electronics

Article

Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling

Georgiana Stănescu (Nicolaie) and Simona-Vasilica Oprea *

Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania; georgiana.nicolaie@csie.ase.ro

* Correspondence: simona.oprea@csie.ase.ro

Abstract: This research aims to investigate the roles of ontology and Semantic Web Technologies (SWT) in modern knowledge representation and data management. By analyzing a dataset of 10,037 academic articles from Web of Science (WoS) published in the last 6 years (2019–2024) across several fields, such as computer science, engineering, and telecommunications, our research identifies important trends in the use of ontologies and semantic frameworks. Through bibliometric and semantic analyses, Natural Language Processing (NLP), and topic modeling using Latent Dirichlet Allocation (LDA) and BERT-clustering approach, we map the evolution of semantic technologies, revealing core research themes such as ontology engineering, knowledge graphs, and linked data. Furthermore, we address existing research gaps, including challenges in the semantic web, dynamic ontology updates, and scalability in Big Data environments. By synthesizing insights from the literature, our research provides an overview of the current state of semantic web research and its prospects. With a 0.75 coherence score and perplexity = 48, the topic modeling analysis identifies three distinct thematic clusters: (1) *Ontology-Driven Knowledge Representation and Intelligent Systems*, which focuses on the use of ontologies for AI integration, machine interpretability, and structured knowledge representation; (2) *Bioinformatics, Gene Expression and Biological Data Analysis*, highlighting the role of ontologies and semantic frameworks in biomedical research, particularly in gene expression, protein interactions and biological network modeling; and (3) *Advanced Bioinformatics, Systems Biology and Ethical-Legal Implications*, addressing the intersection of biological data sciences with ethical, legal and regulatory challenges in emerging technologies. The clusters derived from BERT embeddings and clustering show thematic overlap with the LDA-derived topics but with some notable differences in emphasis and granularity. Our contributions extend beyond theoretical discussions, offering practical implications for enhancing data accessibility, semantic search, and automated knowledge discovery.

Keywords: ontology; semantic web; data; knowledge representation; artificial intelligence; machine learning (ML); data interoperability; knowledge graphs

Academic Editor: Fabio Grandi

Received: 28 February 2025
Revised: 21 March 2025
Accepted: 25 March 2025
Published: 26 March 2025

Citation: Stănescu, G.; Oprea, S.-V. Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling. *Electronics* **2025**, *14*, 1313. https://doi.org/10.3390/electronics14071313

1. Introduction

1.1. General Context

Semantic Web (SW) has become more popular, with the goal of creating a future where machines can easily understand and process information [1]. Ontologies, which are key parts of the SW, are essential for defining the terms and meanings within a specific field. By using ontologies, researchers build structured, machine-readable knowledge representations, making it easier to integrate and connect data from different sources [2]. The SW expands on the concept of ontology. This vision for a semantically enriched web is enabling intelligent, interoperable systems that dynamically interpret data, providing a richer user experience and laying the groundwork for innovations in areas such as AI, bioinformatics, education, and healthcare [3,4]. The SW's use of ontologies and standardized vocabularies transforms the web into a vast, interconnected knowledge network, allowing users and systems alike to make sense of complex data landscapes. This facilitates cross-domain research, enabling scientists and analysts to access data across various sectors, thereby accelerating discovery and the development of novel solutions [5].

Our research aims to highlight the importance of ontology and SW in research on technologies over the last decade. They are increasingly growing as they offer innovative solutions for knowledge representation and data management across various domains. By analyzing a large collection of 10,037 academic articles, we aim to explore the changing approaches to data and the important role of knowledge representation, especially in fields like computer science, engineering, and telecommunications. The constant growth of Big Data [6] has led to a time when organizing, analyzing, and connecting information efficiently is more important than ever. Ontology and SWT, which help structure and add meaning to data, have become essential tools for managing this data complexity [7]. As the datasets are growing worldwide, the organization and analysis of complex information have become essential. The integration of these frameworks enables more efficient, machine-readable structures for data, promoting interoperability and advancing research capabilities [8] across fields such as computer science, engineering, and telecommunications.

1.2. Objective, Motivation and Contribution

The *goal* of our research is to provide insights into emerging research in ontology and SWT. In this paper, we explore the use and influence of specific keywords, such as "ontology" and "semantic web", within academic literature. By analyzing 10,037 articles across diverse research areas, our research provides insights into the discussions around data and knowledge representation. These findings reveal significant trends in the use of ontologies to structure information, in applying the SW for enhanced information analysis, and in how we can explore the potential for future research. In particular, the use of ontologies and semantic frameworks demonstrates an essential point toward more structured, interconnected knowledge systems, driving innovation in both academic research and industry.

Our research is motivated by the realization that traditional data management methods are frequently unable to keep up with the huge volume, speed, and diversity of data available today. Ontology and SWT present a promising solution by offering a structured way to represent knowledge and define relationships between information. This allows machines to interpret and work with data more intelligently, opening new possibilities for smarter and more automated systems.

In comparison with previous works, the novelty of this research is the integrative approach to advancing the application of ontologies and SWT across various fields. The novelty of this research lies in several aspects:

(1) Cross-domain integration of ontology and SW applications. Unlike many previous studies that focus on a single domain (e.g., IoT, healthcare, or business intelligence), our research provides a comprehensive, cross-domain analysis of how SWT and ontologies impact various disciplines, including computer science, engineering,

telecommunications, and mathematical sciences. By analyzing a large dataset of 10,037 articles, it identifies interdisciplinary connections and emerging applications;

(2) Advanced bibliometric and NLP-driven insights. Previous research primarily relied on traditional bibliometric techniques, whereas our research employs an integrated approach using NLP, topic modeling using LDA and BERT-clustering approach, and keyword frequency analysis to uncover deeper semantic relationships between research trends;

(3) Visualization and graph-based representation of research trends. Unlike conventional bibliometric studies, our research employs graph-based visualizations, including knowledge graph associations and word representations, to depict relationships between research areas and keywords;

(4) Novelty in the use of LDA and BERT-clustering approach for topic modeling. While topic modeling has been applied in various studies, this research uniquely integrates LDA coherence score analysis (0.74846) to evaluate the quality of extracted topics. This approach refines topic classification within the SW domain and offers an objective measure of how well themes align with real-world research trends.

Additionally, it provides insights into areas that need further exploration to push the boundaries of ontology and SWT. Our research provides insights into the integration of ontologies with AI and ML techniques, which can further enhance the ability of systems to reason over data. It also examines the role of SWT in improving data interoperability and supporting more effective collaboration across different disciplines.

To structure our investigation and provide a focused analysis, we define the following Research Questions (RQ):

RQ1: In recent years, how have ontologies and SWT evolved across multiple disciplines?
RQ2: What are the main associations in ontology-driven knowledge representation and semantics?
RQ3: What are the latent topics in ontology-based research, as revealed through topic modeling analysis?

This paper is structured into several sections. The first section introduces the SW and ontology concepts, discussing their relevance in structuring data and fostering innovation. The second section reviews prior research on ontology-driven knowledge representation and its applications. The third section details the research process, including data collection, preprocessing, analysis using Python 3.10, and visualization techniques like NLP and topic modeling. The fourth section presents findings from keyword analysis, publication trends, and thematic associations within the dataset, highlighting core themes like knowledge systems, AI, and data management. In the fifth section, several implications for future research are debated. The last section summarizes insights, emphasizing the significance of ontologies and SW frameworks for future research and application.

2. Literature Review

Advances in peer-to-peer systems and the SW exposed a critical issue: the lack of semantic interoperability in the research by A. Rejeb et al. [9]. Businesses required improved supply chain automation, interoperability, and data governance. Despite growing research on the SW, the link between IT and interoperability remained underexplored. A review of 3511 Scopus papers from the past two decades analyzed bibliometric indicators, keyword co-occurrence, and co-citation networks. Findings highlighted the dominance of conference papers, significant contributions from developed nations, and research themes

like IoT, SW services, and ontology mapping. Furthermore, ontology has been a popular topic in information science, with applications in information sharing, system integration, and knowledge-based software development. A bibliometric analysis of ontology literature from 1986–2020, using tools like citation analysis and knowledge graphs, revealed that the SW (linked to NLP) and gene ontology (linked to bioinformatics) were two major research areas, as mentioned by A. Wu and Y. Ye [10]. This study tracked research outcomes in ontology over the past 25 years.

The rapid growth of data made traditional processing applications struggle. The rise of the Internet led to an explosion in accessible information, while the World Wide Web (WWW) promoted the SW to improve how information is searched, reused, and shared. Businesses increasingly incorporated SWT with Big Data for enhanced value, offering benefits like better data management, adaptability to change, and improved connections across sources. New methods combining Big Data and SWT were needed for social network analysis. The evolving business environment required flexible solutions for business intelligence, which could be supported by distributed ontologies in data warehousing. Ahmed, J., and Muqeem Ahmed, Dr [11] explored the integration of Big Data with the SW, its benefits, challenges, and future directions.

Moreover, the SW offered a theoretical framework and technologies applicable beyond the web, demonstrating its consistency in an article written by C. S. Coneglian, et al. [12]. Big Data projects could benefit from these principles, especially in adding semantic features for data contextualization. In this article, an exploratory qualitative methodology was identified. It discussed four key points: (1) using linked data as a source for Big Data (2) applying ontologies in data analysis; (3) promoting interoperability in Big Data with SWT; and (4) utilizing ML to convert data to SW standards. The research concluded that the SW could support Big Data by providing a new approach to data analysis.

Sometimes, complex data systems are not able to process large volumes of data, making it difficult to retrieve relevant information, as described by S. V. Amanoul, et al. [13]. The rise of the internet and organizations like W3C spurred the development of semantic technologies to improve information accessibility and web functionality. These technologies have become important for managing Big Data, offering better data management, adaptability, and integration. The need for flexible information strategies has led to the use of distributed corporate ontologies in data warehousing. This research examined how the SW can enhance Big Data intelligence, exploring the challenges and opportunities of their integration. K. Eldahshan, et al. [14] demonstrate that integrating technologies like cloud computing, semantic technology, Big Data, data visualization, and the Internet of Things led to providing meaning for Big Data information. This framework was tested in a case study predicting air quality and weather, revealing a connection between air pollution and weather conditions. The framework performed well with diverse Big Data. Additionally, SWT has shown promise for managing and sharing research data in Materials Science and Engineering (MSE), but a comprehensive overview of their applications was lacking, as described by A. Valdestilhas, et al. [15]. This work categorized the primary uses of SWT in MSE, highlighting opportunities like improving experiments, enriching data with context in knowledge graphs, and enabling specific queries on semantically structured data. While interdisciplinary collaboration is still developing, there is a need for user-friendly tools and broader adoption by the MSE community to fully realize the potential of SWT. Ultimately, these technologies facilitated data-driven approaches and enhanced knowledge generation.

Through the past years, the SW aimed to structure web data for both human and computer processing, using ontologies and annotations. However, ontologies could be incomplete or contain errors, impacting data quality. M. Barati, et al. [16] addressed these issues by focusing on correcting incorrect instance-class assignments and discovering new classes for ontologies. It proposed ACE (Automated Class Corrector and Enricher), an entropy-based approach, and demonstrated its effectiveness on a SW dataset. Moreover, in the analysis by M. Chang, et al. [17], the pedagogical models were crucial for Intelligent

Tutoring Systems (ITS), defined by tutoring rules. SW ontologies offered a good way to represent these rules, enabling interoperability. However, building these ontologies required significant effort. This research proposed a new approach that used data mining to automatically extract rules from tutoring sessions and represent them in OWL, maintaining the benefits of SW representation while reducing manual effort.

Nevertheless, the SW extended the existing web, providing information with well-defined meanings to enable global cooperation. It played a key role in describing content and services in a machine-readable format, relying on ontologies as its foundation. Ontologies annotated semantics and provided a common base for resources. The SW, along with Web 3.0, created a smarter web, simplifying information retrieval on e-commerce platforms. Traditional recommendation systems, based on static ontologies, struggled with evolving user preferences. R. Alaa, et al. [18] proposed a new recommendation system architecture using ontology evolution, including a semi-automatic ontology-building technique and an ontology reasoning method for personalized product recommendations, aiming to improve consumer purchase decisions. Furthermore, data science combines data inference, algorithm development, and technology to solve complex problems. Students new to the field often lacked knowledge of all its technological and algorithmic aspects, requiring them to seek expert advice. V. Shah and S. Shridevi [19] proposed a data science recommendation system using SW data-science ontology and service-oriented architecture to suggest relevant resources to students based on their queries or exam scores. The system recommended various resources, including books, documentation, software tools, code repositories, and experts. An ontology-based, service-oriented web platform with a conversational NLP interface was proposed and shown to be efficient.

The growth of the IoT led to widespread deployment, with devices generating massive amounts of heterogeneous data, creating interoperability issues. Data modeling and knowledge representation using SWT were proposed to address it. In the article by F. Z. Amara, et al. [20], there are reviewed challenges, prospects, and recent work on semantic interoperability in IoT systems, exploring the use of ontologies and frameworks to solve heterogeneity and interoperability problems. For the past decade, knowledge graphs have become prominent for large-scale data analysis. The SW, using linked data and ontologies, was crucial for data sharing. V. Ryen, et al. [21] reviewed knowledge graph creation from structured and semi-structured data using SWT, focusing on key publications. It covered tools, methods, data sources, and ontologies, also highlighting challenges and lessons learned.

Research landscape analysis relies on ontologies, but while fields like Biology and Physics possessed comprehensive taxonomies, Computer Science lacked them. A. A. Salatino, et al. [22] presented the Computer Science Ontology (CSO), a large-scale, automatically generated ontology with 14,000 topics and 162,000 relationships, created using the Klink-2 algorithm on 16 million articles. CSO offered both comprehensiveness and automatic updates, powering tools at Springer Nature for publication classification and research community detection. The CSO classifier and portal were released for use and feedback, with planned regular updates. Moreover, the fourth industrial revolution impacted industries, but project outcome prediction and resource estimation remained weak due to a lack of unified data definitions. For this issue P. Zangeneh and B. McCabe [23] proposed and evaluated UPonto, a unified ontology for project knowledge representation across megaproject lifecycles. UPonto supported data collection and utilization, providing a data infrastructure for project analytics. It enabled logical deductions, data expansion, cost normalization, and queries using linked data and the SW, defining universal semantics for project risks. UPonto also formed a project knowledge graph foundation and provided semantic definitions for smart IoT agents. In the article by A. L. Antunes, et al. [24] we can see that SW techniques, such as ontologies, were used in Information Systems to address the growing need for data sharing. While unstructured data analysis became a focus, structured data analysis remained crucial. This review analyzed the use of ontologies in Data Warehouse/Business Intelligence (DW/BI) systems, classifying studies by field, SW techniques, and motivations. Ontologies, typically defined using OWL, supported tasks like dimensional modeling and BI application design. Motivations for using ontologies included solving data heterogeneity and improving interoperability. Also, SW and linked data enhanced data accessibility and interoperability, but their use in sustainability assessments was limited, as mentioned by A. Ghose, et al. [25]. This research presented a core ontology for life cycle sustainability assessments, integrating datasets like EXIOBASE and the Yale database into the SW and making them accessible via SPARQL. This work laid the foundation for broader SW use in sustainability assessments.

In the article by M. Barker [26], it is mentioned that ontology-based semantic technologies can be used in geospatial mapping, 3D virtual space networking tools, visual object tracking, deep learning algorithms, and cloud computing technologies. The article highlighted the need for sensor fusion, 3D path planning algorithms, synthetic biometric data, and ontology-based semantic technologies in extended reality environments. A quantitative literature review of Web of Science (WoS), Scopus, and ProQuest was conducted in April 2023, using relevant search terms. From 181 eligible articles, 30 empirical sources were selected. Data visualization tools like Dimensions and VOSviewer were used, and quality assessments were conducted using PRISMA, AMSTAR, Dedoose, Distiller SR, and SRDR. Additionally, F. Gandon [27] conducted a survey of research topics within the interconnected fields of the SW, linked data, and the broader web of data, examining the accumulated contributions from the first two decades. Through the comprehensive compilation of a wide range of bibliographical sources, including journal articles, conference proceedings, and book chapters, and the analysis of various bibliometric indicators such as citation counts, publication frequencies, and author collaborations, the research successfully identified several research trends that shaped the early development of these domains. F. Gandon [27] pointed out these trends and highlighted major publications that were particularly influential during this formative period, providing valuable insights into the evolution of the field. Finally, a thoughtful discussion of the future research challenges that lay ahead for the SW, linked data, and the web of data communities was presented, offering potential avenues for exploration and innovation in the years to come. Moreover, how SW and linked data could transform information management was investigated by interlinking data for machine understanding, thereby improving identification, classification, sharing, and reuse, in an article by M. St-Germain and P. Mongeon [28]. Driven by the potential benefits championed by organizations like the WWW Consortium, information professionals began implementing SW-based initiatives. This multidisciplinary research has applications across numerous fields. A bibliometric analysis of 6438 articles published between 2001 and 2016 (from WoS) was conducted to examine the topic's interdisciplinary evolution and discourse. A citation analysis was also performed. SW research has been extensive, with a recent emphasis on semantic enrichment.

In another research by M. J. Shayegan and M. M. Mohammad [29], bibliometrics is used to assess the field's current analyzing metadata from Scopus. Results showed increased article publication since 2018, with keywords like "ontology", "semantics", "semantic web" and "semantic enrichment" gaining prominence. The United States, Germany, the United Kingdom, France, Italy, and Brazil led in publications. Further analysis and illustrations were provided by M. J. Shayegan [30], to support researchers. Driven by the need for interaction and cooperation in the IoT, semantic interoperability became a key concept, leading to the development of the SW and its associated tools and reasoners for personal information management. To investigate the growing use of semantic techniques in this area, M. J. Shayegan [30] provided a bibliometric analysis of semantic reasoning in the IoT was conducted. It examined 799 articles from the WoS database, analyzing topic categories, influential authors, publication language, publishers, geographical distribution, frequently cited articles, and keyword trends. The analysis revealed that ten countries were responsible for 84% of the publications, with China leading the publication efforts.

Technology convergence was crucial for creating new value and services. Predicting this convergence, however, was challenging due to the dynamic environment. To address this, T. S. Kim and S. Y. Sohn [31] proposed an ML framework, integrating semantic analysis with methods like link prediction and bibliometric analysis to identify convergence patterns. The framework used patent text, employing a document-to-vector method to assess semantic relevance. Application to motor vehicle and signal transmission convergence demonstrated that incorporating text information improved prediction accuracy. Knowledge representation using ontologies and SWT was explored in the article by T. Hagedorn, et al. [32], to enable AI in systems engineering. As digital engineering evolved, AI and ML were increasingly integrated into new methods and tools. While ML techniques supported classification and clustering, they struggled with explaining decision-making. Conversely, multi-domain semantic modeling and rule-based reasoning excelled in this area. Ontologies played a significant role in knowledge representation, supporting domain modeling and reasoning in digital thread domains instantiated in digital system models. They evolved over time as digital twins, co-evolving with their physical counterparts. Semantic technologies and ontologies formalized knowledge to enable reasoning and interoperability across systems engineering domains.

The need for intelligent system software in English teaching and learning, to overcome the limitations of traditional methods, was further addressed in an article by Y. Dong [33]. SWT and AI were combined to develop an English learning and teaching system. Speech data was divided into frames for processing, and the autocorrelation function method in the time domain was used to extract the pitch for each frame corresponding to English sentences. The system's function modules were designed based on the needs of English teaching and learning, and experiments were conducted to evaluate system performance and user satisfaction. The results indicated that the system effectively met the needs of autonomous English learning and teaching. H. H. Guedea Noriega and F. Garcia Sanchez [34] mentioned that for many years, companies used transactional systems to extract useful information from their everyday operations to assist in decision-making. With the increasing volume, variety, and velocity of data, Big Data processing mechanisms were developed to handle such large amounts of information. The main challenges in Big Data management involve collection, storage, search, sharing, analysis, and visualization. SWT, with its sophisticated inference and reasoning techniques, enabled automated data processing. These technologies were applied in various scenarios for data integration, analysis at the knowledge level, and linked data visualization. This research reviewed the literature on the integration of semantic technologies in Big Data analysis, highlighting the benefits and the remaining challenges. Furthermore, Big Data has recently attracted considerable attention across various fields, as mentioned by D. Beneventano and M. Vincini [35]. While much research has focused on the challenges of data volume and velocity, other crucial aspects: variety, velocity, and veracity, are equally important due to the data's heterogeneity and complexity. These challenges necessitate advanced solutions, and semantic technologies offer a potential approach. It explored emerging approaches from academia and industry, highlighting innovative solutions leveraging semantics for Big Data. It discussed the challenges and opportunities of adapting semantic technologies to the Big Data domain, emphasizing the potential benefits of this combination.

T. Georgieva-Trifonova and M. Galabov [36] investigated how semantic technologies, particularly SWT, can aid analysts in selecting, building, and explaining Big Data models. Motivated by the lack of a comprehensive review of semantic technology use in Big Data modeling for analysis, it explored research interest in the topic, identified target Big Data models, applied semantic technologies, and addressed analytics tasks. It also aimed to identify trends and provide future research guidelines. Forty-four scientific papers from prominent digital libraries (2011–early 2021) were reviewed. The findings summarized frequently studied Big Data models, utilized semantic technologies, and research tasks solved through their application. To enhance quality, educational institutions needed regular self-evaluations, assessing both internal and external conditions to inform future work. A challenge was integrating data scattered across various manual, computer-based, and online systems, a process often time-consuming and error-prone. To address this, a SW-based self-evaluation system was developed in five stages: analysis, design, development, implementation, and evaluation. Authors M. Ali and F. M. Falakh [37] focused on the system's analysis and design, demonstrating that SWT has a role in improving its performance. However, the rapid growth of information and communication technology has increased the availability of heterogeneous web information, as mentioned by D. Panneer, et al. [38]. Effective retrieval requires various technologies, including the SW, which is designed for distributed and heterogeneous data. The SW extends the current web with standardized content and relationship representation. Ontologies are important for interoperability within distributed environments and SW applications, capturing domain knowledge generically for deeper understanding. Ontology mapping combines different ontologies. A. Patel and S. Jain [39] also discussed the existing ontology matching tools, and their constraints and provided a brief classification and analysis. The SW and its associated technologies have attracted interest across numerous fields because of their capacity to organize and link web data consistently and coherently. These technologies, such as RDF schema, OWL, and query languages like SPARQL, offer solutions to problems in various domains. This review analyzed the nature and requirements of the SW, considering ten related domains. Its three main contributions are: analyzing the SW and domains driving its growth; discussing domains where SWT is significant; and highlighting domains closely aligned with these technologies.

The reviewed literature highlights significant advancements in the field of ontology and SWT, with a strong emphasis on knowledge representation, data interoperability, and integration with emerging digital infrastructures. Several studies have explored bibliometric trends and research evolution within the SW domain, particularly its role in addressing interoperability challenges in various fields, including IoT, business intelligence, and Big Data analytics. Table 1 summarizes SW studies (covering Big Data, ontologies, interoperability, knowledge representation, and AI), outlining their objectives, methods, data sources, article counts, and findings. It identifies research gaps in real-time processing, dynamic ontologies, and interdisciplinary applications, suggesting future work on scalability, real-world implementation, and hybrid AI-SW frameworks.

Table 1. Comparison of the main aspects of analyzed references.

Ref.	Objective	Methods	Platform	No. of Papers	Main Findings
[9]	Examine the link between the SW and interoperability	Bibliometric analysis, keyword co-occurrence, co-citation networks	Scopus	3511	Conference papers dominate; research focuses on IoT, SW services, and ontology mapping
[10]	Analyze ontology-related research trends	Metrology citation analysis, knowledge graphs, co-citation analysis	Web of Science	2500	Key ontology topics: SW (NLP) and gene ontology (bioinformatics)

[11]	Explore the integration of Big Data and the SW	Literature review, bibliometric indicators	Multiple academic sources	1200	SW improves data management, adaptability, and business intelligence
[12]	Investigate the role of SW in Big Data analysis	Exploratory qualitative methodology, Linked Data integration	Not specified	850	SW supports Big Data via Linked Data, ontologies, and ML-based conversion
[13]	Study the impact of SW on Big Data intelligence	Systematic review, ontology-based knowledge representation	Academic research repositories	950	Distributed corporate ontologies improve data management and integration
[14]	Develop a framework integrating semantic technology, cloud computing, and IoT for air quality prediction	Case study, experimental evaluation	Environmental data platforms	400	Framework effectively handles heterogeneous Big Data and predicts air quality trends
[15]	Evaluate the use of SWT in MSE	Literature compilation, classification	Materials Science datasets	600	SWT enhances experimental processes, enriches data, and improves accessibility
[16]	Improve ontology-based data structuring for the SW	Entropy-based method (ACE), ontology enrichment	SW dataset	1100	ACE improves ontology correctness and enriches class structures
[17]	Enhance Intelligent Tutoring Systems (ITS) using SW ontologies	Data mining, ontology-based modeling, OWL representation	Intelligent Tutoring Systems (ITS) data	500	Automated ontology-based pedagogical rule extraction improves tutoring systems
[18]	Develop a dynamic ontology-based recommendation system for e-commerce	Ontology evolution, semi-automatic ontology building, reasoning techniques	E-commerce platforms	750	Improved product recommendations via ontology evolution and reasoning
[19]	Create an ontology-based recommendation system for Data Science education	Ontology-based service-oriented architecture, NLP	Data Science educational resources	900	System efficiently recommends learning resources based on student needs
[20]	Address semantic interoperability issues in IoT	Literature review, framework analysis, ontology modeling	IoT interoperability platforms	799	Ontologies and SW frameworks improve data integration in IoT systems
[21]	Review knowledge graph creation methods using SWT	Systematic literature review, ontology mapping	SW research venues	1300	Linked Data and ontologies enhance knowledge graph-based data integration
[22]	Develop an evolving ontology for Computer Science research	Automatic ontology generation (Klink-2 algorithm)	16M scientific articles	~14 K topics, 162 K relationships	CSO Classifier & Portal support research trend analysis and community detection
[23]	Develop UPonto, a unified ontology for project knowledge in industrial megaprojects	Ontology-based modeling, SPARQL queries	Linked Data, SW	1250	UPonto improves project analytics and risk management through semantic definitions and linked data

[24]	Explore the use of ontologies in Data Warehouse/Business Intelligence (DW/BI) systems	Systematic literature review, classification	DW/BI industry data	700	Ontologies enhance interoperability, data heterogeneity resolution, and semantic content for BI applications
[25]	Integrate life cycle sustainability assessments with the SW	Ontology design, dataset integration, SPARQL queries	EXIOBASE, Yale Stocks and Flow Database	1100	SW enables enhanced data accessibility and interoperability in sustainability assessments
[26]	Review the application of SW in geospatial mapping and extended reality	Quantitative literature review, data visualization, deep learning	Web of Science, Scopus, ProQuest	181 articles reviewed, 30 empirical sources	Importance of ontology-based semantic technologies for 3D mapping, visualization, and deep learning in extended reality
[27]	Examine research trends in SW, Linked Data, and Web of Data	Bibliometric compilation, research trend identification	Web of Science	1500	Identified key research trends and publications from the first 20 years of SW research
[28]	Analyze the transformation potential of SW and Linked Data	Bibliometric analysis, citation network analysis	Web of Science	6438	SW's multidisciplinary impact on data identification, sharing, and reuse; future research challenges
[29]	Assess the state of research in semantic enrichment on the SW	Bibliometric analysis, Scopus metadata extraction	Scopus	1050	Semantic enrichment in the SW shows increased research and application
[30]	Investigate semantic reasoning in the Internet of Things (IoT)	Bibliometric investigation, keyword analysis	Web of Science	799	Significant research focus on semantic reasoning for IoT interoperability
[31]	Predict technological convergence using ML and semantic analysis	ML, semantic analysis, bibliometric analysis	Patent databases, SW corpora	1250	Framework improves prediction of technological convergence in fields like motor vehicles and signal transmission
[32]	Enable AI in Systems Engineering using ontologies and SWT	Ontology-based modeling, rule-based reasoning	Digital Twin Models, SW tools	1400	Ontologies support knowledge representation, reasoning, and interoperability in digital system models (digital twins)
[33]	Develop an intelligent English learning and teaching system with AI and SW	Speech data processing, autocorrelation function method	English learning datasets, AI corpora	1200	System meets autonomous English learning needs using SW and AI technologies
[34]	Integrate SWT with Big Data analysis	Systematic review, Big Data knowledge-level analysis	Big Data repositories, SW tools	1600	SWT enhance Big Data processing, enabling automated data integration and analysis
[35]	Explore the use of semantic technologies in Big Data to	Industry and academic reviews,	Enterprise and academic Big Data platforms	44	Semantic technologies help address challenges like

	address volume, velocity, and variety	knowledge representation			heterogeneity and complexity in Big Data
[36]	Review the use of semantic technologies in Big Data modeling	Systematic literature review, digital library analysis	Digital libraries (2011–2021)	2200	Identified trends and semantic technologies applied in Big Data modeling for analytics
[37]	Develop a self-evaluation system for educational institutions using SWT	System development, ontology-based design	Educational institutions' self-evaluation systems	1300	SWT improves performance of self-evaluation systems for educational institutions
[38]	Address challenges in retrieving heterogeneous data on the web using SWT	Ontology matching, semantic integration	Web-based ontology repositories	900	Focused on ontology matching tools for effective data retrieval in heterogeneous environments
[39]	Review the role of SWT in various domains	Domain analysis, literature review, SPARQL-based evaluation	SW applications in multiple domains	1000	SWT play a crucial role in linking and organizing data across multiple domains

Our proposal addresses several gaps identified in the previous research:

1. Dynamic ontology updates and scalability. While earlier studies often focus on static ontology structures, our research targets the challenge of updating and scaling ontologies in real-time, particularly within heterogeneous Big Data environments.
2. Interdisciplinary integration and advanced integration with AI and ML. Prior work tends to be domain-specific: focusing on areas like IoT, bioinformatics, or educational systems, without a comprehensive, cross-domain perspective. Our study spans computer science, engineering, telecommunications, and more, aiming to uncover universal trends and interdisciplinary applications.
3. Holistic analysis of research trends. Existing research often relies on traditional bibliometric methods or isolated literature reviews. By analyzing a large dataset of 10,037 academic articles using bibliometric analysis, NLP, and topic modeling, our research aims to map the evolution of semantic technologies comprehensively, revealing core themes such as ontology engineering, knowledge graphs, and linked data.

3. Methodology

The flowchart in Figure 1 illustrates the step-by-step process of analyzing scientific publications from the Web of Science (WoS) database, with a focus on SW and ontology-related research. The workflow begins with data collection that took place in September 2024, specifically by accessing databases like the WoS and formulating a targeted search query (e.g., "semantic web" and "ontology") for publications from the past six years. Only the relevant data regarding publications is identified, downloaded, and imported into a suitable format for analysis.

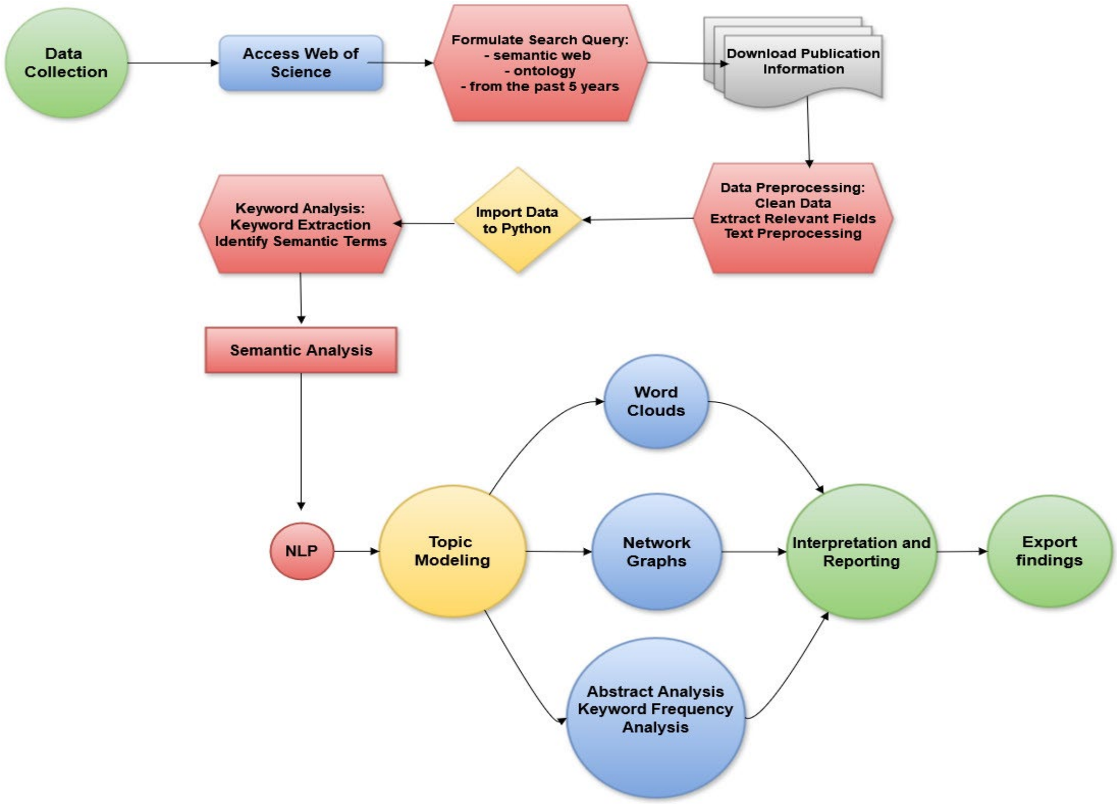


Figure 1. Methodology flowchart.

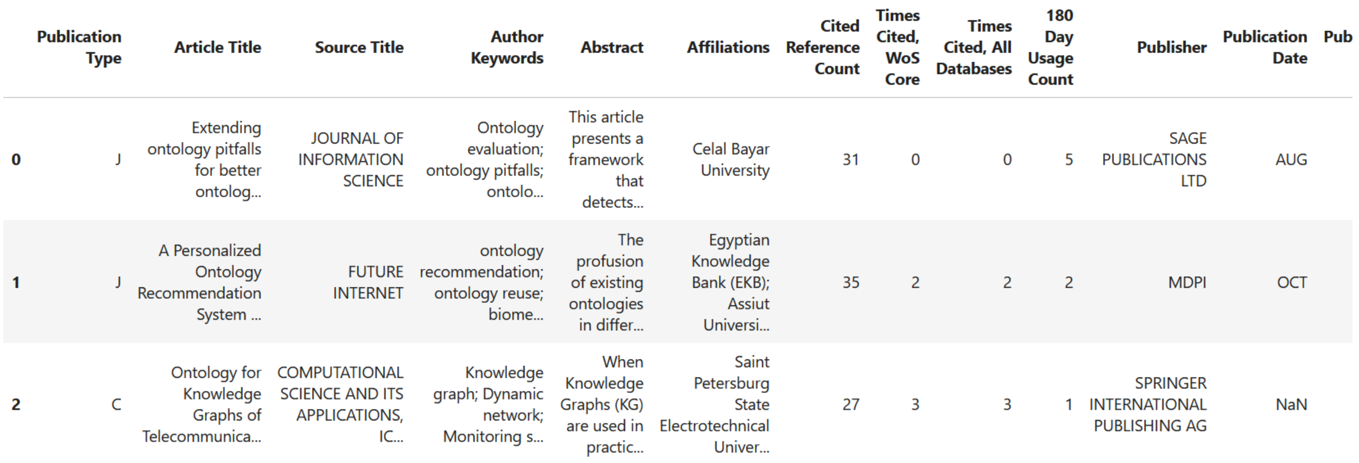
To extract relevant articles from WoS, our data collection process involved formulating a targeted search query using keywords such as “semantic web”, “ontology”, a publication date range between 2019 to September 2024, language of publication English, and included document types like research articles and conference proceedings. The obtained dataset consists of 10,037 academic articles, covering various disciplines including computer science, engineering, and telecommunications.

The retrieved data was exported in CSV files of 1000 full records each and then integrated into one file to maintain compatibility with different data processing tools. We have processed the data and kept specific data fields, such as Publication Type, Article Title, Source Title, Keywords, Affiliations, Abstract, Publication Year, Research Area, Cited Reference Count, and Publication Date.

The data preprocessing stage involves cleaning and preparing the retrieved data, including tasks like extracting relevant fields, cleaning text, and removing stop words. Next, the data is imported into Python 3.10 (Google Colab) for data analysis. This data is then analyzed to identify their frequency and co-occurrence patterns. The semantic analysis stage utilizes advanced techniques like NLP and topic modeling to discover the semantic relationships between keywords and identify relevant themes in publications. NLP techniques result in outputs like word clouds, network graphs, and abstract keyword frequency analysis. Finally, the data visualization, interpretation, and reporting stages involve presenting the findings through various visualizations and drawing conclusions about the trends and insights in the field of SW and ontology research.

Usually, the abstract part of an article starts by introducing the central issue that the study addresses and making use of the important keywords that should be included. By analyzing the most common words in academic abstracts and author keywords, we identify several key themes that dominate the research direction in fields like computer science, management science, and mathematics. Figure 2 presents a sample of publications related to ontologies and semantics. It includes Publication Type, Article Title, Source

Title, Author, Keywords, Abstract, Affiliations, Reference Count, Times Cited, Times Cited All Databases, 180 Day Usage Count, Publisher, and Publication Date. The input data is analyzed to identify trends in publication types and keywords, which can be used in the scope and direction of our research.



	Publication Type	Article Title	Source Title	Author Keywords	Abstract	Affiliations	Cited Reference Count	Times Cited, WoS Core	Times Cited, All Databases	180 Day Usage Count	Publisher	Publication Date	Pub
0	J	Extending ontology pitfalls for better ontolog...	JOURNAL OF INFORMATION SCIENCE	Ontology evaluation; ontology pitfalls; ontolo...	This article presents a framework that detects...	Celal Bayar University	31	0	0	5	SAGE PUBLICATIONS LTD	AUG	
1	J	A Personalized Ontology Recommendation System ...	FUTURE INTERNET	ontology recommendation; ontology reuse; biome...	The profusion of existing ontologies in differ...	Egyptian Knowledge Bank (EKB); Assiut Universi...	35	2	2	2	MDPI	OCT	
2	C	Ontology for Knowledge Graphs of Telecommunica...	COMPUTATIONAL SCIENCE AND ITS APPLICATIONS, IC...	Knowledge graph; Dynamic network; Monitoring s...	When Knowledge Graphs (KG) are used in practic...	Saint Petersburg State Electrotechnical Univer...	27	3	3	1	SPRINGER INTERNATIONAL PUBLISHING AG	NaN	

Figure 2. Sample data.

To extract meaningful insights from the dataset, we use NLP techniques and topic modeling using LDA [40]. Before processing, the text data was cleaned by removing stopwords, abbreviations, punctuation, and special characters, using Python's NLTK library. LDA is a generative probabilistic model used for topic modeling, which involves uncovering hidden thematic structures within a collection of documents (abstracts) [41]. The idea behind LDA is that abstracts are composed of a mixture of topics and each topic is characterized by a distribution over words. LDA operates under the assumption that there are a fixed number of topics in the corpus. Each document is then represented as a probabilistic distribution over these topics and each topic, in turn, is represented as a probabilistic distribution over words.

The model follows a generative process, where it imagines how a document was created: first by choosing a distribution over topics for that document, and then, for each word in the document, selecting a topic according to that distribution and picking a word from the chosen topic's word distribution. Both the topic distribution per document and the word distribution per topic are drawn from Dirichlet distributions, which allow for modeling uncertainty and variation in how topics and words are distributed.

In practice, LDA takes a set of documents and tries to infer the hidden topic structure: the likely set of topics, the composition of each topic in terms of words, and the topic mixture that makes up each document. This inference is done through algorithms like variational Bayes or Gibbs sampling, which iteratively adjust the model parameters to best explain the observed data [42].

The output of LDA includes topic-word distributions that highlight the most representative words for each topic and document-topic distributions that show how much each topic contributes to each document. These results can be used to summarize large text corpora [43], detect trends, or build recommender systems [44,45]. The term "latent" refers to the hidden nature of the topics, "Dirichlet" denotes the type of distribution used to model probabilities, and "allocation" reflects the way LDA assigns words to topics based on these probabilities.

Another approach to topic modeling is by combining BERT embeddings with clustering [46,47]. The model SentenceTransformer ('all-MiniLM-L6-v2') is part of the Sentence-Transformers library and is designed to produce dense vector representations of sentences or documents/abstracts. Once each document is embedded into a vector using the transformer model, clustering algorithms are applied directly to the resulting embeddings. This allows us to group semantically similar texts together based on their positions in the vector space. Each resulting cluster can then be interpreted as a coherent topic. Unlike traditional models like LDA, which rely on word co-occurrence and bag-of-words representations, this method leverages contextualized language representations, making it particularly effective for short or noisy texts.

Clustering can be performed using various algorithms. HDBSCAN automatically determines the number of clusters and even identifies outliers or noise [48]. Once the clusters are formed, topic interpretation is performed. This BERT-based clustering approach often yields topics that are more semantically coherent and human-readable than those generated by LDA.

4. Results

The analysis is done on 10,037 articles from fields like computer science, engineering, mathematics, and telecommunication. The plots in Figure 3 synthesize insights from this keyword analysis and highlight the research topics in data, ontology, and semantic knowledge. The author's keywords have the purpose of pointing out the study's main objective. One can see that the most common words in abstracts are "data", "ontology", "knowledge", "learning", "graph", "information" and "analysis", which highlights the impact of these concepts in the research field. The identified topics may offer a roadmap for researchers to explore new possibilities in data management, ontology, and knowledge systems, shaping the future of data-driven research.

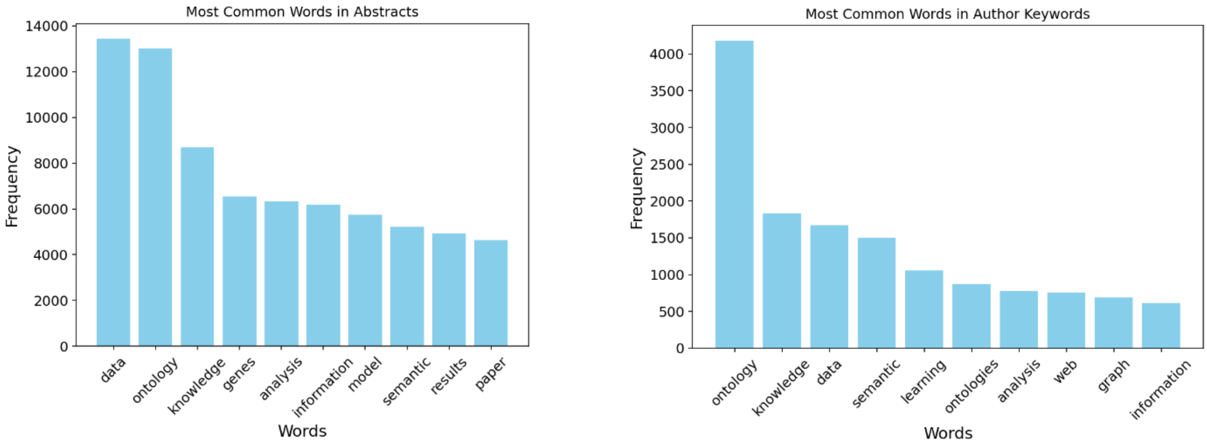


Figure 3. Keywords analysis.

The mentioned keywords form the backbone of modern research and are foundational to numerous applications, from business intelligence to scientific discovery. The frequent use of the term "data" in abstracts (13,435 occurrences) underscores the universal focus on managing and interpreting data efficiently. Researchers are continuously exploring methodologies to analyze large datasets (Big Data), using statistical, ML, and hybrid techniques to extract important insights. This leads to using keywords such as data growth, data storage, complex datasets, and advanced techniques for analyzing the data. The reason why the "data" keyword is at the top of our analysis, is because it acts as the backbone of both ontology and SW initiatives. The volume of data available today has brought both tremendous opportunities and significant challenges. In the absence of structured ontologies and semantic technologies, this data is often locked, inaccessible, and difficult to analyze. However, with the application of ontologies and the principles of the SW, data becomes more accessible, integrative, and valuable.

Another mentioned term, “ontology” is among the most frequent keywords in both abstracts and author keywords, which evidences his relevance in structuring data for complex domains. Ontologies serve as frameworks for organizing information and facilitating semantic interoperability. The third most recurrent term keyword is “knowledge”. This area explores frameworks for converting diverse knowledge sources into integrated databases that support decision-making in complex systems. Also, ontology-based knowledge systems may build models that utilize ontologies to enhance knowledge discovery and sharing, particularly in collaborative environments. The rising frequency of “learning” in research keywords reflects the integration of ML in data analysis, particularly where ontologies enable structured learning from vast datasets. During our research, we could also see that the integration of web technologies with semantic frameworks is an emerging research area, indicated by the frequent use of “web” in keywords.

Figure 4 displays the distribution of publications across various research areas. Computer Science has the highest number of publications, followed by Science & Technology—Other Topics, and Engineering. While analyzing Figure 4, there is a significant distribution of scientific publications across various research fields, with a clear predominance of computer science, which has almost twice the number of publications as the second most popular field. It follows at a distance such fields as the sciences and technologies with other themes, engineering, and computational biology. A gradual decrease in the number of publications is observed as we move towards more specialized fields such as linguistics or automation and control systems. This distribution suggests a strong interest of the scientific community in digital and technological research, as well as a diversity of topics covered, from basic science to practical applications in fields such as biomedicine, economics, and computer science.

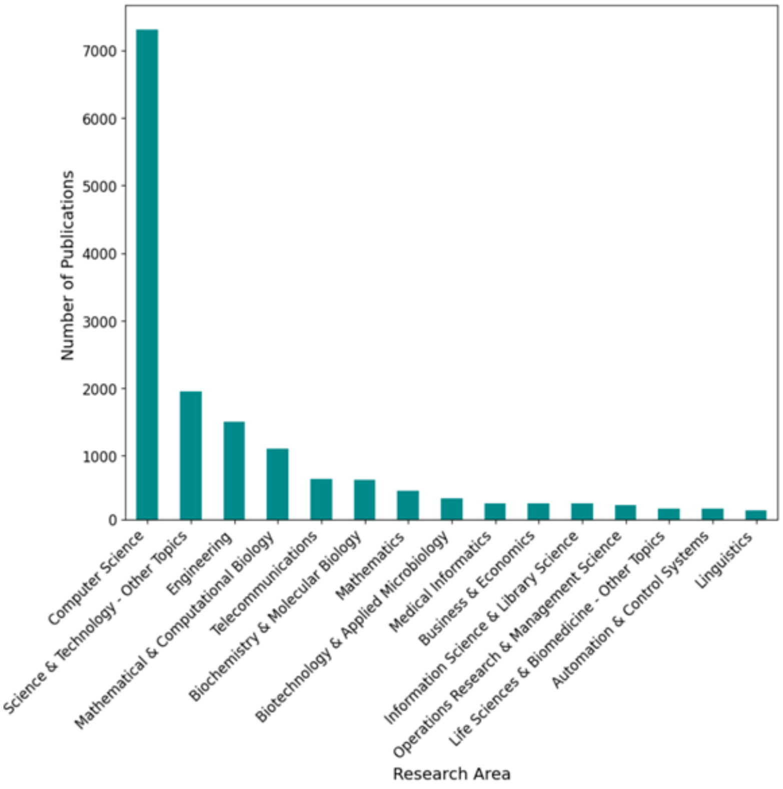


Figure 4. Distribution of publications across research areas.

The opportunities for research for ontologies, SW principles, and data management techniques have increased in the past few years. Figure 5 presents a line graph defining the temporal trend in the number of publications related to five key concepts within the field of semantic technologies: “ontology”, “ontologies”, “semantic web”, “knowledge graph” and “machine learning”, within the publication year, ranging from 2019 to partial 2024. Analyzing the articles from our dataset, we notice that these interconnected elements may address complex, interdisciplinary challenges and promote deeper understanding across specified fields.

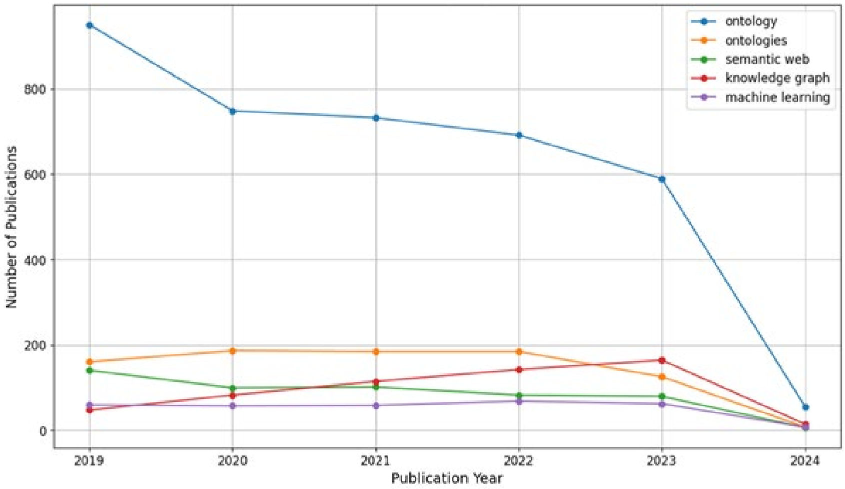


Figure 5. Publication trends for top 5 keywords over time.

Figure 6 highlights relationships between the top research areas and the top keywords, showcasing the most associated topics.

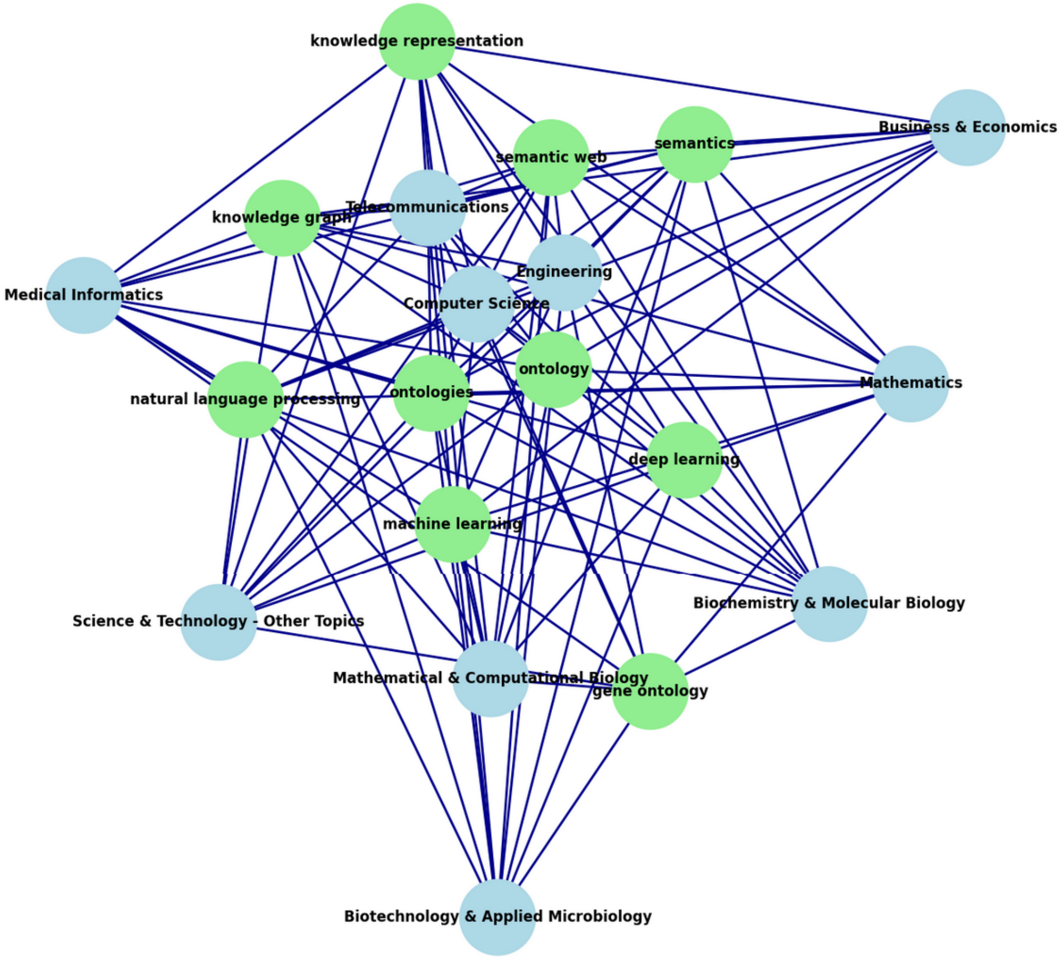


Figure 6. Knowledge graph association for research areas and keywords.

The nodes represent the research areas and keywords, while the edges indicate their association. Computer Science, Engineering, and Mathematics emerge as central hubs, connecting to a wide range of keywords and research areas. Semantic Web and Ontologies play an important role in bridging different domains, as evidenced by their numerous connections. The prominence of keywords like “deep learning” and “machine learning” highlights the growing influence of AI in knowledge representation.

The word cloud generated from the abstracts of the papers in the dataset (in Figure 7) highlights the key concepts and themes frequently approached in research. The words “model”, “data” and “ontology” appear prominently, indicating a focus on knowledge representation and semantic understanding. Other terms include “gene”, “system”, “approach”, “information” and “knowledge”, suggesting a multidisciplinary approach that incorporates research perspectives. The presence of words like “research”, “development”, and “method” also suggests the ongoing research and development efforts within the field.

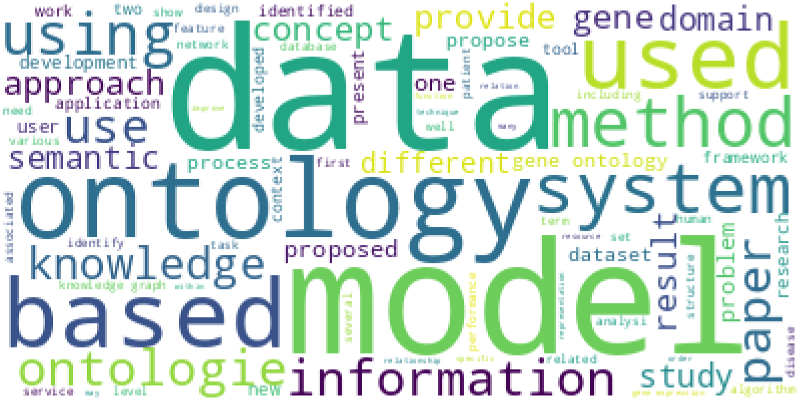


Figure 7. Dataset highlights of keywords.

In the context of evaluating semantic technologies and ontologies, the coherence score of 0.74846, derived from the LDA topic model, provides information about the quality of the topics generated. This score essentially measures how well the top words within each topic relate to each other semantically, indicating how meaningful and interpretable those topics are. The score indicates a moderate to high level of interpretability.

A reason why we used the LDA topic model in this research is that this is a well-known probabilistic generative model for topic modeling. It assumes that each abstract in a collection is a mixture of a set of topics. Each topic is characterized by a distribution of words. LDA aims to discover the underlying thematic structure within a collection of documents by identifying groups of words that frequently occur within the same documents. These groups of words then represent the topics. The obtained coherence score reflects how clearly defined and distinct the analyzed topics are and reveals some semantic alignment.

For topic modeling, we apply LDA using the Gensim library in Python 3.10. The topic range is optimized by testing the number of topic values from 3 to 11 using coherence score analysis, with the final model chosen based on achieving an LDA coherence score of 0.74846. Parameter tuning is applied, setting Alpha = 0.01 to ensure sparsity in topic distribution per document and Beta = 0.1 to refine topic-word distribution. Alpha is the parameter of the Dirichlet prior to the per-document topic distributions. It controls how many topics are likely to appear in a single document. Beta is the parameter of the Dirichlet prior to the per-topic word distribution. It controls how many words are likely to be associated with each topic.

The coherence score is a metric used to evaluate the quality of topics generated by topic modeling algorithms like LDA. It measures how semantically interpretable and meaningful topics are to humans. It combines several steps, including a sliding window, word co-occurrence counts, and a measure of confirmation between word pairs [49]. Typically, it ranges from 0 to 1, where 1 indicates perfect coherence (i.e., all topic words appear together consistently in the corpus) and 0 indicates no semantic similarity among the topic words.

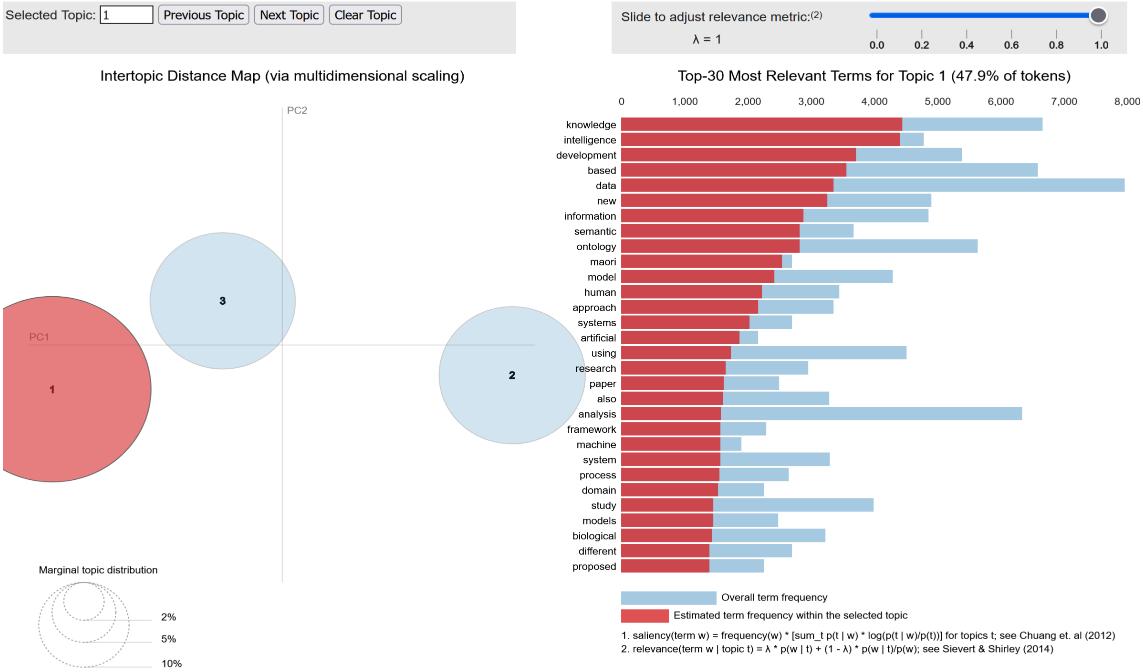
Valuable insights into the key themes and concepts are identified in Figure 8 which provides an overview of the topic modeling results obtained during the analysis of SW and ontology-driven knowledge representation trends, generated using LDA. It consists of two main components:

(1) Intertopic Distance Map (IDM). It visualizes the relationships between different topics and is generated via multidimensional scaling, where three distinct topics are visualized based on their proportional sizes and relationships. Topic 1, represented by the red cluster, is dominant, comprising 47.9% of the token distribution.

(2) Top 30 most relevant terms for Topic 1. It displays the 30 most frequent and relevant terms associated with Topic 1. The length of each bar represents estimated term frequencies both within the selected topic (red) and across the entire corpus (blue). The relevance metric slider (λ) allows adjustment of term importance between saliency and distinctiveness.

Continuing to analyze Figure 8, we discover that the size of Topic 1 in the IDM suggests it is the most frequent topic within the corpus analyzed. This indicates that the research landscape is significantly influenced by this theme, likely around core areas of knowledge representation, AI, and the development of ontologies and knowledge-based systems, as evidenced by the prominence of terms like “knowledge”, “intelligence”, “development”, “ontology”, “information” and “models” in the Top-30 most relevant terms for Topic 1 bar chart.

Furthermore, the IDM visually represents the relationship between Topic 1 and other identified topics (Topics 2 and 3), illustrating the connections and distinctions within the field. The positioning of Topic 1 suggests its centrality, while the distances to Topics 2 and 3 imply varying degrees of thematic divergence. This spatial representation offers a quick grasp of how different research themes relate to each other, highlighting clusters of closely related work as well as more distinct areas of inquiry.



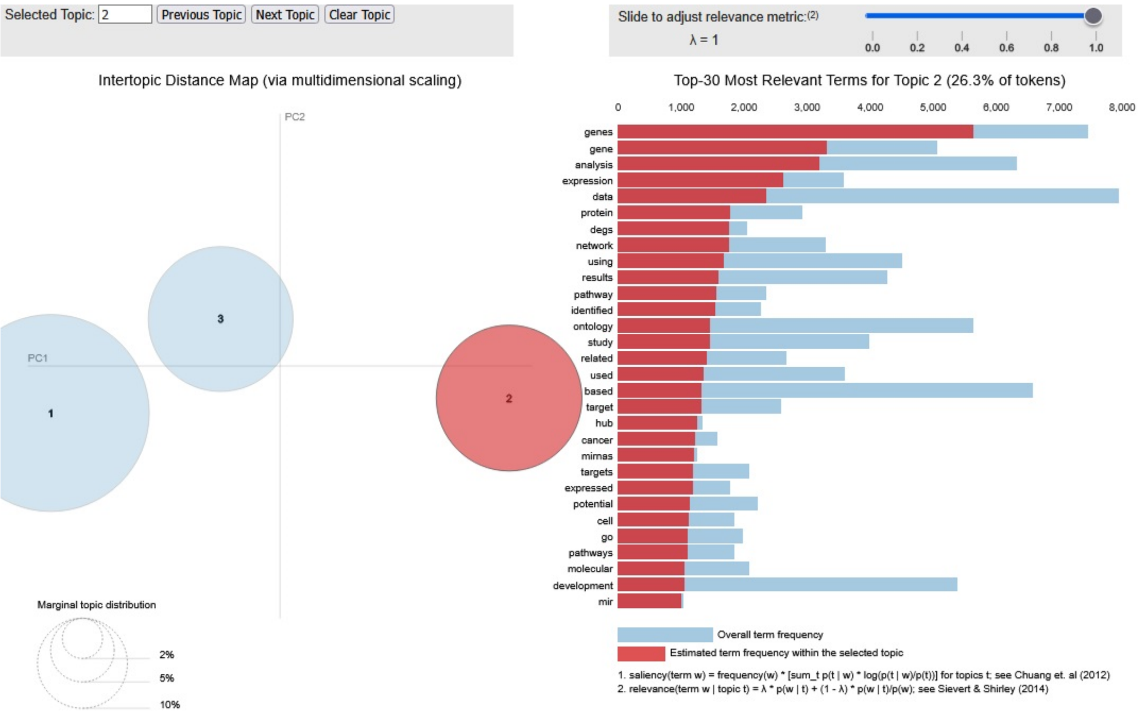
Selected Topic: 1

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. Topic 1 graphical representation.

In Figure 9, we analyze the specifics of Topic 2, which the interface indicates comprises 26.3% of the analyzed tokens. This suggests that Topic 2 is a significant thematic component within the dataset. The Top 30 most relevant terms for Topic 2 bar chart reveal a distinct vocabulary associated with this topic. Terms such as “genes”, “gene”, “analysis”, “expression”, “protein”, “network”, “pathway” and “ontology” highlight a strong focus on genomics, bioinformatics, and molecular biology. The presence of terms like “cancer”, “targets” and “molecular” suggests that biomedical applications are a key part of this topic, related to disease research and targeted therapies. Furthermore, the frequent appearance of “data” and “results” suggests a methodological or computational approach, emphasizing the role of data-driven research in this area.



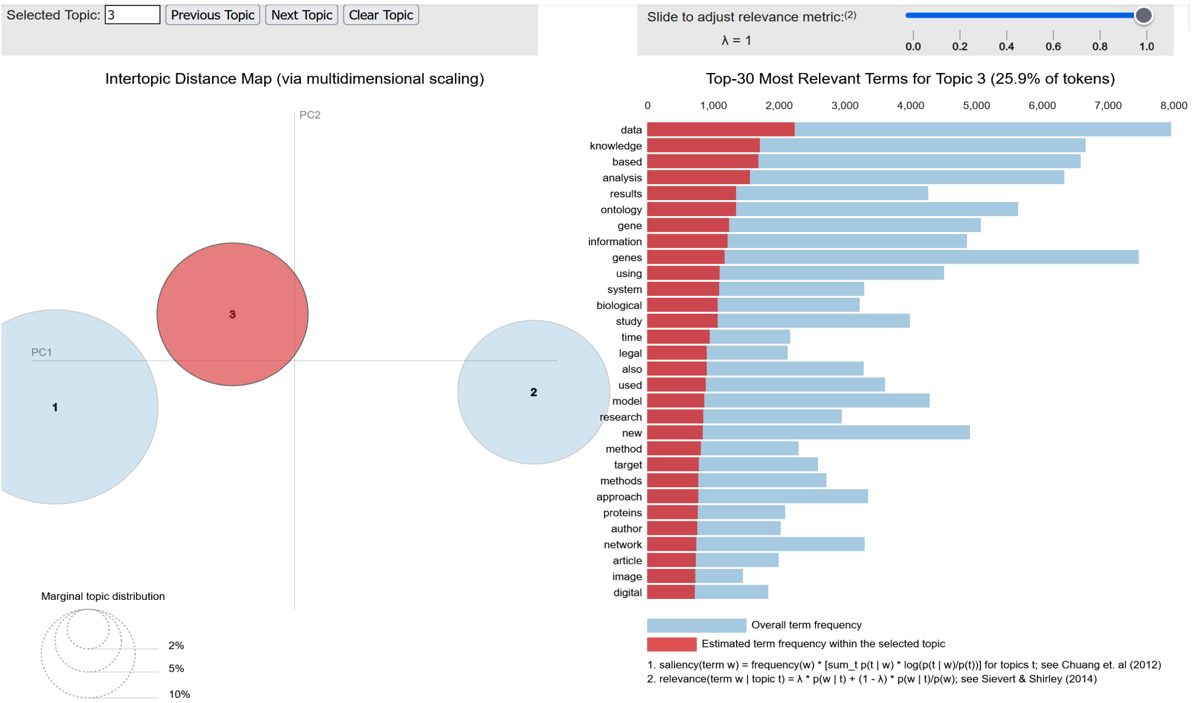
Selected Topic: 2

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9. Topic 2 graphical representation.

Shifting our focus to Figure 10, we now analyze the specifics of Topic 3, which the interface indicates comprises 25.9% of the analyzed tokens. This suggests that while Topic 3 is not the most dominant theme (as Topic 1 was in Figure 8), it still represents a substantial portion of the research discourse. The Top 30 most relevant terms for Topic 3 bar chart reveal a distinct vocabulary associated with this topic. While there is still a presence of core terms, we noticed the emergence of terms strongly indicative of a specific domain focus. The high relevance of “data”, “gene”, “genes”, “biological”, “study”, “model”, “research”, “proteins”, “network”, “image” and “digital” clearly points toward applications in bioinformatics, systems biology or related life sciences fields. The presence of “legal” alongside these terms suggests research exploring the ethical, legal, and social implications of these technologies.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10. Topic 3 graphical representation.

Perplexity is a statistical measure also used to evaluate the quality of probabilistic models such as LDA in topic modeling. Perplexity can be understood as a measure of how “surprised” the model is by the test data. A lower perplexity score indicates that the model is less surprised, meaning it assigns a higher probability to the actual words in the documents, which implies a better generalization performance. Mathematically, perplexity is defined as the exponentiation of the negative normalized log-likelihood of the test corpus. It calculates the average uncertainty per word, taking into account how likely the model thinks each word in the document is, based on the topic distributions it learned during training.

One of the strengths of perplexity is that it provides an objective, quantitative means of comparing different LDA models, particularly when selecting the optimal number of topics. Models with lower perplexity values are generally considered better in terms of statistical fit. However, perplexity has its limitations. While it may reflect the model’s predictive accuracy, it does not necessarily correlate with human interpretability. A model that achieves a very low perplexity might still produce topics that are incoherent or not meaningful to human readers. This is because perplexity evaluates likelihoods without considering whether the words grouped into a topic make sense together semantically. Due to this limitation, researchers often complement perplexity with coherence measures. Unlike perplexity, coherence scores evaluate how semantically related the top words in each topic are, which tends to align more closely with human judgment. Therefore, while perplexity helps assess how well the model fits the data, coherence helps determine how useful or interpretable the topics are.

A perplexity score of 48 in LDA topic modeling indicates that the model is quite confident in its predictions. It is validated by also checking the topic coherence score.

The three clusters obtained with the BERT-clustering approach represent semantically distinct but thematically related topics in the broader area of ontologies and their applications in various domains, especially within semantic technologies, multimedia, biomedical research, and virtual environments. Cluster 0 revolves around the application of ontologies in the music domain and cultural data. The presence of terms like *music*, *musical*, *audio*, *jazz*, *cultural*, and *representation* suggests that this cluster focuses on how ontologies are used to structure, describe, or reason about musical or cultural data. Terms like *semantic*, *knowledge*, *representation,* and *model* indicate that these works are grounded in Semantic Web technologies or knowledge modeling. Additionally, specific terms like *SMIS*, *IO-MUST* (likely names of systems or projects) and *things* (linked to the Internet of Things or linked data) point toward interdisciplinary and data-driven research in music or culture using ontologies. Cluster 1 represents a more general or biomedical domain, specifically bioinformatics and gene-related ontology applications. Words like *genes*, *gene*, *analysis*, *results*, *study* and *different* strongly suggest that this cluster involves using ontologies in the analysis and interpretation of biological or genomic data. This is supported by terms like *ontology*, *knowledge*, *semantics,* and *domain*, which show that formal knowledge structures are central. The cluster is rooted in ontology-driven data analysis and modeling in life sciences, leveraging ontologies like the Gene Ontology (GO) or other domain-specific frameworks for scientific research. Cluster 2 focuses on the use of ontologies in virtual and augmented reality (VR/AR) contexts. The frequent presence of *virtual*, *vr*, *training*, *reality*, *surgical*, *surgery*, *3D,* and *ar* indicates applications in simulation-based learning or medical training environments. Ontology-related terms (*knowledge*, *ontology*, *semantic*, *content*, *domain*) show that these VR/AR environments are semantically enriched, for instance, to create intelligent, context-aware simulations or to structure learning content. This cluster includes research in semantic modeling for immersive environments, particularly for education, healthcare, and training.

5. Implications for Future Research

Our research shows that the development of ontologies and SWT influences studies in different fields and we have identified three main topics: (1) Ontology-Based Knowledge Representation and Intelligent Systems, (2) Bioinformatics, Gene Expression, and Biological Data Analysis, and (3) Advanced Bioinformatics, Systems Biology, and Ethical-Legal Issues. These topics show both the variety and connections in this area of research. Future research could improve by focusing on:

(1) Collaboration between fields—Since computer science, bioinformatics, and systems biology are connected, an interdisciplinary partnership across these areas is becoming more important. Encouraging collaborations across fields would help fill knowledge gaps and create better ontologies.

(2) Addressing research gaps—Even though SWT is being used more, there are still some gaps, like updating ontologies in real-time, making different systems work together, and combining semantic frameworks with AI. Our research shows that flexible ontologies are needed to keep up with changing data.

(3) Standardization—A major challenge in ontology research is the lack of common methods to measure how well models work. Future studies should focus on setting clear benchmarks to test how scalable and effective knowledge representation systems are.

Besides its theoretical value, our study has real-world benefits for both research institutions and industry:

Better knowledge discovery—Using NLP, for instance, helps organize research trends more effectively. Researchers can apply these tools to speed up literature reviews and predict future developments.

Optimizing research—Organizations or academic institutions can use our findings to focus on key research areas.

Improving data interoperability in industry—The adoption of ontology-based frameworks in fields like healthcare, business intelligence, and IoT improves data compatibility and makes information exchange more efficient across different systems.

Future research in this field should focus on how AI and machine learning can further help create ontologies that update and adapt in real-time. Also, since ontology-based systems are used in fields like healthcare and law, future research should explore the ethical and legal dimensions of ontology-driven systems, particularly concerning data privacy and security.

6. Conclusions

This research provides a comprehensive examination of ontology and SWT, highlighting their increasing role in knowledge representation, data interoperability, and intelligent applications. By analyzing 10,037 academic articles across multiple disciplines, including computer science, engineering, telecommunications, and mathematical sciences, this study uncovers significant trends, challenges, and opportunities in the field.

This research stands out by offering a comprehensive, cross-domain analysis of ontology and SWT, moving beyond single-domain studies to explore their impact across multiple disciplines. By integrating NLP, topic modeling, and bibliometric analysis, it uncovers deeper semantic relationships and evolving trends in knowledge representation. Unlike previous studies that primarily focused on isolated applications of SWT, this research takes a holistic, interdisciplinary approach, analyzing 10,037 academic articles across various domains. Prior works have explored the role of ontologies in specific sectors, such as IoT, business intelligence, and healthcare, but have not provided a broad, comparative analysis of their evolution and cross-domain applications.

Based on the LDA topic modeling analysis, a set of descriptive names for the three latent topics is provided:

Ontology-Driven Knowledge Representation and Intelligent Systems. This topic is dominant (47.9% of tokens) and centers on how ontologies support machine interpretability, AI integration, and the development of knowledge-based systems, as indicated by terms like “knowledge”, “intelligence”, “ontology” and “models”.

Bioinformatics, Gene Expression, and Biological Data Analysis. Topic 2 (26.3% of tokens such as genes, analysis, expression, biology, protein, networks, cancer, and pathways) is focused on bioinformatics and biological data analysis, especially gene expression, protein interactions, and related biomedical applications.

Advanced Bioinformatics, Systems Biology, and Ethical-Legal Implications. This theme (25.9% of tokens) is characterized by domain-specific vocabulary such as “gene”, “biological”, “proteins” and “network”, pointing to applications in bioinformatics and life sciences, while also incorporating ethical, legal, and social considerations.

The clusters derived from BERT embeddings and clustering show thematic overlap with the LDA-derived topics but with some notable differences in emphasis and granularity. Cluster 0, which centers on ontologies in music, audio, and cultural domains, aligns only partially with LDA Topic 1 (Ontology-Driven Knowledge Representation and Intelligent Systems). While both highlight structured knowledge representation and semantic technologies, Cluster 0 emphasizes application in cultural and multimedia contexts, which are not prominent in the LDA topics. Cluster 1, focused on genes, semantic data, and biological analysis, closely matches LDA Topic 2 (Bioinformatics, Gene Expression, and Biological Data Analysis). Both highlight the role of ontologies in modeling and analyzing biological data, particularly around gene-related research. Cluster 2, dealing with virtual reality, training, and surgery through semantic modeling, shares some conceptual ground with LDA Topic 1 in terms of intelligent systems and semantic integration but introduces a distinct application area—VR/AR environments—which is not explicitly addressed in the LDA topics. It does not align directly with LDA Topic 3, which focuses more on systems biology and the ethical-legal dimensions of emerging biotechnologies.

Overall, BERT-based clustering captures more application-specific nuances (like music and VR), while LDA topics present broader thematic categories centered around AI integration and bioinformatics.

Our findings confirm that ontology-driven knowledge representation has evolved into a critical component for structuring and linking heterogeneous data sources, enabling intelligent automation and enhancing decision-making in both academic and industrial settings. The research underscores the growing integration of ontologies with AI and ML, demonstrating their potential in improving semantic search, recommendation systems, and predictive analytics.

This research highlights the transformative impact of ontology and SWT on data management, intelligent systems, and AI-driven applications. By structuring knowledge and interoperability, these tools address the growing demands of data-driven research. The findings highlight a strong focus on integrating semantic frameworks with AI, which has broad implications for advancing interdisciplinary collaboration, innovation, and efficiency. Future work could further investigate applications in underexplored fields, emphasizing the scalability and adaptability of these frameworks to evolving datasets and technologies.

Author Contributions: S.-V.O.: Conceptualization, Validation, Formal analysis, Investigation, Writing—Original Draft, Writing—Review and Editing, Visualization, Project administration. G.S.: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data will be made available upon reasonable request.

Acknowledgments: This work was supported by a grant from the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV.

Conflicts of Interest: The authors have no relevant financial or non-financial interests to disclose.

References

1. Bibi, N.; Rana, T.A.; Maqbool, A.; Mirza, A.; Iqbal, Z.; Khan, M.A.; Alhaisoni, M.; Tariq, U.; Damaševičius, R. Web semantics and ontologies-based framework for software component selection from online repositories. *Int. J. Web Grid Serv.* **2023**, *19*, 318–349. https://doi.org/10.1504/ijwgs.2023.133503.

2. Balaji, V.; Acharjee, P.B.; Elangovan, M.; Kalnoor, G.; Rastogi, R.; Patidar, V. Developing a semantic framework for categorizing IoT agriculture sensor data: A machine learning and web semantics approach. *Sci. Temper* **2023**, *14*, 1332–1338. https://doi.org/10.58414/scientifictemper.2023.14.4.40.
3. Khaled, A.; Ouchani, S.; Chohra, C. Recommendations-based on semantic analysis of social networks in learning environments. *Comput. Hum. Behav.* **2018**, *101*, 435–449. https://doi.org/10.1016/j.chb.2018.08.051.
4. Jain, S.; Jain, V.; Balas, V.E. *Web Semantics: Cutting Edge and Future Directions in Healthcare*; Academic Press: Cambridge, MA, USA, 2021. https://doi.org/10.1016/B978-0-12-822468-7.00020-1.
5. Wu, J.; Wang, S.; Shen, S.; Peng, Y.-H.; Nichols, J.; Bigham, J.P. WebUI: A Dataset for Enhancing Visual UI Understanding with Web Semantics. In Proceedings of the CHI '23: CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–14.
6. Barbu, D.C. Big Data Processing Solutions. *Rom. J. Inf. Technol. Autom. Control* **2023**, *29*, 35–48. https://doi.org/10.33436/v29i2y201903.
7. Zhou, M.; Peng, L. English Article Style Recognition and Matching by Using Web Semantics. *Int. J. Mob. Comput. Multimedia Commun.* **2022**, *13*, 1–13. https://doi.org/10.4018/ijmcmc.293751.
8. Singh, A.; Dey, N.; Ashour, A.S.; Santhi, V. *Web Semantics for Textual and Visual Information Retrieval*; IGI Global: Hershey, PA, USA, 2017; ISBN: 10.4018/978-1-5225-2483-0.
9. Rejeb, A.; Keogh, J.G.; Martindale, W.; Dooley, D.; Smart, E.; Simske, S.; Wamba, S.F.; Breslin, J.G.; Bandara, K.Y.; Thakur, S.; et al. Charting Past, Present, and Future Research in the Semantic Web and Interoperability. *Futur. Internet* **2022**, *14*, 161. https://doi.org/10.3390/fi14060161.
10. Wu, A.; Ye, Y. Bibliometric Analysis on Bibliometric-based Ontology Research. *Sci. Technol. Libr.* **2021**, *40*, 435–453. https://doi.org/10.1080/0194262x.2021.1920555.
11. Ahmed, J.; Ahmed, M. Big data and semantic web, challenges and opportunities a survey. *Int. J. Eng. Technol.* **2018**, *7*, 631. https://doi.org/10.14419/ijet.v7i4.5.21174.
12. Coneglian, C.S.; Dieger, R.; Segundo, J.E.S.; Capretz, M. The role of semantic web in the big data process. *Encontros Bibli-Rev. Eletronica Bibliotecon. Cienc. Inf.* **2018**, *23*, 53.
13. Amanoul, S.V.; Abdulrahman, L.M.; Abdullah, R.M.; Qashi, R. Orchestrating Distributed Computing and Web Technology with Semantic Web and Big Data. *J. Smart Internet Things* **2023**, *2023*, 174–192. https://doi.org/10.2478/jsiot-2023-0019.
14. ElDahshan, K.; Elsayed, E.K.; Mancy, H. Semantic Smart World Framework. *Appl. Comput. Intell. Soft. Comput.* **2020**, *2020*, 081578. https://doi.org/10.1155/2020/8081578.
15. Valdestilhas, A.; Bayerlein, B.; Torres, B.M.; Zia, G.A.J.; Muth, T. The Intersection Between Semantic Web and Materials Science. *Adv. Intell. Syst.* **2023**, *5*, 2300051. https://doi.org/10.1002/aisy.202300051.
16. Barati, M.; Bai, Q.; Liu, Q. Automated Class Correction and Enrichment in the Semantic Web. *J. Web Semant.* **2019**, *59*, 100533. https://doi.org/10.1016/j.websem.2019.100533.
17. Chang, M.; D'Aniello, G.; Gaeta, M.; Orciuoli, F.; Sampson, D.; Simonelli, C. Building Ontology-Driven Tutoring Models for Intelligent Tutoring Systems Using Data Mining. *IEEE Access* **2020**, *8*, 48151–48162. https://doi.org/10.1109/access.2020.2979281.
18. Alaa, R.; Gawish, M.; Fernández-Veiga, M. Improving Recommendations for Online Retail Markets Based on Ontology Evolution. *Electronics* **2021**, *10*, 1650. https://doi.org/10.3390/electronics10141650.
19. Shah, V.; Shidrevi, S. Data Science Recommendation System using Semantic Technology. *Int. J. Eng. Adv. Technol.* **2019**, *9*, 2592–2599. https://doi.org/10.35940/ijeat.a9375.109119.
20. Amara, F.Z.; Hemam, M.; Djezzar, M.; Maimor, M. Semantic Web and Internet of Things: Challenges, Applications and Perspectives. *J. ICT Stand.* **2022**, *10*, 261–292. https://doi.org/10.13052/jicts2245-800x.1029.
21. Ryen, V.; Soylu, A.; Roman, D. Building Semantic Knowledge Graphs from (Semi-)Structured Data: A Review. *Futur. Internet* **2022**, *14*, 129. https://doi.org/10.3390/fi14050129.
22. Salatino, A.A.; Thanapalasingam, T.; Mannocci, A.; Birukou, A.; Osborne, F.; Motta, E. The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. *Data Intell.* **2020**, *2*, 379–416. https://doi.org/10.1162/dint_a_00055.
23. Zangeneh, P.; McCabe, B. Ontology-based knowledge representation for industrial megaprojects analytics using linked data and the semantic web. *Adv. Eng. Inform.* **2020**, *46*, 101164. https://doi.org/10.1016/j.aei.2020.101164.
24. Antunes, A.L.; Cardoso, E.; Barateiro, J. Incorporation of Ontologies in Data Warehouse/Business Intelligence Systems—A Systematic Literature Review. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100131. https://doi.org/10.1016/j.jjimei.2022.100131.

25. Ghose, A.; Lissandrini, M.; Hansen, E.R.; Weidema, B.P. A core ontology for modeling life cycle sustainability assessment on the Semantic Web. *J. Ind. Ecol.* **2021**, *26*, 731–747. https://doi.org/10.1111/jiec.13220.
26. Barker, M. Context Awareness and Deep Learning Algorithms, Immersive Visualization and Autonomous Cognitive Systems, and Natural Language Processing and Digital Twin Modeling Tools in the Blockchain-based Metaverse. *Rev. Contemp. Philos.* **2023**, *22*, 102–118. https://doi.org/10.22381/rcp2220236.
27. Gandon, F. A survey of the first 20 years of research on semantic web and linked data. *Ingnierie Systmes D'Inf.* **2018**, *23*, 11–38. https://doi.org/10.3166/ISI.23.3-4.11-56.
28. St-Germain, M.; Mongeon, P. The contribution of information science in the Semantic Web research landscape. *Proc. Assoc. Inf. Sci. Technol.* **2018**, *55*, 470–477. https://doi.org/10.1002/pra2.2018.14505501051.
29. Shayegan, M.J.; Mohammad, M.M. Bibliometric of Semantic Enrichment. In Proceedings of the 7th International Conference on Web Research (ICWR), Tehran, Iran, 19–21 May 2021; pp. 202–205.
30. Shayegan, M.J. A bibliometric investigation into the literature of semantic reasoning in internet of things. *Internet Technol. Lett.* **2022**, *6*, e401. https://doi.org/10.1002/itl2.401.
31. Kim, T.S.; Sohn, S.Y. Machine-learning-based deep semantic analysis approach for forecasting new technology convergence. *Technol. Forecast. Soc. Change* **2020**, *157*, 120095. https://doi.org/10.1016/j.techfore.2020.120095.
32. Hagedorn, T.; Bone, M.; Kruse, B.; Grosse, I.; Blackburn, M. Knowledge Representation with Ontologies and Semantic Web Technologies to Promote Augmented and Artificial Intelligence in Systems Engineering. *Insight* **2020**, *23*, 15–20. https://doi.org/10.1002/inst.12279.
33. Dong, Y. Application of Artificial Intelligence Software based on Semantic Web Technology in English Learning and Teaching. *J. Internet Technol.* **2022**, *23*, 145–154. https://doi.org/10.53106/160792642022012301015.
34. Guedea-Noriega, H.H.; García-Sánchez, F. Semantic (Big) Data Analysis: An Extensive Literature Review. *IEEE Lat. Am. Trans.* **2019**, *17*, 796–806. https://doi.org/10.1109/tla.2019.8891948.
35. Beneventano, D.; Vincini, M. Foreword to the Special Issue: "Semantics for Big Data Integration". *Information* **2019**, *10*, 68. https://doi.org/10.3390/info10020068.
36. Georgieva-Trifonova, T.; Galabov, M. Semantic Web Technologies for Big Data Modeling from Analytics Perspective: A Systematic Literature Review. *Balt. J. Mod. Comput.* **2021**, *9*, 377–402. https://doi.org/10.22364/bjmc.2021.9.4.01.
37. Ali, M.; Falakh, F.M. Design of Vocational Education Self-Evaluation System Based-on Semantic Web Ontology. *J. Phys. Conf. Ser.* **2021**, *1737*, 012024. https://doi.org/10.1088/1742-6596/1737/1/012024.
38. Panneer, D.; Ragunathan, K.; Ramalingam, M.; Narayanan, L.K. Comparative study on ontology matching tools and methods. In Proceedings of the 7th IUPAP International Conference on Women in Physics, Melbourne, VIC, Australia, 11–16 July 2021; p. 030002.
39. Patel, A.; Jain, S. Present and future of semantic web technologies: A research statement. *Int. J. Comput. Appl.* **2019**, *43*, 413–422. https://doi.org/10.1080/1206212x.2019.1570666.
40. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082.
41. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X. Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211.
42. Ogunwale, Y.E.; Ajinaja, M.O. Application Research on Semantic Analysis Using Latent Dirichlet Allocation and Collapsed Gibbs Sampling for Topic Discovery. *Asian J. Res. Comput. Sci.* **2023**, *16*, 445–452. https://doi.org/10.9734/ajrcos/2023/v16i4404.
43. Delcea, C.; Oprea, S.-V.; Dima, A.M.; Domenteanu, A.; Bara, A.; Cotfas, L.-A. Energy communities: Insights from scientific publications. *Oeconomia Copernic.* **2024**, *15*, 1101–1155. https://doi.org/10.24136/oc.3137.
44. Pion-Tonachini, L.; Makeig, S.; Kreutz-Delgado, K. Crowd labeling latent Dirichlet allocation. *Knowl. Inf. Syst.* **2017**, *53*, 749–765. https://doi.org/10.1007/s10115-017-1053-1.
45. Tran, B.X.; Nghiem, S.; Sahin, O.; Vu, T.M.; Ha, G.H.; Vu, G.T.; Pham, H.Q.; Do, H.T.; A Latkin, C.; Tam, W.; et al. Modeling Research Topics for Artificial Intelligence Applications in Medicine: Latent Dirichlet Allocation Application Study. *J. Med. Internet Res.* **2019**, *21*, e15511. https://doi.org/10.2196/15511.
46. George, L.; Sumathy, P. An integrated clustering and BERT framework for improved topic modeling. *Int. J. Inf. Technol.* **2023**, *15*, 2187–2195. https://doi.org/10.1007/s41870-023-01268-w.
47. Subakti, A.; Murfi, H.; Hariadi, N. The performance of BERT as data representation of text clustering. *J. Big Data* **2022**, *9*, 1–21. https://doi.org/10.1186/s40537-022-00564-9.

48. Eklund, A.; Forsman, M.; Drewes, F. Empirical Configuration Study of a Common Document Clustering Pipeline. *North. Eur. J. Lang. Technol.* **2023**, *9*, 1–14. https://doi.org/10.3384/nejlt.2000-1533.2023.4396.
49. Albalawi, R.; Yeap, T.H.; Benyoucef, M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell.* **2020**, *3*, 42. https://doi.org/10.3389/frai.2020.00042.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.