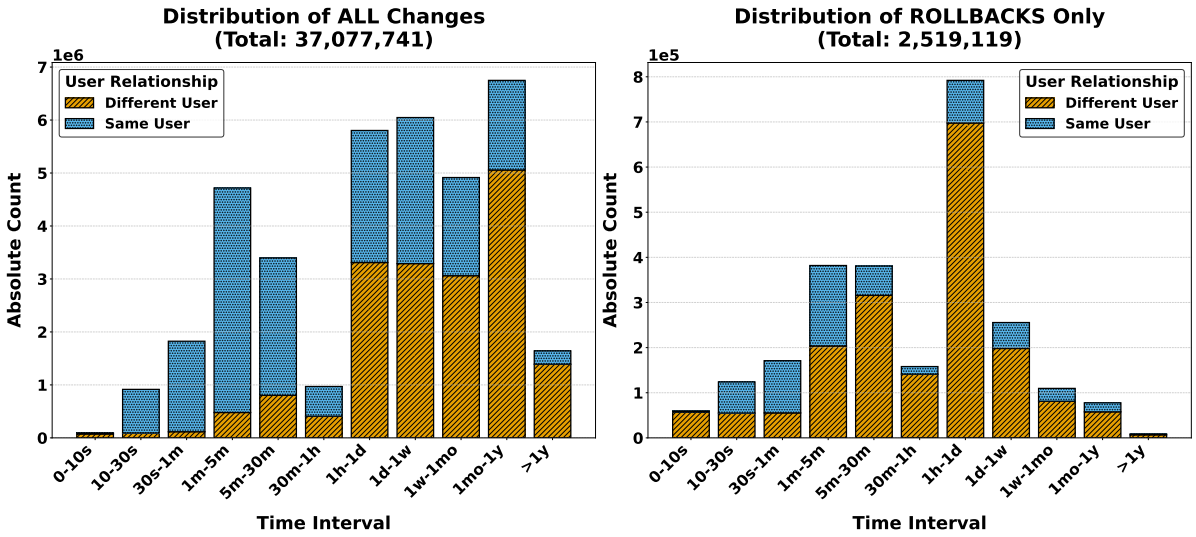

Figure 7: Change interval distribution for all changes vs. rollbacks.

5. Discussion and Future Work

The exploratory analysis of the Wiki Lake yields several key insights and future directions:

Not all change is evolution. We observe that raw revision counts significantly overestimate meaningful table evolution. While tables exhibit a high volume of revisions, a substantial fraction of these changes occur within very short intervals and often reflect non-atomic edits, self-corrections, vandalism, or moderation-driven rollbacks rather than genuine data evolution. In contrast, long-term intervals reflect the natural evolution of the data. Future discovery systems must differentiate between these "maintenance" intermediate table revisions and "evolutionary" stable ones to avoid being overwhelmed by them. It will be interesting to see how these insights carry over to other temporal data lakes where changes are not made through Wiki pages, but rather through data cleaning or integration operations.

The persistence of structural integrity. Despite the high volume of content updates, the fundamental structure of tables remains stable. Our experiments show that in 80% of the changes the schema of tables remains unchanged. For data lakes that may not be as entity-based as the Wikipedia tables, we hypothesize that there will nonetheless be some structural integrity as tables evolve.

Future systems. Our findings suggest that temporal information is not merely supplementary metadata but a foundational signal for reasoning about data lakes. Metrics such as change interval regularity and rollback frequency provide strong indicators of table reliability, maturity, and semantic intent. These signals are not entirely orthogonal to traditional discovery features such as schema similarity or value overlap, and hence should be considered for downstream tasks of temporally-valid data discovery.

Data discovery over temporal data lakes. One future direction is the development of a data discovery architecture that can reason over the semantics of evolving tables during the retrieval process. By incorporating this reasoning, a data discovery system could move beyond value and structure matching to understand the temporal validity of the retrieved results according to the user-provided query.

Semantic evolution analysis. Another angle for future research is to characterize the semantic trajectories of changes, determining whether tables evolve according to predictable patterns. Current advancements in Tabular Foundation Models (TFM) [38] and table embeddings [39] provide the tools to project tables into high-dimensional semantic spaces. This allows for a robust analysis that moves beyond simple syntax or schema changes to capture the underlying evolution of temporal semantics.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Gemini3 to: Grammar and spelling check. Furthermore, the authors used Gemini3 for Figures to: Generate images. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Chepurko, R. Marcus, E. Zgraggen, R. C. Fernandez, T. Kraska, D. R. Karger, ARDA: automatic relational data augmentation for machine learning 13 (2020) 1373–1387. URL: http://www.vldb.org/pvldb/vol13/p1373-chepurko.pdf. doi:10.14778/3397230.3397235.

[2] M. Esmailoghli, J. Quiané-Ruiz, Z. Abedjan, MATE: multi-attribute table extraction, Proc. VLDB Endow. 15 (2022) 1684–1696. URL: https://www.vldb.org/pvldb/vol15/p1684-esmailoghli.pdf.

[3] E. Zhu, D. Deng, F. Nargesian, R. J. Miller, JOSIE: overlap set similarity search for finding joinable tables in data lakes, in: SIGMOD, ACM, 2019, pp. 847–864. URL: https://doi.org/10.1145/3299869.3300065. doi:10.1145/3299869.3300065.

[4] G. Fan, J. Wang, Y. Li, D. Zhang, R. J. Miller, Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning, Proc. VLDB Endow. 16 (2023) 1726–1739. URL: https://www.vldb.org/pvldb/vol16/p1726-fan.pdf.

[5] M. Esmailoghli, J. Quiané-Ruiz, Z. Abedjan, COCOA: correlation coefficient-aware data augmentation, in: EDBT, OpenProceedings.org, 2021, pp. 331–336. URL: https://doi.org/10.5441/002/edbt.2021.30. doi:10.5441/002/edbt.2021.30.

[6] A. S. R. Santos, A. Bessa, C. Musco, J. Freire, A sketch-based index for correlated dataset search, in: 38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022, IEEE, 2022, pp. 2928–2941. URL: https://doi.org/10.1109/ICDE53745.2022.00264. doi:10.1109/ICDE53745.2022.00264.

[7] J. Becktepe, M. Esmailoghli, M. Koch, Z. Abedjan, Demonstrating MATE and COCOA for data discovery, in: S. Das, I. Pandis, K. S. Candan, S. Amer-Yahia (Eds.), Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023, ACM, 2023, pp. 119–122. URL: https://doi.org/10.1145/3555041.3589716. doi:10.1145/3555041.3589716.

[8] M. Esmailoghli, C. Schnell, R. J. Miller, Z. Abedjan, BLEND: A unified data discovery system, in: 41st IEEE International Conference on Data Engineering, ICDE 2025, Hong Kong, May 19-23, 2025, IEEE, 2025, pp. 737–750. URL: https://doi.org/10.1109/ICDE65448.2025.00061. doi:10.1109/ICDE65448.2025.00061.

[9] R. Shraga, R. J. Miller, Explaining dataset changes for semantic data versioning with explain-da-v, Proceedings of the VLDB Endowment 16 (2023).

[10] M. Esmailoghli, M. Weidlich, The past still matters: A temporally-valid data discovery system, CoRR abs/2510.13662 (2025). URL: https://doi.org/10.48550/arXiv.2510.13662. doi:10.48550/ARXIV.2510.13662. arXiv:2510.13662.

[11] L. Bornemann, T. Bleifuß, D. V. Kalashnikov, F. Nargesian, F. Naumann, D. Srivastava, Efficient discovery of temporal inclusion dependencies in wikipedia tables, in: EDBT, OpenProceedings.org, 2024, pp. 399–411. URL: https://doi.org/10.48786/edbt.2024.35. doi:10.48786/EDBT.2024.35.

[12] T. Bleifuß, L. Bornemann, D. V. Kalashnikov, F. Naumann, D. Srivastava, Structured object matching across web page revisions, in: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021, IEEE, 2021, pp. 1284–1295. URL: https://doi.org/10.1109/ICDE51399.2021.00115. doi:10.1109/ICDE51399.2021.00115.

[13] F. Tschirschnitz, T. Papenbrock, F. Naumann, Detecting inclusion dependencies on very many tables, ACM Trans. Database Syst. 42 (2017) 18:1–18:29. URL: https://doi.org/10.1145/3105959. doi:10.1145/3105959.

[14] C. S. Bhagavatula, T. Noraset, D. Downey, Methods for exploring and mining tables on wikipedia, in: D. H. Chau, J. Vreeken, M. van Leeuwen, C. Faloutsos (Eds.), Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, IDEA@KDD 2013, Chicago, Illinois, USA, August 11, 2013, ACM, 2013, pp. 18–26. URL: https://doi.org/10.1145/2501511.2501516. doi:10.1145/2501511.2501516.

[15] T. Bleifuß, L. Bornemann, D. V. Kalashnikov, F. Naumann, D. Srivastava, The secret life of wikipedia tables, in: D. Mottin, M. Lissandrini, S. B. Roy, Y. Velegrakis (Eds.), Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA-Data 2021) co-located with 47th International Conference on Very Large Data Bases (VLDB 2021), Copenhagen, Denmark, August 20, 2021, volume 2929 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 20–26. URL: https://ceur-ws.org/Vol-2929/paper4.pdf.

[16] A. Kittur, B. Suh, B. A. Pendleton, E. H. Chi, He says, she says: conflict and coordination in wikipedia, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI07, ACM, 2007, p. 453–462. URL: http://dx.doi.org/10.1145/1240624.1240698. doi:10.1145/1240624.1240698.

[17] T. Yasseri, R. Sumi, A. Rung, A. Kornai, J. Kertész, Dynamics of conflicts in wikipedia, PLoS ONE 7 (2012) e38869. URL: http://dx.doi.org/10.1371/journal.pone.0038869. doi:10.1371/journal.pone.0038869.

[18] A. G. West, S. Kannan, I. Lee, Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata?, in: Proceedings of the Third European Workshop on System Security, EuroSys '10, ACM, 2010, p. 22–28. URL: http://dx.doi.org/10.1145/1752046.1752050. doi:10.1145/1752046.1752050.

[19] B. T. Adler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, A. G. West, Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features, Springer Berlin Heidelberg, 2011, p. 277–288. URL: http://dx.doi.org/10.1007/978-3-642-19437-5_23. doi:10.1007/978-3-642-19437-5_23.

[20] J. Daxenberger, I. Gurevych, Automatically classifying edit categories in Wikipedia revisions, in: D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, S. Bethard (Eds.), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 578–589. URL: https://aclanthology.org/D13-1055/.

[21] T. Bleifuß, T. Johnson, D. V. Kalashnikov, F. Naumann, V. Shkapenyuk, D. Srivastava, Enabling change exploration: Vision paper, in: M. A. Sharaf, Y. Velegrakis (Eds.), Proceedings of the ExploreDB'17, Chicago, IL, USA, May 19, 2017, ACM, 2017, pp. 1:1–1:3. URL: https://doi.org/10.1145/3077331.3077340. doi:10.1145/3077331.3077340.

[22] T. Bleifuß, L. Bornemann, D. V. Kalashnikov, F. Naumann, D. Srivastava, Dbchex: Interactive exploration of data and schema change, in: 9th Biennial Conference on Innovative Data Systems Research, CIDR 2019, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings, www.cidrdb.org, 2019. URL: http://cidrdb.org/cidr2019/papers/p65-bleifuss-cidr19.pdf.

[23] A. P. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, A. G. Parameswaran, Datahub: Collaborative data science & dataset version management at scale, in: Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings, www.cidrdb.org, 2015. URL: http://cidrdb.org/cidr2015/Papers/CIDR15_Paper18.pdf.

[24] S. Bhattacherjee, A. Chavan, S. Huang, A. Deshpande, A. G. Parameswaran, Principles of dataset versioning: Exploring the recreation/storage tradeoff, Proc. VLDB Endow. 8 (2015) 1346–1357. URL: http://www.vldb.org/pvldb/vol8/p1346-bhattacherjee.pdf. doi:10.14778/2824032.2824035.

[25] S. Huang, L. Xu, J. Liu, A. J. Elmore, A. G. Parameswaran, Orpheusdb: Bolt-on versioning for relational databases, Proc. VLDB Endow. 10 (2017) 1130–1141. URL: http://www.vldb.org/pvldb/vol10/p1130-huang.pdf. doi:10.14778/3115404.3115417.

[26] M. E. Schüle, J. Schmeißer, T. Blum, A. Kemper, T. Neumann, Tardisdb: Extending sql to support versioning, in: Proceedings of the 2021 International Conference on Management of Data, SIGMOD/PODS '21, ACM, 2021, p. 2775–2778. URL: http://dx.doi.org/10.1145/3448016.3452767. doi:10.1145/3448016.3452767.

[27] G. S. Yilmaz, T. Wattanawaroon, L. Xu, A. Nigam, A. J. Elmore, A. Parameswaran, Datadiff: User-interpretable data transformation summaries for collaborative data analysis, in: Proceedings of the 2018 International Conference on Management of Data, SIGMOD/PODS '18, ACM, 2018, p. 1769–1772. URL: http://dx.doi.org/10.1145/3183713.3193564. doi:10.1145/3183713.3193564.

[28] M. Armbrust, T. Das, L. Sun, B. Yavuz, S. Zhu, M. Murthy, J. Torres, H. van Hovell, A. Ionescu, A. Łuszczak, M. Świtakowski, M. Szafrański, X. Li, T. Ueshin, M. Mokhtar, P. Boncz, A. Ghodsi, S. Paranjpye, P. Senster, R. Xin, M. Zaharia, Delta lake: high-performance acid table storage over cloud object stores, Proceedings of the VLDB Endowment 13 (2020) 3411–3424. URL: http://dx.doi.org/10.14778/3415478.3415560. doi:10.14778/3415478.3415560.

[29] Z. Abedjan, M. Esmailoghli, S. Galhorta, Data discovery in data lakes: Operations, indexes, systems, Proc. VLDB Endow. 18 (2025) 5455–5459. URL: https://www.vldb.org/pvldb/vol18/p5455-abedjan.pdf.

[30] R. C. Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, M. Stonebraker, Aurum: A data discovery system, in: 34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018, IEEE Computer Society, 2018, pp. 1001–1012. URL: https://doi.org/10.1109/ICDE.2018.00094. doi:10.1109/ICDE.2018.00094.

[31] N. Heist, H. Paulheim, Caligraph: A knowledge graph from wikipedia categories and lists, Semantic Web 16 (2025). URL: https://doi.org/10.1177/22104968251361349. doi:10.1177/22104968251361349.

[32] E. Muñoz, A. Hogan, A. Mileo, Dreta: Extracting RDF from wikitables, in: E. Blomqvist, T. Groza (Eds.), Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013, volume 1035 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2013, pp. 89–92. URL: https://ceur-ws.org/Vol-1035/iswc2013_demo_23.pdf.

[33] G. Karagiannis, I. Trummer, S. Jo, S. Khandelwal, X. Wang, C. Yu, Mining an "anti-knowledge base" from

wikipedia updates with applications to fact checking and beyond, Proc. VLDB Endow. 13 (2019) 561–573. URL: http://www.vldb.org/pvldb/vol13/p561-karagiannis.pdf. doi:10.14778/3372716.3372727.

[34] M. Mahdavi, Z. Abedjan, Baran: Effective error correction via a unified context representation and transfer learning, Proc. VLDB Endow. 13 (2020) 1948–1961. URL: http://www.vldb.org/pvldb/vol13/p1948-mahdavi.pdf.

[35] T. Bleifuß, L. Bornemann, T. Johnson, D. V. Kalashnikov, F. Naumann, D. Srivastava, Exploring change - A new dimension of data analytics, Proc. VLDB Endow. 12 (2018) 85–98. URL: http://www.vldb.org/pvldb/vol12/p85-bleifuss.pdf. doi:10.14778/3282495.3282496.

[36] L. Bornemann, T. Bleifuß, D. V. Kalashnikov, F. Naumann, D. Srivastava, Natural key discovery in wikipedia tables, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 2789–2795. URL: https://doi.org/10.1145/3366423.3380039. doi:10.1145/3366423.3380039.

[37] R. Sumi, T. Yasseri, A. Rung, A. Kornai, J. Kertész, Edit wars in wikipedia, in: PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011, IEEE Computer Society, 2011, pp. 724–727. URL: https://doi.org/10.1109/PASSAT/SocialCom.2011.47. doi:10.1109/PASSAT/SOCIALCOM.2011.47.

[38] P. Papotti, C. Binnig, Panel on neural relational data: Tabular foundation models, llms... or both?, Proc. VLDB Endow. 18 (2025) 5513–5515. URL: https://www.vldb.org/pvldb/vol18/p5513-paolo.pdf. doi:10.14778/3750601.3760519.

[39] G. Shrestha, C. Jiang, S. Akula, V. Yannam, A. Pyayt, M. N. Gubanov, Tabular embeddings for tables with bi-dimensional hierarchical metadata and nesting, in: A. Simitsis, B. Kemme, A. Queralt, O. Romero, P. Jovanovic (Eds.), Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025, OpenProceedings.org, 2025, pp. 92–105. URL: https://doi.org/10.48786/edbt.2025.08. doi:10.48786/EDBT.2025.08.