Revista de Ingenierías Interfaces, Vol 9 (2) Julio-Diciembre 2026. ISSN 2619-4473 E-ISSN 2619-4465, pp 1-23.

# Evolution and Research Dynamics of Metadata Extraction in Academic Documents *

## Evolución y dinámica de investigación de la extracción de metadatos en documentos académicos.

Recibido: diciembre 21 de 2025 - Evaluado: febrero 18 de 2026 - Aceptado: abril 11 de 2026

Anderson Joseph Ochoa-Trujillo**
Cristian Andrés Villarreal-Orozco***
Juan Pablo Hoyos-Sánchez****

**Para citar este artículo / To cite this Article**
A. J. Ochoa-Trujillo, C. A. Villarreal-Orozco, J. P. Hoyos-Sanchez "Evolution and Research Dynamics of Metadata Extraction in Academic Documents" Revista de Ingenierías Interfaces, vol. 9, no. 2, pp.1-26, 2026.

**Abstract**

Metadata extraction is a key technique for organizing, retrieving, and analyzing unstructured information in academic, clinical, and legal documents. It enables the transformation of raw text into structured knowledge, supporting applications such as text mining, recommendation systems, and semantic ontologies. This field has seen substantial growth, driven by advancements in Natural Language Processing (NLP). This paper analyzes the scientific production on metadata extraction from 2001 to 2025, highlighting three phases of development: a slow emergence (2001–2009, with a 2.25% annual growth rate), consolidation (2010–2016, with a 12.25% annual growth rate), and exponential expansion (2017–2024, with a 24.98% annual growth rate). The highest publication output occurred in 2023 (141 articles), while citation impact peaked in 2020 (2,922 citations), partly due to health-related research during the COVID-19 pandemic. A systematic search in Web of Science and Scopus yielded 1,177 unique articles. The analysis, conducted following the PRISMA framework, employed scientometric methods in conjunction with the Tree of Science approach to map the structural dynamics of knowledge in the field. The United States emerged as the leading contributor, accounting for 272 publications (20.25%) and 8,075 citations (30.81%). Notable researchers such as Zhang J. (14 publications, 974 citations, H-index = 10) demonstrate significant individual influence. The findings further highlight a methodological shift toward deep learning and transformer-based models, underscoring the field's strategic importance in contemporary digital ecosystems.

**Keywords**: Metadata Extraction, Scientometric Analysis, Natural Language Processing, Tree of Science, Deep Learning.

---

*Artículo inédito: "Evolution and Research Dynamics of Metadata Extraction in Academic Documents".

**Estudiantes Ingenieria Mecatrónica, Universidad Nacional de Colombia, Correo electrónico: aochoat@unal.edu.co, ORCID: 0009-0003-9778-0874, Colombia.

**Estudiantes Ingenieria Mecatrónica, Universidad Nacional de Colombia, Correo electrónico: cvillarrealo@unal.edu.co, ORCID: 0009-0000-8016-6825, Colombia.

** Doctor en Ciencias de la Electrónica, Ingeniero en Electrónica y Telecomunicaciones, Correo electrónico: jhoyoss@unal.edu.co, Universidad Nacional de Colombia, La Paz, ORCID: 0002-1825-0097, Cesar, Colombia.

**Resumen**

La extracción de metadatos es una técnica clave para organizar, recuperar y analizar información no estructurada en documentos académicos, clínicos y legales. Permite transformar texto sin procesar en conocimiento estructurado, dando soporte a aplicaciones como la minería de textos, los sistemas de recomendación y las ontologías semánticas. Este campo ha experimentado un crecimiento sustancial, impulsado por los avances en el procesamiento del lenguaje natural (PLN). El presente artículo analiza la producción científica sobre la extracción de metadatos entre 2001 y 2025, destacando tres fases de desarrollo: un surgimiento lento (2001–2009, con una tasa de crecimiento anual del 2,25 %), una consolidación (2010–2016, con una tasa del 12,25 %) y una expansión exponencial (2017–2024, con una tasa del 24,98 %). El mayor volumen de publicaciones se registró en 2023 (141 artículos), mientras que el impacto de las citas alcanzó su punto máximo en 2020 (2922 citas), en parte debido a las investigaciones relacionadas con la salud durante la pandemia de COVID-19. Una búsqueda sistemática en Web of Science y Scopus arrojó 1177 artículos únicos. El análisis, realizado siguiendo el marco PRISMA, empleó métodos cienciométricos junto con el enfoque *Tree of Science* para mapear la dinámica estructural del conocimiento en este campo. Estados Unidos surgió como el principal contribuyente, con 272 publicaciones (20,25 %) y 8075 citas (30,81 %). Investigadores destacados, como Zhang J. (14 publicaciones, 974 citas, índice H = 10), demuestran una influencia individual significativa. Los hallazgos también resaltan un cambio metodológico hacia el aprendizaje profundo (*deep learning*) y los modelos basados en *transformers*, subrayando la importancia estratégica del campo en los ecosistemas digitales contemporáneos.

**Palabras clave:** Extracción de metadatos, análisis cienciométrico, procesamiento de lenguaje natural, Tree of Science, aprendizaje profundo.

## 1. Introduction

In the information age, the ability to efficiently manage large volumes of digital data has become critically important in various scientific, technical, and professional disciplines. Among the associated challenges, one of the most relevant is the transformation of unstructured documents into useful and actionable knowledge. In this context, metadata extraction has established itself as an essential tool for the organization, retrieval, and automated analysis of information contained in academic, clinical, legal, and industrial documents [1]. Metadata are structured descriptions of the content, authorship, context, and semantics of documents that enable the transformation of unstructured data into actionable knowledge, enabling applications ranging from text mining to recommendation systems and specialized ontologies [2]. The growing interest in this topic has led to a notable expansion in scientific production over the last two decades, particularly in areas such as Natural Language Processing (NLP), deep learning (DL), and intelligent systems [3], [4], [5]. Traditional NLP techniques, such as grammatical tagging, entity extraction, and coreference resolution, have evolved into more sophisticated models based on DL, especially with the development of transformer-type architectures such as BERT, RoBERTa, and GPT. These technologies enable a contextualized representation of

language, substantially improving accuracy in tasks such as content segmentation, topic classification, author identification, and semantic linking between concepts [6], [7].

Despite growing interest in automatic metadata extraction in academic documents—a key line of research within NLP—existing reviews tend to focus on specific aspects, such as comparing software tools, analyzing algorithms in specific domains such as biomedicine—identified in this study as one of the main thematic clusters—or managing scientific documents. This specialization has left a significant gap: the absence of a comprehensive characterization that simultaneously addresses the structure, evolution, and overall dynamics of the field, from both a quantitative and structural perspective.

This research seeks to fill this gap through three fundamental contributions. First, it proposes a methodological hybridization that combines a systematic review based on the PRISMA model [8], a scientometric analysis, and the Tree of Science (ToS) approach [9], which allows us to capture both the evolution of the field and its internal structure. Second, it adopts an ecosystemic perspective that is not limited to extraction techniques, but also considers the key actors, collaboration networks, and publication platforms that shape the scientific environment. Thirdly, it offers a structural analysis that identifies and quantifies the thematic macrostructures that organize scientific production, thus contributing to a dynamic and in-depth map of available knowledge.

Based on these objectives, a systematic literature search was conducted in the Web of Science (WoS) and Scopus databases, covering the period from 2001 to 2025 and considering only academic articles. The methodological strategy followed the PRISMA protocol, and the results were analyzed from two complementary approaches: a scientometric analysis to evaluate productivity, collaboration, and impact, and the ToS method to examine the structural organization of accumulated knowledge [10].

The findings reveal a rapidly expanding field, with the United States leading production (20.25% of articles and 30.81% of total citations), and IEEE Access as the most productive journal. Thematically, the analyzed corpus is organized into three main areas: (1) the application of machine learning (ML) and data classification techniques to manage and facilitate access to academic publications in digital environments (45.61% of the papers); (2) entity recognition and information extraction from scientific documents, with a strong influence from computational linguistics (31.28%); and (3) natural language extraction in bioinformatics and clinical information contexts (23.11%).

The rest of the article is organized as follows: first, the methodology used for the collection and analysis of the documents is presented; second, the results are presented from a scientometric and structural perspective; and finally, the main conclusions, practical implications, and future directions for research in the area are discussed.

## 2. Methodology

The search was conducted using the WoS and Scopus databases, widely recognized for their comprehensive coverage of high-impact scientific publications worldwide [11], [12]. The parameters used for the search are summarized in Table I. The analysis focused on the term "Metadata Extraction," considering only documents classified as articles within the time range of 2001 to 2025. The search was conducted on April 7, 2025. A total of 507 documents were retrieved from WoS and 665 from Scopus, resulting in a combined dataset of 1,144 results.

**Table I.** Bibliometric summary of publications in scientific databases

| Parameter | Web of Science | Scopus |
|---|---|---|
| Range | 2001-2025 | |
| Date | April 7th, 2025 | |
| Document Type | Paper | |
| Words | "Metadata Extraction" | |
| Results | 507 | 665 |
| Total (Wos+Scopus) | | 1144 |

Once the results from both databases were obtained, a process of cleaning and harmonization was carried out, thus ensuring a robust and representative set for further analysis. The PRISMA model guided the systematic review process, structuring the collection, cleaning, and analysis of the data, as illustrated in Figure 1.

After the identification and screening phases, which included the elimination of duplicate records from WoS and Scopus, a final set of 1,177 publications was consolidated. All this information was organized into a 22-sheet Excel file to facilitate processing.

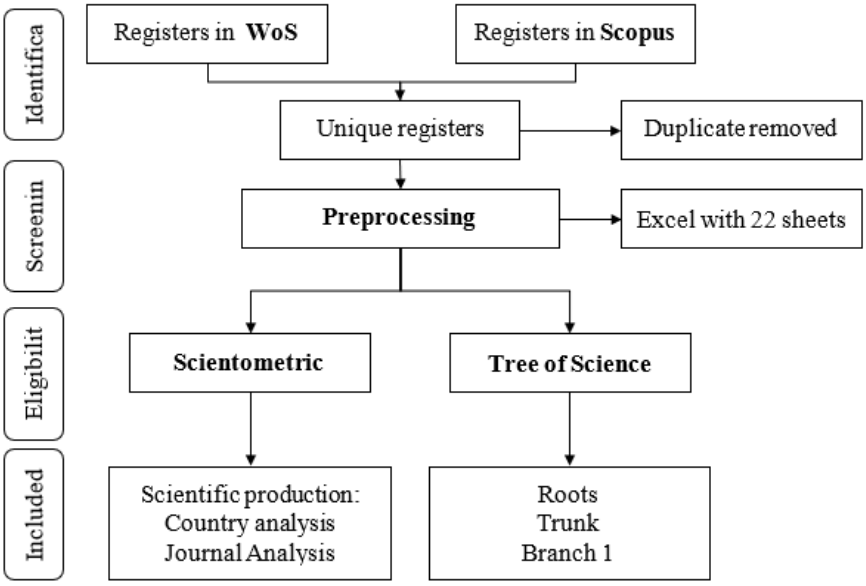


**Figure 1.** Flowchart based on the PRISMA model for document analysis

Subsequently, the refined data underwent preprocessing using advanced techniques such as text mining and automated information extraction (web scraping) in order to structure the metadata of each publication for subsequent analysis. This process prepared the dataset for two analytical approaches: scientometric analysis and the ToS methodology. Scientometric analysis made it possible to evaluate scientific output from three fundamental dimensions: countries, journals, and authors, while ToS organized the literature according to its structural contribution to the field, facilitating the identification of foundational, consolidated, and emerging documents.In this way, the process of searching for and extracting documents is complemented by a detailed analysis that allows for the evaluation of the global research landscape, as well as an understanding of its emerging lines and evolution. This approach is in line with a growing tradition of high-impact scientometric analyses that use similar techniques to map the intellectual structure and development of complex scientific fields, demonstrating their effectiveness and replicability in different areas of knowledge [13], [14], [15], [16].

## 3. Results

### Scientometric Analysis

### Scientific production

Figure 2 shows the evolution of scientific output and its impact, measured through citations, from 2001 to 2025. During this period, scientific output shows sustained growth, with a turning point around 2016, after which the total number of publications begins to increase more rapidly. This trend peaks in 2023 with 141 publications. At the same time, publications indexed in Scopus and WoS also show positive growth, especially between 2017 and 2023, a period that accounts for 56.4% of all historical output, indicating a progressive consolidation of academic output in high-impact international databases.

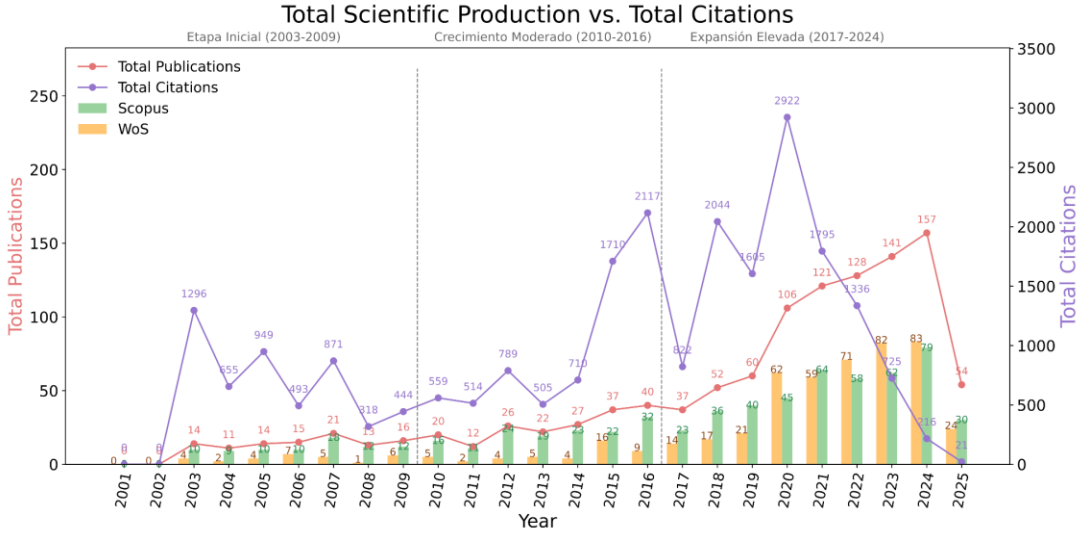


**Figure 2.** Evolution of scientific output and citation trends over time

The behavior of citations in research on metadata extraction, as illustrated by the red line, reveals an evolution marked by different stages of development. In the first stage, between 2001 and 2009, scientific production was still in its infancy, resulting in modest average annual growth of 2.25%. there were occasional peaks, such as in 2003, with 1,296 citations, suggesting the emergence of foundational works that laid the groundwork for the future of the field. A key example is the creation of the BioInfer corpus[17], a public, annotated resource designed to facilitate the development and evaluation of relationship extraction methods in the biomedical domain.

Subsequently, between 2010 and 2016, the area entered a phase of consolidation and sustained growth. This period of maturation is evident not only in a steady increase in publications and a milestone of 2,117 citations in 2015, but also in a notable acceleration quantified by an annual growth rate that skyrocketed to 12.25%. This figure confirms the transition from an emerging phase to one of consolidated expansion and greater thematic interest on the part of the scientific community. Exploring areas such as the application of machine learning to large data sets, a representative study [18] conducted an exhaustive meta-analysis on microbial profiles as biomarkers, concluding that models improved with the use of strain-specific markers and complementary control samples.

From 2017 to 2024, the field experienced a period of massive impact and exponential growth, reaching an impressive annual growth rate of 22.93%. This boom materialized in a steady increase in publications, with a maximum of 141 articles in 2023 and an absolute peak of 2,922 citations in 2020. This impact was further boosted by the COVID-19 pandemic, which spurred research on mental health analysis and digital interaction patterns, demonstrating the relevance and applicability of the field. In this context, the publication by Trotzek et al. [19] demonstrated that a hybrid model, combining convolutional neural networks and linguistic metadata, achieved cutting-edge results in identifying signs of depression in text sequences.

However, the decrease observed in the 2025 period does not reflect an actual contraction in the field. This is because the data is partial, as it is the current year, and the figures only show what has been recorded to date.

## Country Analysis

Table II presents a comparative analysis of the scientific output of the ten most active countries in the subject area studied, incorporating impact indicators (citations) and quality indicators (according to the Scimago quartile classification). Notably, the United States leads on all fronts, with the highest proportion of publications (20.25%) and a superior impact, accounting for 30.81% of total citations. This dominance is also reflected in a high concentration of articles in Q1 quartile journals, demonstrating not only abundant scientific output but also high quality.

**Table II.** Scientific Contributions by Country

| Country | Production | | Citation | | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|---|---|---|
| Usa | 272 | 20,25% | 8075 | 30,81% | 173 | 40 | 7 | 7 |
| China | 154 | 11,47% | 2185 | 8,34% | 75 | 28 | 17 | 13 |
| India | 89 | 6,63% | 896 | 3,42% | 22 | 22 | 16 | 14 |
| United Kingdom | 69 | 5,14% | 2689 | 10,26% | 47 | 10 | 1 | 1 |
| Germany | 62 | 4,62% | 1210 | 4,62% | 37 | 13 | 2 | 0 |
| Korea | 50 | 3,72% | 515 | 1,96% | 18 | 10 | 9 | 0 |
| Spain | 47 | 3,5% | 1001 | 3,82% | 31 | 5 | 2 | 2 |
| Italy | 38 | 2,83% | 860 | 3,28% | 21 | 9 | 2 | 0 |
| Canada | 36 | 2,68% | 337 | 1,29% | 21 | 7 | 3 | 0 |
| France | 35 | 2,61% | 548 | 2,09% | 23 | 3 | 1 | 0 |

China and India are emerging as significant players in terms of volume, although their relative impact is significantly lower, especially in the case of India, whose production is more dispersed among the lower quartiles (Q3 and Q4), suggesting challenges in terms of visibility and quality. In contrast, the United Kingdom represents a notable case, as, with a smaller output, it achieves a higher citation level than more productive countries, reflecting a strategy geared toward research excellence. This trend is also observed in countries such as Germany and Spain, whose scientific contribution, although more limited, maintains a positive correlation between quality and impact.

On the other hand, Korea, Canada, and France have a more modest profile in terms of both the number of publications and citations, with a lower presence in high-impact quartiles. Overall, the data reveal that international scientific relevance does not depend solely on the volume of production, but also on strategic positioning in high-impact journals, underscoring the importance of promoting policies that prioritize quality over quantity.

One of the most cited articles in the United States is the work of [20], which analyzes the interpretation of hypernymic propositions in biomedical texts using natural language processing techniques. The study employs unspecified syntactic analysis and leverages structured domain knowledge through the Unified Medical Language System (UMLS) to extract taxonomic relationships between concepts. It focuses on the use of semantic types and hierarchical relationships to improve the accuracy of biomedical information extraction, with relevant applications in information retrieval, indexing, and specialized ontology generation.

In China, the most cited article is that of Luo et al. [21], which proposes a neural network-based approach for the recognition of chemical entities in biomedical texts. The model presented, Att-BiLSTM-CRF, incorporates an attention mechanism that improves consistency in labeling and reduces dependence on manual feature engineering, achieving outstanding performance in the extraction of chemical metadata at the document level.

Meanwhile, the most influential work from India [22] introduces the SCPSO algorithm, a combination of spectral clustering with particle swarm optimization. This technique improves the classification and grouping of text documents by generating similarity graphs, which leads to more effective metadata extraction and information retrieval.

Figure 3 shows an international collaboration network built from scientific publications linked to an emerging field within information and computer sciences. In the central graph, the nodes represent countries, and the size of each node is determined by its degree of connectivity, i.e., the number of collaborative links with other countries. The thickness of the links indicates the intensity of these collaborations.

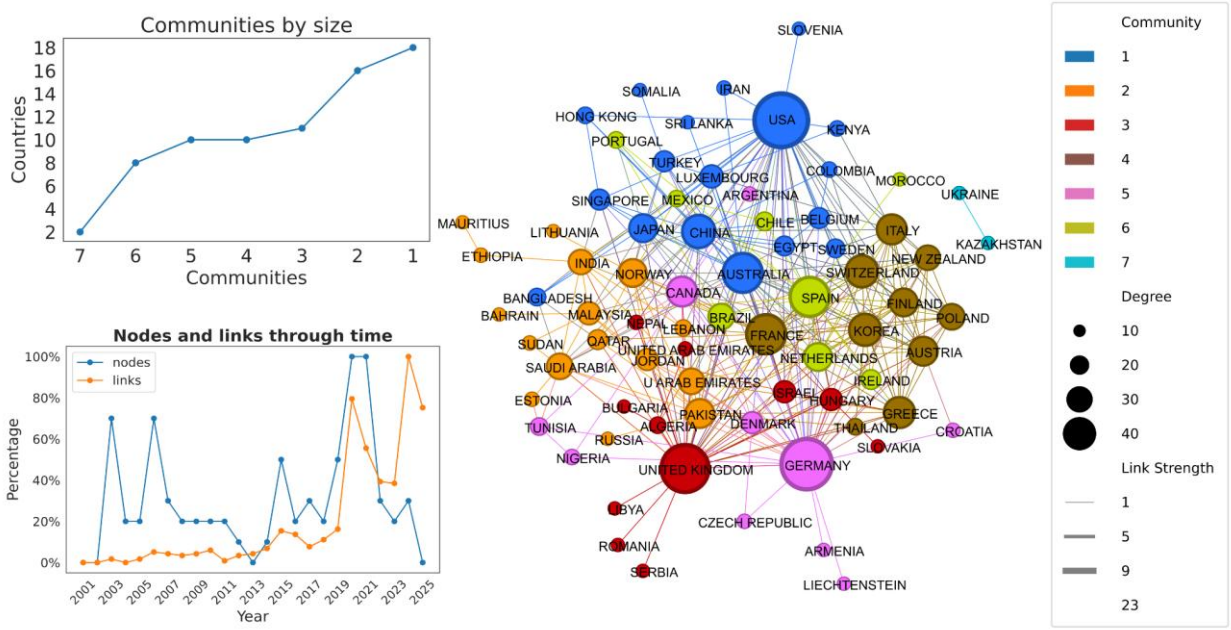


**Figure 3.** International Collaboration Network: Communities and Dynamics

A highly centralized structure can be observed, with countries such as the United States, the United Kingdom, Germany, France, and China acting as key nodes, suggesting that these actors not only lead scientific production in the area but also play an important role in the articulation of knowledge and the transfer of methodologies. Their central position may be linked to the development and adoption of technologies that enable the efficient management of large volumes of information.

The modular analysis shows that community 1 is the largest, with 18 countries, followed by other communities of decreasing size. This pattern reflects a significant concentration of technical capabilities and infrastructure in a few blocks, which could be related to the

availability of computational resources and advanced information systems.The graph in the lower left shows the evolution over time of the percentage of active countries (nodes) and international collaborations (links) between 2001 and 2024. Until approximately 2011, activity remained low and stable; however, from 2012 onwards, there has been progressive growth, with pronounced peaks in recent years. This trend may be linked to the rise in the production of scientific and digital data, as well as the need to develop more efficient mechanisms for its organization, description, and interoperability [23], [24], [25].

## Journal Analysis

Table III Journals provides a detailed overview of ten scientific journals relevant to the field of study, comparing them using six bibliometric indicators: number of publications indexed in WoS and Scopus, SJR (SCImago Journal Rank), H-Index, and quartile position (Quantile). In terms of scientific output, IEEE ACCESS leads with 33 articles in WoS and 28 in Scopus, followed at a distance by Applied Sciences-Basel (12 in WoS and 0 in Scopus) and BMC Bioinformatics (4 in WoS and 29 in Scopus). Several medium-sized journals, such as Journal of Biomedical Informatics and Expert Systems with Applications, show mixed results: only 3 publications in WoS compared to 35 and 17 in Scopus, respectively.

**Table III.** Scientific output and impact of the most relevant journals in the field of study

| Journal | WoS | Scopus | Sjr | H Index | Quantile |
|---|---|---|---|---|---|
| IEEE ACCESS | 33 | 28 | 0.849 | 290 | Q1 |
| Journal Of Biomedical Informatics | 3 | 35 | 1.257 | 137 | Q1 |
| Bmc Bioinformatics | 4 | 29 | 1.190 | 251 | Q1 |
| Expert Systems With Applications | 3 | 17 | 1.854 | 290 | Q1 |
| Artificial Intelligence In Medicine | 1 | 14 | 1.396 | 118 | Q1 |
| Information Processing And Management | 0 | 15 | 2.062 | 137 | Q1 |
| Journal Of Biomedical Semantics | 3 | 9 | 0.491 | 48 | Q2 |
| Applied Sciences-Basel | 12 | 0 | -- | -- | -- |
| Bioinformatics | 0 | 10 | 2.451 | 486 | Q1 |
| Multimedia Tools And Applications | 4 | 7 | 0.777 | 116 | Q1 |

In terms of impact and visibility, the SJR ranges from 0.491 (Journal of Biomedical Semantics) to 2.451 (Bioinformatics). Information Processing and Management (SJR 2.062) and Expert Systems with Applications (1.854) stand out as the most prestigious according to this indicator, even though the former does not have any articles indexed in WoS. Bioinformatics has the highest SJR and an H-Index of 486, the highest value in Table 3, and is indicative of a solid citation history. The H-index reflects both productivity and cumulative citations: IEEE ACCESS and Expert Systems with Applications share an H-index of 290, while other journals, such as Journal of Biomedical Informatics and Information Processing and Management, reach an H-index of 137. Only the Journal of Biomedical Semantics is ranked in Q2, suggesting an intermediate impact, and Applied Sciences-Basel lacks SJR, H-Index, and quartile data, possibly because it is newer or less established.

An example of the prestige of this journal is the work of [20], which explores how domain knowledge and linguistic structure combine to interpret hypernymic propositions in biomedical texts. This work has been cited more than 400 times.

Another high-profile journal is BMC Bioinformatics, which also appears in the Q1 quartile. This journal has been an important channel for research focused on natural language processing and information extraction in biomedicine. For example, the article [17] proposes an annotated corpus specifically designed for text mining tasks, which has been instrumental in training and evaluating NLP models in biomedical contexts.

The most prominent journal is IEEE Access, with a total of 65 publications, an H-index of 290, an SJR of 0.840, and a position in the Q1 quartile. One of its representative articles is the work by [26], which addresses the improvement of medical entity recognition in Spanish clinical narratives through the use of contextualized word embeddings. The study proposes a deep learning-based system that uses models such as BERT and Flair, evaluating different textual representation strategies to extract mentions of diseases, symptoms, and medications from electronic clinical records.

Figure 4 shows a citation network between scientific journals, organized into three distinct thematic communities, represented by different colors. Each node corresponds to a journal, whose size indicates its degree of connectivity (number of links to other journals), while the thickness of the lines represents the intensity of shared citations. This structure allows us to identify centers of knowledge production and circulation that share common methodological and thematic approaches.

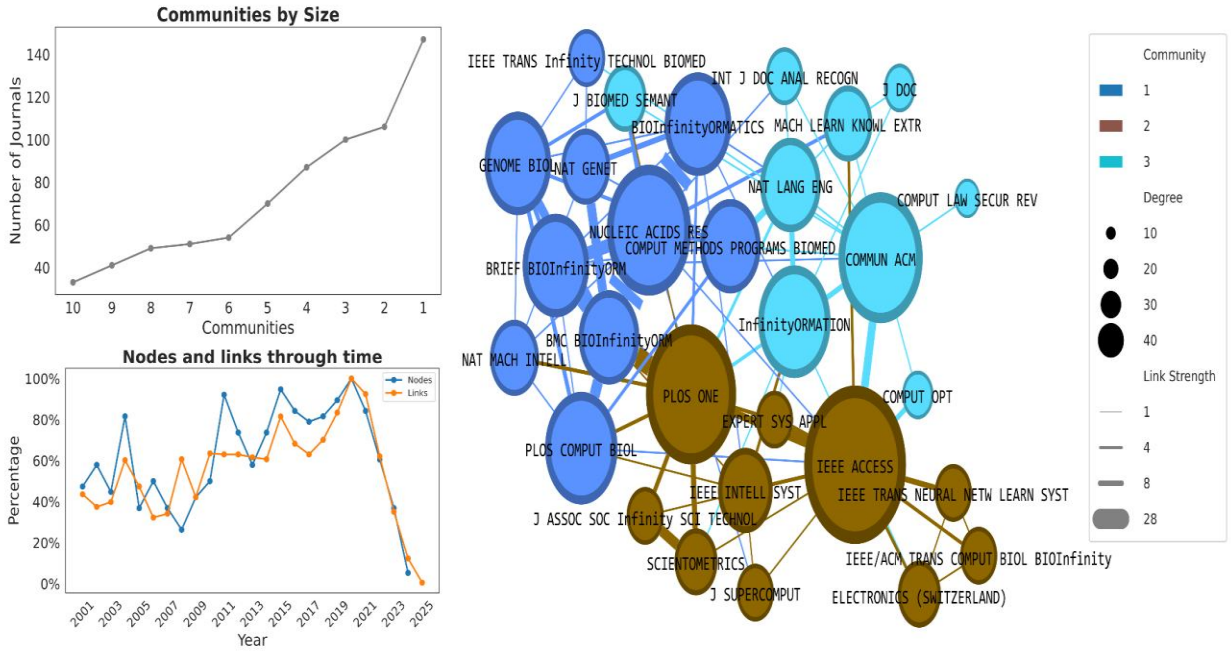


**Figure 4.** Citation network between journals: Communities and Dynamics

Community 1 is distinguished by its focus on the use of NLP techniques, machine learning, and text mining in the biomedical domain [17], [18], [27]. Community 2 brings together work focused on the automated management of scientific, legal, and financial documents using artificial intelligence [28], [29], [30]. For its part, community 3 focuses on the classification and semantic enrichment of complex texts with applications in health, digital libraries, and public administration [31], [32], [33].

The auxiliary figures reinforce this interpretation. The graph "Communities by Size" indicates that community 1 is the largest, with more than 140 journals, which demonstrates its consolidation and thematic leadership. Meanwhile, the graph "Nodes and links through time" shows sustained growth in the participation of journals (nodes) and their links (citations) until recent years, although there is evidence of a sharp decline since 2022.

## Author Collaboration Network

Table IV summarizes the productivity and impact of ten researchers by number of articles, total citations, and H-index. Zhang Y, with 21 publications, leads in volume. One of his key contributions is the Att-BiLSTM-CRF model for chemical entity recognition, with outstanding results in biomedical tasks [21].

**Table IV.** Leading authors by scientific output and impact metrics

| No | Researcher | Papers Total | Total Citations | H-Index |
|---|---|---|---|---|
| 1 | Zhang Y | 21 | 507 | 6 |
| 2 | Wang J | 16 | 480 | 6 |
| 3 | Zhang J | 14 | 974 | 10 |
| 4 | Afzal M | 10 | 173 | 7 |
| 5 | Kim S | 9 | 464 | 4 |
| 6 | Li Y | 9 | 108 | 3 |
| 7 | El-Gohary N | 8 | 724 | 8 |
| 8 | Li Q | 8 | 64 | 4 |
| 9 | Liu X | 8 | 16 | 2 |
| 10 | Wang L | 8 | 628 | 4 |

Zhang J, with only 14 articles, has the highest number of citations (974) and the highest H-index (10), reflecting a consistent impact. His joint work with El-Gohary N proposes an automated semantic extension of BIM models using NLP and ML. Both stand out for combining quality and influence in their publications [34]. Afzal M also shows a solid balance (10 articles, H=7)[35], thanks to his research on document classification using BERT and genetic algorithms . In contrast, Liu X, with 8 articles but only 16 citations (H=2), reflects a low impact. Cases such as Wang L (628 citations, H=4) and Kim S (464 citations, H=4) show an impact concentrated in a few publications. The table highlights that, beyond volume, scientific influence is measured by the sustained resonance of the works.

The scientific collaboration network reflects the emergence of interconnected academic communities. Figure 5 was created based on the ego networks of the main authors (see Table 4). For example, Chen Y. has collaborated with Xu N., Zhang Z., Shen Y., Zhang Q., Liu Z., Yu Y., Wang Y., Lei C., Ke M., Qiu D., Lu T., Xiong J., and Qian H. In their study [34], they evaluated the performance of different binary classification models applied to high-throughput microbial sequencing datasets. The authors compared machine learning and deep learning approaches—including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and a Backpropagation Neural Network (BPNN)—to predict microbial response to environmental changes. The results showed that deep learning models are capable of capturing complex relationships in microbiome data and, in many cases, outperform traditional algorithms in predictive accuracy.

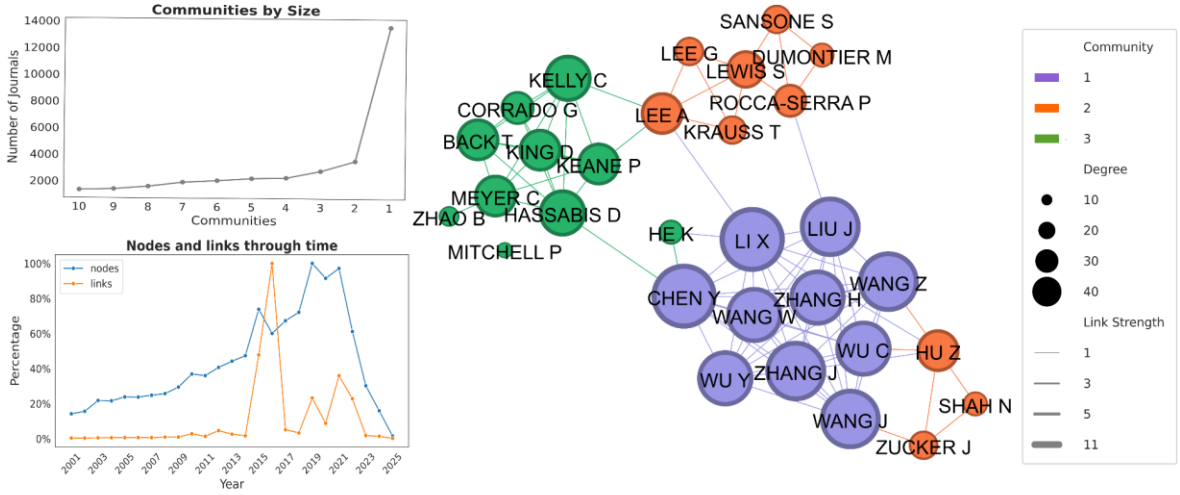


**Figure 5.** Co-authorship network: Communities and Dynamics

## Tree of Science

### Root

Over the last decade, research into language representations has undergone a remarkable evolution, moving from methods based on local co-occurrences to deep attention architectures that capture broad, bidirectional contexts. In 2013, [36] introduced Word2Vec, a pair of architectures (CBOW and Skip-gram) capable of learning large-scale word vectors in a matter of hours. Thanks to their simplicity and computational efficiency, these vectors demonstrated, for the first time, that it is possible to model syntactic and semantic relationships at a low cost, thus laying the foundation for NLP systems that require measuring similarity between terms.

Shortly thereafter, [37] introduced GloVe, which enriched embeddings by incorporating global co-occurrence statistics through weighted matrix factorization. This approach demonstrated that considering the entire corpus—and not just local context windows—improves the quality of lexical analogies and word similarity tasks, reinforcing the idea that higher-level semantic relationships are best captured with global information.

Meanwhile, in 2011, the development of Scikit-learn [38] offered a Python library for classic ML algorithms with a consistent API and optimized performance in C++. Although not directly related to deep neural networks, its presence was key to establishing a unified environment where preprocessing techniques, feature selection, and lightweight models could be combined—in many cases using Word2Vec or GloVe embeddings as input—thus accelerating both research and industrial adoption.

The real turning point came in 2017 with “Attention Is All You Need” [6], in which Vaswani et al. demonstrated that a model based solely on self-attention mechanisms could outperform recurrent and convolutional architectures in machine translation, while also

facilitating massive parallelization. With BLEU scores of 28.4 from English to German and 41.8 from English to French, Transformer redefined the design of NLP networks and paved the way for a whole family of attention-based models.

Building on that foundation, Devlin et al. introduced BERT in 2019 [7], a pre-trained bidirectional Transformer model with token masking and consecutive sentence prediction objectives. By simultaneously "reading" the context to the left and right of each term, BERT achieved state-of-the-art results in classification, question answering, and other tasks, and demonstrated that fine-tuning a pre-trained model at relatively low cost was sufficient to adapt it to a wide variety of applications.

## Trunk

The development of advanced language models has enabled unprecedented advances in the automated analysis of specialized texts. A relevant case in the clinical field in Spanish is the study by[26], which demonstrates that the use of specialized embeddings in medicine, together with stacking techniques, significantly improves performance in named entity recognition (NER) in medical reports.

This approach has been adopted in other technical fields. In the area of manufacturing, Kumar and Starly propose FabNER [39], a BiLSTM-CRF-based system that classifies technical terms (such as materials, methods, or production parameters) with 88% accuracy. Their model not only automates the extraction of knowledge from scientific articles but also facilitates the creation of structured databases to optimize industrial processes.

Beyond NER, document classification also benefits from these techniques. [40] introduce a hierarchical attention capsule network capable of processing long and complex biomedical texts, analyzing both individual words and sentence structures. When evaluating the model on public datasets, it matches or outperforms the most advanced methods, demonstrating that the combination of multilevel attention and capsules improves semantic representation.

However, not all embeddings work equally well in all contexts. [41] compare representations trained on different types of corpora (clinical, scientific, Wikipedia, and news) and conclude that medical and scientific embeddings better capture specialized terminology. However, they also point out that there is no one-size-fits-all solution: generalist embeddings may be more useful for certain tasks, while specialized embeddings are essential for specific applications.

The versatility of these methods is demonstrated in multilingual applications. [42] use a multilingual variant of BERT to extract structured information from resumes in several languages, identifying sections such as education, experience, and personal details with high accuracy. Furthermore, their analysis of attention weights provides transparency to the process, confirming that techniques developed in biomedicine can be successfully adapted to other domains.

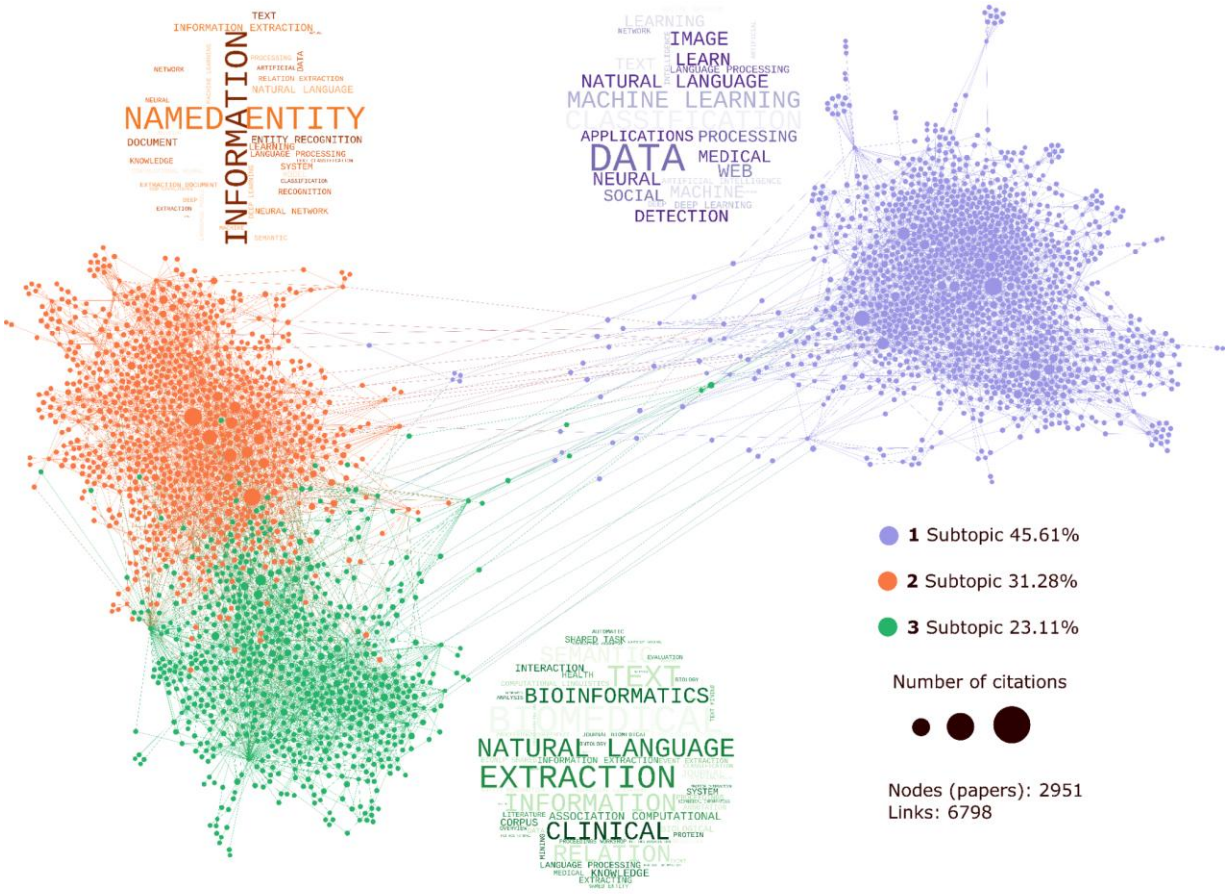


**Figure 6**. Thematic network of concepts in publications on metadata extraction

**Branch 1 - Trends and Applications of Machine Learning and Deep Learning**

The field of artificial intelligence shows a clear emphasis on the application of ML and DL techniques to solve complex problems, with a particular interest in computer vision and its uses in the industrial sector. This trend is evident in the development of advanced models designed to process both textual and visual data, to create increasingly sophisticated and accurate automated analysis systems.

In the field of text processing and systems management, [43] present an approach that combines NLP with multidimensional classifiers to correct metadata in computerized maintenance management systems (CMMS) in the pharmaceutical industry. By treating fault notifications with GRU networks and semantic embeddings, their method improves the accuracy of error codes and illustrates how automated analysis and management techniques optimize workflows in industrial environments.

For their part, [44] develop a system for detecting and classifying malware in executable files that merges graphical representations of function calls FCGV with statistical features.

Although it does not use computer vision, its combination of machine learning and binary data analysis shows the cross-cutting scope of these methods beyond the visual domain.

On the other hand, the work of Tajjour [45] and [45], [46] focuses on the area of image vision, specifically Segmentation and Detection. For example, Tajjour et al. design a hybrid CNN and MLP model for the early diagnosis of skin cancer, applying color space conversions and convolutional networks to extract visual features from lesions. With an accuracy of 86% and an AUC of 96%, it exemplifies how DL is advancing in medical applications of Image Analysis. Similarly, [46] use object detectors and geometric projections on monocular video for counting, classifying, and estimating the speed of vehicles in traffic environments. Their pipeline is capable of processing 60 fps and generating high-quality metadata.

**Branch 2 - Intelligent Systems for Semantic Information Extraction**

The focus on the design, implementation, and evaluation of advanced computer systems is a cross-cutting theme in current research, with applications that address highly complex problems in the real world. A paradigmatic example of this trend is the study by [47] that focuses on the legal domain, a sector where accuracy and efficiency are critical. The authors identify a fundamental bottleneck in judicial processes: the extraction of crucial information from massive and heterogeneous lists of evidence, a task that, when performed manually, is "complex and inaccurate." To overcome this challenge, they propose a sophisticated multimodal network model based on semantic enhancements (SEBM). The architecture of this system is distinguished by its ability to explicitly integrate expert knowledge. Before training, they construct a "semantic entity graph" that encodes the relationships between different categories of legal information, serving as a guide for the model.

Complementarily, if Luo and Yu's work exemplifies how a system can be guided by explicit knowledge, Benfer and Müller's review [48] explores the opposite challenge: how to build that knowledge from scratch based on empirical data. The solution they analyze is the automated creation of "semantic digital twins": virtual, dynamic replicas of physical systems that are updated in real time. Instead of relying on a predefined graph, the methods they review are based on "metadata inference", using AI to discover patterns and relationships directly in the raw data and thus build the semantic model automatically.

Similarly, the article by [49] focuses on a different but related challenge: the extraction of information from "Visually Rich Documents". In documents such as invoices, reports, or forms, meaning resides not only in the text but also in its spatial layout, and traditional methods that ignore visual design fail. Their study reviews and categorizes DL approaches designed for this task, highlighting that the key to automated document understanding lies in the effective fusion of textual and geometric information to interpret the logical and semantic structure of the document.

On the other hand, the hybrid model proposed by [50], which combines CNNs with SVMs to classify office documents, is directly related to the focus on DL and linguistic processing techniques. The use of attention mechanisms and Word2Vec in its methodological pipeline reflects the refinement of current models for textual categorization tasks.

Finally, the application of these predictive systems extends to the clinical field, an area where accuracy is vital and training data is often scarce. The study [51] addresses this challenge in computer-assisted detection for medical images. Given the difficulty of obtaining large annotated medical datasets, their research explores various training strategies for deep neural networks.

**Branch 3 - Information Extraction and Processing in Biomedical and Clinical Texts**

The current academic ecosystem shows an intense focus on data science and machine learning, with a strong interest in the development of predictive, intelligent, and adaptive models. This emphasis translates into the construction of complex systems that integrate DL and NLP techniques to address challenges in critical domains.

A particularly relevant and high-impact area of application is the extraction of clinical information from medical texts, where these technological advances offer considerable potential to transform health analysis and management. For example,[52] developed a system based on Recurrent Neural Networks (RNN) to identify clinical events in cardiology reports written in Italian. Their model, trained with a corpus of 75 annotated reports, significantly improves accuracy compared to traditional approaches by combining DL with classic NLP techniques.

From an unsupervised perspective, Zhang and Elhadad propose an alternative methodology based on entity identification using noun phrases and contextual semantic analysis, without the need for predefined rules [53]. This approach is particularly useful in contexts where annotated data is scarce, yielding highly competitive results. Along the same lines, but in the context of Spanish, [54] present a multi-head CRF classifier capable of simultaneously identifying five types of biomedical entities: symptoms, procedures, diseases, chemicals, and proteins. This system avoids label overlap and is based on the SNOMED CT ontology, ensuring structured and accurate extraction from Spanish clinical texts.

The temporal dimension has also been addressed by [55], who integrate dynamic representations of multiple sentences to improve the chronological interpretation of medical events. Their model, validated on the THYME corpus, demonstrates high performance, underscoring the importance of time in automated clinical analysis.

The use of advanced language models such as BERT is evident in the work of [56], who apply data mining and artificial intelligence techniques to construct knowledge graphs and clinical ontologies. Their proposal facilitates the detection of complex semantic entities and relationships and has been successfully validated in pediatric records.

Finally, [57] propose a software architecture geared toward the structured extraction of medical knowledge. Their system is based on ontologies and complies with semantic interoperability standards, consisting of four stages: linguistic analysis, annotation extraction, ontological population, and knowledge representation. Validated with more than 200 documents, the model achieves high levels of accuracy, recall, and F-measure.

## Conclusions

Metadata extraction has established itself as a fundamental discipline for the organization, retrieval, and automated analysis of information in academic, clinical, and legal contexts. Based on a systematic review of the literature and a scientometric analysis of the period 2001-2025, this research has characterized the evolution of the field, revealing exponential growth in its scientific output. This development has occurred in three clear phases: a slow emergence between 2001 and 2009 (2.25% annual growth), a consolidation phase between 2010 and 2016 (12.25%), and an accelerated expansion between 2017 and 2024 (22.93%). This last stage accounts for 56.4% of all historical production, reaching a maximum of 141 articles in 2023 and a peak impact of 2,922 citations in 2020, driven in part by research related to the COVID-19 pandemic.

Sciometric analysis has made it possible to quantify the structure of the field, identifying the United States as the undisputed leader, with 20.25% of total publications and 30.81% of citations, also standing out for the high quality of its output concentrated in Q1 quartile journals. China ranks as the second most productive country in terms of volume, with 11.47% of articles. Thematically, knowledge is organized into three broad areas: (1) the application of machine learning techniques for the management of academic publications, which accounts for 45.61% of the papers; (2) entity recognition and information extraction influenced by computational linguistics (31.28%); and (3) natural language extraction in bioinformatics and clinical information contexts (23.11%).

The methodological approach of the ToS confirmed a clear transition from traditional NLP approaches to advanced DL models. There is evidence of an evolution from early vector representations, such as Word2Vec, to Transformer architectures and contextualized embeddings such as BERT, which have become the industry standard. Emerging trends point to the design of intelligent systems for semantic extraction, application in specialized domains such as law or manufacturing, and the integration of multimodal data. These findings demonstrate not only a quantitative expansion of the field but also a growing methodological sophistication. Ultimately, research in metadata extraction will continue to play a strategic role in the digital ecosystem, driving semantic interoperability and the construction of structured knowledge in increasingly dynamic and complex environments.

This study has limitations inherent to its scope. The multidisciplinary nature of the field, which ranges from biomedicine to the legal sector, requires that the evaluation of model performance be complemented by expert validation, going beyond purely technical metrics. Furthermore, the analysis is restricted to scientific output identified under the term

"Metadata Extraction," which may have omitted relevant work that uses different terminology.

Future research should focus on three key areas: developing multimodal models that combine text and visual elements from documents, expanding their application to new complex fields such as law, and resolving important challenges such as scalability, model interpretability, and adaptation to more languages.

Finally, it is concluded that research into metadata extraction will continue to play a strategic role in the contemporary digital ecosystem, driving semantic interoperability, document automation, and the construction of structured knowledge in increasingly complex and dynamic environments.

**References**

[1]M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents," in Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, in JCDL '13. New York, NY, USA: Association for Computing Machinery, Jul. 2013, pp. 385–386. doi: 10.1145/2467696.2467753.

[2]R. Abascal, B. Rumpler, and J.-M. Pinon, "Information Retrieval in Digital Theses Based on Natural Language Processing Tools," in Advances in Natural Language Processing, J. L. Vicedo, P. Martínez-Barco, R. Muńoz, and M. Saiz Noeda, Eds., Berlin, Heidelberg: Springer, 2004, pp. 172–182. doi: 10.1007/978-3-540-30228-5_16.

[3]Z. Boukhers and A. Bouabdallah, "Vision and natural language for metadata extraction from scientific PDF documents: a multimodal approach," in Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, in JCDL '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 1–5. doi: 10.1145/3529372.3533295.

[4]J. Du et al., "Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks," Sci. Rep., vol. 15, no. 1, p. 19379, Jun. 2025, doi: 10.1038/s41598-025-03979-5.

[5]P. Raval and H. Bhaidasna, "A Review of Extracting Metadata from Scholarly Articles using Natural Language Processing (NLP)," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), Dec. 2024, pp. 1355–1359. doi: 10.1109/ICACRS62842.2024.10841656.

[6]A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[7]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," presented at the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[8]M. J. Page et al., "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," BMJ, vol. 372, p. n160, Mar. 2021, doi: 10.1136/bmj.n160.

[9]M. Zuluaga, S. Robledo, O. Arbelaez-Echeverri, G. A. Osorio-Zuluaga, and N. Duque-Méndez, "Tree of Science - ToS: A Web-Based Tool for Scientific Literature Recommendation. Search Less, Research More!," Issues Sci. Technol. Librariansh., no. 100, Art. no. 100, Aug. 2022, doi: 10.29173/istl2696.

[10]S. Robledo, L. Valencia, M. Zuluaga, O. A. Echeverri, and J. W. A. Valencia, "tosr: Create the Tree of Science from WoS and Scopus," J. Scientometr. Res., vol. 13, no. 2, pp. 459–465, Jul. 2025, doi: 10.5530/jscires.13.2.36.

[11]L. Ganti, N. A. Persaud, and T. S. Stead, "Bibliometric analysis methods for the medical literature," Acad. Med. Surg., Jan. 2025, doi: 10.62186/001c.129134.

[12]W. M. To and B. T. W. Yu, "Artificial Intelligence Research in Tourism and Hospitality Journals: Trends, Emerging Themes, and the Rise of Generative AI," Tour. Hosp., vol. 6, no. 2, Art. no. 2, Jun. 2025, doi: 10.3390/tourhosp6020063.

[13]S. Valencia, M. Zuluaga, M. C. Florian Pérez, K. F. Montoya-Quintero, M. S. Candamil-Cortés, and S. Robledo, "Human Gut Microbiome: A Connecting Organ Between Nutrition, Metabolism, and Health," Int. J. Mol. Sci., vol. 26, no. 9, Art. no. 9, Jan. 2025, doi: 10.3390/ijms26094112.

[14]A. Vargas-Hernández, S. Robledo, and G. R. Quiceno, "Virtual Teaching for Online Learning from the Perspective of Higher Education: A Bibliometric Analysis," J. Scientometr. Res., vol. 13, no. 2, pp. 406–418, Jul. 2025, doi: 10.5530/jscires.13.2.32.

[15]I. Irwanto and T. D. S. Rini, "Research trends in blended learning in chemistry: A bibliometric analysis of scopus indexed publications (2012–2022)," J. Turk. Sci. Educ., vol. 21, no. 3, pp. 566–578, Sep. 2024, doi: 10.36681/tused.2024.030.

[16]H. Ma and L. Ismail, "Bibliometric analysis and systematic review of digital competence in education," Humanit. Soc. Sci. Commun., vol. 12, no. 1, p. 185, Feb. 2025, doi: 10.1057/s41599-025-04401-1.

[17]S. Pyysalo et al., "BioInfer: a corpus for information extraction in the biomedical domain," BMC Bioinformatics, vol. 8, pp. 1–24, 2007.

[18]E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," PLOS Comput. Biol., vol. 12, no. 7, p. e1004977, Jul. 2016, doi: 10.1371/journal.pcbi.1004977.

[19]M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences," IEEE Trans. Knowl. Data Eng., vol. 32, no. 3, pp. 588–601, Mar. 2020, doi: 10.1109/TKDE.2018.2885515.

[20]T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," Unified Med. Lang. Syst., vol. 36, no. 6, pp. 462–477, Dec. 2003, doi: 10.1016/j.jbi.2003.11.003.

[21]L. Luo et al., "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," Bioinformatics, vol. 34, no. 8, pp. 1381–1388, 2018.

[22]R. Janani and S. Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimization," Expert Syst. Appl., vol. 134, pp. 192–200, 2019, doi: 10.1016/j.eswa.2019.05.030.

[23]N. Asadova et al., "T-matrix representation of optical scattering response: Suggestion for a data format," J. Quant. Spectrosc. Radiat. Transf., vol. 333, p. 109310, Mar. 2025, doi: 10.1016/j.jqsrt.2024.109310.

[24]H. Turki et al., "MeSH2Matrix: combining MeSH keywords and machine learning for biomedical relation classification based on PubMed," J. Biomed. Semant., vol. 15, no. 1, p. 18, Oct. 2024, doi: 10.1186/s13326-024-00319-w.

[25]C. A. Primiero et al., "A protocol for annotation of total body photography for machine learning to analyze skin phenotype and lesion classification," Front. Med., vol. 11, p. 1380984, 2024, doi: 10.3389/fmed.2024.1380984.

[26]L. Akhtyamova, P. Martínez, K. Verspoor, and J. Cardiff, "Testing contextualized word embeddings to improve NER in Spanish clinical case narratives," IEEE Access, vol. 8, pp. 164717–164726, 2020.

[27]S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, a natural language processing engine for MEDLINE abstracts," Bioinformatics, vol. 19, no. 13, pp. 1699–1706, 2003, doi: 10.1093/bioinformatics/btg207.

[28]L. Shi, C. Jianping, and X. Jie, "Prospecting Information Extraction by Text Mining Based on Convolutional Neural Networks-A Case Study of the Lala Copper Deposit, China," IEEE Access, vol. 6, pp. 52286–52297, 2018, doi: 10.1109/ACCESS.2018.2870203.

[29]L. Chen, S. Xu, L. Zhu, J. Zhang, X. Lei, and G. Yang, "A deep learning based method for extracting semantic information from patent documents," Scientometrics, vol. 125, no. 1, pp. 289–312, 2020, doi: 10.1007/s11192-020-03634-y.

[30]J.-L. Seng and J. T. Lai, "An intelligent information segmentation approach to extract financial data for business valuation," Expert Syst. Appl., vol. 37, no. 9, pp. 6515–6530, 2010, doi: 10.1016/j.eswa.2010.02.134.

[31]I. Malashin, I. Masich, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports," Mach. Learn. Knowl. Extr., vol. 6, no. 2, pp. 1361–1377, 2024, doi: 10.3390/make6020064.

[32]D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," Int. J. Doc. Anal. Recognit., vol. 18, no. 4, pp. 317–335, 2015, doi: 10.1007/s10032-015-0249-8.

[33]R. K. Srihari, W. Li, T. Cornell, and C. Niu, "InfoXtract: A customizable intermediate level information extraction engine," Nat. Lang. Eng., vol. 14, no. 1, pp. 33–69, 2008, doi: 10.1017/S1351324906004116.

[34]N. Xu et al., "Compare the performance of multiple binary classification models in microbial high-throughput sequencing datasets," Sci. Total Environ., vol. 837, p. 155807, 2022.

[35]G. Mustafa et al., "Optimizing document classification: Unleashing the power of genetic algorithms," IEEE Access, vol. 11, pp. 83136–83149, 2023.

[36]T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," ArXiv Prepr. ArXiv13013781, 2013.

[37]J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[38]F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[39]A. Kumar and B. Starly, "'FabNER': information extraction from manufacturing process science domain literature using named entity recognition," J. Intell. Manuf., vol. 33, no. 8, pp. 2393–2407, 2022.

[40]J. Wang, M. Li, Q. Diao, H. Lin, Z. Yang, and Y. Zhang, "Biomedical document triage using a hierarchical attention-based capsule network," BMC Bioinformatics, vol. 21, pp. 1–20, 2020.

[41]Y. Wang et al., "A comparison of word embeddings for the biomedical natural language processing," J. Biomed. Inform., vol. 87, pp. 12–20, Nov. 2018, doi: 10.1016/j.jbi.2018.09.008.

[42]D. Vukadin, A. S. Kurdija, G. Delač, and M. Šilić, "Information extraction from free-form CV documents in multiple languages," IEEE Access, vol. 9, pp. 84559–84575, 2021.

[43]A. Deloose, G. Gysels, B. De Baets, and J. Verwaeren, "Combining natural language processing and multidimensional classifiers to predict and correct CMMS metadata," Comput. Ind., vol. 145, p. 103830, 2023.

[44]Y. Zhang, X. Chang, Y. Lin, J. Mišić, and V. B. Mišić, "Exploring function call graph vectorization and file statistical features in malicious PE file classification," IEEE Access, vol. 8, pp. 44652–44660, 2020.

[45]S. Tajjour, S. Garg, S. S. Chandel, and D. Sharma, "A novel hybrid artificial neural network technique for the early skin cancer diagnosis using color space conversions of original images," Int. J. Imaging Syst. Technol., vol. 33, no. 1, pp. 276–286, 2023.

[46]C. Liu, D. Q. Huynh, Y. Sun, M. Reynolds, and S. Atkinson, "A vision-based pipeline for vehicle counting, speed estimation, and classification," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 12, pp. 7547–7560, 2020.

[47]S. Luo and J. Yu, "A semantic enhancement-based multimodal network model for extracting information from evidence lists," Neural Netw., vol. 187, p. 107387, 2025.

[48]R. Benfer and J. Müller, "Semantic digital twin creation of building systems through time series based metadata inference–A review," Energy Build., p. 114637, 2024.

[49]H. Gbada, K. Kalti, and M. A. Mahjoub, "Deep learning approaches for information extraction from visually rich documents: datasets, challenges and methods," Int. J. Doc. Anal. Recognit. IJDAR, pp. 1–22, 2024.

[50]S. Yuan, “Design of Intelligent Document Categorization System for Office Software Combined with Neural Networks,” Appl. Math. Nonlinear Sci., vol. 9, no. 1, 2024, doi: 10.2478/amns-2024-3357.

[51]H.-C. Shin et al., “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[52]N. Viani et al., “Supervised methods to extract clinical events from cardiology reports in Italian,” J. Biomed. Inform., vol. 95, p. 103219, 2019.

[53]S. Zhang and N. Elhadad, “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts,” J. Biomed. Inform., vol. 46, no. 6, pp. 1088–1098, 2013.

[54]R. A. Jonker, T. Almeida, R. Antunes, J. R. Almeida, and S. Matos, “Multi-head CRF classifier for biomedical multi-class named entity recognition on Spanish clinical notes,” Database, vol. 2024, p. baae068, 2024.

[55]S. Zhao and L. Li, “Temporal information extraction with the scalable cross-sentence context for electronic health records,” J. Biomed. Inform., vol. 128, p. 104052, 2022, doi: 10.1016/j.jbi.2022.104052.

[56]R. K. Zulkarneev, N. I. Yusupova, O. N. Smetanina, M. M. Gayanova, and A. M. Vulfin, “Method and models of extraction of knowledge from medical documents,” Inform. Autom., vol. 21, no. 6, pp. 1169–1210, 2022.

[57]D. C. Moreno and M. Vargas-Lombardo, “Design and construction of a NLP based knowledge extraction methodology in the medical domain applied to clinical information,” Healthc. Inform. Res., vol. 24, no. 4, pp. 376–380, 2018.