S-PIC4CHU: Semantics-Enriched Techniques for Data Preparation in Data Science

Gianvincenzo Alfano[1], Ilaria Bartolini[2], Diego Calvanese[3], Paolo Ciaccia[2], Sergio Greco[1], Davide Lanti[3], Pasquale Leonardo Lazzaro[5], Emilia Lenzi[4], Davide Martinenghi[4], Cristian Molinaro[1], Marco Patella[2], Letizia Tanca[4], Riccardo Torlone[5] and Irina Trubitsyna[1]

[1]*University of Calabria, DIMES, Rende (CS), Italy*
[2]*Alma Mater Studiorum University of Bologna, DISI, Bologna, Italy*
[3]*Free University of Bozen-Bolzano, Faculty of Engineering, Bolzano, Italy*
[4]*Politecnico di Milano, DEIB, Milano, Italy*
[5]*Roma Tre University, DICITA, Roma, Italy*

Abstract

The S-PIC4CHU project deals with the crucial issue of data preparation for Data Science and Machine Learning, and aims to offer new models and techniques for fighting inaccuracy, noise, uncertainty, bias, and incompleteness of data. While, at the core, the project embraces a semantics-based approach, the proposed data preparation pipeline includes data cleaning —also from the ethical viewpoint—, transformation, reduction as well as deduplication, error detection, missing value imputation, and space transformations for multimedia data. This paper illustrates the advancements on all these fronts, achieved during the first months of work on the project, and sets out the forthcoming actionable objectives.

Keywords

Data Science, data preparation, data quality, semantics, ontologies, inconsistency, incompleteness, knowledge graphs, provenance, explanation, bias.

1. Introduction

The effectiveness of data-driven applications critically depends on the quality of the data they consume. Yet, in most real-world scenarios, data is rarely "clean": it often suffers from inaccuracy, noise, incompleteness, duplication, ethical problems, and inconsistencies, which can significantly hinder downstream analytical or learning tasks. Despite the growing sophistication of machine learning algorithms and data analytics platforms, the preparatory steps that bring raw data to a usable state remain complex, labor-intensive, and error-prone. As a result, data preparation has emerged as one of the most resource-demanding and mission-critical stages in modern data science workflows.

Traditional data preparation techniques are typically tailored to specific tasks, such as missing value imputation, outlier detection, or deduplication, and often rely on heuristic or statistical models that lack generalizability and transparency. Moreover, the increasing diversity and heterogeneity of data sources — including structured, semi-structured, unstructured, and multimedia formats — further complicate the design of unified, scalable preparation pipelines. In this context, semantics-based approaches offer a promising direction by enabling a higher-level understanding of the meaning and structure of data, allowing for more robust and explainable interventions.

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy
g.alfano@dimes.unical.it (G. Alfano); ilaria.bartolini@unibo.it (I. Bartolini); diego.calvanese@unibz.it (D. Calvanese); paolo.ciaccia@unibo.it (P. Ciaccia); greco@dimes.unical.it (S. Greco); davide.lanti@unibz.it (D. Lanti); pasqualeleonardo.lazzaro@uniroma3.it (P. L. Lazzaro); emilia.lenzi@polimi.it (E. Lenzi); davide.martinenghi@polimi.it (D. Martinenghi); c.molinaro@dimes.unical.it (C. Molinaro); marco.patella@unibo.it (M. Patella); letizia.tanca@polimi.it (L. Tanca); riccardo.torlone@uniroma3.it (R. Torlone); i.trubitsyna@dimes.unical.it (I. Trubitsyna)
0000-0002-7280-4759 (G. Alfano); 0000-0002-8074-1129 (I. Bartolini); 0000-0001-5174-9693 (D. Calvanese); 0000-0002-1794-6244 (P. Ciaccia); 0000-0003-2966-3484 (S. Greco); 0000-0003-1097-2965 (D. Lanti); 0000-0003-4475-9994 (E. Lenzi); 0000-0002-2726-7683 (D. Martinenghi); 0000-0003-4103-1084 (C. Molinaro); 0000-0003-2655-0759 (M. Patella); 0000-0003-2607-3171 (L. Tanca); 0000-0003-1484-3693 (R. Torlone); 0000-0002-9031-0672 (I. Trubitsyna)

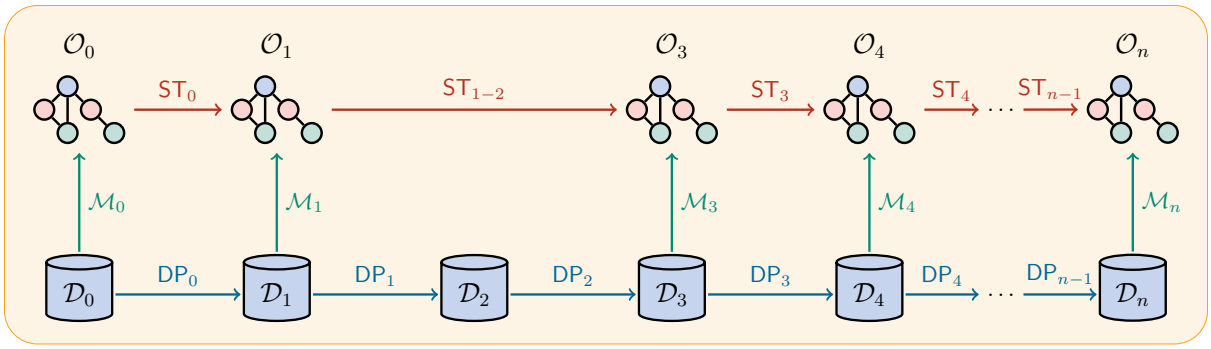

Figure 1: Conceptual architecture of S-PIC4CHU

This paper presents the S-PIC4CHU project, which addresses the challenge of building a comprehensive and semantically-informed data preparation pipeline that systematically tackles various quality issues across multiple data modalities. The proposed pipeline incorporates modules for data cleaning, transformation, dimensionality reduction, deduplication, error detection, missing value imputation, and space transformations for multimedia data. The key innovation lies in the integration of semantic models and domain knowledge across the entire preparation workflow, which allows for more accurate resolution of inconsistencies and a principled handling of uncertainty.

We report on the initial advancements and design decisions resulting from the early development phases of this approach. These include novel strategies for schema-driven imputation, semantics-aware record linkage, and embedding-based error detection in multimedia datasets. We also outline the forthcoming objectives of our work, which include the extension of the framework to support adaptive preparation strategies and the evaluation of its effectiveness across diverse real-world domains.

By laying the groundwork for a unified semantics-oriented data preparation framework, this work contributes to the broader effort of making data science pipelines more transparent, automated, and reliable — particularly in high-stakes settings where data quality directly impacts the analytical outcomes.

The paper is organized as follows. Section 2 introduces a Semantic Transformation Pipeline that enriches the traditional Data Preparation Pipeline. Section 3 discusses how ontologies can be profitably used to deal with inconsistent and/or incomplete data. Section 4 builds on a conceptual model for expressing contexts and ethical requirements, aimed at achieving a context-aware approach to be adopted in the pipeline. Section 5 presents a provenance infrastructure for capturing and querying detailed data lineage at the level of individual dataset elements. Section 6 discusses the use of ranking techniques to retrieve results that are balanced with respect to user preferences and fairness, while also offering insight into parallel execution. Finally, Section 7 concludes the paper.

2. Semantics-Enriched Data Preparation

The S-PIC4CHU project aims to deliver scalable, semantically-aware data preparation techniques that improve the quality, traceability, and explainability of machine learning workflows. Central to this vision is the development of a *Semantic Transformation Pipeline* (STP), which complements and enriches the traditional *Data Preparation Pipeline* (DPP). The STP captures, at a semantic level, both the data and the transformations occurring at each DPP stage that is conceptually significant. This is achieved by linking through semantic mappings each such DPP stage to a corresponding ontology. The resulting conceptual architecture is illustrated in Figure 1, where we have assumed that, e.g., $\mathcal{D}_2$ is a DPP stage that is *not* conceptually significant, and therefore it is *not* mapped to a corresponding stage in the STP. The mappings allow each STP stage to be viewed as a *Virtual Knowledge Graph* (VKG) over the data, enabling powerful capabilities for provenance tracking, data quality management, and bias detection.

A major challenge in realizing this vision has been the automation of ontology construction and mapping derivation across pipeline stages. To address this, S-PIC4CHU is building on a catalog of

mapping patterns, inspired by well-established database design practices (e.g., primary-foreign key structures, naming conventions, and schema normalization principles). These patterns provide reusable templates for systematically transforming relational database schemas into semantically-rich ontological representations, therefore aligning intermediate DPP stages to their semantic counterparts in the STP.

A key milestone of the project has been the development and validation of LLM4VKG, a framework to automate VKG construction [1]. LLM4VKG leverages Large Language Models (LLMs) to operationalize the pattern catalog and assist in two critical tasks:

- *Mapping Suggestion*: Given a relational schema and an initial target ontology, LLM4VKG uses cataloged patterns to guide the LLM in proposing instance-level mappings between relational elements and ontology concepts (both classes and properties), especially in the presence of naming ambiguities or incomplete ontologies.
- *Ontology Enrichment*: When the initial ontology lacks relevant concepts to describe the data, LLM4VKG employs the LLM to suggest ontology extensions that align with common design patterns and the underlying data semantics.

Importantly, the LLM is not learning the patterns but acts in the role of aligning the ontology prescribed by the curated pattern catalog to the target ontology. This hybrid design ensures both pattern fidelity and flexibility in mapping generation.

The effectiveness of LLM4VKG has been demonstrated through an extensive evaluation on the RODI benchmark, a standard suite for testing relational-to-ontology mapping tools [2]. LLM4VKG achieved an average F1-score improvement of +17% over state-of-the-art baselines, with peak gains reaching +39% in the most challenging scenarios. Moreover, LLM4VKG exhibited robustness to incomplete ontologies and could handle complex mapping tasks that are representative of real-world DPP stages in S-PIC4CHU.

By integrating LLM4VKG into the S-PIC4CHU architecture, we have already advanced some key objectives of the project:

1. semi-automatic ontology construction for STP stages;
2. derivation of semantic mappings from relational DPP stages using a pattern-based, LLM-assisted approach;
3. support for the automated generation of provenance links and explanations, by formally connecting intermediate and final data stages back to their semantic definitions and source data.

These achievements have laid the foundation and mark a significant step towards delivering open-source, semantically-driven tools for scalable data preparation in data science and machine learning workflows, fully aligned with the goals of the S-PIC4CHU project. A key step that still needs to be explored in the project remains the abstraction of data preparation operations into semantic transformation steps within the STP.

3. Incompleteness and Inconsistency

In the context of real-world data preparation, incompleteness and inconsistency are not exceptions, they are the norm. Incompleteness in datasets arises when certain values or observations are missing, partially recorded, or entirely unavailable, while inconsistency refers to data not complying with constraints expressing the application semantics. In such settings, ontologies can provide valuable knowledge to guide how incompleteness and inconsistency should be resolved in a meaningful way. Our goal is to take into account different kinds of knowledge on the application domain expressed in terms of user preferences, ontologies, data constraints, and data imputation rules (that is, rules that guide how missing values are filled out).

3.1. Ontology-enriched Data Imputation Rules

An approach to deal with incomplete data is the imputation of missing data, which consists in replacing missing values with concrete ones [3, 4]. This approach proves very useful when the subsequent

tasks regard data analytics or aggregate queries. Algorithms in this area can be divided into two main categories: statistical algorithms and machine learning algorithms. However, current algorithms deal only with raw datasets, that is, without additional knowledge that comes with them. We propose missing data imputation techniques in the presence of ontologies, leveraging data imputation rules that incorporate semantic knowledge on the application at hand to guide the process of replacing missing values in a meaningful way.

3.2. Preferences to Resolve Data Inconsistencies

Expressing preferences is natural and desirable in many applications, e.g., when one data source is more reliable than another one, or when more recent facts are preferred over earlier ones. In the presence of inconsistent data, preferences help refine the "consistent" information we can extract from inconsistent data sources, ruling out undesirable results. When data sources are accompanied by ontologies providing knowledge on the application domain, it is relevant to be able to express preferences on information not directly available in the data, but derivable from it via the ontology. An additional important aspect stems from the observation that, in the real world, preferences may not hold always, but may depend on several underlying factors. Most often, users have different preferences under different circumstances (think, e.g., of personalized e-commerce applications, where one's preferences may change based on location, time, weather, etc.). The inclusion of contextual preferences is particularly useful in ontological settings [5], where part of the knowledge is not known in advance, but it can well affect which preferences should be applied. We envisage a framework able to manage inconsistent information under user preferences that incorporates the aforementioned features. Such goals require introducing suitable formalisms to express preferences, balancing expressiveness and complexity, and establishing their impact on query answering.

4. Ethics and Context-Awareness

Ethics has become a major concern to the information management community, as both algorithms and data should satisfy ethical rules that guarantee not to generate dishonourable behaviours when they are used. However, we should also take into account that the ethical rules may vary according to the situation — i.e., the context — in which the application programs must work. Therefore, we are working on a bipartite conceptual model, composed of the *Context Dimensions Tree* (CDT) [6], a conceptual model for describing the possible contexts, and the *Ethical Requirements Tree* (ERT), which describes the ethical rules necessary to tailor and preprocess the datasets that should be fed to data analysis and learning systems in each possible context. The results of this research are collected in a paper that is currently under review.

4.1. Other Uses of Context-awareness

We plan to apply context-awareness to other aspects of the data preparation pipeline, related to data format and data quality. In this direction, we plan to use a similar framework beyond ethics, to include data quality dimensions, both at a theoretical level and through the application to concrete case studies. A first use case will be developed with the IMM Design Lab at PoliMI[1], aiming to support urban policy-makers using the Integrated Modification Methodology (IMM) [7], aligning with European Sustainable Development Goals (SDGs)[2]. The goal is to create open-source tools that span data acquisition, curation, integration, and analysis addressing semantic and spatial inconsistencies, and serve as a testbed for the S-PIC4CHU architecture, where context-aware ethical and quality requirements can be operationalized and evaluated.

When multimedia data are concerned, CDT and ERT can be effectively exploited to pre-process data according to the task at hand. Indeed, the ubiquitous use of deep neural embedding models to describe

[1] http://www.immdesignlab.com/
[2] https://unric.org/it/agenda-2030/

multimedia data produces representations consisting of hundreds, if not thousands, of dimensions. However, it is common practice to select a reduced number of such dimensions (*feature selection*) or to map representations to a subspace (*feature engineering*) with the goal of simplifying the next steps in the pipeline and removing redundancy and/or irrelevance among embedding dimensions [8]. As it can be easily argued, the choice of the actual feature selection/engineering technique to be used depends on both the dataset at hand *and* the task to be applied (classification, retrieval, and so on), for example, because a specific dimension can be relevant for a particular task and irrelevant for another one. In this light, the goal of the data analytics pipeline represents the context that can guide the application of different data-preparation algorithms on multimedia data embeddings, thus justifying the use of CDT and ERT for multimedia data. In addition, the data analytics task can be enriched to use not only the (appropriately pre-processed) embedded model, but also existing (orthogonal) semantic dimension trees, so as to improve its effectiveness (to bridge the semantic gap) and efficiency (as an additional filtering step) [9, 10].

During the project, we plan to apply the CDT framework to a heterogeneous multimedia case study, where different data modalities (e.g., images, video, and audio) and use cases will require context-specific selection of relevant features and priorities. This will allow us to validate the CDT as a guiding structure not only for ethical filtering but also for adapting the pipeline to domain-specific constraints in multimodal data preparation.

5. Data Provenance

Reliable data-driven science relies heavily on data pipelines that transform raw inputs into machine learning-ready datasets. Each transformation step may influence the outcome significantly, yet current explainability research mostly focuses on models, not on how the data was shaped before training. This is problematic, as preprocessing may inadvertently introduce bias or distort patterns, undermining trust in results. Our goal is to enable a fine-grained understanding of how each preprocessing step impacts the data, aligning with the S-PIC4CHU project's focus on semantics, quality, and explainability.

5.1. Data and Provenance Models

To support this, we designed a provenance infrastructure for capturing and querying detailed data lineage at the level of individual dataset elements [11]. Data is modeled as two-dimensional tabular structures (dataframes), with features as columns and records as rows. Preprocessing operations are categorized into four groups:

- *Data reductions*: These operations decrease the size of a dataset by eliminating rows (e.g., instance selection) or columns (e.g., feature selection).
- *Data augmentations*: These operations increase the size of a dataset by adding rows (e.g., record augmentation) or columns (e.g., feature augmentation).
- *Data transformations*: We define these as operations that modify existing elements in the dataset without altering its overall size or schema (e.g., imputation, normalization, binarization).
- *Data fusions*: These operations combine two or more datasets (e.g., join, append).

Provenance is captured using a graphical model in which the nodes are called Entities and denote specific data elements, uniquely identified by dataset, row, and column. Activities represent transformations, and their relationships are modeled using links between entities like `wasGeneratedBy`, `wasDerivedFrom`, and `wasInvalidatedBy`. Each transformation generates a compact "provlet" document per data element, which can be assembled into a full provenance graph.

5.2. Provenance Generation and Implementation

Our solution observes data changes to infer provenance, without requiring internal access to transformation logic, which is ideal for black-box or composite operations. The algorithm detects structural

and value changes between input and output dataframes, then applies PROV templates to model dependencies. For example, one-hot encoding is captured as a vertical augmentation followed by a projection.

We implemented this approach in Python using `pandas` for data handling and Neo4j as the graph store [12]. Dataframes are wrapped using an Observer pattern to enable automatic provenance capture during transformations. Expensive tasks are parallelized using `multiprocessing`. Efficient join tracking is achieved with hash-based indexing to avoid costly scanning.

A primary limitation often encountered in traditional data provenance systems, is the excessive volume of provenance data, which can severely complicate the readability and interpretability of the provenance graph. We effectively addressed this challenge by enabling customization of the level of granularity at which provenance is collected and queried. This multi-granular view balances between flexibility, efficiency, and detail in provenance analysis.

Finally, using a Retrieval-Augmented Generation approach, an LLM [13] is used to translate user-defined natural language questions into executable queries over the collected data provenance graph [14]. The query results are then contextualized by the LLM to generate user-friendly, textual narratives over the provenance.

5.3. Evaluation and Insights

We validated our approach on both real and synthetic pipelines (e.g., German Credit, COMPAS, Census, TPC-DI) [11]. The system supports a rich suite of provenance queries, including classic "Why," "How," and "Why Not" questions, as well as new ones like generate: "All Transformations," "Item History," and "Impact on Feature/Dataset Spread". These help detect changes in data distribution that may affect fairness or introduce bias.

Performance-wise, overhead introduced by provenance capture s modest: about 1.4–1.8 seconds for medium pipelines and under 30 seconds for larger ones. Complex operations like one-hot encoding produce more entities, increasing overhead, but overall, the scalability is good. Simple graph queries execute rapidly, while complex lineage traversals require more processing. A web-based interface allows users to explore transformations and inspect before/after data states, aiding debugging and trust.

In sum, our system enables detailed introspection of data pipelines by linking each output element back to its origin. This supports bias detection, fairness auditing, and transparency, key aspects of responsible AI. By tracing data derivation through every transformation, we empower users to understand not just how models behave, but why, improving accountability and trust in data science workflows.

6. Ranking

In the context of the activities concerning data reduction, we are also focusing on the relationship between top-k queries (aka ranking queries, based on scoring functions) and skyline queries (based on Pareto-dominance), both aiming at selecting relevant objects. In particular, we have addressed dataset partitioning strategies for the computation of such operators, the balance of the tuples in the retrieved result set, and the fairness thereof.

6.1. Partitioning Strategies for Computing the Skyline

We have analyzed and experimented with alternative partitioning strategies for computing the skyline of large datasets, which may turn out to be a challenging computational task [15]. Our results suggest that even partitioning by using the values of a single attribute is highly effective in reducing costs, in particular when coupled with a parallelization of the final phase in which the skyline is obtained by combining the "local" skylines obtained in the different partitions. The work [16] shows that this strategy also proves to be effective for the computation of operators, known as *flexible skylines*, that extend the skyline and hybridize it with ranking queries [17, 18]. Another line of research regards the

vertical partitioning of data for the so-called *middleware scenario*, which resulted in classical algorithms such as Fagin's Algorithm [19] and the Threshold Algorithm [20]. These algorithms have been extended and adapted to the case of flexible skylines [21], recently including the relevant scenario in which indices for data access are not available (No Random Access scenario) [22].

6.2. Directional Queries

We have studied how much top-k queries, in particular those using a linear scoring function, are effective in retrieving non-dominated objects, i.e., those in the skyline [23, 24]. To this end, we introduced four indicators to measure the difficulty of retrieving skyline points as well as their interestingness (some of which even allow for a parallelized computation that exploits partitioning [25]). We observed that relevant, yet hard to retrieve, objects occur in all the many real-world datasets we analyzed. A practical way to circumvent this problem is represented by a novel type of scoring functions, which yield the so-called *directional queries*. Such queries, besides considering the score/utility of an object, also take into account how the attribute values of an object are "balanced" with respect to the stated user preferences. Experimental results obtained on both synthetic and real-world datasets demonstrate that directional queries consistently outperform linear queries (as well as queries using some form of non-linearity) in terms of *cumulative recall*, i.e., the fraction of skyline points that are retrieved by a set of top-k queries. Along this line of research, we are also considering the problem of how to retrieve all skyline points using a *minimal* number of top-k queries, which has practical interest for characterizing the difficulty of implementing an effective data exploration process.

Finally, we remark that favoring more balanced results is not in contrast with the requirement of *diversity* of the tuples in the result set. Indeed, since the very notion of diversity is independent of the family of scoring functions one is going to use, we plan to combine it with directional queries into an integrated framework.

6.3. Fairness of a Top-k Set

A prominent aspect of Ethics (see Section 4), which is also very relevant in the context of ranking queries, is that of ensuring that the result of a top-k query also respects some *fairness* criteria. Since there is an intrinsic trade-off between the overall utility of a set of objects and the fairness requirements (maximizing the utility may lead to a biased result, and vice versa), we are now studying models and algorithms for computing what we call the "UF-skyline" of a dataset, i.e., the skyline consisting of all the k-sets of objects that are non-dominated with respect to the utility and fairness measures. The advantage of this approach with respect to the solutions available in the literature is that it provides the decision maker with the possibility of exploring all the possible trade-offs between utility and fairness and, therefore, of making a more informed choice.

7. Conclusions

In this work, we presented the latest advancements of the S-PIC4CHU project, which proposes a semantic-driven approach to data preparation addressing core challenges such as noise, incompleteness, inconsistency, bias, and lack of explainability. By integrating semantic layers, ontologies, and contextual knowledge into traditional pipelines, the project enhances the transparency, traceability, and adaptability of data workflows. A central milestone is the development of LLM4VKG, a framework that leverages LLMs to support the semi-automatic construction of VKGs and the derivation of semantic mappings. Its integration enables scalable, ontology-aware, and explainable data transformations, outperforming state-of-the-art tools.

The project also highlights the importance of domain knowledge and user preferences in guiding imputation and resolving inconsistencies, especially when ethical and contextual dimensions are relevant. The inclusion of ethical requirements and the use of the Context Dimension Trees ensure alignment with societal values and application-specific constraints.

In parallel, S-PIC4CHU delivers a fine-grained provenance system that captures and explains data transformations at the element level. Successfully tested on benchmark datasets, it enables inspection queries — such as transformation tracing and fairness impact assessment — while maintaining low computational overhead. Built on this infrastructure, the project also explores fairness-aware ranking techniques, such as directional queries and UF-skylines, which support data reduction decisions that balance utility and equity.

As future work, we plan to evaluate the entire pipeline — semantic enrichment, provenance tracking, and fairness-aware ranking — within the domain of multimedia data. In particular, leveraging the CDT, we will tailor the selection of relevant dimensions to specific applications, testing the robustness, fairness, and transparency of the S-PIC4CHU architecture in multimodal settings.

Acknowledgments

This work was supported by the Italian Ministry of University and Research (MUR) PRIN 2022 grant 2022XERWK9 "S-PIC4CHU- Semantics-based Provenance, Integrity, and Curation for Consistent, High-quality, and Unbiased data science".

This work was partially supported by the PNRR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI, under the NRRP MUR program funded by the NextGenerationEU.

References

[1] G. Xiao, L. Ren, G. Qi, H. Xue, M. D. Panfilo, D. Lanti, LLM4VKG: Leveraging large language models for virtual knowledge graph construction, in: Proc. of the Int. Joint Conf. on Artificial Intelligence, 2025. To appear.

[2] C. Pinkel, C. Binnig, E. Jiménez-Ruiz, E. Kharlamov, W. May, A. Nikolov, A. Sasa Bastinos, M. G. Skjæveland, A. Solimando, M. Taheriyan, C. Heupel, I. Horrocks, RODI: Benchmarking relational-to-ontology mapping generation quality, Semantic Web 9 (2018) 25–52. doi:10.3233/SW-170268.

[3] R. Shahbazian, I. Trubitsyna, DEGAIN: Generative-adversarial-network-based missing data imputation, Information 13 (2022) 575.

[4] R. Shahbazian, S. Greco, Generative adversarial networks assist missing data imputation: A comprehensive survey and evaluation, IEEE Access 11 (2023) 88908–88928.

[5] M. Calautti, S. Greco, C. Molinaro, I. Trubitsyna, Preference-based inconsistency-tolerant query answering under existential rules, in: Proc. of the Int. Conf. on Principles of Knowledge Representation and Reasoning, 2020, pp. 203–212.

[6] C. Bolchini, E. Quintarelli, L. Tanca, CARVE: Context-aware automatic view definition over relational databases, Information Systems 38 (2013) 45–67. doi:10.1016/J.IS.2012.05.004.

[7] T. Massimo, Integrated Modification Methodology (IMM): A phasing process for sustainable urban design, WASET World Academy of Science Engineenering and Technology. 77 (2013).

[8] H. Liu, Feature selection, in: C. Sammut, G. I. Webb (Eds.), Encyclopedia of Machine Learning, Springer, 2010, pp. 402–406. doi:10.1007/978-0-387-30164-8_306.

[9] I. Bartolini, P. Ciaccia, Scenique: a multimodal image retrieval interface, in: S. Levialdi (Ed.), Proc. of the working conference on Advanced Visual Interfaces (AVI), ACM Press, 2008, pp. 476–477. doi:10.1145/1385569.1385664.

[10] I. Bartolini, M. Patella, C. Romani, SHIATSU: tagging and retrieving videos without worries, Multim. Tools Appl. 63 (2013) 357–385. doi:10.1007/S11042-011-0948-1.

[11] A. Chapman, L. Lauro, P. Missier, R. Torlone, Supporting better insights of data science pipelines with fine-grained provenance, ACM Trans. Database Syst. 49 (2024) 6:1–6:42. doi:10.1145/3644385.

[12] P. L. Lazzaro, M. Lazzaro, P. Missier, R. Torlone, PROLIT: Supporting the transparency of data preparation pipelines through narratives over data provenance, in: Proc. of the Int. Conf. on Extending Database Technology (EDBT), OpenProceedings.org, 2025, pp. 1138–1141.

[13] A. Matarazzo, R. Torlone, A survey on large language models with some insights on their capabilities and limitations, CoRR abs/2501.04040 (2025). doi:10.48550/arXiv.2501.04040. arXiv:2501.04040.
[14] L. Gregori, P. L. Lazzaro, M. Lazzaro, P. Missier, R. Torlone, An LLM-guided platform for multi-granular collection and management of data provenance, J. Big Data (2025). To appear.
[15] P. Ciaccia, D. Martinenghi, Optimization strategies for parallel computation of skylines, Distributed and Parallel Databases (2025). To appear.
[16] E. D. Lorenzis, D. Martinenghi, Partitioning Strategies for Parallel Computation of Flexible Skylines, Algorithms 18 (2025). doi:10.3390/a18030141.
[17] P. Ciaccia, D. Martinenghi, Reconciling skyline and ranking queries, Proc. of the VLDB Endowment 10 (2017) 1454–1465.
[18] P. Ciaccia, D. Martinenghi, Flexible skylines: Dominance for arbitrary sets of monotone functions, ACM Transactions on Database Systems 45 (2020) 18:1–18:45. doi:https://doi.org/10.1145/3406113.
[19] R. Fagin, Combining fuzzy information from multiple systems, in: Proc. of the ACM Symp. on Principles of Database Systems, 1996, pp. 216–226. doi:10.1145/237661.237715.
[20] R. Fagin, A. Lotem, M. Naor, Optimal aggregation algorithms for middleware, in: Proc. of the ACM Symp. on Principles of Database Systems, 2001. doi:10.1145/375551.375567.
[21] P. Ciaccia, D. Martinenghi, FA + TA < FSA: Flexible score aggregation, in: Proc. of the ACM Int. Conf. on Information and Knowledge Management (CIKM), 2018, pp. 57–66. doi:10.1145/3269206.3271753.
[22] D. Martinenghi, Computing non-dominated flexible skylines in vertically distributed datasets with no random access, Data 10 (2025). doi:10.3390/data10050076.
[23] P. Ciaccia, D. Martinenghi, Directional Queries: Making Top-k Queries More Effective in Discovering Relevant Results, Proc. ACM Manag. Data 2 (2024). doi:10.1145/3698807.
[24] P. Ciaccia, D. Martinenghi, Relevant, yet hard to find: Directional queries to the rescue, in: Proc. of the Italian Symposium on Advanced Database Systems, CEUR Workshop Proceedings, CEUR-WS.org, 2025. To appear.
[25] D. Martinenghi, Parallelizing the Computation of Grid Resistance to Measure the Strength of Skyline Tuples, Algorithms 18 (2025). doi:10.3390/a18010029.