S-PIC4CHU: Semantics-based Provenance, Integrity, and Curation for Consistent, High-quality, and Unbiased Data Science

Gianvincenzo Alfano[1], Ilaria Bartolini[2], Diego Calvanese[3], Paolo Ciaccia[2], Sergio Greco[1], Davide Lanti[3], 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, 87036 Rende (CS), Italy*
[2]*Alma Mater Studiorum University of Bologna, DISI, 40100 Bologna, Italy*
[3]*Free University of Bozen-Bolzano, Faculty of Engineering, 39100 Bolzano, Italy*
[4]*Politecnico di Milano, DEIB, 20133 Milano, Italy*
[5]*Roma Tre University, DICITA, 00146 Roma, Italy*

Abstract

This paper presents the vision of the S-PIC4CHU project, which aims to develop innovative models and techniques for scalable data preparation in Data Science and Machine Learning. The project focuses on leveraging data semantics throughout all data preparation stages to improve data quality and ensure unbiased results. The proposed approach involves a novel data preparation pipeline semantically enriched with domain knowledge from ontologies and knowledge graphs, along with novel, semantic-based techniques for data cleaning, integration, provenance, explanation, and quality management. The validation of the approach relies on use cases from different domains, with the goal of releasing open-source tools.

Keywords

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

1. Introduction

The increasing reliance on Data Science (DS) and Machine Learning (ML) techniques across various sectors highlights the critical importance of data quality [1, 2]. Real-world data is often characterized by inaccuracies, noise, uncertainties, and inconsistencies, which can significantly

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16–19, 2025, Ischia, 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); 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)

affect the results of DS and ML tasks. There are indeed various libraries and tools available to correct erroneous values, impute missing ones, eliminate duplicate records, or disambiguate conflicting data. However, the activities of data preparation are difficult to fully automate and there is no solution today for composing, analyzing, and explaining end-to-end pipelines that transform data from raw input into training sets ready to be used for learning. In addition, these data processing activities merely refer to the syntax of the data and are rarely explained or characterized in terms of their meaning. The S-PIC4CHU project addresses these challenges by introducing a semantics-based approach to data preparation.

The core idea is to develop a *Data Preparation Pipeline* (DPP) where data is annotated with semantic information derived from ontologies and knowledge graphs [3]. This semantic enrichment plays a crucial role in subsequent data preparation steps, including data cleaning, integration, transformation, reduction, deduplication, error detection, missing value imputation, and space transformations. Furthermore, semantic techniques assist in reconciling conflicts among different data quality dimensions, a well-known challenge [4] in the field of data quality.

The project's main objectives are to develop novel models and techniques for data preparation, focusing on the semantic enrichment of data and ensuring data quality and fairness, also providing provenance information [5] to enable explainable AI [6]. The challenges raised by this paradigm shift in data preparation requires formalizing and addressing new aspects that are relevant to the data preparation process, and investigating key challenging scientific issues aimed at delivering novel solutions to well-known problems. The project contributes to the development of open-source software tools and promotes awareness of data fairness, engaging with various stakeholders.

The project is validated on two selected use cases from different domains to showcase the generality and effectiveness of the proposed solutions. S-PIC4CHU seeks to provide solutions that are not only scientifically sound but also have significant societal and economic impact due to the increasing importance of DS and ML in various sectors.

2. Related Work

The problem of developing a quality-aware pipeline of data preparation operations has been the subject of recent research, see, e.g., [7, 8]. Here we discuss more in detail the state of the art in the areas that are crucial for the stages of data preparation in DS, noting that the literature has so far paid little attention to the risks of ignoring potential conflicts among different data quality dimensions.

Data Semantics. A clear understanding of the data semantics plays a key role in all stages of the data processing pipeline, therefore data semantics needs to be taken into account explicitly. However, technologies and tools that provide semantics-based solutions, notably those relying on ontologies [9] and automated reasoning [10], currently are considered either in a limited way or not at all in the steps of the pipeline. A critical aspect in the adoption of such technologies is their scalability, w.r.t. both the size of the data and the size and complexity of the ontology. Such aspects have been studied in the Semantic Web community, but mostly restricted to the specific case of RDF (linked) data [11], thus scalability of semantic technologies in the context

of a data processing pipeline is an open research challenge.

Data Imputation. An approach to deal with the problem of missing values is data imputation, i.e., replacing unspecified values with concrete ones [12, 13], for which statistical and machine learning (e.g., Adversarial Networks) algorithms have been proposed. However, all algorithms proposed so far deal only with raw datasets, i.e., not equipped with additional knowledge.

Preference-based Inconsistency Management. While the management of inconsistent data has been studied extensively, incorporating preferences has received less attention. Even if a few preference-based approaches have been proposed, e.g., [14, 15], they neither allow users to express preferences on knowledge derived from the data, nor preconditions for preferences to hold. The first recent proposal to overcome these limitations is [16].

Data Dependencies. Functional dependencies (FDs) and their variants [17] have already been used to enforce data quality. For instance, (approximate) conditional FDs, i.e., FDs that hold only over a portion of the data, can be searched via data mining techniques, like association rules [18]: these rules, and the possible non-conformant records, are returned to the user to decide which of them must be fixed to improve data quality.

Bias. When data is used to build models impacting people's lives, we have to be sure that data and the accordingly trained models do not introduce bias [19]. Thus, in such applications, data can be considered of good quality only if they respect fairness requirements [1]. Methods based on FDs and their variants can also be adopted to discover discrimination and bias in a dataset [20], to avoid (possibly unintentional) unfair behavior and consequences.

Data Reduction. Before being fed into DS algorithms, data is usually "reduced", to improve the quality of results, using activities such as feature selection [21], object selection [22], data aggregation, and clustering. The novel notion of $\mathcal{F}$-dominance [23, 24] allows for expressing object and feature selection through a family of ranking functions with constraints on parameters (e.g., weights). Considering a family rather than a single function improves the robustness and flexibility of the selection process.

Multimedia Data Curation. Although DS technologies allow for extracting value from large conventional data repositories, their application to multimedia (MM) data is still an open research issue [25]. This is mainly due to the very complex nature of MM data whose content and semantics still lack an appropriate methodology for accurate and efficient characterization [26]. The problem is further aggravated in the case of real-time analysis of MM streams [27].

Provenance and Explanation. Explaining positive/negative query answers, i.e., knowing why a query result was/was not obtained, falls into the broad topic of Explainable AI. For relational DBs, provenance [5] is used in systems like Perm and ProvSQL [28] to keep track of the specific DB tuples responsible for deriving an answer. For DS, provenance has been applied, e.g., in data preparation [7]. For ontologies, explaining query answers has been studied

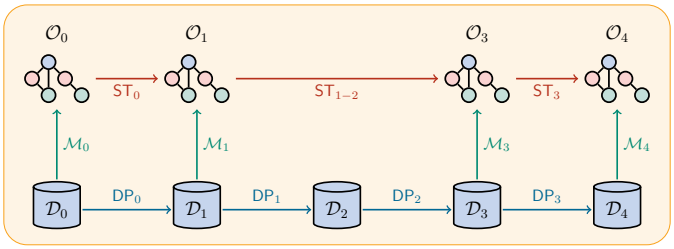

Figure 1: Conceptual architecture of S-PIC4CHU

for different description logics (DLs) [29, 30, 31] under existential rules [32, 33], considering inconsistency [34], and for Ontology-based Data Access [35], but several problems still need to be investigated.

3. Proposed Approach

We refer to the typical *Data Preparation Pipeline* (DPP), organized in a sequence of steps aimed at transforming raw data into a clean, structured, and meaningful format that improves the accuracy, efficiency, and robustness of machine learning models. A distinguishing feature of our approach is in fact the *semantic enrichment* of data, whose aim is to annotate all data that is involved in the various stages of the DPP with semantic information capturing domain knowledge and coming from ontologies and knowledge graphs. The methodology for semantic enrichment that we adopt in the project builds on Ontology-based Data Access (OBDA) and extends it to cover multiple versions of the data corresponding to the pipeline stages. Specifically, we capture at the semantic level, into a so-called *Semantic Transformation Pipeline* (STP), the data and its transformations that constitute the DPP. We illustrate the resulting conceptual architecture in Figure 1.

As mentioned, the various stages $\mathcal{D}_i$ of the DPP are linked to each other through data preparation steps DP_i. Those stages $\mathcal{D}_i$ that are (semantically) meaningful for the application domain have a corresponding semantic stage counterpart in the STP, provided as an ontology $\mathcal{O}_i$ to which they are linked through a semantic mapping $\mathcal{M}_i$. This correspondence induces, in turn, a correspondence between sequences of data preparation steps (e.g., DP_1–DP_2 in the figure) and single semantic transformation steps (e.g., ST_{1-2}).

Our semantics-centered architecture serves as the basis for realizing the specific contributions of the S-PIC4CHU approach, in which we concentrate on the key issues that follow.

Semantic Enrichment. To achieve semantic enrichment, we advance the state of the art in several directions:

1) We establish mechanisms to construct ontologies for STP stages, representing the data handled in DPP stages. Our (semi-)automatic approach leverages both data and metadata while identifying ontology elements that remain unchanged and those evolving with data transformations.

2) We develop novel methods to abstract data preparation operations of the DPP into ontology transformation operations of the STP, balancing retention of detail vs. abstraction, based on the granularity of semantic representations from Item 1.

3) We (semi-)automatically derive semantic mappings between DPP stages and their STP counterparts. For these mappings, in which queries provide the detailed correspondences between the data layer and the semantic layer, we rely on OBDA mapping patterns [36], which exploit relational constraints reflected in ontology constructs. We also consider operations at both data and semantic levels that generate the various stages of the pipelines (cf. Item 2).

4) We explore the automatic derivation of additional semantic mappings between STP stages and intermediate DPP stages, including the original data sources, building on [37].

The resulting semantic enrichments of DPPs aid in designing semantically-aware data preparation tasks and in deriving provenance information and explanations, linking back to original or intermediate data.

Data Quality. Real-world data from multiple sources are often inconsistent or incomplete, lowering the quality of DS tasks. Domain knowledge (e.g., ontologies, preferences) can enhance data quality by addressing inconsistency, incompleteness, and fairness while ensuring efficient data reduction. We approach these issues by leveraging domain knowledge expressed through user preferences, ontologies, functional dependencies, and Data Imputation Rules (DIRs), that is rules that guide how missing values are filled out. Our goals include:

1) Ontology-enriched usage of DIRs for missing data imputation. Many data analysis algorithms require complete datasets, often discarding incomplete records. Data imputation replaces missing values, aiding analytics and aggregate queries. Existing methods rely solely on raw data. We propose missing data imputation techniques within OBDA, leveraging DIRs to incorporate semantic knowledge that guides the replacement of missing values in a meaningful way.

2) Preference-based resolution of inconsistencies. Preferences help extract meaningful information from inconsistent data, e.g., when one source is more reliable than another one or when more recent facts are preferred over earlier ones. When coupled with ontologies, one should be able to express preferences also on information derivable from the data via the ontology. Also, it is important to consider that users have different preferences under different circumstances, and in ontological settings knowledge not known in advance can affect which contextual preferences should be applied. We develop a framework to manage inconsistencies with contextual preferences, which requires formalizing preferences while balancing expressiveness and computational complexity, and establishing their impact on query answering.

3) Detecting and correcting bias in data. Fairness is essential in training data for ML models, to ensure that one can trust the outcome of the process. Most existing work assesses the fairness of the employed analysis algorithms, but coherently with the S-PIC4CHU aim, we focus on ensuring fair input data. Approximate Conditional Functional Dependencies (ACFDs) [17] and data mining techniques [18] help uncover biases, identifying attribute correlations that influence decisions (e.g., job recommendations based on gender or ethnicity). We propose a framework for bias discovery based on mining ACFDs and possibly other kinds of dependencies that take into account the domain semantics. This allows the system to discover bias that would not be found otherwise. We also propose new evaluation metrics for bias in input data. For

multi-media (MM) data, biases arise from spatial (e.g., images) and temporal (e.g., audio/video) correlations, which cause a violation of the assumption of independently identically distributed data and lead to overfitting training data [38]. We investigate semantic enrichment of MM data via features characterizing their content, to exploit alternative DSAs (like 3D CNNs) able to deal with existing data correlations.

Data Reduction. Feature selection [21] removes irrelevant features that introduce noise, bias, and computational overhead, a critical issue especially for MM data, as illustrated by the US Army's neural network experiment on camouflaged tanks [39]. Given the variety of object descriptions, optimal feature sets vary by media type and DS task. Indeed, raw MM data is inappropriate for DSA algorithms without a careful selection to prevent biased learning. We propose a software framework for (semi-automated) feature selection tailored to specific use cases. Moreover, we consider the relevant example of real-time massive MM stream analysis, where very low latency for real-time results is essential.

Object selection [22] ensures high-quality, error-free input to DSAs, preventing wasted resources and biased models. We explore $\mathcal{F}$-dominance-based methods to analyze ranking functions' impact on diversity and fairness [40]. A key challenge is leveraging semantic information for cross-source object matching, to enable distributed $\mathcal{F}$-dominance algorithms [41]. Our techniques also support advanced analysis of classifier outputs, revealing correlations between data variables and predictions. Additionally, we examine the relationship between top-k queries (based on ranking functions) and skyline queries (using Pareto-dominance), both aimed at selecting relevant objects. Understanding the "discovery power" of top-k queries, i.e., their ability to retrieve skyline objects, helps assess dataset suitability for analysis tasks, providing decision-makers with valuable insights.

Data Provenance and Explanation. Data provenance and explanation serve two purposes: explaining positive query answers by enriching results with provenance details (e.g., creation as a new derived feature, transformation, or deletion) and explaining negative query answers via abduction, identifying information to be added to derive some missing answer. In the presence of inconsistency, such tasks rely on a repair-based semantics.

For positive query answers in OBDA, existing methods extend the semiring approach for relational DBs [5], taking into account both the ontology and mappings [35]. Such methodology, however, imposes strong assumptions on the provenance semirings, which restrict the forms of derivable explanations. We aim to enhance the expressiveness of explanations by incorporating ontological deduction steps, considering diverse data sources (e.g., graph-structured data, temporally annotated data), and also the stages in the DPP and the corresponding stages in the STP. For the latter, we exploit the techniques developed in Item 4 of semantic enrichment, to provide explanations also in terms of the data at any stage of the DPP, including the original data.

For negative query answers, existing explanations are limited. We plan to explore new notions based on alternative minimality criteria and preference structures, also addressing different inconsistency-tolerant semantics. In addition, we properly take into account semantic mappings to STP stages, building on Item 4 of semantic enrichment and on [37].

4. Case Studies and Experimentation

The S-PIC4CHU approach will be validated on two use cases drawn from different domains to ensure generality and effectiveness.

4.1. Health Data Curation

Nowadays, the healthcare organizations are increasingly adopting data-centric architectures, to be more effective in clinical research: this is achievable by going beyond the conventional hospital boundaries, to tap into datasets made available by other hospitals. With a focus on the Italian reality, we aim to leverage semantic-based data management and the S-PIC4CHU Data Preparation Pipeline to support multicentric clinical trials: the Policlinico Universitario Agostino Gemelli, the second-largest hospital in Italy, provides various health data sources, including ambulatories, hospitalizations, and drug data, and also contributes with feedback from their domain experts. Specifically, attention should be paid to the pipeline for preparing the data used by data lake platforms for data sharing among different hospitals according to the FAIR principles (Findable, Accessible, Interoperable, Reusable) [42]. This requires *(i)* data acquisition and curation to mediate the different data representations used by the various hospitals, *(ii)* accurate metadata management for enabling subsequent data discovery and comparison, *(iii)* an access control model satisfying the constraints imposed by personal data protection regulations, *(iv)* novel analysis tools, such as AI-based approaches, pursuing increasingly flexible and scalable solutions. The first two steps are the ones taken into account within S-PIC4CHU. Indeed, health data comes in various forms, requiring specific processing to extract valuable information. Macro categories include biosignals, bioimages, "omic" data (e.g., genomics and proteomics), and textual documents. Differences in structure and content pose several challenges to the design of a platform performing data integration from disparate hospitals or research centers, and to reach this goal data interoperability is a key target. Syntactic interoperability is already a challenge, due to a variety of formats and structures of the data, but Semantic Interoperability is much more important to ensure data trustworthiness and reliable interpretation of the analysis. The various, already existing Medical Ontologies or Vocabularies, such as SNOMED CT[1] or LOINC[2], help unambiguously identify variable meanings, along with harmonization pipelines to minimize variability in measurements collected across different centers. The S-PIC4CHU DPL is a precious support in this work, since the envisaged semantic techniques and the improved the data preparation tools provided by the project assist in reconciling conflicts among different medical data standards.

4.2. Architecture and Sustainable Development

This use case, developed with the IMM Design Lab at PoliMI[3], aims to support urban policy-makers using the Integrated Modification Methodology (IMM) [43], aligning with European Sustainable Development Goals (SDGs)[4]. The goal is to create open-source tools compatible

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

with the S-PIC4CHU reference architecture, serving as a testbed for the techniques discussed in Section 3. The experimental phase evaluates the applicability, robustness, efficiency, and effectiveness of the proposed solutions.

Data collection for urban environments is challenging due to the lack of unified standards, semantic ambiguities, varied data collection scales, and the absence of agreed-upon ontologies and benchmarks. The proposed pipeline addresses these issues, focusing on dataset creation for urban analysis. **Data Acquisition:** Data on urban parameters, including geospatial layouts and environmental conditions are collected from municipalities, open-source platforms, and IMM files. Using platform APIs and custom parsers, data collection is standardized, formats are unified, and georeferencing ensures spatial alignment and compatibility through PostgreSQL, PostGIS, and GIS tools. **Data Curation:** Raw datasets are refined through feature engineering, introducing derived variables and spatial indices using PostGIS functions. Ontology-based transformations address semantic inconsistencies, aligning terms and metrics for coherent analysis. **Data Integration:** This phase harmonizes data by resolving granularity and semantic discrepancies, ensuring spatial and temporal coherence. Techniques include temporal alignment of datasets with different resolutions, spatial integration of demographic and environmental layers, and ontology-based management of multi-scale datasets from district to city levels. **Database Implementation:** Data is stored in a scalable, relational database optimized for geospatial analysis. The database supports advanced queries, such as identifying low-accessibility areas or mapping urban heat islands, with ontology-based enhancements for improved adaptability and usability. **Utilization and Analysis:** The pipeline uses the database for predictive modeling, feature analysis, and visualization. Models analyze relationships between variables, such as traffic congestion's impact on air quality. Feature-importance analysis highlights key drivers of urban trends, offering actionable insights.

In conclusion, the pipeline standardizes data formats, harmonizes semantics, and integrates datasets across scales, ensuring cohesive and contextual relevant outputs. It supports IMM indicator computation and integrates findings with other research domains, leveraging ontology-based methods and geospatial tools for semantic alignment and precision.

5. Conclusions

The S-PIC4CHU project aims to revolutionize the field of data preparation by incorporating semantic techniques throughout the entire process. The project addresses the critical issues of data quality, bias, and explainability, and intends to deliver innovative models and techniques to overcome the limitations of existing tools. By combining research from different domains, the S-PIC4CHU project is set to make a significant contribution to both the scientific community and society as a whole, with the potential to have an economic impact. The project is committed to dissemination, and a range of activities undertaken to ensure the widest possible dissemination of our results, reaching a broad cross-section of computer science researchers and IT practitioners. By promoting awareness of data fairness, the project intends to ensure its impact on society with fairness-aware tools and methods.

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".

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Firmani, L. Tanca, R. Torlone, Ethical dimensions for data quality, ACM J. Data Inf. Qual. 12 (2020) 2:1–2:5. doi:10.1145/3362121.

[2] Z. Shang, J. Yang, M. Yang, C. Yu, J. Han, Democratizing data science through interactive curation of ML pipelines, in: Proc. of the 40th ACM Int. Conf. on Management of Data (SIGMOD), ACM, 2019, pp. 1171–1188.

[3] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2022) 71:1–71:37.

[4] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, Data-Centric Systems and Applications, Springer, 2006. URL: https://doi.org/10.1007/3-540-33173-5. doi:10.1007/3-540-33173-5.

[5] T. J. Green, V. Tannen, The semiring framework for database provenance, in: Proc. of the ACM Symp. on Principles of Database Systems, 2017, pp. 93–99. doi:10.1145/3034786.3056125.

[6] V. C. Storey, R. Lukyanenko, W. Maass, J. Parsons, Explainable AI, Communications of the ACM 65 (2022) 27–29.

[7] A. Chapman, P. Missier, G. Simonelli, R. Torlone, Capturing and querying fine-grained provenance of preprocessing pipelines in data science, Proc. of the VLDB Endowment 14 (2021) 507–520.

[8] Z. Shang, E. Zgraggen, B. Buratti, F. Kossmann, P. Eichmann, Y. Chung, C. Binnig, E. Upfal, T. Kraska, Democratizing data science through interactive curation of ML pipelines, in: Proc. of the ACM Int. Conf. on Management of Data (SIGMOD), 2019, pp. 1171–1188.

[9] P. Hitzler, M. Krötzsch, S. Rudolph, Foundations of Semantic Web Technologies, Chapman & Hall/CRC, 2009.

[10] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation and Applications, 2nd ed., Cambridge University Press, 2007.

[11] J.-M. Herrera, A. Hogan, T. Käfer, BTC-2019: The 2019 Billion Triple Challenge dataset, in: Proc. of the Int. Semantic Web Conf. (ISWC), 2019, pp. 163–180. doi:10.1007/978-3-030-30796-7_11.

[12] W. Lin, C. Tsai, Missing value imputation: A review and analysis of the literature (2006-2017), Artificial Intelligence Review 53 (2020) 1487–1509.
[13] S. Song, A. Zhang, L. Chen, L. Wang, Enriching data imputation with extensive similarity neighbors, Proc. of the VLDB Endowment 8 (2015) 1286–1297.
[14] M. Bienvenu, C. Bourgaux, F. Goasdoue, Querying inconsistent description logic knowledge bases under preferred repair semantics, in: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI), 2014, pp. 996–1002.
[15] S. Staworko, J. Chomicki, J. Marcinkowski, Prioritized repairing and consistent query answering in relational databases, Ann. of Mathematics and Artificial Intelligence 64 (2012) 209–246.
[16] 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.
[17] L. Caruccio, V. Deufemia, G. Polese, Relaxed functional dependencies - a survey of approaches, IEEE Trans. on Knowledge and Data Engineering 28 (2016) 147–165.
[18] M. Mazuran, E. Quintarelli, L. Tanca, S. Ugolini, Semi-automatic support for evolving functional dependencies, in: Proc. of the 19th Int. Conf. on Extending Database Technology (EDBT), 2016, pp. 293–304.
[19] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Penguin Books, 2016.
[20] J. Stoyanovich, B. Howe, H. Jagadish, Responsible data management, Proc. of the VLDB Endowment 13 (2020) 3474–3488.
[21] M. Bedo, P. Ciaccia, D. Martinenghi, D. de Oliveira, A k-Skyband approach for feature selection, in: Proc. of the 12th Int. Conf. on Similarity Search and Applications (SISAP), volume 11807 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 160–168.
[22] P. Ciaccia, D. Martinenghi, Directional queries: Making top-k queries more effective in discovering relevant results, Proc. of ACM Management of Data 2 (2024) 232:1–232:26. doi:10.1145/3698807.
[23] P. Ciaccia, D. Martinenghi, Reconciling skyline and ranking queries, Proc. of the VLDB Endowment 10 (2017) 1454–1465.
[24] P. Ciaccia, D. Martinenghi, Flexible skylines: Dominance for arbitrary sets of monotone functions, ACM Trans. on Database Systems 45 (2020) 1–45.
[25] J. Han, M. Kamber, J. Pei, Data mining: Concepts and techniques, in: Data Mining Trends and Research Frontiers, Elsevier, 2016.
[26] I. Bartolini, M. Patella, Windsurf: The best way to SURF (and SIFT/BRISK/ORB/FREAK, too), Multimedia Systems 24 (2018) 459–476.
[27] I. Bartolini, M. Patella, A general framework for real-time analysis of massive multimedia streams, Multimedia Systems 24 (2018) 391–406.
[28] P. Senellart, L. Jachiet, S. Maniu, Y. Ramusat, ProvSQL: Provenance and probability management in PostgreSQL, Proc. of the VLDB Endowment 11 (2018) 2034–2037.
[29] A. Borgida, D. Calvanese, M. Rodriguez-Muro, Explanation in the DL-Lite family of description logics, in: Proc. of On the Move to Meaningful Internet Systems: Confederated Int. Conf. (OTM), 2008, pp. 1440–1457.
[30] C. Bourgaux, A. Ozaki, Querying attributed DL-Lite ontologies using provenance semirings,

in: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI), 2019, pp. 2719–2726.
[31] D. Calvanese, M. Ortiz, M. Simkus, G. Stefanoni, Reasoning about explanations for negative query answers in DL-Lite, J. of Artificial Intelligence Research 48 (2013) 635–669.
[32] I. Ceylan, T. Lukasiewicz, E. Malizia, A. Vaicenavicius, Explanations for query answers under existential rules, in: Proc. of the Int. Joint Conf. on Artificial Intelligence, 2019, pp. 1639–1646.
[33] I. Ceylan, T. Lukasiewicz, E. Malizia, C. Molinaro, A. Vaicenavicius, Explanations for negative query answers under existential rules, in: Proc. of the Int. Conf. on Principles of Knowledge Representation and Reasoning, 2020, pp. 223–232.
[34] T. Lukasiewicz, E. Malizia, C. Molinaro, Explanations for inconsistency-tolerant query answering under existential rules, in: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI), 2020, pp. 2909–2916.
[35] D. Calvanese, D. Lanti, A. Ozaki, R. Penaloza, G. Xiao, Enriching ontology-based data access with provenance, in: Proc. of the Int. Joint Conf. on Artificial Intelligence, 2019, pp. 1616–1623.
[36] D. Calvanese, A. Gal, D. Lanti, M. Montali, A. Mosca, R. Shraga, Conceptually-grounded mapping patterns for Virtual Knowledge Graphs, Data and Knowledge Engineering 145 (2023) 102157. doi:10.1016/j.datak.2023.102157.
[37] D. Calvanese, T. Kalayci, M. Montali, A. Santoso, W. van der Aalst, Conceptual schema transformation in ontology-based data access, in: Proc. of the Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW), 2018, pp. 50–67.
[38] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence 35 (2012) 221–231.
[39] H. Dreyfus, S. Dreyfus, What artificial experts can and cannot do, AI & Society 6 (1992) 18–26.
[40] A. Gale, A. Marian, Explaining ranking functions, Proc. of the VLDB Endowment 14 (2021) 640–652.
[41] 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.
[42] R. D. Kush, et al., FAIR data sharing: The roles of common data elements and harmonization, J. of Biomedical Informatics 107 (2020) 103421. doi:10.1016/j.jbi.2020.103421.
[43] T. Massimo, Integrated Modification Methodology (IMM): A phasing process for sustainable urban design, WASET World Academy of Science Engineenering and Technology. 77 (2013).