Mapping Research Trends with the CoLiRa Framework: A Computational Review of Semantic Enrichment of Tabular Data

Luis Omar Colombo-Mendoza *, Julieta del Carmen Villalobos-Espinosa, María Elisa Espinosa-Valdés *, Elías Beltrán-Naturi

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1353.v1

Keywords: computational literature review; topic modeling; scientometrics; unsupervised clustering; semantic enrichment; tabular data

 doi:10.20944/preprints202603.1353.v1

Article

Mapping Research Trends with the CoLiRa Framework: A Computational Review of Semantic Enrichment of Tabular Data

Luis Omar Colombo-Mendoza [1,*], Julieta del Carmen Villalobos-Espinosa [1], María Elisa Espinosa-Valdés [2,*] and Elías Beltrán-Naturi [3]

[1] Instituto Tecnológico Superior de Teziutlán, Teziutlán, Puebla, Mexico
[2] Instituto Tecnológico de Minatitlán, Minatitlán, Veracruz, Mexico
[3] Universidad Autónoma de Chiapas, Tuxtla Gutiérrez, Mexico
* Correspondence: luis.cm@teziutlan.tecnm.mx (L.O.C.-M.); maria.ev@teziutlan.tecnm.mx (M.E.E.-V.)

Abstract

This article proposes a novel and replicable computational methodology named CoLiRa (Computational Literature Review & Analysis) Framework to quantitatively analyze and map the evolution of a scientific field. As a multi-stage approach, the CoLiRa Framework first uses Latent Dirichlet Allocation (LDA) to identify core research topics from a body of literature. Second, it applies cluster analysis (K-Means and Multidimensional Scaling) to map the conceptual structure of the field's key terms. Finally, it uses linear regression analysis to quantitatively assess the development trends of these topics over time. We demonstrate our proposal through a semi-systematic literature review on the semantic enrichment of tabular data, which covers studies (up to 2024) that utilize Semantic Web ontologies, Linked Data, and knowledge graphs. The analysis of this case study revealed three core research topics and found no statistically significant evidence of a shift in topic prevalence, indicating a stable research ecosystem. This work thus offers a validated computational approach for conducting literature reviews and mapping research trends.

Keywords: computational literature review; topic modeling; scientometrics; unsupervised clustering; semantic enrichment; tabular data

1. Introduction

Tabular datasets, understood as datasets composed of semi-structured data tables, are typically available on the web in formats such as CSV, JSON, or XML. Semantically enriching this kind of data source is a potentially beneficial task. Its applicability extends beyond simple decentralized data management through integration with existing datasets to specific data mining activities such as feature selection and data cleaning.

In this context, recent years have seen a rise in popularity of competitions aimed at generating knowledge graphs from tabular datasets from the web, such as SemTab (Semantic Web Challenge on Tabular Data to Knowledge Graph Matching). These competitions typically focus on three specific data table-knowledge graph mapping tasks: (1) assigning semantic types to columns, (2) assigning knowledge graph entities to cells, and (3) assigning properties to relationships between columns. These tasks are generally conceptualized as semantic annotation, formally defined as the process of identifying concepts and relationships within documents [1].

This article proposes a replicable computational methodology for analyzing scientific literature named CoLiRa (Computational Literature Review & Analysis) Framework. This methodology is a multi-stage approach based on topic modeling, cluster analysis, and linear regression analysis. Topic modeling is used to identify core themes from the target body of literature, cluster analysis to map the conceptual structure of the field from its key terms, and linear regression for the quantitative

 doi:10.20944/preprints202603.1353.v1

analysis of the research trends of the field over time. For demonstration, this article offers a semi-systematic literature review on semantic enrichment approaches, updating existing reviews such as [2] with publications up to 2024 and focusing on research that utilizes Semantic Web ontologies, Linked Data, and knowledge graphs.

The main motivation of this work is multifactorial. First, manual literature reviews become increasingly impractical as bodies of literature grow exponentially [3]. Second, these traditional literature reviews are susceptible to selection biases [4], emphasizing the need for reproducible methods [5]. Third, literature reviews are primarily qualitative and therefore do not focus on quantitatively assessing research trends [6], revealing the need for approaches to understand the evolution of scientific fields [7].

2. Materials and Methods

2.1. Literature Review

The literature review in this study was conducted using the methodology presented in (Liu et al., 2023), which essentially consists of the activities described below. This methodology was extended with some features of systematic literature reviews as outlined in [8].

2.1.1. Key Terms Definitions

A list of relevant terms within the scope of the literature review was established. Two key terms representing the objectives of the studies under consideration were identified: “semantic enrichment” and “semantic interpretation”. Additionally, two key terms related to the technologies employed for these objectives were specified: “semantic web” and “ontology”. Finally, the term “tabular data” was selected to denote the subject of the studies of interest.

2.1.2. Search String Definition

Once the key terms were established, they were combined using the boolean operators AND and OR to define a search string. In this combination, some search criteria were considered synonyms of others, leading to the use of the logical operator OR. The result of this activity was the following search string:

"tabular data" AND ("semantic enrichment" OR "semantic interpretation") AND ("semantic web" OR "ontology").

2.1.3. Definition of Inclusion/Exclusion Criteria

A set of inclusion and exclusion criteria was established to filter the studies identified through the search process. These criteria excluded studies that: (1) were not published in English as journal articles or conference proceedings in scientific research between 2015 and 2025 (inclusive); (2) did not fully align with the objective of this review; or (3) had a focus differing from that of this review. The criteria pertaining to the objective and focus of the review are detailed in Table 1.

Table 1. Initial revision Inclusion/Exclusion criteria.

Type of Criteria	Revision Objective	Focus of the Review
Inclusion	Semantic Enrichment of Tabular Data	Transformation of Flat Tabular Data into Semantically Enriched Data.
Exclusion	Publication of Linked Data on the Web	• Creation of Domain Ontologies. • Construction of Knowledge Graphs. • Linking Data on the Web.

2.1.4. Search and Inclusion/Exclusion of Initial Studies

The resulting search string was used in the academic search engine Google Scholar to manually retrieve relevant studies for this review, yielding a total of 444 results. These search results were

 doi:10.20944/preprints202603.1353.v1

filtered by selecting only those that corresponded to the 10 most-cited articles that strictly met the previously defined inclusion/exclusion criteria, thus establishing the set of initial studies for the review. For this purpose, each article was manually reviewed, starting with the abstract and introduction and, when necessary, proceeding to the content, results, and conclusions in cases where a decision could not be reached during the initial reading stage.

2.1.5. Expansion of the Initial Set of Studies

The initial set of studies was expanded using the same procedure as in the previous stage of the review. Specifically, the ten most-cited studies among those referenced by each initial study were selected, provided they met the review's inclusion and exclusion criteria. This process yielded a total of 48 studies, comprising 11 journal articles and 37 conference papers. Figure 1 presents the distribution of publication years for all selected studies, categorized as either journal articles or conference papers.

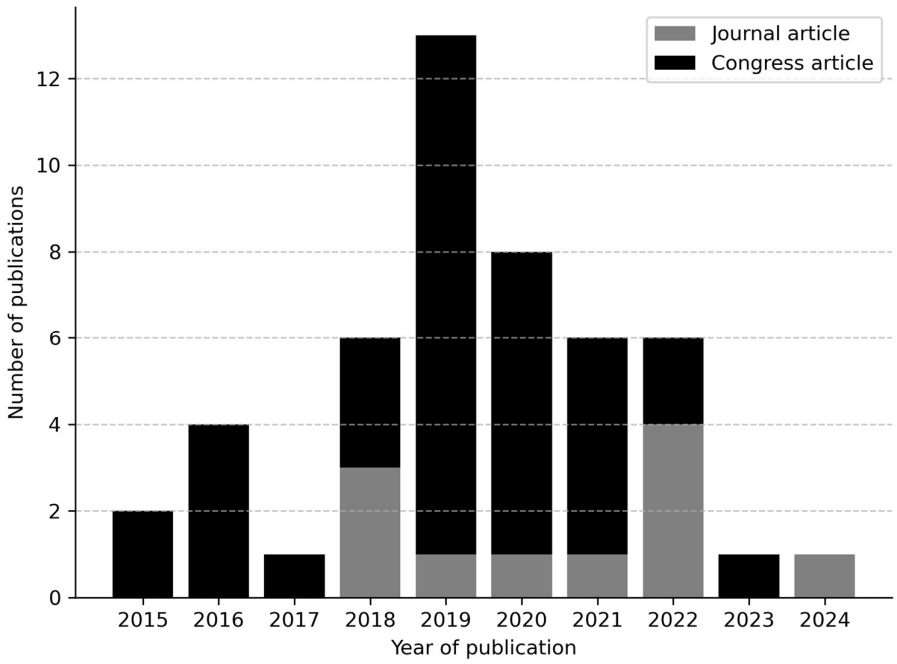

Figure 1. Distribution of publication years and document types for the selected corpus (n=48). The bar chart illustrates the chronological progression of research, categorized by journal articles and conference proceedings, highlighting the increased scholarly activity in the last five years.

Figure 1 demonstrates that the selected corpus (n=48) exhibits a significant concentration of publications from 2018 onwards, with a pronounced peak in 2021 and 2022. This trend indicates that the CoLiRa framework assesses a period characterized by heightened activity and recent consolidation in advancements related to the semantic enrichment of tabular data. Furthermore, the distribution reveals a predominant focus on conference proceedings, a pattern typical of rapidly evolving computer science disciplines, complemented by peer-reviewed journal articles. This combination ensures both the relevance and the quality of the literature underpinning the framework.

 doi:10.20944/preprints202603.1353.v1

2.2. *Topic Modeling*

The initial phase of the computational analysis methodology applied a topic modeling approach using the Latent Dirichlet Allocation (LDA) technique to automatically analyze the selected studies [9]. A Python script was developed for this purpose, primarily employing the Scikit-learn machine-learning library (version 1.2.2), the Matplotlib visualization library (version 3.7.1), and the Natural Language Toolkit (NLTK, version 3.7). The analysis was restricted to the abstracts of the articles, which were manually collected and supplied as input documents for the script.

Specifically, from this corpus of documents, an LDA model was trained, enabling the identification of three abstract topics within the abstracts of the analyzed articles by calculating the probabilities of words belonging to one of the three topics. In this study, the number of topics was experimentally set to three, corroborated by domain experts: different parameters were tested, and the coherence of the resulting topics in each case was analyzed, determined by their 10 most representative words. Additionally, the balance in the classification of abstracts across the resulting topics was taken into account. It is worth noting that the implementation of the LDA technique in the Scikit-learn package is based on the Online Variational Bayes algorithm [10]. For the doc_topic_prior (alpha) and topic_word_prior (betha) parameters the library's defatult values were used; similarly, the random_state parameter was set to 42 for reproducibility.

Once the LDA model was trained, it was possible to classify the abstracts of the articles in the final set of studies for this review, estimating a document-topic distribution.

The literature was limited to articles published up to December 2024 to ensure statistical consistency across complete annual cycles, which is essential for the linear regression analysis of historical development trends. The CoLiRa framework relies on unsupervised machine learning algorithms such as LDA and K-Means to map the conceptual structure. Consequently, the dataset must represent a fixed and consolidated snapshot of the field. Including literature from 2025 onward would alter the mathematical vector space and shift cluster centroids, thereby complicating methodological validation. The primary contribution of this study is the introduction and reproducibility of the computational framework, with the semantic enrichment of tabular data serving as a case study.

2.3. *Cluster Analysis*

Furthermore, to analyze the conceptual structure of the research field, a cluster analysis was performed on the most frequent key terms from the document corpus using the Scikit-learn and Matplotlib libraries. For this purpose, a dissimilarity matrix (1 - cosine similarity) was constructed between the most frequent terms.

Subsequently, Multidimensional Scaling (MDS) was employed to project each key term into a 2D space, setting the random_state parameter to 0 for reproducibility. Then, the K-Means algorithm was applied to group these points into clusters. The number of clusters was experimentally set to five (k=5) and the coherence of each cluster was corroborated by domain experts. For the n_init and max_iter parameters, the values 10 and the library's default were used, respectively. Similarly, the random_state parameter was set to 0 for reproducibility.

2.4. *Analysis of Development Trends*

To identify development trends in the modeled topics, an ordinary least squares (OLS) linear regression model was fitted to each topic. The years under study were used as the independent variable, while the dependent variable was the log-transformed topic prevalence (log(y+1)), which stabilized variance and addressed heteroscedasticity observed in the initial model residuals. Model validity was assessed by examining residual versus fitted value plots for homoscedasticity, Q-Q plots for normality, and the Durbin-Watson statistic for autocorrelation. All analyses were conducted using the statsmodels library in Python.

 doi:10.20944/preprints202603.1353.v1

3. Results

The results of the document-topic distribution estimation are summarized in Table 2, which classifies the 48 final studies across the three identified topics.

Table 2. Classification of the final studies in the revealed topics.

Topic	Number of studies	Belonging studies
Topic 1	12	[11,17,22,23,27,40], [43,45,48,55,56,58]
Topic 2	21	[14–16], [18–21,24,36], [37,38,41,44,46,47], [50–54,57]
Topic 3	15	[12,13,25], [26,28–32], [33–35,39,42,49],

To visualize the temporal development of these themes—a key component of our trend analysis— a stacked bar chart illustrating the annual distribution of topics is presented in Figure 2. This chart plots the number of publications per year, segmented by the corresponding topic.

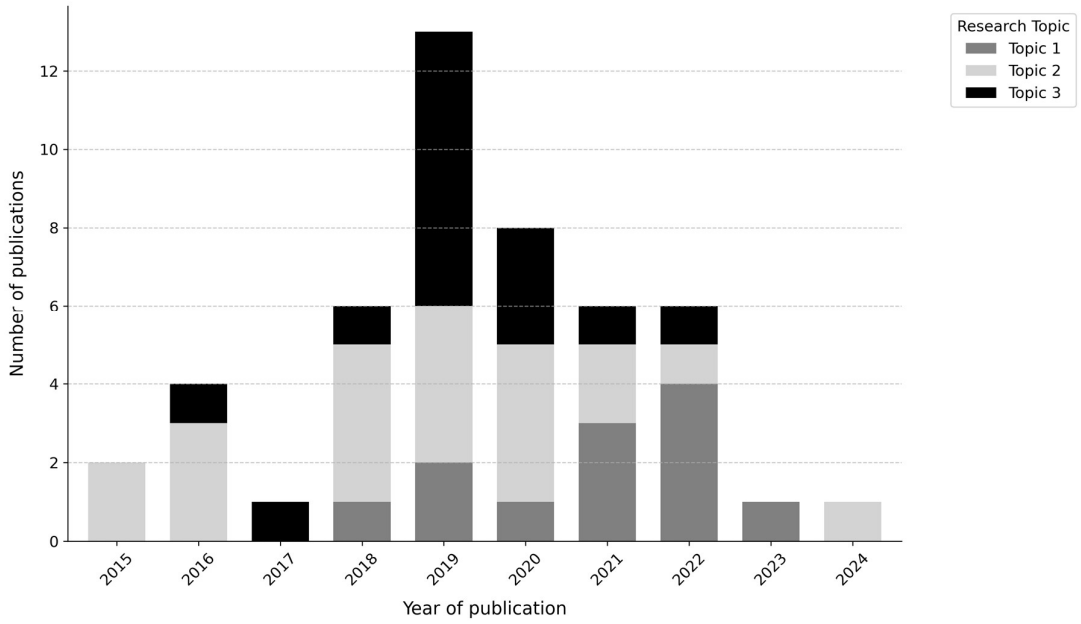

Figure 2. Temporal distribution of research topics. The stacked bar chart shows the absolute number of publications (Y-axis) for each of the three identified research topics (Topic 1, Topic 2, and Topic 3) per year of publication (X-axis).

It can be observed that the total volume of publications changed over time, peaking in 2019 (consistent with the data in Figure 1). However, the relative prevalence of the three topics remains visibly stable throughout the period analyzed. This visualization clearly supports the findings of the OLS regression (Table 4), which found no statistically significant shift in research focus.

To provide a holistic view of the field's conceptual topology and assess the independence of the identified research themes, we generated an Inter-topic Distance Map (see Figure 3).

 doi:10.20944/preprints202603.1353.v1

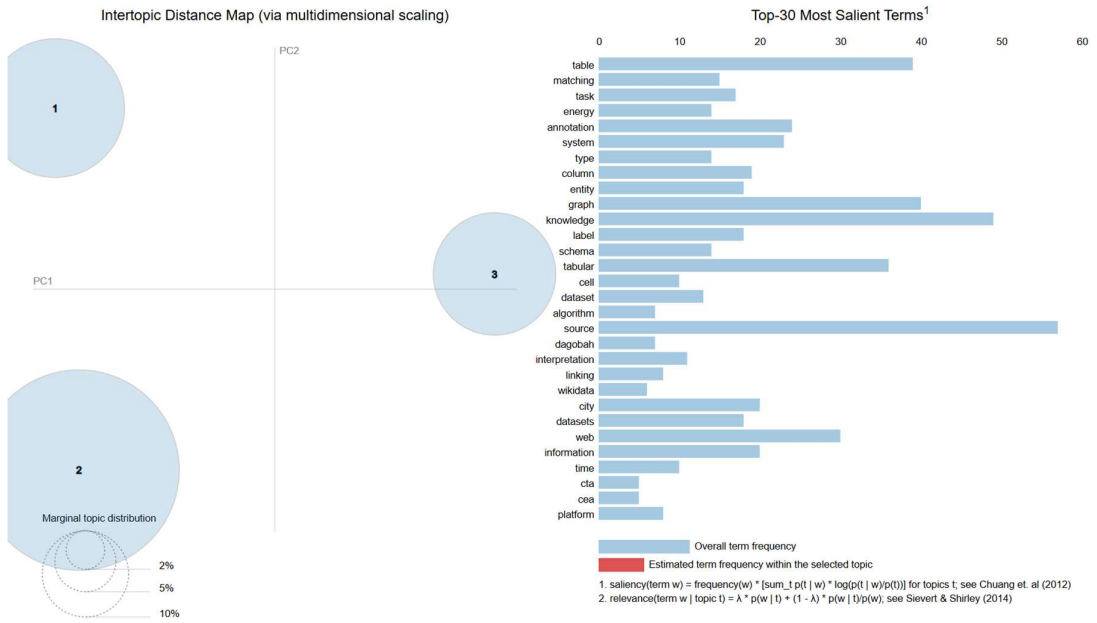

Figure 3. Inter-topic distance map. The left panel visualizes the semantic relationship between the identified topics in a two-dimensional space using MDS. The size of each bubble represents the marginal topic distribution (prevalence) within the corpus. The distance between bubbles indicates semantic dissimilarity. The right panel displays the top 30 most relevant terms for the entire corpus, ranked by their saliency, providing an overview of the field's aggregate vocabulary before topic-specific segmentation.

As illustrated in the Inter-topic Distance Map (Figure 3), the analysis identified three distinct research topics. The spatial separation between the bubbles in the left panel shows that these topics are semantically distinct (with minimal conceptual overlap). Topic 2 represented by the largest bubble dominates the corpus and makes up 54% of the tokens. Topic 1 and Topic 3 show strong independence, accounting for 25.9% and 20.1% of the tokens, respectively.

Additionally, Figure 4 shows the 10 most representative words selected for each topic found by the trained LDA model, specifically, the 10 words with the highest probabilities of belonging to each of the three topics identified.

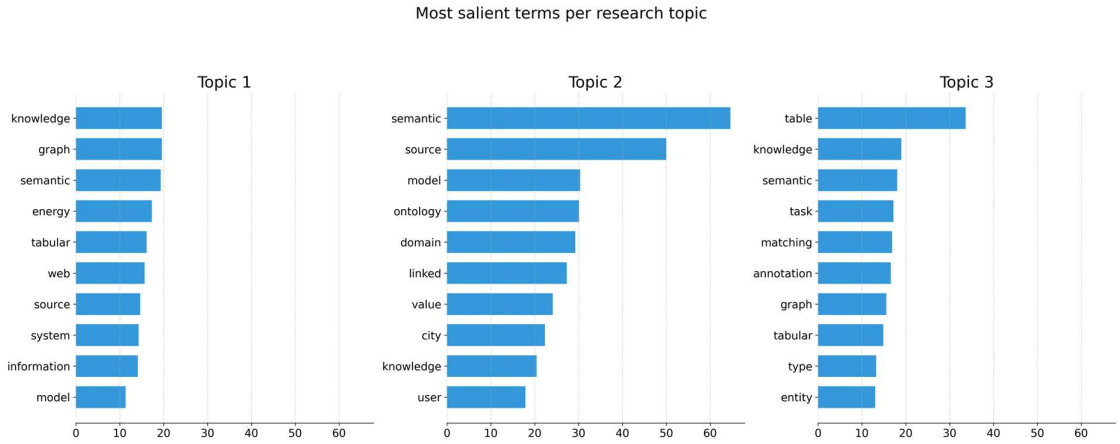

Figure 4. Top-10 most salient terms for each research topic. Horizontal bars represent the importance weight of each term within its respective topic.

 doi:10.20944/preprints202603.1353.v1

It can be observed that “Topic 1” (left) is mainly characterized by terms like “knowledge”, “graph” and “semantic”, “Topic 2” (center) is built upon terms such as “source”, “model” and “ontology”, and “Topic 3” (right) is characterized by terms like “table”, “knowledge” and "task".

Based on the domain experts' interpretation of the key terms presented in Figure 4, the following conclusions were reached:

- The core concepts and systems for information modeling appear to be represented by the first category of studies, labeled as “Topic 1”. Key terms such as “model”, “information”, “system”, “semantic”, “graph” and “knowledge” point to the foundational elements of representing and managing structured data. The inclusion of key terms like “source”, “web” and “tabular” places these concepts within the context of web and tabular data sources. The appearance of the key term “energy” is particularly notable, since it indicates that a significant piece of these studies applies these concepts to a specific domain, namely, the energy sector. In essence, this topic captures the foundational aspects of semantic enrichment.
- The second category of studies, labeled as “Topic 2”, is clearly focused on domain-specific applications and value generation. The inclusion of key terms like “user”, “city” and “value” indicates a strong emphasis on the practical utility and the end-user of these technologies, often in contexts like smart cities. The combination of key terms like “linked”, “ontology” and “domain” suggests that this topic encompasses research that uses Linked Data principles and domain-specific ontologies to create tangible value for users. Thus, this topic highlights the applied aspect of semantic enrichment and shows its direct impact in specific areas.
- The technical processes and tasks of tabular data annotation appear to be represented by the third category of studies, labeled as “Topic 3”. The key terms “entity”, “type”, “annotation”, “matching”, “task”, “table” and “tabular” form a cohesive and highly-specialized body of literature that describes the “how-to” of the semantic enrichment process. In essence, this topic is about the specific and granular operations involved in linking entities and assigning types to data within data tables, often with the goal of mapping them to some knowledge “graph”.

Thus, while Topic 1 lays the conceptual groundwork and Topic 2 explores practical applications, Topic 3 provides a detailed view of the methodological and algorithmic challenges of implementing semantic annotation.

In addition to the topic modeling analysis, a conceptual map was generated to visualize the relationships among individual key terms. While LDA identified broad thematic categories, the cluster analysis, mapped using multidimensional scaling (MDS) in Figure 5, reveals greater granularity within the field's vocabulary. This representation enables the identification of specific “thematic neighborhoods” where semantically related terms cluster, thereby verifying the coherence of terminology across the analyzed literature.

 doi:10.20944/preprints202603.1353.v1

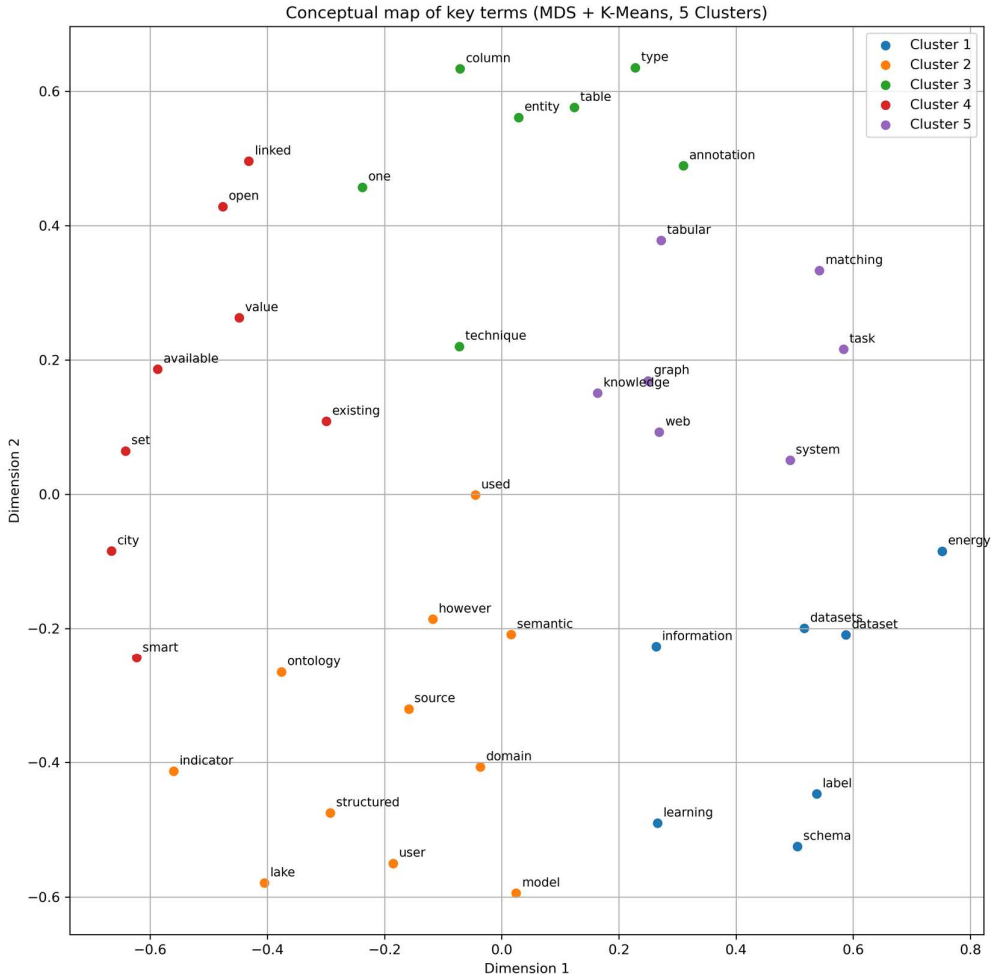

Figure 5. Conceptual map of key cluster terms generated by K-Means and MDS. Each point represents a key term. The spatial distance reflects semantic dissimilarity (1 - cosine similarity), meaning closer terms appear in similar zones. Colors indicate the cluster membership (k=5), depicting the field into distinct conceptual zones.

Based on the domain experts' interpretation of the resulting conceptual map (Fig 5), the following conclusions were reached:

- Cluster 1, which includes the key concepts “linked”, “value”, “city”, “open”, “existing”, “smart”, “used” and “available”, strongly points to the practical application and utilization of Open and Linked Data. The key terms “linked”, “open”, “smart”, and “city” suggest applications in the context of smart cities and linked open data ecosystems. The key terms “value”, “used” and “available” reinforce the idea of data exploitation and utility, connecting this conceptual structure to practical infrastructure, data publication, and its socio-economic impact.
- Cluster 2, which includes the key terms “information”, “datasets”, “label”, “energy”, “schema”, “learning” and “dataset”, appears to be focused on data preparation and machine learning for information and data systems. Key terms like “datasets”, “dataset”, “information” and “label” are central to the preparation of data for analysis. The key term “learning” is a clear indicator of the use of algorithms, while “schema” relates to the structure of the data. The presence of the key term “energy” suggests a notable application domain related to these concepts.
- Cluster 3, which includes the key terms “knowledge”, “graph”, “table”, “web”, “system”, “task”, “matching” and “technique” is a clear indicator of research on Knowledge

 doi:10.20944/preprints202603.1353.v1

Representation and Linking, especially concerning Knowledge Graphs. The key terms "knowledge", "graph", "web", "system" and "table" (as a source for tabular data) are central. The key terms "matching" and "technique" point to the specific methods for constructing or utilizing these graphs, standing this as a foundational pillar of semantic enrichment and the Semantic Web.

- Cluster 4, which includes the key terms "semantic", "source", "model", "ontology", "domain", "lake", "user", "structured", "indicator" and "set", represents the conceptual core of semantic enrichment and data modeling. The key terms "semantic", "model", "ontology", "domain" and "structured" are fundamental concepts. The terms "source" and "lake" likely refer to data sources or data lakes where enrichment is applied, while the key terms "user" and "indicator" suggest the purpose of this enrichment is to generate value for end-users or to create meaningful indicators.
- Cluster 5, which includes the key terms "tabular", "annotation", "column", "entity and "type" is a highly specific cluster focused on the nature of tabular data and its annotation. These key terms precisely refer to the technical process of adding semantic information to tabular data by identifying data types and linking values to known entities. This conceptual structure is very specific to the core techniques of tabular data enrichment.

The integrated analysis of topics and clusters revealed a strong coherence, with clusters both validating and providing a more detailed view of the LDA topics. Topic 1 (concepts and modeling) broadly overlaps with Cluster 4 and Cluster 3, representing the foundational concepts of the field. Topic 2 (applications and value) strongly aligns with Cluster 1, thereby confirming the focus on practical application and Linked Data. Finally, Topic 3 (annotation techniques) corresponds directly to the combination of Cluster 5 (the specific process of annotation) and Cluster 3 (the techniques for constructing knowledge graphs).

The results of the OLS regression analysis on the log-transformed topic prevalence are summarized in Table 3. These models, which were validated for statistical assumptions (see Methods 2.1), were used to assess the development trend of each topic.

Table 3. Classification of the final studies in the revealed topics.

Topic	Trend (slope coefficient, x1)	p-value (unadjusted)	Durbin-Watson
Topic 1	0.049	0.066	1.850
Topic 2	-0.022	0.466	2.158
Topic 3	-0.027	0.301	1.872

† Adjusted significance threshold: $p < 0.0167$ (0.05 / 3 topics).

Furthermore, the Durbin-Watson statistics for all models fell within the acceptable range (1.5-2.5), indicating that temporal autocorrelation was not a confounding factor.

Although Topic 1 (foundational concepts) showed a slight positive slope (0.049) and Topics 2 and 3 showed slight negative slopes (–0.022 and –0.027, respectively), these observed trends are not statistically distinguishable from random fluctuation. This suggests that the research focus across these three core areas has remained statistically stable during the period analyzed.

As shown in Table 3, no topic showed a statistically significant development trend over the last decade. After applying a Bonferroni correction for multiple comparisons (significance threshold $p < 0.0167$), the p-values for all three topics (Topic 1: p=0.198; Topic 2: p=1.000; Topic 3: p=0.903) were well above the threshold.

It is important to note that these trend results are not contradictory to the distribution of publications shown in Fig 1. While Figure 1 illustrates the absolute volume of publications per year, the regression analysis was designed to measure the trend of the relative prevalence of each topic over time. Therefore, the combined findings tell a more complete story: although the overall interest in semantic enrichment has fluctuated, the statistical analysis found no evidence of a significant shift in the research focus. The observed directional trends - a slight rise in foundational modeling (Topic 1) and slight declines in applications (Topic 2) and techniques (Topic 3) - were not statistically

 doi:10.20944/preprints202603.1353.v1

distinguishable from random fluctuation (all p-values > 0.05, see Table 3). This suggests a maturation of the field into a stable state, where foundational principles, practical applications, and technical processes maintain a consistent, balanced focus.

4. Discussion

In this study, we demonstrated the utility of the CoLiRa (Computational Literature Review & Analysis) Framework, a multi-stage computational methodology for analyzing literature named. We achieve a comprehensive and robust analysis by combining topic modeling (LDA) to identify broad research themes, cluster analysis (K-Means) to map the underlying conceptual structure of the field, and linear regression to map temporal trends.

Specifically, this study confirmed that research on semantic enrichment predominantly addresses three areas: Topic 1 (core concepts and systems), Topic 2 (domain-specific applications), and Topic 3 (technical annotation processes). Beyond this finding, the quantitative analysis of temporal trends found no statistically significant evidence of a shift in focus in this scientific field. We observed slight directional trends (a positive slope for Topic 1 and negative slopes for Topics 2 and 3), which were not statistically distinguishable from random fluctuation after applying a Bonferroni correction (all p-values > 0.19).

This deeper understanding, in contrast to the case of one area replacing another, suggests a clear maturation of the field into a stable state. Our interpretation of this statistical stability is that foundational principles, specific techniques, and domain applications coexist in a consistent and stable research ecosystem, which is characteristic of a well-established discipline.

When interpreted from the perspective of previous scientometric studies and systematic literature reviews (SLRs), these results highlight a major improvement in methods. Traditional SLRs offer detailed qualitative insights, but their dependence on manual coding introduces human bias and limits their ability to keep up with the growing body of literature. On the other hand, current computational methods often focus on specific tasks, such as topic modeling without temporal trend evaluation, or clustering without statistical validation. CoLiRa fills this gap by bringing these separate methods together into one clear and cohesive pipeline. By removing subjective human categorization, the framework offers a neutral, scalable, and statistically reliable tool that can be used across various scientific fields to map conceptual trajectories objectively.

For future work, a semi-systematic literature review is planned to explore semantic enrichment approaches for tabular data in the specific context of semantic data mining processes, further testing the adaptability of our computational methodology.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, S1 File. Jupyter Notebook: contains the complete Python code used for the corpus preprocessing, topic modeling, cluster analysis, and regression analysis: the CoLiRa Framework code; S2 File. Requirements file: contains the list of all Python libraries required to execute the code in S1 File and reproduce the analysis; S3 File. Corpus CSV file: contains the titles, years of publication and abstracts of the studies included in the literature review realized as a case study for the CoLiRa Framework. The last material is the input dataset for the Python code in S1 File.

Author Contributions: Conceptualization, C.M.; methodology, C.M.; software, C.M.; validation, V.E.; formal analysis, V.E.; investigation, E.V.; supervision, E.V.; resources, B.N.; data curation, B.N.; writing—original draft preparation, C.M.; writing—review and editing, V.E; visualization, B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All relevant data and code underlying the findings of this study are fully available without restriction. The dataset containing the bibliographic information of the analyzed literature, as well as the Python scripts used for the Latent Dirichlet Allocation (LDA), K-Means clustering, and regression analysis, are available in the GitHub repository at https://github.com/Colombator/CoLiRa.

 doi:10.20944/preprints202603.1353.v1

Acknowledgments: During the preparation of this work, the authors used Gemini Pro in order to improve the readability, grammatical style, and structural flow of the Abstract, Introduction, and Discussion sections. The authors reviewed and edited the content generated by this tool and take full responsibility for the content of the publication. The specific computational methods used for the data analysis (LDA, K-Means, and Linear Regression) were implemented using standard Python libraries (scikit-learn, statsmodels) as detailed in the Methodology section, and did not involve generative AI.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Uren, V.; Buckingham Shum, S.; Bachler, M.; Li, G.; Domingue, J. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *J Web Semant*. **2006**, 4, 14–28.
2. Liu, J.; Chabot, Y.; Troncy, R.; Huynh, V.-P.; Labbé, T.; Monnin, P. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *J Web Semant*. **2023**, 76, 100761.
3. Chen, C.; Song, M. Visualizing a field of research: A methodology of systematic scientometric reviews. *PLoS One* **2019**, 14, e0223994.
4. Gusenbauer, M.; Haddaway, N.R. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Synth Methods* **2020**, 11, 181–217.
5. Nakagawa, S.; Koricheva, J.; Macleod, M.; Viechtbauer, W. Introducing our series: research synthesis and meta-research in biology. *BMC Biol.* **2020**, 18, 20.
6. Börner, K.; Chen, C.; Boyack, K.W. Visualizing knowledge domains. *Annu Rev Inf Sci Technol*. **2003**, 37, 179–255.
7. Fortunato, S.; et al. Science of science. Science 2018, 359, eaao0185.
8. Kitchenham, B. Procedures for Performing Systematic Reviews. Keele University, Keele, UK, Joint Technical Report TR/SE-0401, 2004.
9. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J Mach Learn Res*. **2003**, 3, 993–1022.
10. Hoffman, M.D.; Blei, D.M.; Bach, F. Online learning for Latent Dirichlet Allocation. In Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1; Curran Associates Inc.: Red Hook, NY, USA, 2010; pp. 856–864.
11. Wu, J.; Orlandi, F.; AlSkaif, T.; O'Sullivan, D.; Dev, S. A semantic web approach to uplift decentralized household energy data. *Sustain Energy Grids Netw*. **2022**, 32, 100891.
12. Knap, T. Towards Odalic, a Semantic Table Interpretation Tool in the ADEQUATE Project. 2017, pp. 26–37. Available online: http://ceur-ws.org/Vol-1946/#paper-04
13. Orlandi, F.; et al. Leveraging Knowledge Graphs of Movies and Their Content for Web-Scale Analysis. In Proceedings of the 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2018; pp. 609–616.
14. An, J.; Kumar, S.; Lee, J.; Jeong, S.; Song, J. Synapse : Towards Linked Data for Smart Cities using a Semantic Annotation Framework. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020; pp. 1–6.
15. Wu, J.; Orlandi, F.; Lee, Y.H.; O'Sullivan, D.; Dev, S. Organizing Decentralized Energy Data Using Semantic Approach. In Proceedings of the 2021 Photonics & Electromagnetics Research Symposium (PIERS), 2021; pp. 2213–2216.
16. Rümmele, N. Evaluating Approaches for Supervised Semantic Labeling. 2018, pp. 4–13. Available online: https://ceur-ws.org/Vol-2073/article-04.pdf
17. Chen, Z.; Jia, H.; Heflin, J.; Davison, B.D. Generating Schema Labels through Dataset Content Analysis. In Companion Proceedings of the The Web Conference 2018; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, CHE, 2018; pp. 1515–1522.
18. Ramnandan, S.K.; Mittal, A.; Knoblock, C.A.; Szekely, P. Assigning Semantic Labels to Data Sources. In The Semantic Web. Latest Advances and New Domains; Springer International Publishing: Cham, 2015; pp. 403–417.

 doi:10.20944/preprints202603.1353.v1

19. Knap, T. Increasing Quality of Austrian Open Data by Linking Them to Linked Data Sources: Lessons Learned. In The Semantic Web; Springer International Publishing: Cham, 2016; pp. 243–254.
20. Neumaier, S.; Umbrich, J.; Parreira, J.X.; Polleres, A. Multi-level Semantic Labelling of Numerical Values. In *The Semantic Web – ISWC 2016*; Springer International Publishing: Cham, 2016; pp. 428–445.
21. Azzi, R.; Diallo, G. AMALGAM: A Matching Approach to Fairfy TabuLar Data with KnowledGe GrAph Model. In *Trends and Applications in Information Systems and Technologies*; Springer International Publishing: Cham, 2021; pp. 76–86.
22. Özcan, F.; Lei, C.; Quamar, A.; Efthymiou, V. Semantic enrichment of data for AI applications. In Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–7.
23. Ciavotta, M.; Cutrona, V.; De Paoli, F.; Nikolov, N.; Palmonari, M.; Roman, D. Supporting Semantic Data Enrichment at Scale. In *Technologies and Applications for Big Data Value*; Springer International Publishing: Cham, 2022; pp. 19–39.
24. Bischof, S.; Harth, A.; Kämpgen, B.; Polleres, A.; Schneider, P. Enriching integrated statistical open city data by combining equational knowledge and missing value imputation. *J Web Semant*. **2018**, 48, 22–47.
25. Wu, J.; Orlandi, F.; O'Sullivan, D.; Pisoni, E.; Dev, S. Boosting Climate Analysis With Semantically Uplifted Knowledge Graphs. *IEEE J Sel Top Appl Earth Obs Remote Sens*. **2022**, 15, 4708–4718.
26. Taheriyan, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. Learning the semantics of structured data sources. *J Web Semant*. **2016**, 37-38, 152–169.
27. Nguyen, P.; Kertkeidkachorn, N.; Ichise, R.; Takeda, H. MTab: Matching Tabular Data to Knowledge Graph using Probability Models. 2019, pp. 7–14. Available online: https://ceur-ws.org/Vol-2553/paper2.pdf
28. Oliveira, D. ADOG - Annotating Data with Ontologies and Graphs. 2019, pp. 1–6. Available online: https://ceur-ws.org/Vol-2553/paper1.pdf
29. Thawani, A.; Hu, M.; Hu, E. Entity Linking to Knowledge Graphs to Infer Column Types and Properties. 2019, pp. 22–25.
30. Shigapov, R.; Zumstein, P.; Kamlah, J.; Oberlander, L.; Mechnich, J.; Schumm, I. bbw: Matching CSV to Wikidata via Meta-lookup. 2020, pp. 17–26. Available online: https://ceur-ws.org/Vol-2775/paper2.pdf
31. Cremaschi, M.; Avogadro, R.; Barazzetti, A. MantisTable SE: an Efficient Approach for the Semantic Table Interpretation. 2020, pp. 75–85. Available online: https://ceur-ws.org/Vol-2775/paper8.pdf
32. Vu, B.; Knoblock, C.; Pujara, J. Learning Semantic Models of Data Sources Using Probabilistic Graphical Models. In Proceedings of the The World Wide Web Conference; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1944–1953.
33. Steenwinckel, B.; Vandewiele, G.; Turck, F.D.; Ongenae, F. CSV2KG: Transforming Tabular Data into Semantic Knowledge. 2019, pp. 33–40. Available online: https://ceur-ws.org/Vol-2553/paper5.pdf
34. Morikawa, H. Semantic Table Interpretation using LOD4ALL. 2019, pp. 49–56. Available online: https://ceur-ws.org/Vol-2553/paper7.pdf
35. Alobaid, A.; Kacprzak, E.; Corcho, O. Typology-based semantic labeling of numeric tabular data. *Semantic Web* **2020**, 12, 5–20.
36. Nguyen, P.; Nguyen, K.; Ichise, R.; Takeda, H. EmbNum+: Effective, Efficient, and Robust Semantic Labeling for Numerical Values. *New Gener Comput*. **2019**, 37, 393–427.
37. Pomp, A.; Paulus, A.; Kirmse, A.; Kraus, V.; Meisen, T. Applying Semantics to Reduce the Time to Analytics within Complex Heterogeneous Infrastructures. *Technologies* **2018**, 6.
38. Nikolov, N.; Ciavotta, M.; De Paoli, F. Data wrangling at scale: the experience of EW-shopp. In Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–4.
39. Huynh, V.-P.; et al. DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data. In CEUR Workshop Proceedings; En ligne, Unknown Region, 2021; p. 2. Available online: https://hal.science/hal-04170864
40. Bonfitto, S.; Perlasca, P.; Mesiti, M. Easy-to-use interfaces for supporting the semantic annotation of web tables. 2023, pp. 1–10. Available online: https://ceur-ws.org/Vol-3379/DataPlat_2023_601.pdf

 doi:10.20944/preprints202603.1353.v1

41. Alam, M.; et al. Tab2KG: Semantic table interpretation with lightweight semantic profiles. *Semant web* **2022**, 13, 571–597.
42. Chabot, Y.; Labbé, T.; Liu, J.; Troncy, R. Dagobah: An End-To-End Context-Free Tabular Data Semantic Annotation System. In Proceedings of the SemTab@ISWC, 2019; pp. 41–48. Available online: https://ceur-ws.org/Vol-2553/paper6.pdf
43. Takeoka, K.; Oyamada, M.; Nakadai, S.; Okadome, T. Meimei: an efficient probabilistic approach for semantically annotating tables. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; AAAI Press: Honolulu, Hawaii, USA, 2019; pp. 281–288.
44. Chen, J.; Jiménez-Ruiz, E.; Horrocks, I.; Sutton, C. Learning semantic annotations for tabular data. In Proceedings of the 28th International Joint Conference on Artificial Intelligence; AAAI Press: Macao, China, 2019; pp. 2088–2094.
45. Chen, S.; et al. LinkingPark: An Integrated Approach for Semantic Table Interpretation. 2020, pp. 65–74. Available online: https://ceur-ws.org/Vol-2775/paper7.pdf
46. Huynh, V.-P.; Liu, J.; Chabot, Y.; Labbe, T.; Monnin, P.; Troncy, R. DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. 2020, pp. 27–39. Available online: https://ceur-ws.org/Vol-2775/paper3.pdf
47. Nguyen, P.; Yamada, I.; Kertkeidkachorn, N.; Ichise, R.; Takeda, H. MTab4Wikidata at SemTab 2020: Tabular Data Annotation with Wikidata. 2020, pp. 86–95. Available online: https://ceur-ws.org/Vol-2775/paper9.pdf
48. Bonfitto, S.; Cappelletti, L.; Trovato, F.; Valentini, G.; Mesiti, M. Semi-automatic Column Type Inference for CSV Table Understanding. In *SOFSEM 2021: Theory and Practice of Computer Science*; Springer-Verlag: Berlin, Heidelberg, 2021; pp. 535–549.
49. Gottschalk, S.; Tempelmeier, N.; Kniesel, G.; Iosifidis, V.; Fetahu, B.; Demidova, E. Simple-ML: Towards a Framework for Semantic Data Analytics Workflows. In *Semantic Systems. The Power of AI and Knowledge Graphs*; Springer International Publishing: Cham, 2019; pp. 359–366.
50. Taheriyan, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. Leveraging Linked Data to Discover Semantic Relations Within Data Sources. In *The Semantic Web – ISWC 2016*; Springer-Verlag: Berlin, Heidelberg, 2016; pp. 549–565.
51. Cutrona, V.; Bianchi, F.; Jiménez-Ruiz, E.; Palmonari, M. Tough Tables: Carefully Evaluating Entity Linking for Tabular Data. In *The Semantic Web – ISWC 2020*; Springer-Verlag: Berlin, Heidelberg, 2020; pp. 328–343.
52. Bischof, S.; Martin, C.; Polleres, A.; Schneider, P. Collecting, Integrating, Enriching and Republishing Open City Data as Linked Data. In *The Semantic Web - ISWC 2015*; Springer International Publishing: Cham, 2015; pp. 57–75.
53. Bianchini, D.; De Antonellis, V.; Garda, M. A semantics-enabled approach for personalised Data Lake exploration. *Knowl Inf Syst*. **2024**, 66, 1469–1502.
54. Mami, M.N.; Graux, D.; Scerri, S.; Jabeen, H.; Auer, S.; Lehmann, J. Squerall: Virtual Ontology-Based Access to Heterogeneous and Large Data Sources. *In The Semantic Web – ISWC 2019*; Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., et al., Eds.; Springer International Publishing: Cham, 2019; pp. 229–245.
55. Diamantini, C.; Lo Giudice, P.; Potena, D.; Storti, E.; Ursino, D. An Approach to Extracting Topic-guided Views from the Sources of a Data Lake. *Inf Syst Front*. **2021**, 23, 243–262.
56. Pingos, M.; Andreou, A.S. A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints. In Proceedings of the International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE), 2022.

 doi:10.20944/preprints202603.1353.v1

57. Bagozi, A.; Bianchini, D.; De Antonellis, V.; Garda, M.; Melchiori, M. Personalised Exploration Graphs on Semantic Data Lakes. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*; Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C.A., Meersman, R., Eds.; Springer International Publishing: Cham, 2019; pp. 22–39.
58. Sarramia, D.; Claude, A.; Ogereau, F.; Mezhoud, J.; Mailhot, G. CEBA: A Data Lake for Data Sharing and Environmental Monitoring. *Sensors (Basel)* **2022**, 22, 2733.