SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I9P113

Volume 11 Issue 9, 135-148, September 2024

Review Article

Global Trends in Data Integration Systems: A Bibliometric and Patent Analysis for 2013–2023

Samiya TAMTAM[1], Ahmed LAGUIDI[2], Abderafiaa ELKALAY[3]

[1,3]*Laboratory of Computer Science and Smart Systems (C3S laboratory), ESTC, Hassan II University, Casablanca, Morocco.*
[2]*Laboratory of Systems Analysis, Information Processing and Industrial Management, ESTS, Mohammed V University, Rabat, Morocco.*

[1]*Corresponding Author : tamtam.samiya@gmail.com*

Received: 03 July 2024 Revised: 13 August 2024 Accepted: 04 September 2024 Published: 30 September 2024

Abstract *- Data Integration Systems (DIS) represent the processes and technologies leveraged to combine data from different sources and formats into a unified and coherent view to provide a complete and coherent data set for analysis and reporting, decision making.To evaluate the global trends in DIS. We employed both bibliometric and patent analyses. This research investigates the worldwide output of scientific studies and patented innovations in the field of DIS. By examining databases such as Scopus for academic articles and Lens Patents for patent information from 2013 to 2023, and by answering the following questions: What are the main publications and countries that have contributed to the advancement of DIS? What are the emerging trends in patents in this field? and How are collaboration and innovation networks structured on a global scale?. The examination of 1038 articles and 17 patents reveals a continuous spanning growth trend, offering a comprehensive overview of identifying trends in the fields and general cultures in this disciplinary field, understanding gaps in projects and ideas, identifying the main actors on DIS, Monitoring capturing developments identifying avenues for collaboration, allowing to detect innovative techniques and avoid unnecessary repetition of research work.*

Keywords *- Data integration systems, Patent data analysis, Lens, Scopus, Scientific evolution, Bibliometric analysis.*

1. Introduction

In the age of digital transformation, data integration has emerged as a cornerstone for harnessing the full potential of disparate data sources across organizational technology boundaries. The concept of data integration encompasses the processes, tools and architecture that are used to integrate data from different sources and between them [1]. This integration is essential to enable practitioners and researchers to make informed decisions based on comprehensive data collection, thus fostering innovation and efficiency.

Significant progress has marked the evolution of data integration processes, with the move from state-of-the-art data storage techniques to the latest cloud-based real-time integration solutions revealing responsiveness to the growing complexity and volume of transport data, as well as increasing speed and scalability reflected in data processing and analysis. This reflects a need[2]. Thus, the evolution of data integration processes has coincided with broader developments in information technology, adapting to changes in data types, sources and applications.

Historically, data integration efforts have focused primarily on internal organizational data, aimed at supporting business intelligence and reporting applications, but also big data, the Internet of Things (IoT), and the emergence of cloud computing is expanding the scope of data integration, from external data sources to real-time data streams, with scalable processing capabilities [3].

On the other hand, the evolution of the data ecosystem offers many opportunities to use data integration in new ways. Advanced analytics, machine learning and Artificial Intelligence (AI) applications depend on the seamless integration of datasets to deliver previously unattainable insights and predictions. In addition, data integration is particularly used in collaborative research, interdisciplinary partnerships and intelligent services and applications [4].

Moreover, there is a notable gap in the literature regarding a comprehensive overview of the research evolution in DIS, and the problem lies in the absence of bibliometric studies that would allow us to systematically analyze trends, collaboration networks and major contributions in the field of DIS. So far, most studies focus on technical aspects or specific use cases, neglecting as well as a comprehensive analysis of major players, influential publications, and research flows. In this context, this paper proposes and aims to explore global trends in DIS, starting from the identification of the most influential academic publications, patent mapping, analysis of collaboration networks, which are systematically collated through comprehensive bibliometric and patent research, and the opportunities to capitalize on them.

2. Literature Review

The analysis of data integration frameworks covers a variety of areas, reflecting research efforts and fine-grained theoretical realizations. This body of work describes not only

technical advances in data integration solutions but also the evolving theoretical foundations that characterize current practices and the future roadmap. Current research is part of a continuum from fundamental theories to cutting-edge applications in the field of data integration.

2.1. Data Integration Systems: A Historical Perspective and Case Studies

The practical implementation of scientific research and data integration programs spans several decades, evolving more and more to meet the changing technological environment and organizational needs. This evolution reflects ongoing efforts to address the challenges of redundant and inconsistent data within and across organizational boundaries. Initially, data integration focused on integrating databases within organizational structures. The main objective was to solve the problems of data redundancy and inconsistency, which are common in fragmented work environments. [5] examined the challenges of database integration and possible solutions, emphasizing the importance of data intermediation and federated database management systems as a means of overcoming these obstacles. Their work laid the foundations for subsequent developments in the field of database integration. The advent of the internet, followed by cloud computing, has revolutionized data integration. Organizations are beginning to transcend traditional boundaries, combining data storage on the web and in the cloud. This expansion has necessitated the development of new frameworks and models to meet the challenges posed by the sheer number, diversity and mobility of data sources. [6] Emphasized the importance of using advanced schema mapping and query mediation techniques to integrate heterogeneous data, highlighting the difficulty of achieving seamless data integration in an increasingly interconnected world. The advent of big data and advanced analytics has further expanded the scope of data integration, forcing organizations to process and integrate large amounts of real-time data from various sources, such as IoT devices, social media, and other online platforms has been able to do so provide real-time insights through analytics. Currently, the field of data integration is evolving towards transformative intelligent systems, capable of learning data structures and adapting integration processes. This evolution aims to apply artificial intelligence and machine learning to data integration [7] and overcome some of today's challenges, such as unstructured data and integration. This new era of data integration is characterized by increased automation, greater precision in integration efforts, and a better ability to generate business insights from complex and diverse datasets.

2.2. Data Integration Systems: Case Studies

These case studies will provide a practical perspective on how the challenges and trends identified in the bibliometric analysis are playing out in real projects.

2.1.1. Data Integration in the Healthcare Sector: The OpenMRS Project [8]

OpenMRS (Open Medical Record System) is an open-source system for managing medical records used in several developing countries to improve healthcare. The main objective of the project was to enable the integration of various sources of medical data to facilitate the management of patient records, improve diagnosis and optimize care.

Among the challenges encountered in this project was the heterogeneity of data from several sources, including hospitals, clinics and laboratories, in a variety of formats (paper, digital, etc.). Secondly, the lack of technological infrastructure: in the regions where OpenMRS was implemented, technological infrastructure was often limited, such as Internet access and storage capacity, and thirdly, insufficient security measures.

On the other hand, the team implemented solutions during this project, namely data standardization to convert data into compatible formats, the use of the cloud to enable clinics and hospitals to store data and data encryption.

In addition, the project has helped to improve patient care by providing rapid access to information and avoiding diagnostic errors, as well as reducing the cost of administrative files and enabling medical structures to share data in real-time.

2.1.2. Data Integration in Finance: The JPMorgan Chase Project [9]

The major global bank, JPMorgan Chase, has undertaken a vast data integration project to improve its ability to process and analyze massive volumes of data from its various branches (retail banking, asset management, investment banking). It aims to use the data to improve decision-making, risk management and customer satisfaction through big data and artificial intelligence solutions.

The project faced challenges, including the massive volume of data noting that the data was stored in different silos within the organization, making it difficult to access and analyze in real-time. In addition, data integration became complex due to strict data security and confidentiality requirements.

2.1.3. Data Integration in the Public Sector: The Estonian Government's X-Road Project [10]

To facilitate communication between different public administrations, the Estonian government has implemented a data integration infrastructure called 'X-Road'. This service enables citizens, businesses and government agencies to access and share data securely and in real-time.

The difficulty encountered in this project is the use of different public databases of ministries and administrations, which makes it difficult to share data between them. Secondly, the security of sensitive data.

X-Road has made it possible to reduce administrative procedures and file processing times while simplifying access to public services for citizens and businesses. The system has also served as an example for several other countries in setting up digital government services.

2.1.4. Data Integration in the Retail Sector: The Amazon Project [11]

Amazon uses large-scale data integration systems to manage its vast supply chain and offer personalized recommendations to its customers. The data integration aimed to improve the efficiency of logistics operations while increasing sales through a better understanding of customer behaviour.

One of the challenges faced by the Amazon project was processing billions of data points every day, both from online shopping activities and from the global logistics infrastructure. Data had to be integrated and analyzed in real-time to offer relevant recommendations to users, and warehouses, supply chains and sales platforms had to be synchronized in real-time for smooth management of stocks and deliveries.

To solve these problems, Amazon used a cloud-based system for data management and AI to analyze purchasing behaviour. The integration of this data enabled it to optimize its stock management, demand forecasting and delivery processes.

2.2. Bibliometric Analysis and Patent Study: Methodology and Previous Applications

Bibliometric analysis is an effective tool for inspecting scientific production and research developments in many fields [12]. It is also based on statistical processing, allowing the analysis of large volumes of information to transform them into a set of representative graphs and maps, which facilitate the reading of the results. This method quantifies various elements of the clinical literature, including book frequency, citation styles and co-author networks, providing an overview of the dynamics of research in this field. Ellegaard and Wallin [13] demonstrated the application of bibliometric evaluation to map the medical landscape, revealing how thematic priorities and focal factors of studies have evolved over the years. This approach not only identifies the most prolific authors and influential articles but also uncovers emerging topics and patterns of collaboration between researchers and institutions worldwide. Alongside bibliometric analysis, the study of patents plays an important role in understanding technological progress and innovation pathways in the field of data integration. By examining patent filings, offers and their citations, researchers can gain insights into the pace of technological progress, the regions where innovation is most prevalent, and the key players contributing to the advancement of statistical integration responses. Daim et Rueda [14] have highlighted the value of patent valuation in forecasting technological developments and identifying innovation hotspots, underlining its importance in strategic planning and competitive analysis. This research not only traces the evolution of the times within the discipline but also highlights the institutions and nations at the forefront of the development and use of statistical integration technology. The combination of bibliometric and patent analysis provides a holistic view of the environment of expertise and innovation in the field of data integration systems. While bibliometric analysis highlights pedagogical and theoretical contributions, patent evaluation focuses on sensible and technological improvements. Together, these methodologies offer a dual perspective on how research findings translate into tangible improvements and technological solutions. This comprehensive technique provides more nuanced information on the interaction between academic research and industry trends, giving valuable insights to policy-makers, researchers and practitioners.

Introduced by Peter Chen in 1976, the Entity-Relationship (ER) version revolutionized the way record structures could be conceptualized and represented, particularly in database structures. The ER model provided a graphical technique for database design, taking into account the representation of entities (objects or concepts) and their relationships in a scientific and visually intuitive way. This version has played a key role in improving document integration systems by providing an unusual conceptual framework for knowing and mapping documents from different sources. Its impact extended beyond theoretical models to the design and implementation of relational databases, laying the foundations for complex statistical integration obligations [15].

On the other hand, the conceptualization and implementation of the Extract, Transform and Load (ETL) method represented a major advance in the operationalization of information integration [16], particularly for business intelligence and information warehousing software packages. Introduced in the early 2000s, the ETL procedure encompasses a series of difficult and rapid operations aimed at extracting information from numerous assets, reworking the records into a regular format and loading them into a destination database or data warehouse. Vassiliadis [17] has highlighted the important function of ETL methods in enabling companies to combine, harmonize and examine statistics from multiple sources, thereby facilitating informed decision-making and strategic planning.

ETL has become the cornerstone of data integration, supporting a wide range of analytical and operational sports. At the beginning of the 21st century, important contributions focused on exploiting semantic network technology and ontologies to meet the challenges of integrating statistics in heterogeneous environments. Fensel [18] has indicated how semantic internet technologies, which include the Resource Description Framework (RDF), the Web Ontology Language (OWL) and the SPARQL query language, enable information meanings, relationships, and structures to be explicitly illustrated.

Ontologies play an essential role in this context, providing dependent frameworks that facilitate the understanding and integration of data across disparate assets. By using semantic network technologies and ontologies, researchers and practitioners can improve fact integration tactics, enabling more powerful discovery, access and integration of information from numerous and dispersed assets.

2.3. Gaps in Existing Research

The creation of the Internet of Things (IoT), social media and other virtual structures has led to exponential growth in the real-time information age. This development poses significant challenges for statistics integration systems, primarily in terms of combining these real-time statistics streams with static record repositories. Stonebraker [19] highlights scalability and timeliness issues, as conventional data integration structures are often unprepared to handle the speed and quantity of real-time statistics. Integrating these disparate types of information requires innovative answers capable of processing and analyzing records in real-time, ensuring that the information obtained is both timely and actionable.

Another major gap in cutting-edge data integration research is the effective management of unstructured documents, which account for a massive proportion of facts generated today, particularly from social media platforms, web sources and multimedia content material. Agrawal [20] highlights the complexity of extracting meaningful information from unstructured data and aligning it with established datasets.

Traditional information integration methods, which generally focus on established and semi-structured data, are insufficient in the face of the heterogeneity and ambiguity of unstructured data. This mission requires advanced strategies in Natural Language Processing (NLP) [21], systems-based knowledge acquisition and semantic analysis to interpret and integrate unstructured facts. In addition, frameworks capable of seamlessly combining information from structured, semi-structured and unstructured records are needed to provide a comprehensive view of the panorama.

The gaps diagnosed highlight the need to persevere with innovation, in fact, integration research. Future work could focus on the development of adaptive fact integration systems that leverage synthetic intelligence and gadget-based learning to dynamically process and combine various types of data. In addition, there is a growing need for research into statistics governance and first-class management frameworks that could guarantee the reliability and accuracy of integrated datasets, particularly in the context of unstructured and real-time information.

It should be noted that previous studies on data integration systems mainly focus on the technical aspect, which is often limited to particular case studies or specific sectors such as health or finance. Our research is distinguished by the application of a comprehensive bibliometric approach that has not been applied to this area of DIS so far, which aims to offer a systematic analysis of the entire field of data integration systems by analyzing the growth of publications over time and associated disciplines, identifying the most prolific institutions, as well as their impact through citations, exploring how researchers from different regions collaborate, identifying growing research topics, revealing emerging technologies, business strategies and competitive dynamics in this sector.

3. Methodology

3.1. Analytical Approach

The analytical approach adopted in this study aims to conduct a thorough examination of global trends in data integration systems from 2013 to 2023 (The period 2013-2023 was characterized by a convergence of technological and regulatory factors that have shaped the field of data integration systems, namely: the rise of Big Data, the massive adoption of cloud-based solutions, the emergence of artificial intelligence and machine learning, the introduction of major regulations and the transition to real-time systems.

Analyzing this period allows us to capture the rapid evolution of technologies and practices, as well as to identify trends that will continue to influence the field in the years to come). This analysis leverages a combination of bibliometric methods and patent analyses to assess academic impact and technological advancements within the field. Specific objectives include identifying key contributions, assessing collaboration trends, and detecting emerging areas of research and innovation.

Recognizing the importance of precision and relevance in our dataset, stringent selection criteria were established:

3.1.1. Academic Publications Selection

To curate a dataset of academic publications, the Scopus database was employed [22], renowned for its comprehensive coverage of scholarly literature across various disciplines. Scopus's extensive database covers a wide range of journals and conference proceedings, ensuring that our research encompasses the full spectrum of scholarly discourse on data integration systems.

A search equation was meticulously crafted, incorporating an array of keywords and phrases intimately connected to the realm of data integration systems. These keywords included but were not limited to "data integration system," "hybrid integration system," "information integration," "heterogeneous data sources," and more nuanced terms that encapsulate the breadth and depth of the field. The search was strategically refined to include articles, conference papers, and reviews published between 2013 and 2023, ensuring the dataset's temporal relevance to our study period.

3.1.2. Patent Data Selection

Concurrently, the exploration of technological advancements was facilitated through an exhaustive search of the Lens.org database [23] for patents associated with data integration technologies.

Lens.org was selected for its comprehensive and open-access repository of global patent information, offering global coverage of the patent documents, then relevance and Precision by employing specific codes as search parameters. The analysis of patent data from Lens.org offered valuable insights into how theoretical concepts are translated into practical applications and technological solutions within the realm of data integration.

The selection was guided by keywords related to the data integration systems and a meticulous identification of Cooperative Patent Classification (CPC) [24] and International Patent Classification (IPC) [25] codes that directly align with data integration systems and related technologies. This dual classification system allowed for a comprehensive capture of patents spanning innovative data integration solutions, methodologies, and applications. Key CPC codes such as G06F16/24 (Data processing systems or methods specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes) and relevant IPCR codes served as beacons to distill a dataset of patents that reflect cutting-edge technological progress in data integration.

The combination of academic and patent datasets, each selected through rigorously defined criteria, ensures a holistic view of the advancements, trends, and future directions in data integration systems. This dual-faceted approach not only highlights the academic discourse surrounding data integration but also sheds light on the practical technological innovations that are propelling the field forward.

3.2. Analysis Methods

3.2.1. Indicators

The bibliometric analysis was carefully conducted using a variety of indicators to better map the research landscape in data integration:

- Publication frequency: This indicator was used to identify scholarly output [26] at specific points in time, enabling us to track trends and developments in the field.
- Citation management [27]: Citation analysis highlighted the impact and scope of particular studies and identified key documents that have shaped the development of integration technology.
- International collaboration [28]: This indicator measures the ability to attract international collaborations by a country, an institution or authors.

3.2.2. Visualization Tools and Classification

Advanced tools were used, such as VOSviewer, to visualize these complex datasets [29]. This software facilitated the creation of intuitive web maps, links and thematic groups, visually indicating the relationships between research institutions (authors, institutions, countries) and the thematic orientation of publications. Flourish. studio as used also [30], which allowed the creation of the keyword.

3.2.3. Patent Analysis: Data Extraction and Classification

The patent analysis was structured around a detailed examination of several critical aspects of the patent landscape:

- Filing dates: The distribution of patent dates was examined to understand the temporal dynamics of innovation in data integration systems.
- Classification codes: Each patent was classified according to the Cooperative Patent Classification (CPC) and International Patent Classification (IPC) systems.
- Citation analysis: By studying the citations of each patent, the impact and evolution of technological ideas can be tracked.

The selected data was carefully categorized by relevance and analyzed to identify innovation trends and priorities, as well as the geographical distribution of licensing activity. This comprehensive approach highlighted the many aspects of technological advances in data integration, including where important innovations come from and how they can shape technological advances, including those demonstrated on a global scale.

Together, these processes - catalogs of academic contributions and patent reviews for technological developments - have provided an overview of the state of data integration systems and a better understanding of the theoretical and practical improvements driving the project.

3. Results

3.1. Bibliometric Analysis

Bibliometric analysis of the literature on data integration systems from 2013 to 2023 has highlighted an upward trend in the quantity of guides. This steady growth has peaked in recent years, reflecting the growing enthusiasm of the scientific and technical communities. This boom is linked to the widespread adoption of virtual technology in many sectors, requiring modern document integration solutions to control increasingly complicated document landscapes. The increase in the number of courses is also indicative of the widening scope of studies, as new situations and opportunities for statistics integration continue to emerge as cloud computing, mass records analysis, and machine learning improve.

3.1.1. General Trends in Publications Volume and Growth of Publications

The Scopus database, consulted in December 2023, has listed 2359 publications since 1991, with 1038 on DIS between 2013 and 2023, representing 44% of data integration systems research articles.

Thus, between 2013 and 2023, only 31% of the corpus was open access (Figure 1). Conference proceedings (54%) followed by Journal articles represented 38%, Book chapters (4%), and other types of documents (e.g., reviews, conference reviews, etc.). English (97%) was the predominant language, with Chinese coming next (1%) then French (0.58%).

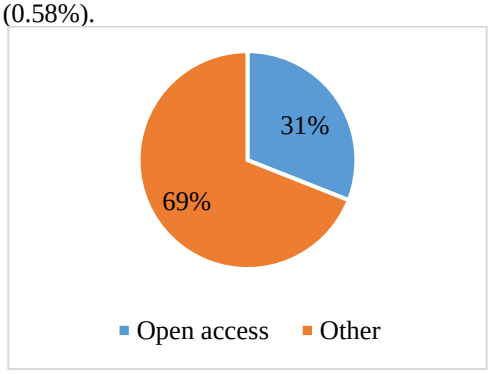

Fig. 1 Open access -DIS 2013-2023

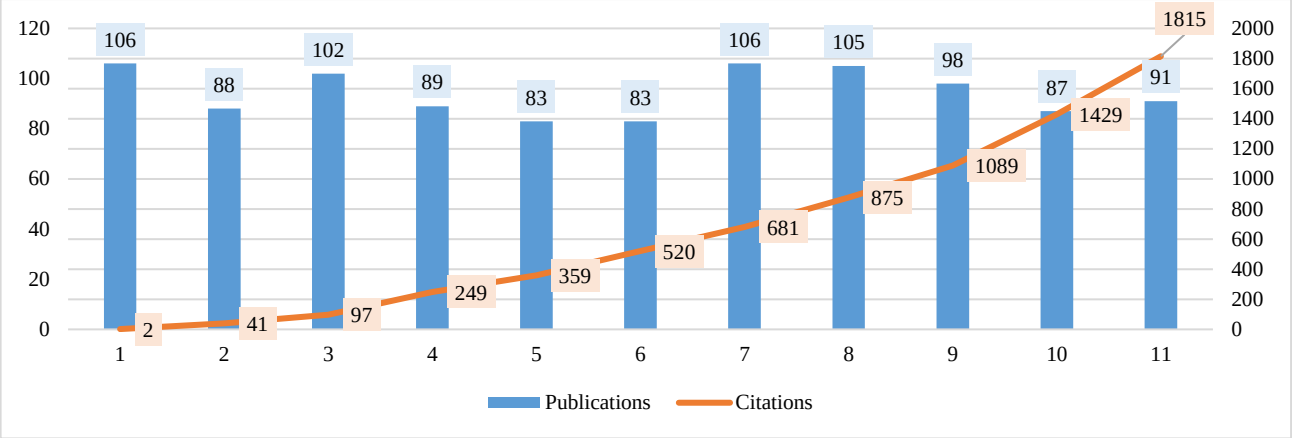

Fig. 2 Publications & citations DIS 2013-2023

3.1.2. Evolution of Scientific Output on DIS Between 2013 and 2022

The total production between 2013 and 2023 amounted to 1038. Academic production is an indicator of productivity. In this analysis, all types of publications were considered, including articles, journals, conference communications, book chapters, etc. This corpus received 7157 citations. The number of publications on data integration systems shows a fluctuating pace over the same period. These variations may reflect the scientific community's fluctuating interest in data integration systems during this period (0).

3.1.3. Main Contributors: Countries

The USA dominates the publishing landscape with 133 publications, of which 41.4 % are in collaboration, while European countries and China express their strategic commitment to the development of data integration systems (Table 1).

The United States ranks first, with the highest scientific output (133) and citations (2912), demonstrating the country's significant influence in the academic world. Their international collaboration rate is 41.4%, suggesting a high level of international engagement, but not the highest. China ranks second in terms of output (127), but its number of citations (1,416) and international collaboration rate (21.3%) are lower than those of the United States. India ranks third in terms of scientific output (79), but the number of citations (388) and the percentage of international collaboration (10.1%) are lower than in the United States. This suggests that India produces a significant amount of research but that its global impact may be more limited than that of the top-ranked countries. France has a high percentage of international collaboration (58.4%), indicating that much of its research is carried out in collaboration with international partners, which may contribute to its higher number of citations.

Table 1. Countries with the highest publications on DIS (2013-2023) and international collaboration -DSI 2013–2023.

Rank	Country	Scholarly Output	Citation Count	International Collaboration (%)
1	United States	133	2912	41,4
2	China	127	1416	21,3
3	India	79	388	10,1
4	France	77	624	58,4
5	Germany	54	1165	46,3
6	Morocco	43	145	4,7
7	Spain	42	919	59,5
8	Italy	39	475	35,9
9	Australia	36	780	50
10	Indonesia	36	144	16,7
11	United Kingdom	32	1108	46,9
12	Portugal	31	340	32,3
13	Tunisia	29	124	58,6
14	Algeria	27	157	74,1
15	SaudiArabia	23	158	87

Germany follows with a good number of citations (1,165) and a good percentage of collaboration (46.3%), showing its balanced presence in terms of global collaboration and influence. French countries with high collaboration percentages: Spain (59.5%), France (58.4%), Tunisia (58.6%) and Australia (50%) have higher collaboration percentages, suggesting that they rely heavily on international partnerships, which can enhance the global impact of their research. Low collaboration: Morocco (4.7%) and India (10.1%) have low percentages of international collaboration, which could mean that their research is more country-focused, limiting global recognition and citations. Finally, countries such as the USA, China and India lead the way in terms of research output. However, nations such as France, Germany and Spain stand out for their high citation rates and collaborative efforts, which improve the visibility and impact of their research on a global scale.

3.1.4. Top Contributing Institutions

Regarding the ranking of the most productive affiliations in the field of DIS, the CNRS in France took the lead, followed by Mohammed V University in Rabat. Next were the University of Minho in Portugal and in fifth position, École Nationale Supérieure d'Informatique in Algeria (Table 2).

Table 2. Top contributing institutions

Institution	Country	Scholarly Output	International Collaboration (%)	Citation Count	Citations per Publication
CNRS	France	36	44,4	311	8,6
Mohammed V University in Rabat	Morocco	18	0	49	2,7
University of Minho	Portugal	18	11,1	172	9,6
École Nationale Supérieure d'Informatique	Algeria	16	81,2	106	6,6
PolytechnicUniversity of Catalonia	Spain	13	61,5	174	13,4
University of Sfax	Tunisia	13	61,5	91	7
Poznań University of Technology	Poland	12	33,3	94	7,8
Université Fédérale Toulouse Midi-Pyrénées	France	12	50	45	3,8
École nationale supérieure de mécanique et d'aérotechnique	France	11	90,9	76	6,9
Université Lumière Lyon 2	France	11	81,8	44	4
Université Paul Sabatier Toulouse III	France	11	45,5	45	4,1
Université Toulouse 1 Capitole	France	11	45,5	45	4,1
Aalborg University	Denmark	10	50	143	14,3
Curtin university	Australia	10	60	48	4,8
Guru Gobind Singh Indraprastha University	India	10	10	14	1,4
Anna University	India	9	0	144	16
Chinese Academy of Sciences	China	9	44,4	102	11,3
IBM	United States	8	50	215	26,9
INRAE	France	8	37,5	117	14,6
Université libre de Bruxelles	Belgium	8	75	138	17,2

Table 3. Number of publications by quartile

Publications	Publication	%
Q1 top 25%	176	29,93%
Q2 top (26%-50%)	156	26,53%
Q3 top (51%-75%)	132	22,45%
Q4 top (76%-100%)	124	21,09%

3.1.5. Journals' Performances

The journals are classified into four groups according to their reputation. Thus, there are four quartiles: Q1 for the top 25%, Q2 for the top 26% to 50%, Q3 for the top 51% to 75%, and Q4 for the top 76% to 100%.

As shown in 0, 29.93% of publications on DSI have appeared in Q1 journals, 26.53% in Q2, 22.45% in Q3, and 21.09% in Q4 journals. The distribution of publications in Q1 and Q2 (56.46% in total) shows a high quality of production in this area.

3.1.6. Topic Clusters and Subject Dispersion

A Topic is a dynamic group of documents with a common focused intellectual focus [31], [32]. SciVal contains around 94000 Topics and 1500 Topic Clusters, which are created by direct citation linking. Each publication is assigned to one topic and one topic Cluster. Between 2013 and 2023, the composition of the various

disciplines and their scientific execution show different levels of interest, and the analyses focus on the decade. (0) [33].

3.1.7. The Distribution of Subject Categories [34]

As shown in 0, "Computer Science" held the first position during the study period (655 publications), representing 37.67% of all articles on DIS. The second most dominant thematic area was Engineering (275 articles, or 15.81%), followed by Mathematics (182 articles, or 10.47%), Decision Sciences (131 articles, or 7.53%), Social Sciences (91 publications or 5.23%) and Business Management and Accounting (75 publications or 4.31%). The six main thematic areas covered 81.02% of the studied corpus.

3.1.8. Global Collaboration Map and Network

As illustrated in Figure 4, the United States is in the first cluster of collaborations, followed by Croatia, Singapore and Taiwan.

The second cluster can be seen between China, India, Canada and Indonesia. Moreover, a third one between France, Algeria and Lebanon.

Table 4. Top topics in scholarly output (2013-2023)

Topic	Scholarly Output
Data Warehouse; Decision-Making; Business Intelligence	106
Data Warehouse; Big Data; Information System	70
Ontology Matching; KnowledgeBasedSystems; Semantic Web	22
Big Data; Data Warehouse; Information System	21
Data Warehouse; Online Analytical Processing; Semantic Web	17
Information Analysis; Decision-Making; Business Intelligence	13
Big Data; Decision-Making; Data Analytics	12
Entity Resolution; Record Linkage; Data Integration	12
Data Warehouse; MaterializedView; Query Processing	12
Big Data; MongoDB; Structured Query Language	12

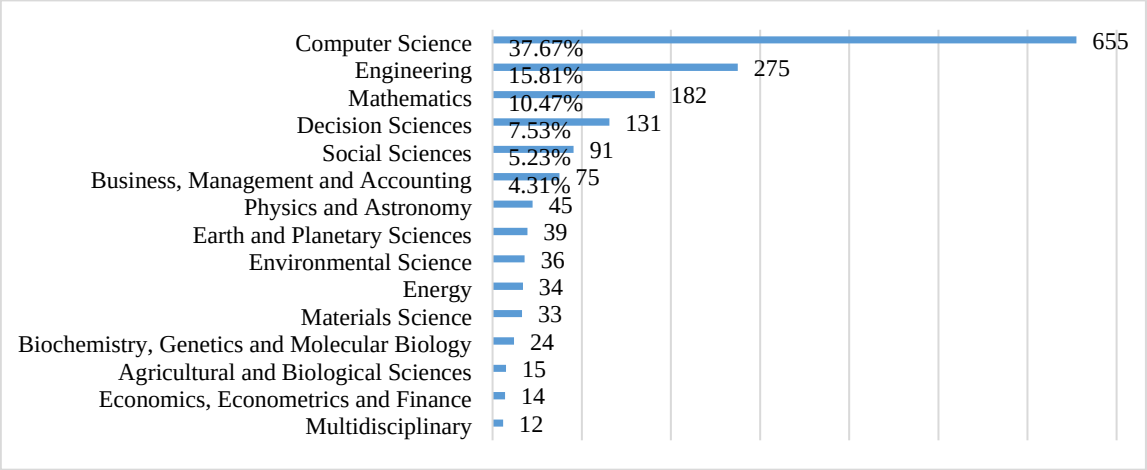

Fig. 3 The distribution of subject categories DIS (2013-2023)

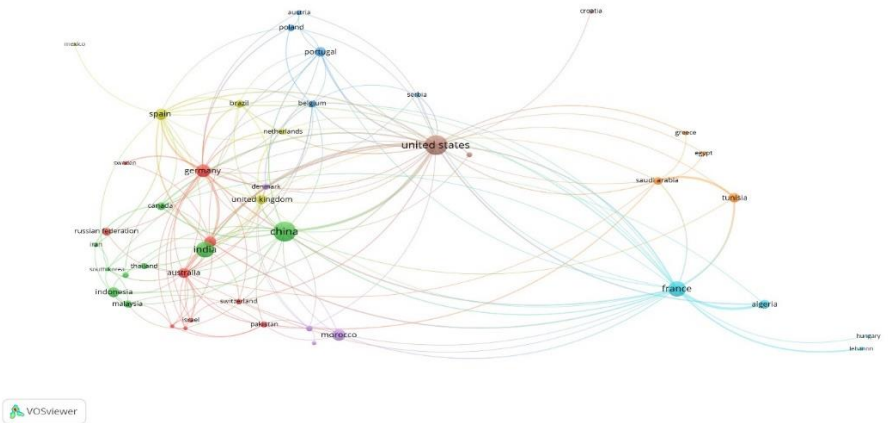

Fig. 4 World map country collaboration generated with VOSViewer

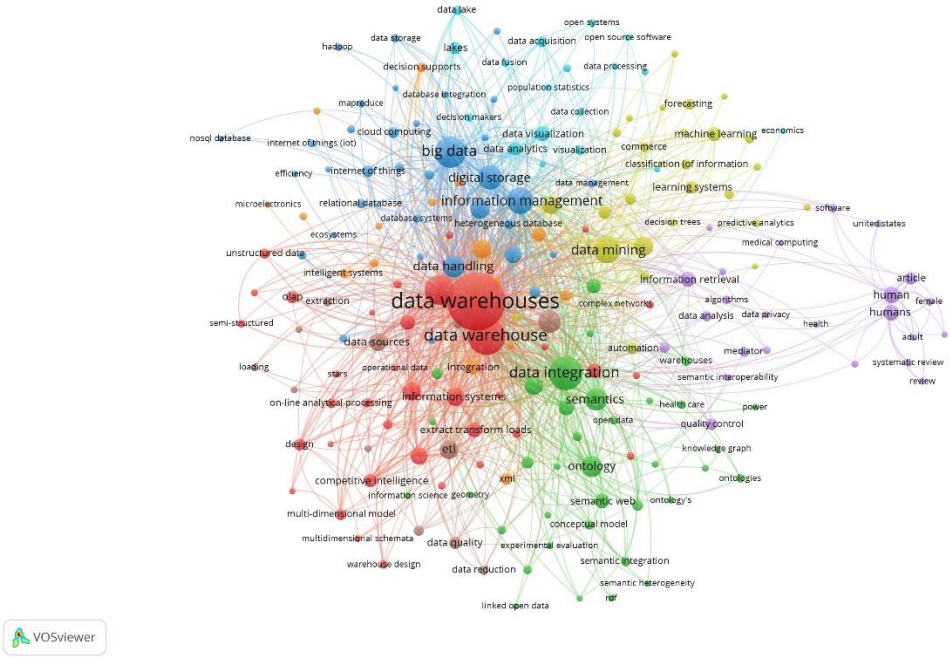

Fig. 5 Keywords co-occurrence network via VOS Viewer.

3.1.9. Co-Occurrence Network of Keywords

The network of keywords is displayed in Figure 5. Each keyword is depicted by a node in the network. The size of the node indicates the frequency of the keyword (representing how often the keyword appears) [35]. The co-occurrence of keywords is shown by the connection between the nodes (connections that occur together), and the thickness of the lines between nodes represents the frequency of keyword co-occurrences (indicating how often two or more terms occur together). Each color represents a thematic cluster. Data warehouse, data integration, information integration, and heterogeneous databases are the most common terms found in various clusters.

3.2. Patent Analysis

3.2.1. Volume and Evolution of Patents

40 % of these patents are patent applications, the peak year 2023 with 3 granted patents, and 60% are granted patents peak year 2022 with 5 applications (Table 5).

This data suggests a dynamic landscape where patent applications have seen a rise, particularly in the last few years, potentially indicating increased research and development activities. The granted patents, while fluctuating, do not show a clear trend, implying that the approval process might not directly correlate with the number of applications each year.

It is important to note that a patent application does not imply that the patent has been granted [36].

3.2.2. Dispersion of Patents by CPC code

The figure shows that 60% corresponds to G06F21/10 (Physics protecting distributed programs or content), 55% corresponds to G06F16/27 (Replication, distribution, or synchronization of data between databases or within a distributed system) and G06F16/9535 (Search customization based on user profiles and personalization), 38% corresponds to G06F16/9038 (Presentation of query results).

CPC codes such as G06F21/10, G06F16/27, and G06F16/9535 received the most attention.

The distribution of these CPCs reflects a focused effort to refine the fundamental components of data integration technologies, processing power efficiency, information access and effectiveness (0).

3.2.3. Filing Countries

According to the Lens platform, patent filing activity in the field of DIS is carried out in 3 countries. The main countries issuing patents within their territories include the United States, European patent and wipo (0).

Table 5. Volume and evolution of patents

Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
Granted Patent	2	1	0	0	1	2	2	3	2	1	3	17
Patent Application	0	2	1	0	2	3	1	3	4	5	4	25

Fig. 6 Dispersion of patents by CPC code

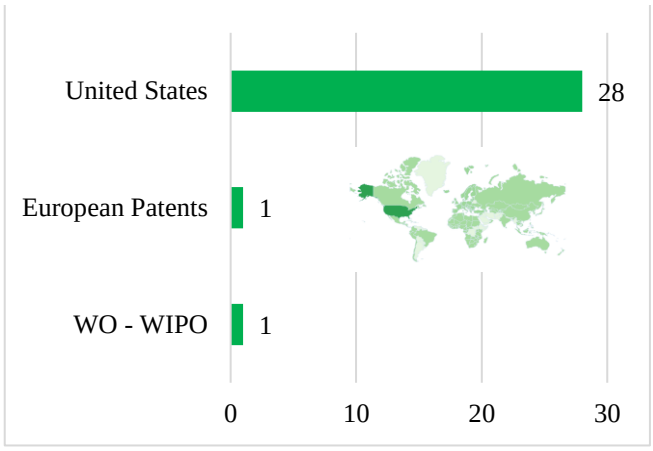

Fig. 7 Countries depositing patents in the field of DIS (2013-2023)

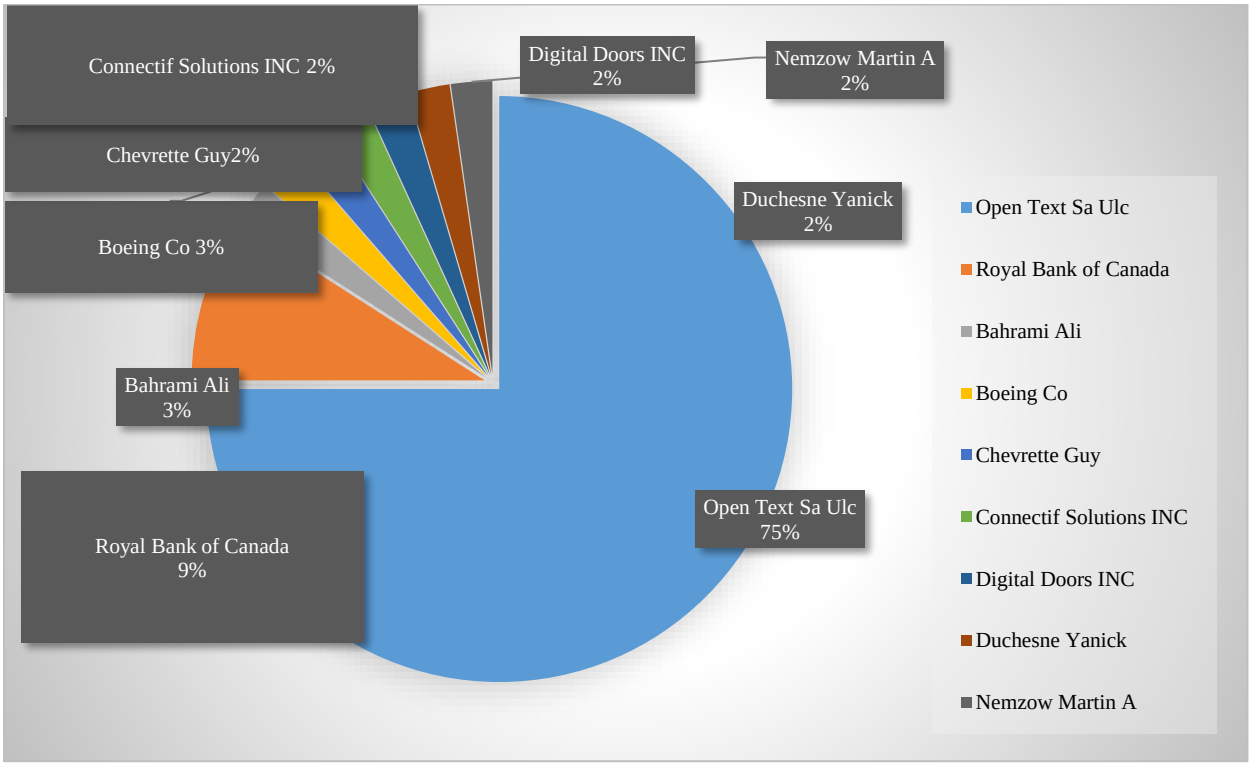

Fig. 8 Companies filing the most patents on DIS

3.2.4. Companies Filing the Most Patents on DIS

As for businesses, the international company Open Text SA ULC is the leader in DSI with 75% of patents, followed by the Royal Bank of Canada with 9%, then Bahrami Ali and Boeing Co with 3%, respectively (0).

3.2.5. Patent Citations and Technological Impact

There are notable differences in the number of cases filed per patent between age groups and according to legal status. For example, patents filed in 2017 accumulated a significant number of citations. A number of 2015 licensees received subpoenas, which may indicate a major renewal or foundation license.

The average size of patent families also varies, with some patents having larger families, as in 2015, with an average size of 4 licensees. The size of patent families applied for in 2021 and 2022 is larger, which may reflect patent applications in different sectors or the extension of existing patents (0).

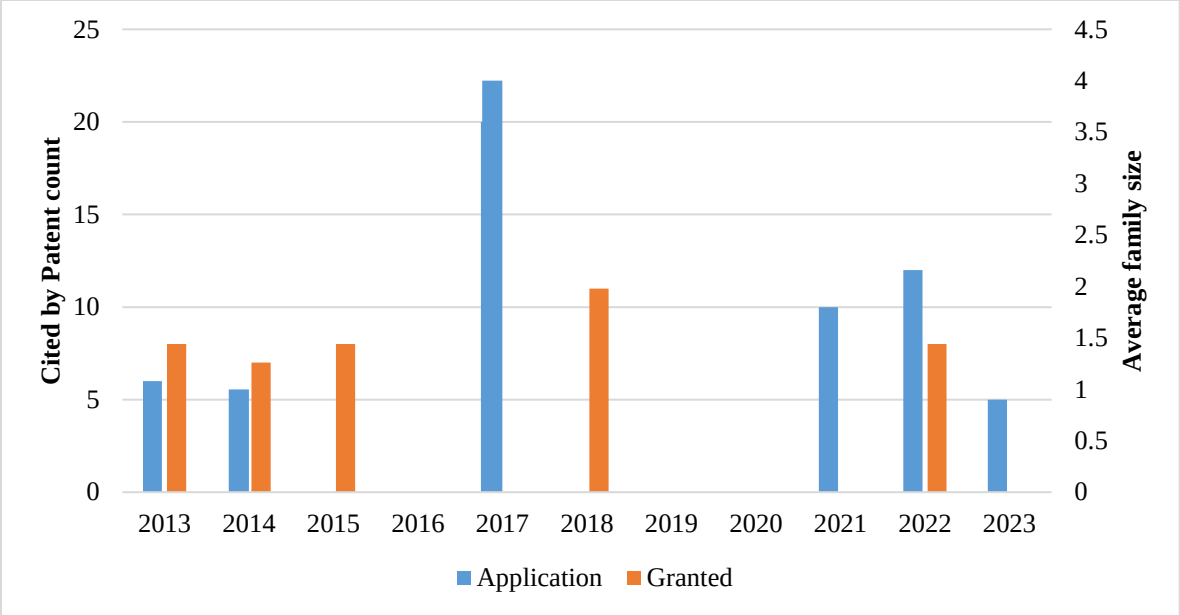

Fig. 9 Cited by patent count and average family size over the years by legal status

4. Discussion

4.1. Comparison of Bibliometric and Patent Trends

Comparative analysis of bibliometric data and patent applications reveals correlations and divergences in the field of data integration. As the number of scientific articles and patent applications continues to rise, a rapid pace of innovation is being noticed.

4.1.1. Correlations

Bibliometrics and patent analysis place greater emphasis on cutting-edge technologies like cloud computing [37], Artificial Intelligence (AI) [38] and real-time data processing [39]. This framework reflects a convergent approach to R&D development and emphasizes this technology to meet the growing demand for sophisticated data processing and analysis capabilities. It reflects a concerted effort across all academic and technical disciplines to develop solutions that will enable timely self-assessment and demonstrate the professional excellence needed to facilitate today's business environment.

4.1.2. Divergences

Despite these commonalities, the geographic distribution of R&D efforts reveals striking differences. The United States maintains a leading position in terms of the number of books and patents, reflecting its important role in the technological frontiers of data integration, but bibliometric data reveals more international collaborations than indicated by s patents.

This comprehensive review not only highlights the current state of data integration technologies but also provides insight into the dynamics of academic research and technology applications. Understanding these developments helps guide future research and development directions, informing academic research and technical strategies in terms of data integration and implications for research and development.

The changing landscape of data integration, highlighted by growing trends in bibliometrics and patent analysis, has many implications for Research and Development (R&D) in various industries. Understanding these implications can be useful in guiding future research efforts and designing optimal development strategies to exploit the full potential of data integration technologies.

By focusing on these key areas, R&D can significantly enhance the capabilities of data integration technologies, making them more robust, secure and adaptable to the needs of today's digital landscapes. This approach not only fosters innovation but also ensures that developments are appropriate and applicable across all sectors, with companies ultimately contributing to the growth of the technology economy.

4.2. Potential Challenges and Opportunities Recognized Through this Analysis

According to the analysis, the evolution of publications and patents in DIS is trending upwards and rapidly due to the increasing integration of artificial intelligence and machine learning into DIS as well as the processing of continuous streams in real-time, all of which can lead to the results that data will be better processed and provide effective analyses and insights in the future this will allow DIS to be more intelligent and autonomous and adapt to changing data sources and business requirements.

The first challenge is the management and collection of data, as DIS covers several domains such as Computer Science, Engineering, Mathematics, and Decision Sciences. Gathering this vast and multidisciplinary data can be complex. Understanding complex collaboration networks between countries requires advanced analysis tools. Another challenge for this analysis is that some articles and patents will not be freely accessible, which limits the bibliometric analysis.

For opportunities, This analysis can be used to identify scientific and technological trends in the DIS field. Depending on collaboration patterns, this study can highlight partnerships between academics and industry.The analysis can identify publications and patents key authors for establishing references for future research.

4.3. Ethical and Societal Implications.

Data integration raises important questions about confidentiality. Increased access to information increases the risk of users' private data being exploited, hence the need to include robust protection mechanisms. The second element is transparency regarding the collection and use of data. The automation of data can have an impact on employment by reducing the workforce, thereby affecting the social economy. In some contexts, data integration can be used to implement mass surveillance systems, affecting the freedom and privacy of individuals. DIS indeed bring technological advantages but sound ethical frameworks must accompany their adoption.

4.4. Practical Implications and Future Directions

4.1.1. For Researchers and Practitioners

The results of this study underline the crucial importance of integrating advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML) into data integration processes. These technologies not only make these systems more effective and efficient but also provide adaptive and predictive data processing capabilities. Researchers and practitioners are invited to pursue the development of systems that are not only functional but also anticipatory, adapting to ongoing changes in data types, volumes and integration requirements. Analysis of these flexible configurations can lead to robust data integration strategies, compatible with the dynamic evolution of the business environment and technological landscape. In addition, there are significant opportunities for innovation in areas such as real-time data processing and data source integration, which are increasingly stimulated by the Internet of Things (IoT) and other digital transformations.

4.1.2. For Policy and Technological Strategy

The lessons learned from this study provide a valuable basis for policymakers and technology strategists to formulate investment and development strategies for the advancement of data integration systems. By identifying and supporting key technologies such as cloud solutions, AI and machine learning, policymakers can help establish technological leadership and stimulate economic growth. In addition, policy plays an important role in terms of international standards and best practices to encourage data integration. This can ensure consistency and interoperability across systems and geographies, improving data management strategies on a global scale. Collaborative policies and research aimed at fostering an open innovation ecosystem can promote these developments and bridge the gap between academic research and industrial applications.

4.1.3. Future Directions

In the future, the data integration process is evolving towards intelligent, automated and connected systems. The continued integration of AI and machine learning could drive the next wave of innovation, enabling systems to learn from data structures and automatically adjust processes for optimal performance. In addition, Cloud computing and Edge computing technologies are expected to play a key role in delivering flexible and agile data integration solutions that support a variety of computing and storage requirements across all industries, as well as managing massive data with improved performance and reduced latency. The use of blockchain in data integration systems will enable transparency, traceability and security of exchanges. For IoT, its application in DIS enables the processing of massive data flows from connected objects, making the automation of real-time integration processes crucial. As far as Big Data technologies are concerned, they enable DISs to process massive data sets in real-time for more accurate and informed analysis and decision-making.

In conclusion, the research community and policy-makers are in a position to play a key role in shaping the future of data integration technologies. The potential to revolutionize data management practices and advance technological innovation through targeted research and strategic planning is enormous.

5. Conclusion

A comprehensive study of global trends in data integration systems through bibliometric analysis of scientific articles and patents looked at a comprehensive review of 1038 publications from 2013 to 2023. The corpus of articles presented commendable quality in terms of citations, with 7157 citations. The CNRS in France emerged as the most productive institution, and the top three countries contributing to scientific output were the United States, China and India." In addition, it was foundthat 56.46% of DIS publications were indexed in Q1 and Q2 journals. Moreover, the six main thematic areas (Computer Science, Engineering, Mathematics, Decision Sciences, Social Sciences, Business Management and Accounting) covered 81.02% of the publications.

The survey of patent applications in the DIS field reveals significant trends with 17 patents only 60% were granted

patents worldwide in the field of AI from 2013 to 2023. For companies, « Open Text SaUlc ranked first with 75% ». The main country filing patents was the United States, with 28 patents. On the other hand, this study reveals important developments and the way forward for this important technology segment. Our findings highlight the key role that DIS play in facilitating business and strategic innovation.

The evolution of data integration infrastructure was documented, which shows a clear shift from traditional database management to advanced infrastructure, including cloud computing, AI and real-time data processing. These developments are intensifying because, in certain areas, the amount of data required is obvious. Based on our results, the following areas for future data integration systems research were recommended. Future research should explore how to integrate emerging technologies such as quantum computing and advanced neural networks into data integration processes. These technologies promise to dramatically increase the speed and security of data processing and eliminate the current problem of data silos and ensure a continuous flow of data.

In conclusion, this study not only highlights the current state of data integration systems but also provides innovations and future developments in this important field. As data is one of an organization's most important assets, the importance of an effective and secure data integration strategy cannot be overstated. The recommendations presented here are aimed at guiding researchers and practitioners towards next-generation data integration solutions that can meet high demands.

References

[1] Nigel Martin et al., "A Methodology and Architecture Embedding Quality Assessment in Data Integration," *Journal of Data and Information Quality*, vol. 4, no. 4, pp. 1-40, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[2] Alon Halevy, Anand Rajaraman, and Joann Ordille, "Data Integration: The Teenage Years," *Proceedings of the 32nd International Conference on Very Large Data Bases VLDB Endowment*, Seoul, Korea, pp. 9-16, 2006. [Google Scholar] [Publisher Link]

[3] Dangzhi Zhao, and Andreas Strotmann, *Analysis and Visualization of Citation Networks*, Morgan&Claypool Publishers, pp. 1-207, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Min Chen, Shiwen Mao, and Yunhao Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, pp. 171-209, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Amit P. Sheth, and James A. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," *ACM Computing Surveys*, vol. 22, no. 3, pp. 183-236, 1990. [CrossRef] [Google Scholar] [Publisher Link]

[6] Maurizio Lenzerini, "Data Integration: A Theoretical Perspective," *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Madison Wisconsin, pp. 233-246, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[7] Marcus Birgersson, Gustav Hansson, and Ulrik Franke, "Data Integration Using Machine Learning," *2016 IEEE 20th International Enterprise Distributed Object Computing Workshop*, Vienna, Austria, pp. 1-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[8] OpenMRS Medical Record System. [Online]. Available: https://openmrs.org/fr/

[9] JPMorganChase. [Online]. Available: https://www.jpmorganchase.com/

[10] Gregorio Robles, Jonas Gamalielsson, and Björn Lundell, "Setting Up Government 3.0 Solutions Based on Open Source Software: The Case of X-Road," *18th IFIP WG 8.5 International Conference, EGOV 2019*, San Benedetto Del Tronto, Italy, pp. 69-81, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] About Amazon. [Online]. Available: https://www.aboutamazon.com/

[12] Alan Pritchard, "Statistical Bibliography or Bibliometrics," *Journal of Documentation*, vol. 25, no. 4, pp. 348-349, 1969. [Google Scholar]

[13] Ole Ellegaard, and Johan A. Wallin, "The Bibliometric Analysis of Scholarly Production: How Great is the Impact?," *Scientometrics*, vol. 105, pp. 1809-1831, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Tugrul U. Daim et al., "Forecasting Emerging Technologies: Use of Bibliometrics and Patent Analysis," *Technological Forecasting and Social Change*, vol. 73, no. 8, pp. 981-1012, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[15] Peter Pin-Shan Chen, "The Entity-Relationship Model-Toward a Unified View of Data," *ACM Transactions on Database Systems*, vol. 1, no. 1, pp. 9-36, 1976. [CrossRef] [Google Scholar] [Publisher Link]

[16] Asma Dhaouadi et al., "Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons," *Data*, vol. 7, no. 8, pp. 1-38, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos, "Conceptual Modeling for ETL Processes," *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*, McLean Virginia USA, pp. 14-21, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[18] Dieter Fensel, *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, MIT Press, pp. 1-479, 2003. [Google Scholar] [Publisher Link]

[19] Michael Stonebraker, Uğur Çetintemel, and Stan Zdonik, "The 8 Requirements of Real-Time Stream Processing, *ACM SIGMOD Record*, vol. 34, no. 4, pp. 42-47, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[20] Divyakant Agrawal, Sudipto Das, and Amr El Abbadi, "Big Data and Cloud Computing: New Wine or Just New Bottles?," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1647-1648, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jim Holdsworth, What is NLP (Natural Language Processing)?, IBM, 2024. [Online]. Available: https://www.ibm.com/topics/natural-language-processing

[22] Scopus: Comprehensive, Multidisciplinary, Trusted Abstract and Citation Database, Elsevier. [Online]. Available: https://www.elsevier.com/products/scopus

[23] Lens.org, About Lens. [Online]. Available: https://about.lens.org/

[24] Cooperative Patent Classification (CPC), Epo.org. [Online]. Available: https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/classification/cpc

[25] International Patent Classification Scheme, WIPO. [Online]. Available: https://ipcpub.wipo.int/?notion=scheme&version=20240101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart

[26] SciVal Metric: Scholarly Output, SciVal Support Center, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/28180/kw/Publication+frequency%E2%80%8E/supporthub/scival/related/1/

[27] SciVal Metric: Citation Count, SciVal Support Center, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/28191/supporthub/scival/

[28] Hans Pohl, Guillaume Warnan, and Jeroen Baas, "Level the Playing Field in Scientific International Collaboration with the Use of a New Indicator: Field-Weighted Internationalization Score," *Research Trends*, vol. 1, no. 39, pp. 1-7, 2007. [Google Scholar] [Publisher Link]

[29] Nees van Eck, and Ludo Waltman, "Software Survey: VOSviewer, A Computer Program for Bibliometric Mapping," *Scientometrics*, vol. 84, no. 2, pp. 523-538, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[30] Beautiful and Easy Data Visualization and Storytelling, Flourish. [Online]. Available: https://flourish.studio/

[31] What is a Topic?, SciVal Support Center, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/35048/supporthub/scival/

[32] How Can i Use SciVal Topics?, SciVal Support Center, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/35047/supporthub/scival/

[33] What do the Bubbles on the Wheel of Science Represent?, SciVal Support Center, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/35053/supporthub/scival/

[34] What is the Complete List of ASJC Subject Areas in Scopus?, Scopus: Access and Use Support Center, Scopus, 2024. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/

[35] Nees Jan van Eck, and Ludo Waltman, *Visualizing Bibliometric Networks*, Measuring Scholarly Impact, Springer, Cham, pp. 285-320, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[36] Eric Waltmire, They Have a Patent, Or Do They? Granted Patents & Published Applications, Eric Waltmire's Blog. [Online]. Available: https://www.waltmire.com/2015/03/21/granted-patents-and-published-applications/

[37] Alessio Bottari et al., "Integration of Cloud Computing and Internet of Things: A Survey," *Future Generation Computer Systems*, vol. 56, pp. 684-700, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[38] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić, "Explainable Artificial Intelligence: A Survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia, pp. 0210-0215, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[39] Uğur Kekevi, and Ahmet Arif Aydın, "Real-Time Big Data Processing and Analytics: Concepts, Technologies, and Domains," *Computer Science*, vol. 7, no. 2, pp. 111-123, 2022. [CrossRef] [Google Scholar] [Publisher Link]